INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE MONTERREY

CAMPUS CUERNAVACA

USO DE LA MINERÍA DE DATOS PARA LA ELABORACIÓN DE UN MODELO DE PREDICCIÓN PARA LA OPTIMIZACIÓN DE LA PRODUCCIÓN AGRÍCOLA EN DISTRITOS DE RIEGO

T E S I S

PRESENTADA COMO REQUISITO PARCIAL PARA OBTENER EL GRADO ACADÉMICO DE:

MAESTRO EN CIENCIAS COMPUTACIONALES

POR:

ALBERTO GONZÁLEZ SÁNCHEZ

ASESOR:

DRA. MÓNICA LARRE BOLAÑOS CACHO

CUERNAVACA, MORELOS
JULIO 2007

Dedicatoria

A mi Esposa: por su paciencia y apoyo.

A mi Madre: por la educación y los principios.

A mi Hija: por alegrarme la vida todos los días.

A mi Hermana: por la niñez compartida.

Y a mi Abuela: por consentirme.

Agradecimientos

Primeramente, quiero agradecer a mi asesora, la Dra. Mónica Larre Bolaños Cacho, por su orientación y guía constante durante el desarrollo y la escritura de esta tesis.

- Al Dr. Juan Frausto Solís, por sus consejos y asesoría durante los cursos y la cátedra de optimización combinatoria.

- A mis "Jefes" –dicho sea con cariño- del IMTA, el Dr. Arturo González Casillas y el Ing. José Ángel Guillén González, quienes me dieron la oportunidad de ausentarme del trabajo para dedicarle el tiempo requerido a la Maestría.

- Al Ing. Josafat Caballero de la Comisión Nacional del Agua, y el Dr. Waldo Ojeda Martínez del IMTA, esto por las facilidades otorgadas en cuanto a la recolección de la información utilizada durante el presente trabajo.

- A los especialistas en irrigación y entrañables compañeros del IMTA, Ing. Mauricio Barbosa, Ing. Omar Olivar, e Ing. Julio César Fernández, por sus magníficos consejos en la materia, bibliografía y demás material facilitado, el cual fue de suma utilidad para el desarrollo de esta tesis.

- Al Ing. José Eleno Curiel Gamiño por enseñarme los principios del diseño y la programación orientada a objetos, pero sobre todo, por darle a un joven programador inexperto su primera oportunidad de trabajo (por allá de 1994).

Finalmente, quisiera agradecer al departamento de becas del Tecnológico de Monterrey Campus Cuernavaca por la beca de descuento recibida, y sin la cual un servidor no podría haber accedido a tal nivel de educación.

A todos ustedes, muchas Gracias.

Resumen

A través del tiempo, distintos modelos y procedimientos han sido propuestos con el propósito de optimizar los elementos involucrados en el proceso productivo de cultivos agrícolas. La obtención de mayores ingresos con la inversión mínima de recursos es un problema que ha atraído desde hace mucho tiempo la atención de especialistas propios y ajenos de la materia. Lamentablemente, el inmenso número de factores involucrados en el proceso productivo dificultan enormemente la construcción de modelos eficientes y realistas que reflejen las condiciones que influyen en el rendimiento de la producción. Bajo este panorama, lo mejor es analizar los datos históricos de las cosechas con el fin de inferir comportamientos futuros que permitan realizar una correcta administración de los recursos involucrados, con miras a obtener el mayor ingreso posible.

En los años recientes, un conjunto de técnicas agrupadas bajo el nombre de minería de datos han surgido como propuesta para extraer conocimiento y descubrir relaciones desconocidas entre los datos a partir de grandes volúmenes de información. La minería de datos forma parte del proceso conocido como "*descubrimiento de conocimiento en bases de datos*" (*KDD, Knowledge Discovery in Databases*), y además de la minería de datos, en el *KDD* están involucradas otras tareas, tales como la preparación de la información, la visualización y la exploración de datos y la interpretación de los resultados obtenidos.

De las áreas dedicadas a la producción agrícola en México, los Distritos de Riego favorecen a la aplicación de este tipo de tecnologías, ya que mucha de la información generada en estas áreas es registrada y concentrada por dependencias oficiales de Gobierno.

En esta tesis se presentan los resultados de la aplicación de un ejercicio de minería de datos orientado a la información productiva de un distrito de riego. La información utilizada contempla los principales elementos involucrados en el proceso de producción, como superficies, rendimientos, volumen de agua utilizada e información climática. Se muestra como ciertas técnicas de aprendizaje automático (concretamente, árboles de regresión M5), ofrecen mayores ventajas que las técnicas de análisis de regresión tradicionales, esto al momento de generar modelos de predicción del rendimiento de los cultivos del distrito de riego.

Los modelos de predicción del rendimiento no son suficientes para optimizar el ingreso generado en el distrito. Para proporcionar una solución completa, esta tesis ofrece un algoritmo basado en métodos evolutivos, cuya función de aptitud esta basada en el ingreso calculado por una maximización con el método Simplex de la producción de los cultivos. Esta maximización utiliza la información aportada por los modelos de predicción del rendimiento desarrollados en la etapa de minería de datos. Los resultados indican que la distribución de recursos realizada por el algoritmo mejora los resultados de una estimación realizada por técnicas tradicionales (como el uso de valores promedio y la maximización Simplex).

Finalmente, el algoritmo desarrollado (denominado GenSM5) es implementado en una herramienta de software, la cual puede ser utilizada para la visualización de distintos escenarios en la planeación del proceso de siembra de los cultivos del distrito.

Abstract

Through the time, different models and procedures has been proposed with the purpose of optimizing the elements involved in the productive process of agricultural crops. The obtaining of more revenues with the minimum investment of resources is a problem that has attracted the attention of specialists for a lot of time. The large number of factors involved in the productive process difficulty the construction of efficient and realistic models that reflect the conditions that influence in the yield of the production. Since this point of view, the best action is to analyze the historical data of the crops, with the purpose of inferring future behaviors that allow a correct administration of the involved resources, with the target in obtaining the biggest possible income.

In recent years, a group of techniques contained under the name of "data mining" has arisen as a proposal to extract knowledge and discover unknown relationships among the data, starting from big amounts of information. The data mining is part of the well-known process as "knowledge discovery in databases" (KDD). Besides data mining, other tasks are involved in KDD, such as data preparation, data exploration, data visualization and the interpretation of the obtained results.

Between the areas dedicated to the agricultural production in Mexico, the irrigation districts promotes the application of this type of technologies, since much of the information generated in these areas is registered and concentrated by Government dependences.

This thesis presents the application of data mining oriented to the productive information of an irrigation district. The used information contemplates the main elements involved in the production process, like surfaces, yields, water volume used and climatic information. It demonstrates that certain techniques of machine learning (concretely, M5 regression trees), offer bigger advantages compared to with the traditional analysis regression techniques, this at the moment to generate models of the crops yield prediction in the irrigation district.

The yield prediction models are not enough to optimize the generated income in the district. To provide a complete solution, this thesis offers an algorithm based on evolutionary methods whose aptitude function is based on the maximization of the income by crops production calculated by the Simplex method. This maximization uses the information contributed by the M5 yield prediction models developed in the stage of data mining. It demonstrated that the distribution of resources made by the algorithm improve the results of an estimate made by traditional techniques (like the use of average values and the Simplex maximization).

Finally, the developed algorithm (denominated GenSM5) is implemented in a software tool, which can be used for the visualization of different scenarios in the planning of the landing process of the crops in the district.

CONTENIDO

ANEXOS

Índice de Figuras

Índice de Tablas

No hay un método simple para analizar una serie de datos compleja y desconocida.
(T. Dasu et all, 2003) [54].

T E S I S

USO DE LA MINERÍA DE DATOS PARA LA ELABORACIÓN DE UN MODELO DE PREDICCIÓN PARA LA OPTIMIZACIÓN DE LA PRODUCCIÓN AGRÍCOLA EN LOS DISTRITOS DE RIEGO.

1. Introducción

1.1. Antecedentes

La agricultura es uno de los sistemas de producción esenciales para el ser humano. A nivel mundial, las áreas de tierra disponibles para dicha actividad se dividen en dos debido al origen del recurso hídrico: áreas agrícolas que se abastecen con agua de lluvia (también llamadas de temporal), y áreas agrícolas que son regadas con agua almacenada o extraída por obras hidráulicas (de riego).

Un análisis de la FAO[1] realizado a 93 países en vías de desarrollo informa que se espera que los sistemas de producción agrícolas se incrementen en el periodo de 1998 al 2030 en un 49% para los sistemas de lluvia y en un 81% para los sistemas bajo riego. Por lo tanto, se espera que mucha de la producción alimenticia mundial provenga de tierras bajo riego, cuyas tres cuartas partes se localizan en países en vías de desarrollo [1]. Aunque en las últimas décadas la superficie bajo riego ha ido en aumento (la superficie bajo riego de 1998 casi dobla a la existente en 1962), existen muchas razones por las cuales creer que esta tendencia no continuará en las próximas décadas. Se calcula que, cuando mucho, la superficie bajo riego crecerá en un 34 %. Lo anterior revela la importancia de incrementar la productividad en las áreas bajo riego, ya que el aumento en la superficie no será suficiente.

En México, las áreas agrícolas bajo riego se organizan en forma de **distritos y unidades de riego**. El origen de dicha división obedece básicamente a cuestiones de tamaño y de organización. Los distritos de riego ocupan actualmente 3.385 millones de hectáreas (ha), mientras que las unidades de riego se distribuyen en 2.949 millones de ha. Esto hace un total de 6.334 millones de ha que cuentan con riego [2].

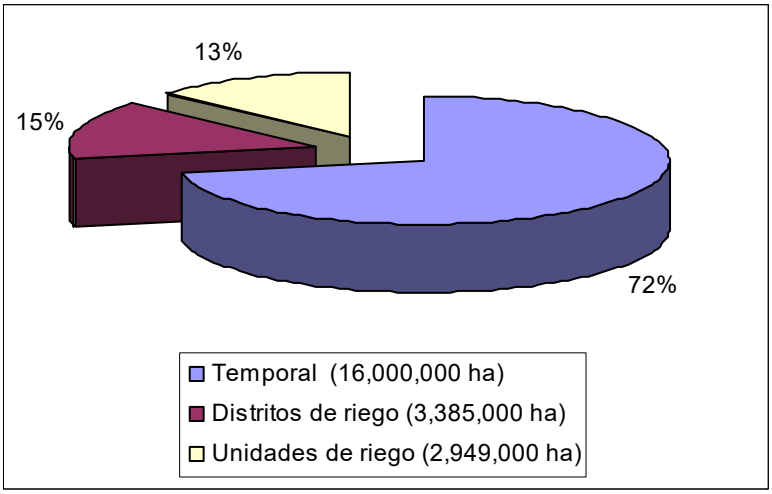


Figura 1. Distribución de superficie para siembra en México

[1] Food and Agriculture Organization of the United Nations, Organización de las Naciones Unidas para la Alimentación y la Agricultura.

De un total de 20 millones de hectáreas que en promedio se cosechan cada año en México, 6 millones son de riego. El valor de la producción en éstas es cercano al 55% del valor total de la cosecha nacional. Es decir, que en las áreas regadas la productividad es 3.7 veces la obtenida en las áreas de temporal [4]. Como ya se mencionó, de las áreas que cuentan con riego, más de la mitad pertenecen a los distritos de riego. Con esta superficie, los distritos contribuyen con casi el 50 % del valor de la producción de las áreas regadas, mientras el restante es proporcionado por las unidades de riego.

Lo que se busca en estas áreas de producción agrícola es obtener cosechas que sean comercialmente productivas junto con rendimientos que sean competitivos en los mercados agrícolas mundiales [5]. El objetivo final perseguido en cada área sembrada es maximizar el rendimiento económico de su producción agrícola. Por esta razón., los responsables de programar la siembra en las áreas agrícolas requieren de herramientas eficientes de pronóstico, que ofrezcan un panorama previo de la producción a obtener bajo ciertos escenarios, y, mejor aún, que permitan identificar aquel escenario que les proporcione el mayor ingreso económico.

Lamentablemente, la elaboración de modelos que consideren fielmente todas las condiciones que influyen en el rendimiento de la producción en los distritos de riego es muy complicada [5]. Una técnica que puede ayudar a suplir esta falta de conocimiento es la Estadística, y en el caso concreto de la producción económica, la Econometría. La Econometría se basa en métodos estadísticos y sirve para estimar el comportamiento de una variable que depende de otras variables [6]. Esta técnica requiere la existencia de cierta "memoria", ya que son los registros históricos los que proporcionan información sobre el comportamiento que sigue una determinada variable bajo ciertas condiciones (por ejemplo, el rendimiento agrícola).

En la temática de obtener conocimiento o inferir comportamientos a partir de información histórica, en los últimos años ha cobrado auge una técnica denominada **minería de datos**. La minería de datos es un conjunto de técnicas y/o algoritmos que permiten extraer conocimiento y descubrir relaciones desconocidas entre los datos a partir de grandes volúmenes de información. La minería de datos forma parte del proceso conocido como "*descubrimiento de conocimiento en bases de datos*" (*KDD, Knowledge Discovery in Databases*) [7]. Además de la minería de datos, en el *KDD* están involucradas otras tareas, tales como la preparación de la información, la visualización y la exploración de datos y la interpretación de los resultados obtenidos. La minería de datos es una tecnología prospectiva que basada en datos del pasado y el presente descubre factores y asociaciones antes desconocidas, llegando incluso a ser predictiva [8]. En la presente tesis, se plantea el uso de este enfoque para la obtención de un modelo de predicción que proporcione información para la optimización de la producción agrícola en los distritos de riego.

1.2. Planteamiento del problema

Existen muchas dificultades para plantear modelos que permitan estimar la cantidad de producción agrícola a obtener en un distrito de riego durante un período de siembra. Más complicado aún, resulta el hecho de optimizar dicha producción de forma tal que se

obtenga el máximo beneficio económico posible. Los productores en un distrito de riego deben decidir periódicamente qué cultivos sembrar y en qué cantidad de superficie hacerlo; la forma en cómo distribuyan sus cultivos influirá en su producción final, y por ende, en el ingreso económico obtenido.

El principal problema que existe para la elaboración de dichos modelos es la inmensa cantidad de factores (condiciones y restricciones) que influyen en la el proceso de producción agrícola en un distrito de riego [5]. Como ejemplos, se pueden mencionar a la disponibilidad de agua, el comportamiento del clima, la disponibilidad de superficie, la predisposición para ciertos cultivos, características topográficas y propiedades particulares del suelo.

Existen varios métodos para investigar las relaciones entre estos factores. Los métodos agronómicos -que incluyen numerosos experimentos con pequeñas parcelas a lo largo de muchos años- son la manera más tradicional y posiblemente mejor para recolectar la información necesaria. Sin embargo, estos métodos son grandes consumidores de tiempo y trabajo intensivo, siendo inviables en un futuro cercano [9]. Otro tipo de métodos [10][11] [12], se enfocan en mejorar cierto factor involucrado en el proceso de producción agrícola, como por ejemplo, las prácticas de manejo de la parcela, la distribución de agua o el uso de fertilizantes, esto con el fin de incrementar la cantidad de producto obtenido al final del proceso. Lamentablemente, estos métodos resultan costosos de implementar tan sólo para un cultivo, por lo que su aplicación a un grupo numeroso de cultivos resulta económicamente inviable.

1.3. Objetivos de la tesis

1.3.1. Objetivo general

Aplicar el proceso de KDD para desarrollar un modelo de predicción que permita optimizar el beneficio económico obtenido por la producción agrícola generada en un distrito de riego.

1.3.2. Objetivos específicos

1. Aplicar el proceso de KDD a la información sobre producción agrícola de distritos de riego obtenida de diversas fuentes (*IMTA, CONAGUA, SEMARNAT* entre otros).
2. Obtener un modelo predictivo como resultado del proceso de descubrimiento del conocimiento.
3. Desarrollar un algoritmo de optimización de la producción agrícola, que utilice como base al modelo predictivo obtenido del proceso de descubrimiento del conocimiento.
4. Implementar el modelo predictivo y el algoritmo de optimización en una herramienta de software que facilite su aplicación y uso.

1.4. Alcances y limitaciones

En la actualidad, existen modelos que se especializan en el proceso de producción de un cultivo en particular, llegando a un nivel de detalle que puede incluir el monitoreo en tiempo real de las etapas de crecimiento del cultivo [12]. Estos modelos son costosos dada la especialización tecnológica requerida, ya que esta no está disponible en la mayoría de los distritos. Además, la mayoría de estos modelos necesitan un estudio especializado para cada tipo de cultivo. Lo anterior implica una dificultad para que dichos modelos sean aplicables a todos los cultivos de un distrito de riego (debido al costo que ello implicaría) y por ende, que puedan ser llevados a la práctica cada año.

El método propuesto por esta tesis no llega a tal grado de especialización. Lo que se buscó, fue crear una herramienta que pueda ser utilizada a nivel gerencial y que sirva de auxilio en la toma de decisiones dentro de la planeación del proceso de producción agrícola en un distrito de riego. Para ello, el método toma como base la información histórica del distrito, por lo que los modelos planteados serán válidos mientras las condiciones en el distrito no cambien significativamente.

Dado que se trabajó con la información exclusiva de un distrito de riego, no se pretende que los modelos desarrollados en esta tesis puedan ser aplicados tales cual a otras áreas agrícolas. Se requiere llevar a cabo una práctica similar de minería de datos como la que se describe en esta tesis.

El algoritmo de optimización desarrollado permite maximizar los ingresos obtenidos al sembrar una serie limitada de cultivos en un determinado distrito de riego. Esta maximización soporta las principales restricciones a las que se ve sujeto el proceso de producción cada año, tales como: el de volumen de agua utilizado, la superficie de tierra disponible, el clima, los costos de producción y el rendimiento. La influencia de otros factores se ve reflejada en la información histórica utilizada. Al final de la aplicación del algoritmo, se obtiene la lista de parejas cultivo-superficie que maximiza el beneficio económico, esto en función de la información suministrada y de ciertos parámetros económicos (como el precio de venta del cultivo y el costo de producción).

1.5. Estructura de la tesis

El capítulo 2 describe el estado del arte de la optimización de la producción agrícola y de la minería de datos. El capítulo 3 indica la metodología empleada en esta tesis. El capítulo 4 muestra el uso de la minería de datos para la obtención de modelos de predicción del rendimiento de los cultivos agrícolas de un determinado distrito de riego. El capítulo 5 se enfoca en el uso de los modelos de predicción generados para proponer un algoritmo de optimización, el cual tiene el objetivo de maximizar el ingreso económico de la producción agrícola de un distrito de riego. El capítulo 6 describe la aplicación desarrollada que implementa el algoritmo de optimización propuesto. Finalmente, el capítulo 7 expone las conclusiones obtenidas del desarrollo de la tesis.
Se proporcionan además, una serie de anexos que sirven de apoyo a la documentación central de esta tesis.

2. Estado del arte

2.1. La maximización de la producción en los distritos de riego

El rendimiento económico de la producción agrícola está dado como la utilidad económica obtenida de la venta de los cultivos que se hayan sembrado durante un periodo agrícola, dentro del cual se considera también el período de gestación de los cultivos involucrados, que abarca a su vez los tiempos de siembra, cuidado y cosecha de los cultivos.

2.1.1. La función para el cálculo del beneficio económico neto

En Palacios et al. (1986) [5], se muestra que el beneficio económico neto obtenido en un período agrícola cualquiera está expresado por:

$$BN = (IB_1 - C_1)S_1 + (IB_2 - C_2)S_2 + ... + (IB_n - C_n)S_n \quad \textbf{(2.1)}$$

Donde:
BN.- Beneficio económico neto expresado en unidad monetaria ($ ó M$).
IB_i.- Ingreso bruto por hectárea (ha*)* del cultivo *i,* obtenido al multiplicar el rendimiento por unidad de superficie (ton/ha) del cultivo *i* por el precio unitario del producto cosechado ($/ha).
C_i.- Costo de producción por unidad de superficie del cultivo *i* ($/ha).
S_i.- Superficie (área) cosechada del cultivo *i* (ha).

Las siguientes son algunas consideraciones para la expresión 2.1:

- $IB_i - C_i$ nos dará la utilidad neta por hectárea del cultivo *i* (se están restando los costos de los ingresos brutos).
- $\sum_{1}^{n} S_i$, es la superficie total en posibilidades de sembrar que abarca la superficie de los *n* cultivos presentes en el distrito.

En Palacios et al. (1986) [5] también se indica que la maximización de la expresión 2.1 está sujeta a varias restricciones. Por ejemplo, la disponibilidad mensual de agua para el distrito, expresada por:

$$\begin{aligned} S_1 a_{1,1} + S_2 a_{2\_1} + ... + S_n a_{n,1} &\le VA_1 \\ S_1 a_{1\_2} + S_2 a_{2\_2} + ... + S_n a_{n\_2} &\le VA_2 \\ &\vdots \\ S_1 a_{1\_12} + S_2 a_{2\_12} + ... + S_n a_{n\_12} &\le VA_{12} \end{aligned} \quad \textbf{(2.2)}$$

Donde:
S_i.- Área total regada y cosechada del cultivo *i* en ha (se asume que la superficie total regada del cultivo *i* también será cosechada).

a_{i_m}.- Requerimiento de riego del cultivo *i* en el mes número *m*.
VAj.- Volumen de agua disponible en el distrito en el mes *j* (en millares de m^3).

Las expresiones (2.2) indican el requerimiento de riego total de los cultivos para un mes dado, que debe ser menor o igual al volumen disponible para dicho mes. Lógicamente, también es necesario considerar otra restricción, tomando en cuenta la disponibilidad anual de agua para los cultivos:

$$S_1 A_1 + S_2 A_2 + ... + S_n A_n \leq VT \qquad \textbf{(2.3)}$$

Donde:
$A_i = \sum_{m=1}^{12} a_{i\_m}$, la suma de los requerimientos mensuales de agua (definidos en la expresión 2.2) del cultivo *i*.
VT.- Disponibilidad total anual de agua del distrito en unidades de volumen.

De manera similar podrían agregarse todas las restricciones involucradas en el proceso de producción agrícola, de tal forma que la maximización de la expresión 2.1 podría plantearse como un problema de programación lineal [5]. Esto podría determinar fácilmente el área de cada cultivo que proporcione el beneficio económico máximo, además de la información relacionada con las restricciones modeladas (por ejemplo, el volumen utilizado).

Lamentablemente, existen un enorme número de restricciones naturales, sociales y económicas involucradas en el proceso de producción de un distrito de riego que impiden que la maximización de la expresión 2.1 pueda ser planteada como un problema puntual y que éste pueda ser resuelto de forma lineal [5]. Otro problema en la expresión 2.1, es que se asume que ya se conocen de antemano algunos datos importantes, tales como los rendimientos de los cultivos (indispensables para calcular el ingreso bruto por cultivo), el volumen requerido en forma mensual y la lista absoluta de cultivos que se van a sembrar en el distrito.

Existe una clara dificultad en el hecho de querer utilizar soluciones lineales en problemas que siguen un comportamiento no lineal. Aunado a esto, se tiene que la producción agrícola es influenciada fuertemente por parámetros locales (como el tipo de superficie disponible para siembra), por lo que los modelos o soluciones globales generalmente no tienen un buen desempeño. Como ejemplo, en Sudduth et al (1996) [13] se demuestra que los métodos lineales generalmente fallan para producir buenas aproximaciones funcionales en la estimación del rendimiento agrícola.

A continuación se listarán algunas técnicas que han sido utilizadas en la estimación de la producción agrícola.

2.1.2. Análisis de regresión para la producción agrícola

El análisis de regresión es probablemente la técnica más utilizada para construir modelos que involucran variables de salida continuas [14].

Fortalezas:

- Fácil de interpretar.
- Ampliamente utilizado, bien documentado.
- Puede construir modelos mixtos de variables continuas y categóricas.
- Permitido para un amplio rango de diagnósticos estadísticos y pruebas de significancia.

Debilidades:

- La regresión no maneja correctamente los valores faltantes en casos variable por variable. En tal caso, se elimina todo el registro ó se reemplaza el valor faltante por un valor predefinido (por ej. la media).
- No es robusto para valores atípicos en los datos.
- Las variables categóricas tienen que ser representadas por variables "dummy".
- El modelo asume incrementos/decrementos fijos en los valores contables.
- Puede no capturar, o al menos no rápidamente, interacciones entre los datos.

Existen muchos trabajos donde han aplicado la regresión lineal en problemas relacionados con la producción agrícola por ejemplo, [15][16][17]. La mayoría de ellos, se enfocan en predecir el rendimiento (ton/ha) o la producción a obtener (ton) en determinados cultivos, o bien, la relación de factores relacionados con éstos (por ej. el clima).

2.1.3. La función de producción agrícola

Otros estudios [18][19] se enfocan en realizar un análisis particular del comportamiento de la producción para cada cultivo con posibilidad de ser sembrado en el distrito de riego. Así, 2.1 se puede descomponer en una suma de funciones de producción, cada una dedicada a un cultivo en particular:

$$\textbf{Producción del cultivo } x = f \textbf{ (superficie, rendimiento, agua, precio)} \qquad (2.4)$$

La función de producción es una relación (función matemática) que especifica la cantidad de producto que puede obtenerse con una cantidad dada o conocida de factores [20]. El concepto usual de una función de producción trata al producto como una variable dependiente de una larga lista de factores de producción. El productor en cualquier punto conoce algunos de los posibles factores que pueden contribuir a la producción de un determinado producto. Por consiguiente, en cierto sentido, siempre hay dos funciones de producción [18]:

- La que utiliza factores conocidos por el hombre y
- La que aun está por ser descubierta en su totalidad

Generalmente, se ha supuesto que una función de producción muestra rendimientos constantes de escala cuando los factores de producción que se utilizan son controlables. Lo que sucede en realidad, es que después que se ha alcanzado un cierto nivel, cada unidad de insumo variable que se agregue añadirá al producto total una cantidad menor que la unidad anterior. Este fenómeno lo conocemos como la “Ley de los rendimientos decrecientes"

[18][5]. En [18] puede verse un ejemplo del uso de funciones de producción del tipo Cobb-Douglas[19][20] para el cultivo del algodón en un distrito de desarrollo rural (que tienen características distintas a un distrito de riego) ubicado en Chihuahua, México.

2.1.4. Técnicas especializadas

Otros trabajos [11] [12] se enfocan en mejorar factores del proceso de producción, con el fin de incrementar la cantidad de producto al final del proceso (las ganancias). Se listan ejemplos de estas técnicas, mencionando que su uso está fuera del alcance de esta tesis, debido a que es inviable económicamente utilizarlas como mecanismo de optimización a nivel distrito (ver sección de alcances), y son mencionadas únicamente con el fin de dejar constancia de su existencia.

Calendarización del riego en tiempo real. El objetivo de la calendarización del riego es aplicar el agua en la cantidad y frecuencia necesaria para reducir la posibilidad de bajos rendimientos por estrés hídrico. Utilizando una base de datos con información sobre los parámetros de los cultivos, suelo, padrón de usuarios, clima, red de distribución, seguimiento de riego de los cultivos y su manejo, se realiza un balance diario del consumo de agua de cada cultivo establecido, desde su fecha de siembra hasta la cosecha y pronostica el momento oportuno y la cantidad del riego [12].

Uso de sistemas de información geográfica (SIG). Los SIG y los sistemas de percepción remota se están aplicando con mucho éxito en los distritos de riego como herramientas para controlar y monitorear la producción. En [11] podemos ver una aplicación de los SIG para la estimación de la superficie sembrada en un distrito de riego.

2.1.5. La producción agrícola y los registros históricos

En la función para el cálculo del beneficio económico neto, como en la función de producción, se presentan dificultades para la estimación y maximización de la producción agrícola debido al desconocimiento del comportamiento que seguirán los factores involucrados.

Sin embargo, utilizando técnicas estadísticas, es posible acercarse a los probables valores de aquellos factores relacionados con (2.1) y con (2.4) que son impredecibles. Éstas técnicas permitirán conocer el comportamiento de algunos factores dado los registros históricos que se tienen de ellos. Por ejemplo, es posible conocer como se comporta el rendimiento del maíz en función de la cantidad de agua y la superficie utilizada. Otra posibilidad, sería la de obtener la lista de los cultivos con mayor probabilidad de ser sembrados en el distrito de riego 005 (por ejemplo), y cuáles serían sus probables requerimientos hídricos para el año 2007.

A este respecto, se han realizado varios trabajos en distritos de riego que utilizan datos históricos de cultivos con el fin de obtener información que permitan mejorar la planeación de la producción. En [21] y [22] se observa como se utilizan indicadores históricos de producción con el fin de emitir recomendaciones para mejorar el rendimiento general de los módulos de distritos de riego, incluyendo indicaciones para incrementar el

beneficio económico. En ambos trabajos fue necesaria la intervención de un experto que aplicara su conocimiento más allá de la estadística con el fin de interpretar la información histórica y establecer u obtener relaciones que no son fáciles de detectar a "simple vista".

En el intento de automatizar el proceso de "aprender de los datos", surge el concepto de aprendizaje automático (*machine learning*), que en [23] se define como "*una técnica que tiene el objetivo de desarrollar métodos computacionales que implementarían varias formas de aprendizaje, en particular, mecanismos capaces de inducir conocimiento a partir de datos*". También se menciona en dicha referencia que la idea de introducir conocimiento por medio de ejemplos parece atractiva al sentido común y que dicha forma de inducción de conocimiento es deseable en problemas que carecen de solución algorítmica eficiente, son vagamente definidos o informalmente especificados (como el problema de producción agrícola).

En los últimos años ha cobrado auge una serie de técnicas que permiten extraer conocimiento y descubrir relaciones desconocidas entre los datos a partir de grandes volúmenes de información que va más allá del uso de la estadística. Este grupo de técnicas, agrupadas bajo el nombre de minería de datos (*data mining*), forman parte del proceso conocido como "descubrimiento de conocimiento en bases de datos".

2.2. La minería de datos

2.2.1. Definición

La minería de datos puede ser vista como resultado de la evolución natural de las tecnologías de la información. El almacenamiento de datos ha evolucionado sistemáticamente desde primitivos sistemas de archivos hasta sofisticados y poderosos sistemas administradores de bases de datos (ver figura 2). La investigación y el desarrollo en bases de datos pasaron de los primeros sistemas jerárquicos y de redes, al desarrollo de sistemas de bases de datos relacionales con herramientas de modelado, indexado y técnicas de organización de datos [25]. Como respuesta a los diferentes paradigmas de programación, han aparecido nuevas tecnologías, como bases de datos orientadas a objetos, bases de datos espacio-temporales, bases de datos geográficas, bases de datos científicas, bases de conocimiento y bases de información de oficina [25]. A fechas recientes, los sistemas de bases de datos heterogéneas y sistemas de información global basados en Internet han adquirido relevancia, jugando un papel vital dentro de la industria de la información.

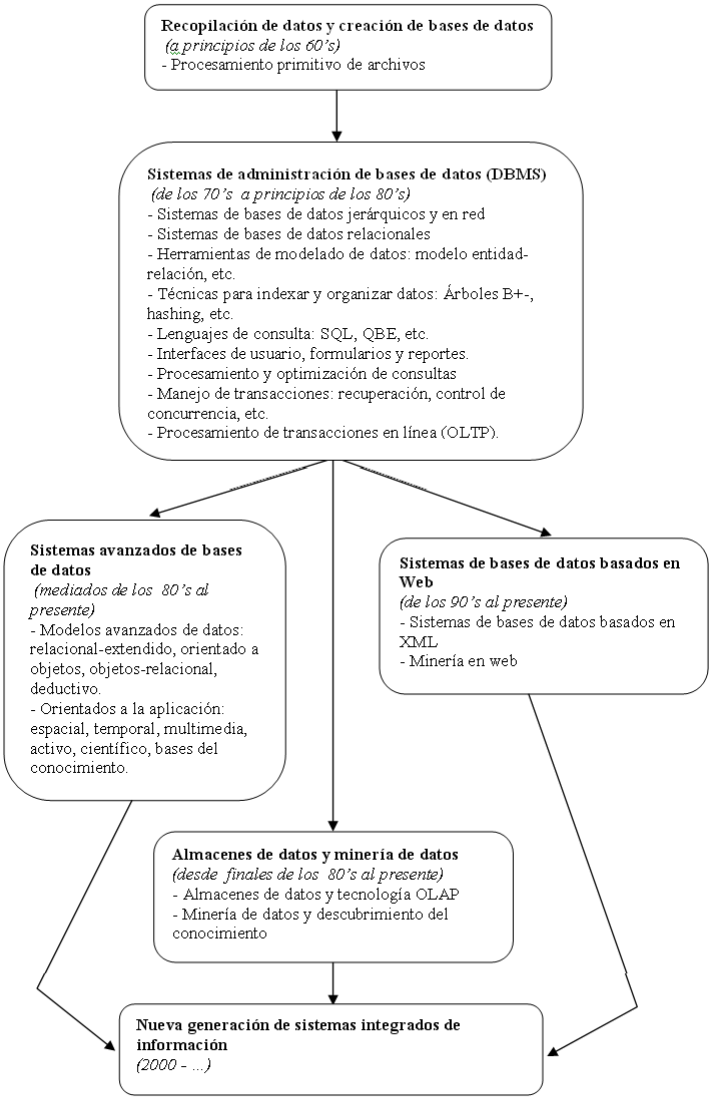


Figura 2. Evolución de los sistemas de bases de datos, por Han et al (2001) [25]

Como parte de esta evolución, las habilidades del ser humano para automatizar los procesos y las capacidades de almacenamiento de la información han ido en aumento. Como consecuencia, la cantidad de datos que se almacenan en las bases de datos también han crecido, haciendo cada vez más complejo su análisis y procesamiento. Esta circunstancia, junto con el hecho de que no siempre se cuenta con las herramientas de análisis de datos requeridas, describen una situación de "*riqueza de datos pero pobreza de información*" [25].

El proceso de descubrir conocimiento en una base de datos ha sido definido por Fayyad et al. (1996) como "*la identificación no trivial de patrones válidos, nuevos, comprensibles y potencialmente útiles en los datos*" [26].

Muchas personas tratan a la minería de datos como sinónimo del proceso de "descubrimiento de conocimiento en bases de datos" (*Knowledge Discovery in Databases,*

KDD). Sin embargo, otros ven a la minería de datos como una etapa esencial del proceso (ver figura 3).

La figura 3 muestra el proceso de etapas interactivo e iterativo de *KDD* [25]:

1. **Limpieza e integración de datos.** Trata de llenar los datos faltantes, eliminar el "ruido" existente y corregir inconsistencias en los datos. Para la integración, se combinan los datos que provienen de diferentes fuentes en una forma de almacenamiento coherente (un almacén integrado).

2. **Selección y transformación.** La selección (reducción) consiste en aplicar una técnica para obtener una representación reducida de los datos originales, que es mucho más pequeña en volumen, pero que mantiene la misma integridad. Los datos necesitan ser transformados para ser sujetos a la minería de datos. Esta transformación puede involucrar a una o más de las siguientes operaciones: pulido, agregación, generalización, normalización y/o construcción de atributos.

3. **Minería de datos.** Los llamados "métodos inteligentes" son aplicados para ajustar modelos o determinar patrones a partir de los datos que permitan hacer predicciones válidas.

4. **Interpretación/evaluación y presentación.** Se identifican los patrones que aportan conocimiento nuevo. Se aplican técnicas para la visualización y representación del conocimiento para mostrar el conocimiento obtenido, "minado", al usuario.

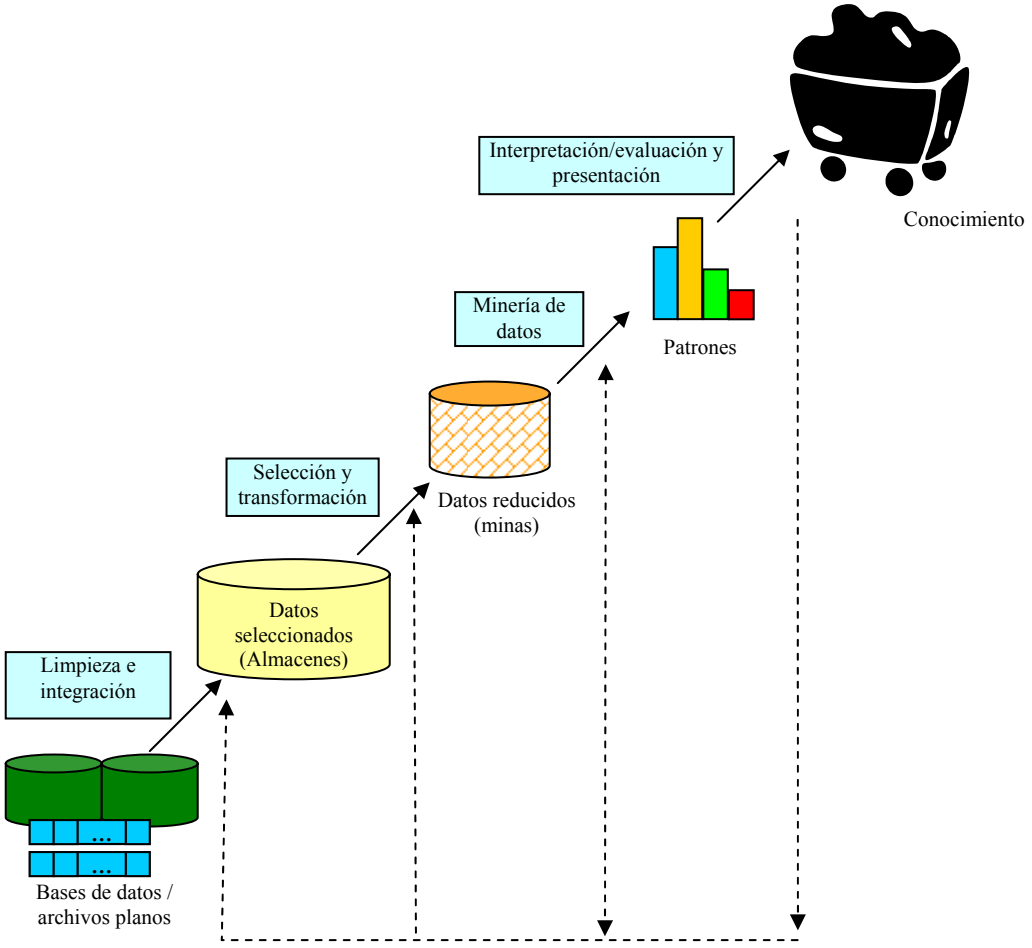


Figura 3. El proceso del descubrimiento del conocimiento

La minería de datos realiza una "*tarea*" específica con los datos que han pasado por las etapas de limpieza y selección. La clasificación de tareas que pueden realizarse a través de la minería de datos según Hand et al (2001) [27] se muestra en la tabla 1.

Tarea	Propósito
Análisis Exploratorio de Datos (*EDA, Exploratory Data Analysis*)	Inspeccionar de forma simple los datos sin tener una idea clara de lo que se está buscando. Las técnicas EDA son interactivas y visuales, existiendo muchos métodos efectivos de despliegue gráfico para mostrar series de datos de pocas dimensiones.
Modelado descriptivo (*Descriptive modeling*)	Describir todos los datos (o el proceso para generar los datos). Algunos ejemplos son modelos para describir la distribución total de datos (estimación de densidad), creación de particiones de un espacio de *p*-dimensiones en grupos (segmentación y análisis de grupos) y modelos para describir la relación entre variables (modelado de dependencias).
Modelado predictivo : clasificación y regresión (*predictive modeling: classification and regression*)	Construir un modelo que permita predecir el valor de una variable en términos de los valores conocidos de otras variables. En la clasificación, la variable a predecir es categórica mientras que en la regresión es cuantitativa. El término "predicción" es utilizada en sentido general y no está incluida la noción de continuidad temporal.
Descubrimiento de patrones y reglas (*Discovering patterns and rules*)	Encontrar relaciones, eventos o secuencias que permiten inferir un comportamiento predecible, aún cuando están presentes medidas de incertidumbre.
Recuperación por contenido (*Retrieval by content*)	Encontrar patrones en los datos que guardan parecido con un patrón de interés predeterminado. Esta tarea es más comúnmente usada para texto e imágenes. Un

	ejemplo de esto es la recuperación de documentos en la Web.

Tabla 1. Tipos de tareas realizadas en la minería de datos, Hand et al (2001) [27].

Fayyad et al (1996) [7] y Hernández et al (2004) [28] indican que en la práctica, las tareas principales en el proceso de minería de datos son el modelado predictivo y el modelado descriptivo.

2.2.2. Algoritmos de minería de datos

Un algoritmo de minería de datos, según Hand et al (2001), se define como un procedimiento bien definido que toma datos como entrada y produce una salida en la forma de modelos o patrones. Un algoritmo de minería está integrado por cinco componentes:

1. **La tarea (*task*).** Meta u propósito del algoritmo. Ésta puede ser de visualización, de descripción, de clasificación, de agrupamiento, de regresión, etc.
2. **La estructura (*structure*).** Forma funcional del modelo o patrón que se ajustará a nuestros datos. Ésta puede ser, por ejemplo, un árbol de decisión, una red neuronal ó en forma de reglas de asociación.
3. **La función de referencia (*score function*)**. Permite evaluar la calidad del modelo o patrón basándose en los datos observados. Ejemplos de funciones de referencia son el error cuadrático (*squared error*), soporte y ajuste (*support/accuracy*), backpropagation, etc.
4. **El método de búsqueda o de optimización (*search or optimization method*)**. Método para buscar entre parámetros y estructuras. Por ejemplo, los procedimientos computacionales y algoritmos que requieren encontrar el valor máximo o mínimo de la función de referencia para ciertos modelos y patrones. Ejemplos: búsqueda avara (*greedy-search*), búsqueda a lo ancho con poda (*breadth-first with pruning*), búsqueda en profundidad (*deep first search*), etc.
5. **La técnica de manejo de datos**. Se requiere para almacenar, indexar y recuperar los datos. Muchos algoritmos estadísticos y de máquinas de aprendizaje no especifican una técnica para manejar los datos, asumiendo que las series de datos son pequeñas y que residen en la memoria principal de manera que el acceso aleatorio a cualquier punto de datos es libre de costo computacional. Pero cuando hay series de datos masivas que exceden la capacidad de la memoria principal y residen en medios secundarios (como el discos duros o cintas) el acceso es más lento, y la localización física de los datos así como la manera en como éstos son accedidos puede ser importante de manera crítica en términos de eficiencia algorítmica.

Fayyad et al (1996) [7] y Cheeseman (1990) [29] proporcionan una "vista reducida" que centra su atención en tres componentes: la representación del modelo, la evaluación del modelo y la búsqueda:

a) La representación del modelo es el lenguaje utilizado para describir los patrones a descubrir. Si la representación es demasiado limitada, ningún tiempo de aprendizaje aplicado o número de ejemplos podrán producir un modelo correcto de los datos. Es importante estar conscientes de las implicaciones de representación una vez seleccionado un determinado algoritmo.

b) El criterio para la evaluación del modelo son parámetros cuantitativos o funciones de ajuste que permiten saber que tan bien un patrón o modelo logra el propósito del proceso de *KDD***.**
c) El método de búsqueda consiste a su vez de dos componentes: la búsqueda del parámetro y la búsqueda del modelo. Una vez seleccionad o el modelo y el criterio de evaluación, el problema de minería de datos es reducido a una tarea de optimización: encontrar los parámetros y el modelo de la familia seleccionada que optimicen el criterio de evaluación.

La figura 4 resume las técnicas y los algoritmos empleados con más frecuencia en la minería de datos. Se agrupan por la clasificación de tareas proporcionada en Hand et al (2001) [27], aunque la lista final de algoritmos proviene de diversas fuentes.

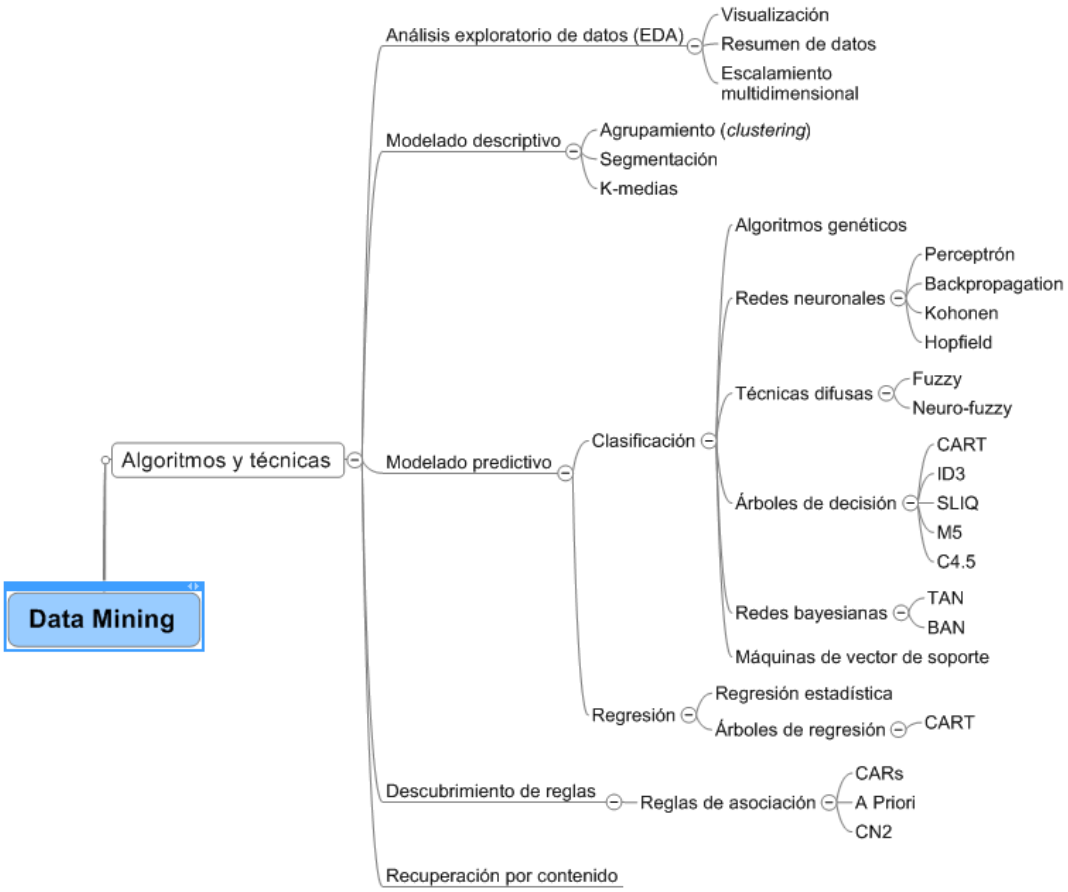


Figura 4. Clasificación de los algoritmos de minería de datos.

La definición de minería de datos estaría incompleta si no existieran metodologías que guiaran al investigador de manera sistemática y organizada a través del proceso.

2.2.3. Metodologías de la minería de datos

Ante la necesidad existente en el mercado de una aproximación sistemática para la realización de los proyectos de explotación de datos, diversas empresas y consultorías han especificado un proceso de modelado diseñado para guiar al usuario a través de una

sucesión de pasos que le dirijan a obtener buenos resultados [30]. SAS (una de las compañías líderes en software de análisis de negocios [32]) propone la utilización de la metodología SEMMA (Sample, Explore, Modify, Model, Assess). En 1999 un importante consorcio de empresas europeas, NCR (Dinamarca), AG (Alemania), SPSS (Inglaterra) y OHRA (Holanda), unieron sus recursos para el desarrollo de la metodología de libre distribución CRISP-DM (*Cross-Industry Standard Process for Data Mining*). Esta metodología, junto con la metodología SEMMA, son las más utilizadas por los analistas en los proyectos de minería de datos [30] .

2.2.3.1. CRISP-DM

La metodología CRISP-DM [33] consta de cuatro niveles de abstracción, organizados de forma jerárquica en tareas que van desde el nivel más general hasta los casos más específicos, figura 5.

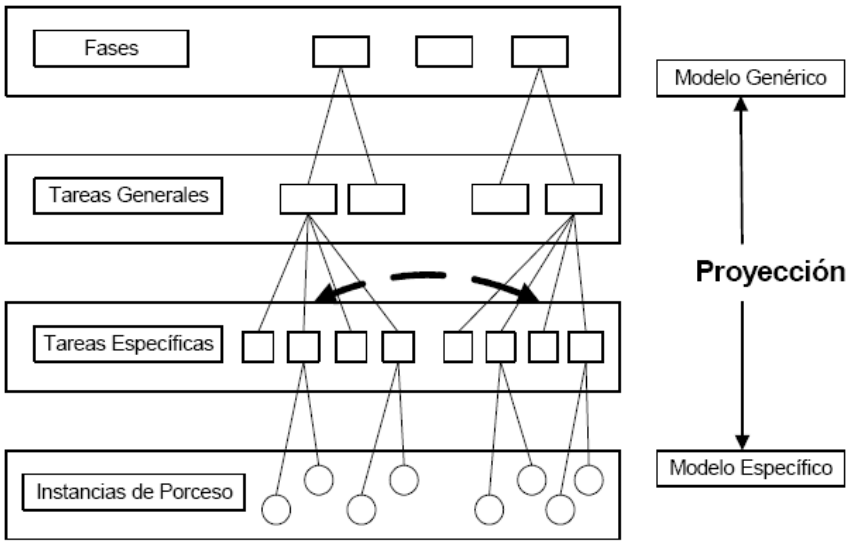


Figura 5. Los cuatro niveles de abstracción de CRISP-DM, Fernández (2006) [30]

De forma general, el proceso está organizado en seis fases y cada fase se divide en varias tareas generales. Las tareas generales se reducen a tareas específicas. Así, si en el segundo nivel se tiene la tarea general "limpieza de datos", en el tercer nivel se dicen las tareas que tienen que desarrollarse para la "limpieza de datos" son: "limpieza de datos numéricos", o "limpieza de datos categóricos". El cuarto nivel, recoge el conjunto de acciones, decisiones y resultados sobre el proyecto de minería de datos específico.

Las seis fases que integran el proceso de modelado con CRISP-DM se muestran en la figura 6. Las flechas indican relaciones más habituales entre las fases, aunque se pueden establecer relaciones entre cualquier fase. El círculo exterior simboliza la naturaleza cíclica del proceso de modelado [34].

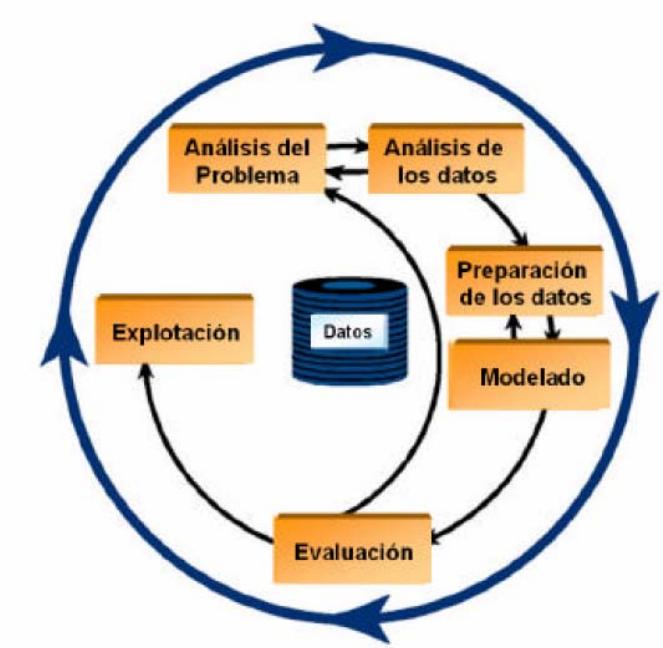


Figura 6. Fases de la metodología CRISP-DM [33].

CRISP-DM proporciona dos documentos como una herramienta de ayuda en el desarrollo del proyecto de minería de datos: el Modelo de referencia y la Guía del usuario. El modelo de referencia describe de forma general las fases, tareas generales y salidas de un proyecto de minería de datos. La guía del usuario proporciona información más detallada sobre la aplicación práctica del modelo de referencia a proyectos de minería de datos específicos, proporcionando consejos y listas de comprobación sobre las tareas correspondientes a cada fase. Además, la metodología define un conjunto de elementos entregables como salidas. La tabla 2 muestra el contenido de cada uno de estos documentos.

The CRISP-DM reference model	The CRISP-DM user guide	The CRISP- DM outputs
1 Business understanding	**1 Business understanding**	**1 Business understanding**
1.1 Determine business objectives	1.1 Determine business objectives	**2 Data understanding**
1.2 Assess situation	1.2 Assess situation	**3 Data preparation**
1.3 Determine data mining goals	1.3 Determine data mining goals	**4 Modeling**
1.4 Produce project plan	1.4 Produce project plan	**5 Evaluation**
2 Data understanding	**2 Data understanding**	**6 Deployment**
2.1 Collect initial data	2.1 Collect initial data	**7 Summary Of dependencies**
2.2 Describe data	2.2 Describe data	**8 Project plan template**
2.3 Explore data	2.3 Explore data	
2.4 Verify data quality	2.4 Verify data quality	
3 Data preparation	**3 Data preparation**	
3.1 Select data	3.1 Select data	
3.2 Clean data	3.2 Clean data	
3.3 Construct data	3.3 Construct data	
3.4 Integrate data	3.4 Integrate data	
3.5 Format data	3.5 Format data	
4 Modeling	**4 Modeling**	
4.1 Select modeling technique	4.1 Select modeling technique	
4.2 Generate test design	4.2 Generate test design	
4.3 Build model	4.3 Build model	
4.4 Assess model	4.4 Assess model	
5 Evaluation	**5 Evaluation**	
5.1 Evaluate results	5.1 Evaluate results	
5.2 Review process	5.2 Review process	
5.3 Determine next steps	5.3 Determine next steps	

6 Deployment 6.1 Plan deployment 6.2 Plan monitoring and maintenance 6.3 Produce final report 6.4 Review project	**6 Deployment** 6.1 Plan deployment 6.2 Plan monitoring and maintenance 6.3 Produce final report 6.4 Review project	

Tabla 2. Contenido de la guía CRISP-DM versión 1.0 [33]

2.2.3.2. SEMMA

La metodología desarrollada por *SAS Institute* define a la minería de datos como el proceso de selección, exploración y modelado de grandes cantidades de datos para descubrir patrones de negocio desconocidos [30]. Las fases de SEMMA se ilustran en la figura 7.

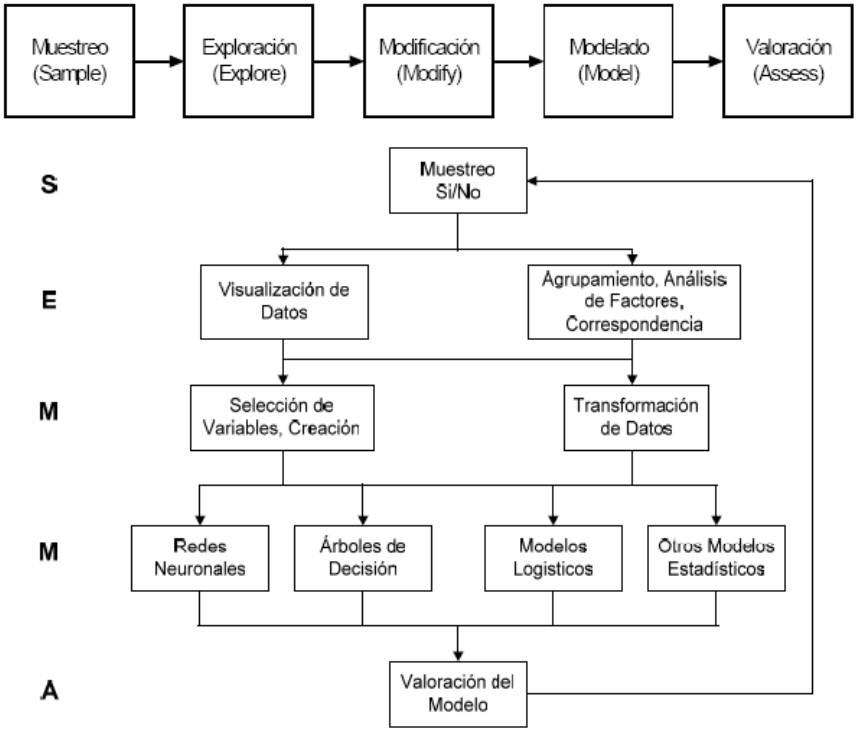


Figura 7. Las fases generales de SEMMA , por Fernández (2006) [30]

El proceso inicia con la extracción de la población muestral a analizar. El objetivo de esta fase es seleccionar una muestra representativa del problema en estudio.

A continuación, se explora la información disponible para simplificar el problema y optimizar la eficiencia del modelo. Para lograr este objetivo, se propone utilizar herramientas de visualización o técnicas estadísticas que ayuden a poner de manifiesto relaciones entre variables.

La tercera fase consiste en manipular los datos, de forma que se definan y tengan el formato adecuado para introducidos al modelo. El objetivo es establecer una relación entre las

variables explicativas y las variables objeto del estudio, que posibiliten inferir el valor de las mismas con un nivel de confianza determinado.

2.2.3.3. Comparando metodologías

Ambas metodologías tienen su grupo de seguidores y han sido probadas con éxito. Generalmente, se encontraron más referencias a CRISP-DM que a SEMMA, quizá por el carácter abierto y público de la primera. SEMMA está más orientada a aplicarse en conjunto con las herramientas SAS [32], lo que limita un poco su uso.

De manera interna, SEMMA y CRISP-DM comparten una estructura similar. Ambas dividen en fases el proyecto de minería, marcando los puntos de interconexión entre las fases. Al final, las metodologías se reducen a un proceso iterativo e interactivo dedicado a extraer modelos desde una o más fuentes de datos. La figura 8 muestra las fases que se relacionan entre ambas metodologías.

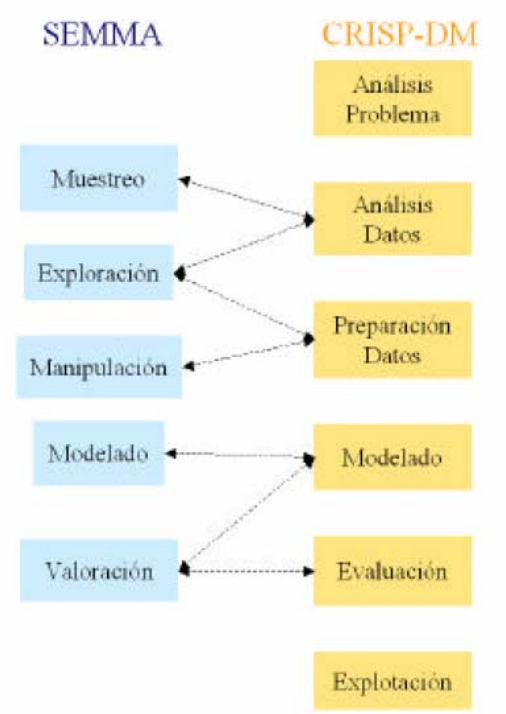


Figura 8. Interrelaciones entre las fases de CRISP-DM y SEMMA [34].

En Rodríguez et al (2003) [34], se señala que "*la metodología SEMMA se centra más en las características técnicas del desarrollo del proceso, mientras que la metodología CRISP-DM mantiene una perspectiva más amplia respecto a los objetivos empresariales del proyecto. Desde esa perspectiva, se puede considerar que la metodología CRISP-DM está más cercana al concepto real del proyecto, pudiendo ser integrada con una Metodología de Gestión de Proyectos específica que complementaría las tareas administrativas y técnicas*".

Se quisiera aclarar, sin embargo, que no es forzoso seguir alguna de estas dos metodologías para llevar a cabo un proyecto de minería. Ciertas encuestas [34] indican el uso de CRISP-DM ha disminuido mientras que el número de usuarios que han optado por utilizar una metodología propia ha aumentado. Básicamente, el trabajo extra derivado del uso de una metodología propia radica en el hecho precisamente de desarrollar dicha metodología.

2.3. La minería de datos y la producción agrícola

Existen dificultades que establece Acosta (2004) [36] para utilizar la minería de datos para el modelado predictivo con procesos relacionados con series de tiempo:

"*La minería de datos es una herramienta exploratoria y no explicativa. Explora los datos para sugerir hipótesis. Es incorrecto aceptar dichas hipótesis como explicaciones o relaciones causa-efecto. Es necesario coleccionar nuevos datos y validar las hipótesis generadas ante los nuevos datos, para después descartar aquellas que no son confirmadas por los nuevos datos. A esto se añaden algunos problemas. El primero, es el de minería de datos con relaciones en el tiempo. Es muy posible que se deseen hacer inferencias y análisis de datos sobre un periodo determinado, pero que durante dicho periodo no se haya registrado el mismo número de variables, o que éstas no tengan la misma precisión, o carezcan de la misma interpretación. En ciertos casos puede que se haya hecho un ejercicio de minería de datos en el pasado y que los datos se hayan descartado o destruido, pero que se desee hacer una comparación con datos más recientes. Nótese que un ejercicio de minería de datos puede traer a la luz relevancia de variables y factores, pero que sea imposible recopilar estas variables y completar adecuadamente conjuntos de datos del pasado*".

Pese a esta dificultad, se han hecho investigaciones importantes que relacionan la minería de datos con la producción agrícola. Algunos de estos trabajos se exponen a continuación.

Witten et al (1993) [37] aplica diversas técnicas de aprendizaje automático (*machine learning*) con problemas relacionados con la producción agrícola. Específicamente, utiliza el aprendizaje basado en similaridad y sistemas para generar reglas para diagnosticar problemas en los cultivos (como la soya), mostrando su potencial aplicación en la agricultura de Nueva Zelanda.

Drummond et al (2000) [38] y Liu et al (1999) [39] desarrollan modelos de redes neuronales (una técnica de minería de datos) y comparan los resultados con los obtenidos con la aplicación de técnicas de regresión lineal. En ambos casos, las redes neuronales lograron obtener un desempeño ligeramente superior contra los modelos de regresión que se compararon.

Thomasson et al (2001) [40] describe un sistema experto que aplica un algoritmo llamado DSS4Ag (*Decision Support System for Agriculture*) que utiliza datos históricos y la minería de datos para la toma de decisiones *in situ* con el fin optimizar de forma económica la aplicación de fertilizantes a campos de cultivos.

Canteri et al (2002) [41] realizó un estudio para determinar si las rutinas de minería de datos son capaces de determinar el comportamiento del rendimiento de un cultivo como una función de las propiedades físico-químicas del suelo. Una base de datos fue analizada utilizando el algoritmo de árboles de decisión. Como resultado, se generaron reglas que describen el rendimiento relacionado y las propiedades físico-químicas del suelo. El análisis por parte de expertos humanos determinó que es posible usar el modelo para determinar el comportamiento del rendimiento de un cultivo como una función de las propiedades físico-químicas del suelo.

Crisler et al (2002) [42] utiliza la minería de datos para automatizar el proceso de extraer información de archivos de datos de rendimientos de cultivos. Este artículo provee ejemplos de técnicas de automatización diseñadas para reducir el tiempo requerido para extraer información útil para el análisis de archivos típicos de datos de rendimientos de cultivos.

Abdullah et al (2004) [43] aplica la minería de datos y herramientas OLAP para analizar datos de cultivos con el fin de encontrar correlaciones entre los pesticidas y el rendimiento de los cultivos.

Lobell et al (2005) [44] utiliza técnicas de minería de datos (análisis exploratorio y árboles de regresión) para “entender” las variaciones en el rendimiento del trigo en un panorama bajo riego.

3. Metodología de la tesis

El desarrollo de la tesis se ha dividido en tres líneas de trabajo:

1. **Minería de datos para modelado predictivo en información de producción de distritos de riego**

   En esta tesis se plantea el uso de la minería de datos como herramienta para generar modelos de predicción de la producción de cultivos agrícolas en un distrito de riego.

   Se considera a la minería de datos como el trabajo más importante de la tesis, ya que es la actividad que permite generar los modelos con los que finalmente se realizará la optimización del ingreso del distrito. La práctica de minería de datos es conducida por la metodología CRISP-DM [33], la cual se describe en el estado del arte, sección de metodologías.

2. **Desarrollo del algoritmo de optimización de la producción**

   El algoritmo de optimización se desarrolló tomando en cuenta principalmente la forma de los modelos generados en la etapa de minería de datos. Una vez disponibles los modelos, se buscó la manera de integrarlos en un contexto de optimización. Lo anterior reveló la necesidad de utilizad una heurística de búsqueda, lo que llevó a proponer un algoritmo genético como mecanismo para encontrar una solución óptima. El algoritmo genético fue seleccionado porque se adaptó de manera muy natural a la forma que manifestaron los modelos dentro del contexto de optimización.

3. **Implementación del algoritmo de optimización en una herramienta de Software.**

   La aplicación que implementa el algoritmo de optimización fue desarrollada utilizando programación orientada a objetos. Para ello, se utilizó un lenguaje de programación que integra dicho paradigma en su estructura.

   Para guiar el desarrollo de la aplicación se utilizó la metodología RUP (*Rational Unified Process*), el cual es un proceso de desarrollo de software que junto con el Lenguaje Unificado de Modelado (*UML*), constituye la metodología estándar más utilizada para el análisis e implementación de sistemas orientados a objetos [45].

4. Desarrollo del modelo de predicción de la producción agrícola con CRISP-DM

Recordando, el objetivo general de la tesis fue:

> *"Aplicar la minería de datos para desarrollar un modelo de predicción que permita optimizar el beneficio económico obtenido por la producción agrícola generada en los distritos de riego del país."*

La labor de minería de datos es la parte medular de esta tesis. No sólo por la relevancia de la práctica en sí[2], si no también por la contribución de los modelos al algoritmo de optimización propuesto en el capítulo 5. Los modelos de predicción de la producción generados durante la práctica de la minería de datos son utilizados durante el algoritmo de optimización como un elemento que contribuye a la estimación del ingreso que se puede obtener por parte de la siembra de los cultivos. Sin la estimación proporcionada por los modelos se carece totalmente de la información necesaria para guiar la labor de la optimización, lo cual dificultaría enormemente el cumplimiento del objetivo principal de la tesis.

Para realizar la labor de minería de datos se propone utilizar la metodología CRISP-DM [33]. CRISP-DM ha sido descrita con anterioridad en el estado del arte, sección de metodologías. Este capítulo es más que nada un resumen de las diferentes salidas que acompañan a la metodología. Se procuró guardar en lo posible el orden y la estructura mostrados en la guía CRISP-DM, por lo que los nombres originales de las etapas indicadas en la guía se mantienen a lo largo del capítulo. Cabe mencionar, que CRISP-DM está orientada principalmente al sector industrial y de negocios, por lo que algunos términos podrán parecer un poco fuera de contexto. Tal es el caso de "comprensión del negocio" (que es la traducción literal de "*bussines understandig*", la primera fase en el modelo de referencia CRISP-DM). Se decidió conservar dichos términos con el fin de mantener la correspondencia con el modelo de referencia.

4.1. Comprensión del negocio

4.1.1. Antecedentes

Los distritos de riego son áreas agrícolas cuyos programas de producción se apoyan básicamente en el servicio de riego que se proporciona a los terrenos de cultivo con las obras de infraestructura hidroagrícola construidas para tal propósito y, por sus múltiples relaciones con los diferentes sectores de la economía regional, extienden su influencia a una amplia zona [2].

[2] No existen trabajos de minería de datos en la información de producción agrícola de los distritos de riego del país.

4. Desarrollo del modelo de predicción de la producción agrícola con CRISP-DM

Recordando, el objetivo general de la tesis fue:

> *"Aplicar la minería de datos para desarrollar un modelo de predicción que permita optimizar el beneficio económico obtenido por la producción agrícola generada en los distritos de riego del país."*

La labor de minería de datos es la parte medular de esta tesis. No sólo por la relevancia de la práctica en sí[2], si no también por la contribución de los modelos al algoritmo de optimización propuesto en el capítulo 5. Los modelos de predicción de la producción generados durante la práctica de la minería de datos son utilizados durante el algoritmo de optimización como un elemento que contribuye a la estimación del ingreso que se puede obtener por parte de la siembra de los cultivos. Sin la estimación proporcionada por los modelos se carece totalmente de la información necesaria para guiar la labor de la optimización, lo cual dificultaría enormemente el cumplimiento del objetivo principal de la tesis.

Para realizar la labor de minería de datos se propone utilizar la metodología CRISP-DM [33]. CRISP-DM ha sido descrita con anterioridad en el estado del arte, sección de metodologías. Este capítulo es más que nada un resumen de las diferentes salidas que acompañan a la metodología. Se procuró guardar en lo posible el orden y la estructura mostrados en la guía CRISP-DM, por lo que los nombres originales de las etapas indicadas en la guía se mantienen a lo largo del capítulo. Cabe mencionar, que CRISP-DM está orientada principalmente al sector industrial y de negocios, por lo que algunos términos podrán parecer un poco fuera de contexto. Tal es el caso de "comprensión del negocio" (que es la traducción literal de "*bussines understandig*", la primera fase en el modelo de referencia CRISP-DM). Se decidió conservar dichos términos con el fin de mantener la correspondencia con el modelo de referencia.

4.1. Comprensión del negocio

4.1.1. Antecedentes

Los distritos de riego son áreas agrícolas cuyos programas de producción se apoyan básicamente en el servicio de riego que se proporciona a los terrenos de cultivo con las obras de infraestructura hidroagrícola construidas para tal propósito y, por sus múltiples relaciones con los diferentes sectores de la economía regional, extienden su influencia a una amplia zona [2].

[2] No existen trabajos de minería de datos en la información de producción agrícola de los distritos de riego del país.

Los distritos de riego, aunque con un objetivo común, tienen características muy variadas; algunos son pequeños, en el orden de 10 mil hectáreas; otros de extensión media, entre 30 y 100 mil hectáreas; y otros más grandes, entre 100 y 270 mil hectáreas [2].

La determinación de una óptima producción agrícola es un problema clave para los distritos de riego del país [5]. Los ingresos del distrito dependen en gran medida de la producción lograda cada año agrícola. Estos ingresos varían de manera significativa de un distrito a otro. Así lo ilustra Soto (1981) [46] con el comentario "*hay distritos como el de la Costa de Hermosillo (051) de la región Noroeste, en donde el ingreso aparente, por predio, de un pequeño propietario, es del orden de 175 mil pesos anuales, mientras que en Tehuantepec, Oaxaca (019), de la región SUR, para el mismo estrato, dicho ingreso no llega a los 5 mil pesos*".

Estas variaciones obedecen a muchos factores locales como propiedades del suelo, clima, prácticas de conservación, etc. y a las relaciones que guardan estos factores entre sí. Algunas relaciones son evidentes, por ejemplo, se sabe que "existe una relación directa entre la lámina de riego (volumen de agua aplicado) y la productividad por hectárea; entre la productividad y el tamaño de la parcela (superficie)" [46]. Pero la cantidad de factores involucrados dificultan la maximización de la productividad en los distritos de riego [5].

Entonces ¿cómo mejorar la producción agrícola?. Quizá la respuesta más obvia sea el cambio del patrón de cultivos, con el fin de utilizar cultivos más remunerativos. Lamentablemente, las restricciones propias de los cultivos (composición del suelo, clima) tienen la consecuencia de que "*los patrones de cultivos no se pueden cambiar a corto plazo, y la única solución para mejorar la productividad es aumentar el rendimiento de los cultivos*" [46].

El rendimiento de los cultivos es la unidad de producto agrícola obtenido por unidad de superficie sembrada (ton/ha). Muchos trabajos se han enfocado en la estimación del rendimiento como una variable dependiente de otras variables conocidas [15][16][17], pero la mayoría de estos trabajos se orientaban a la aplicación de métodos estadísticos simples (como la regresión lineal), que a decir de otros autores no eran particularmente útiles en el estudio de la variabilidad del rendimiento [38][39].

Las herramientas de minería de datos empiezan a verse valiosas en el análisis de conjuntos de datos masivos provenientes de sistemas complejos, a la vez que empiezan a proveer de información de alta calidad [39]. Varios trabajos de minería de datos enfocados en el rendimiento de los cultivos [38][39] [41] [42] [44].

Con tales antecedentes, no es difícil plantear que la meta de la minería de datos será encontrar el tipo de modelación adecuada que permita estimar el rendimiento de los cultivos que pudieran sembrarse en un determinado distrito de riego. Para ello, el o los modelos desarrollados deberán apoyarse en factores cuyos valores son conocidos y están disponibles en una base de datos histórica.

4.1.2. Objetivos del negocio y criterio de éxito

El objetivo del proyecto es la obtención de un modelo predictivo del rendimiento de los cultivos sembrados en un distrito de riego. Un objetivo secundario sería la selección de las técnicas de minería de datos que mejor se adapten al problema en cuestión con el fin de llevarlas a más distritos de riego.

Lo más importante para un usuario de los modelos predicción del rendimiento de cultivos sería la confianza y credibilidad que se pudiera tener a dichos modelos. Seguramente, las estimaciones generadas por los modelos serían una fuente de información muy útil para la toma de decisiones, por lo que la certeza es un factor crítico para el proyecto. Trabajos relacionados han obtenido una certeza de alrededor del 80 % en la efectividad de predicción [39]. Lamentablemente, como se ha mencionado, el problema es afectado por cuestiones locales y la solución es altamente dependiente de la información disponible, por lo que una comparación contra ejercicios similares efectuados en otras partes del mundo sería desigual y podría generar conclusiones erróneas. Es precisamente por esta razón por la que se plantea elaborar primero un modelo "tradicional" estadístico, obtener la certeza proporcionada por dicho modelo, y, posteriormente, utilizar técnicas más recientes de minería de datos con el fin de realizar una comparación más equitativa.

4.1.3. Inventario de recursos

Los recursos se clasifican en tres tipos:

a) Recursos de datos y de conocimiento

Básicamente, los recursos de datos y de conocimiento son las fuentes de información que proporcionan los datos necesarios para la elaboración de los modelos. Éstos son orígenes de datos electrónicos (bases de datos, almacenes, archivos y documentos electrónicos), herramientas de software, técnicas, recursos en línea, expertos, documentación en forma impresa y electrónica. La tabla 3 lista los recursos de datos que fueron obtenidos para el proyecto.

Número de recurso	Recurso	Tipo de recurso	Descripción.
1	Estadísticas agrícolas SINHDR	Base de datos	Segmento dedicado a la producción de la base de datos del Sistema de Información Hidroagrícola para Distritos de Riego (SINHDR), desarrollado por el Instituto Mexicano de Tecnología del Agua. Contiene información de 5 años agrícolas para todos los distritos, más 3 años más de información del distrito de riego 038.
2	Documentos electrónicos de estadísticas agrícolas CONAGUA	Documento electrónico	Documentos electrónicos con las estadísticas de producción de los distritos de riego, emitidos de forma anual por la CONAGUA.
3	ERIC III.	Software y base de datos.	Programa de cómputo con información climatológica extraída de estaciones de medición distribuidas en todo el país. Versión 3.0
4	Hojas de datos del distrito de riego 038	Hoja de cálculo	Hojas electrónicas con un resumen de información del distrito de riego 038 (Río Mayo, SONORA). Consiste en un conjunto de hojas de cálculo que almacena información de tipo económica, técnica y productiva del

			distrito de riego 038. Cada hoja pertenece a un módulo de riego distinto, siendo 16 módulos en total.

Tabla 3. Recursos de datos obtenidos para el proyecto de minería de la tesis

CRISP-DM contempla también como parte de los recursos de conocimiento y datos aquellas herramientas de software que auxilian al usuario durante el proceso. Las herramientas de software utilizadas durante el proyecto de minería de datos de esta tesis fueron:

- Weka 3-4. Suite de minería de datos, versión 3.4
- Excel®. Programa de hoja de cálculo, versión 2003.
- Access®. Programa de base de datos, versión 2003.
- MySQL. Programa de base de datos, versión 5.1.

La sección 4.1.11.2 describe con más profundidad los recursos de software utilizados en el proyecto.

Otras fuentes de conocimiento a las que se tuvo acceso y que no aparecen en la tabla 3 por cuestión de espacio son aquellos expertos a los que se consultó en algún momento, pero por cuestión de espacio son omitidos aquí, más sus nombres aparecen en la sección de agradecimientos.

b) Recursos humanos

Los recursos humanos son el conjunto de personas que apoyarán con la parte técnica y científica del proyecto, tales como: el administrador del sistema, el administrador de bases de datos, el equipo de soporte técnico, los analistas y estadistas. En el caso de este proyecto, la persona que debe ser identificada como el responsable de llevar a cabo dichas actividades es el autor de esta tesis.

c) Recursos de hardware

Los recursos de hardware con los que se contó para el desarrollo del proyecto de minería fueron los siguientes:

Equipo #1. Computadora con procesador Intel Centrino Duo a 1.66 GHz, con 1 GB de memoria RAM y un disco duro de 80 GB.

Equipo #2. Computadora con procesador AMD Semptron a 2.5 GHz, con 512 MB de memoria RAM y un disco duro de 120 GB.

4.1.4. Requerimientos, supuestos y condiciones para el proyecto

a) Requerimientos:

1) Un modelo de predicción del comportamiento del rendimiento de cultivos agrícolas para un distrito de riego que ofrezca mayor eficiencia que la obtenida con los métodos estadísticos.

2) El modelo predicción del rendimiento debe contemplar la influencia de factores de importancia para el distrito, tales como el volumen utilizado, la superficie sembrada, y el tipo de cultivo que soporta.
3) La experiencia del desarrollo del modelo debe ser registrada, con el fin de aplicarla a los demás distritos.
4) La tecnología utilizada para el desarrollo del modelo debe ser accesible a todos los distritos de riego.

b) Supuestos

1) Se asume que los principales atributos involucrados están disponibles entre los recursos de datos obtenidos.

c) Condiciones

1) Se asume que la metodología aplicada en un distrito de riego representativo, puede ser llevada con relativa facilidad a cualquier otro. Para ello se requiere que los distritos donde sea aplicada dicha metodología dispongan de una cantidad similar de información.
2) Es forzoso realizar una inspección a los recursos para validar la calidad de los datos. Pese a que se asume que los datos provienen de fuentes confiables, es necesario la validación de los mismos, con el fin de evitar "ruido" que pueda influir negativamente en el desarrollo del modelo.

4.1.5. Riesgos y contingencias

Los riesgos detectados en el proyecto de minería de datos de esta tesis fueron los siguientes:

a) La calidad de los datos recopilados. Como se verá más adelante, la fuente principal de datos recopilada para el proyecto estaba afectada por una gran cantidad de errores, lo que llevo a evaluar dicha fuente como portadora de datos de mala calidad. Aún después del proceso de limpieza, es posible que errores no detectados prevalezcan en los datos y afecten los modelos resultantes del proceso de minería. Al respecto, en Hand et al (2001) [27] se señala que "*la efectividad de un ejercicio de minería de datos depende críticamente en la calidad de los datos*". En otras palabras, GIGO (Garbage In, Garbage Out). Siempre se corre el riesgo que los modelos más novedosos provengan de datos alterados, imprecisos ó atípicos, siendo estos más perjudiciales que benéficos.

b) La representatividad de la información. Otro riesgo, es que la selección de información realizada para el proyecto genere modelos que carezcan de utilidad práctica para los usuarios.

c) También, es posible que la técnica de modelado seleccionada por esta tesis no funcione en otros distritos de riego, esto debido a que el desempeño de la técnica depende implícitamente de los datos en los que ésta es aplicada.

4.1.6. Terminología

A continuación se proporciona una terminología que incluye los conceptos utilizados con frecuencia a lo largo de la tesis, y que versan sobre todo en el tema de la producción agrícola en distritos de riego.

Año agrícola. Un año agrícola hace referencia a un período de tiempo situado entre dos años comunes. Así, el año agrícola 1999-2000 hace referencia a los meses de octubre de 1999 hasta septiembre del año 2000. Para efectos de esta tesis, cuando se haga referencia al año agrícola en un formato de cuatro dígitos se referirá al año agrícola cuyo intervalo de tiempo está situado entre el año inmediato anterior y el año mencionado. Así, el año agrícola 1999-2000 será mencionado como año agrícola 2000.
Costo de producción: Normalmente, se refiere al costo por sembrar una hectárea de cultivo. Se expresa en M$/ha. Al multiplicarse por toda la superficie sembrada del cultivo de referencia se expresa en M$, y es llamado costo de producción total o costo total.
Cultivo: Hace referencia a un cultivo agrícola.
Lámina de riego: Es una representación lineal del volumen utilizado en una superficie determinada de tierra. Se expresa en milímetros.
Producción agrícola: Normalmente, hace referencia a la cantidad de producto agrícola (cultivo) obtenido en un ciclo agrícola. La producción se mide en toneladas. En ocasiones, se usa para expresar el ingreso obtenido, aunque de manera estricta, el ingreso es obtenido al multiplicar la producción agrícola de un cultivo por su precio de venta.
Rendimiento: La cantidad de producto agrícola obtenido por unidad de superficie. Se expresa en toneladas por hectárea (ton/ha).
Superficie sembrada: Extensión de tierra sembrada con uno o varios cultivos. Se expresa en hectáreas.
Superficie cosechada: Extensión superficie sembrada de la cual finalmente fue extraído el producto agrícola. Cuando se habla del mismo cultivo o grupo de cultivos, ésta no puede ser mayor a la superficie sembrada. Se expresa en hectáreas.
Ingreso: Normalmente hace referencia a la cantidad económica obtenida por la venta del producto agrícola. Se expresa en pesos o miles de pesos. Se obtiene al multiplicar la cantidad de producto agrícola por el precio de venta (ingreso bruto). Si se le restan los costos de producción se está hablando del ingreso neto.

4.1.7. Costos

Siempre que la información esté disponible, el costo de realizar un estudio de minería de datos como el desarrollado por esta tesis es muy bajo. Si la información no está disponible, entonces los costos necesarios para generar la información requerida por la práctica deberá ser añadido al costo total.

En el caso de esta tesis, se contó con la colaboración de instituciones que proporcionaron la información requerida para llevar a cabo el proyecto de minería de datos. Esto redujo enormemente los costos que normalmente deben contemplarse en un proyecto de este tipo. Otro factor que colaboró a la reducción de costos fue el uso de herramientas de software libre, concretamente el caso de la suite de minería de datos utilizada. Las herramientas de software libre se distribuyen con el código fuente sin ningún costo, con la intención de que

puedan ser utilizadas y modificadas por el usuario final con el fin de impulsar la auto-evolución de la herramienta.

4.1.8. Metas de la minería de datos

Las metas del proyecto de minería son las siguientes:

1. Obtener la técnica de modelado que permita desarrollar modelos de predicción del rendimiento que proporcione una certeza superior a la obtenida con los métodos estadísticos tradicionales.
2. Elaborar los modelos de predicción del rendimiento de los cultivos de un distrito de riego, utilizando para ello la información de los factores disponibles en la información recolectada (por ejemplo, superficie, lámina de riego, distribución de temperaturas, etc.).

4.1.9. Criterio de éxito de la minería de datos.

Como medida de éxito, se propone el comparar los modelos resultantes de las técnicas de minería de datos, contra los modelos obtenidos de la aplicación de técnicas estadísticas clásicas (por ej. regresión lineal multivariable).

4.1.10. Plan del proyecto

Para llevar a cabo el proyecto, se elaboró el siguiente plan de trabajo:

Etapa	Descripción	Fases de la metodología CRISP-DM	Salidas
1	Construcción del conjunto de datos base para la aplicación de algoritmos de minería.	Comprensión de los datos Preparación de los datos	Conjunto de datos base.
2	Selección de algoritmos de minería. En base al tipo de problema planteado, se seleccionarán aquellas técnicas de minería que más se adapten al problema.	Selección de la técnica de modelado	Lista de algoritmos de minería a aplicar
3	Aplicación de técnicas estadísticas para el problema de predicción del rendimiento. Durante esta etapa, se desarrollará un modelo estadístico cuya variable dependiente sea el rendimiento. El resultado de esta etapa forma parte del "criterio de éxito" propuesto por esta tesis, ya que es este modelo primario el que nos sirve como punto de comparación para determinar la efectividad de los demás modelos.	Construcción del modelo	Modelo de regresión multivariable de predicción del rendimiento.
4	Aplicación de algoritmos de minería. Los algoritmos seleccionados en la etapa 2 serán aplicados en esta etapa.	Construcción del modelo	Modelo(s) de aprendizaje automático para la predicción del rendimiento.
5	Evaluación de los modelos. Evaluación inicial de cada modelo de manera independiente. Comparación de los modelos de minería contra el modelo primario. Comparación de los modelos entre sí. Selección del mejor modelo.	Evaluación del modelo	Reporte de evaluación de modelos.
6	Interpretación de los modelos y obtención de conclusiones. Corresponde a la etapa de "Despliegue" en CRISP-DM, sólo que adaptada para la elaboración de esta tesis.	Despliegue	Conclusiones

Tabla 4. Etapas en las que fue divido el proyecto de minería para la tesis.

En la tabla 4 se observa que las fases de la metodología CRISP-DM fueron agrupadas por etapas. La razón de este agrupamiento obedece solamente a cuestiones de organización, y no se trata en absoluto de alterar la metodología CRISP-DM. Si se observa, las fases se encuentran en el mismo orden, únicamente fueron agrupadas (como por ejemplo, en la etapa 1) ó divididas (como en las etapas 3 y 4), esto con el fin de optimizar el tiempo y el esfuerzo en el desarrollo del proyecto.

4.1.11. Selección inicial de herramientas y técnicas

4.1.11.1. Técnicas seleccionadas para el proyecto

Para seleccionar las técnicas utilizadas en el proyecto, fue necesario determinar el tipo de problema que se estaba abordando. El punto 2 del apéndice de la guía CRISP-DM [33], se proporciona una clasificación de los problemas que se pueden abordar con la minería de datos. Además de la clasificación, la guía CRISP-DM proporciona una lista de técnicas que pueden ser particularmente útiles para construir el modelo o patrón que se adapte de mejor manera al problema. Un resumen de la clasificación de problemas y las técnicas de solución se proporciona en la tabla A.1 del Anexo 1.

La tabla A1.1 permite determinar el tipo de problema que atañe al proyecto de minería de esta tesis. Dado que lo que se busca es predecir el valor de un atributo en términos de otros atributos conocidos (y que los valores del atributo a predecir son valores continuos), se determinó que problema de modelado era un problema de predicción.

Los problemas de modelado predictivo puede ser tratados como problemas de clasificación o de regresión (como se ve en la figura 4), aunque, como se ve en la tabla A1.1, CRISP-DM hace la distinción entre problemas de clasificación y problemas de predicción. En el caso de este proyecto, se exploraron técnicas relacionadas con ambos tipos de problemas, esto con el fin de encontrar el modelo que mejor se ajustara al problema planteado.

En la sección de antecedentes (punto 4.1.1) se expusieron una serie de trabajos relacionados con la predicción o estimación del rendimiento por medio de diversas técnicas de minería de datos. Durante el desarrollo de esta tesis se utilizaron técnicas similares para tener un punto de comparación en el momento de validar los modelos obtenidos. Las técnicas utilizadas en el proyecto de minería de esta tesis fueron las siguientes:

a) Regresión multivariable.
b) Redes neuronales.
c) Inducción de reglas.
d) Árboles de regresión.

4.1.11.2. Herramientas de software para el proyecto

Dentro de la minería de datos existen actualmente una gran cantidad de herramientas tanto académicas como comerciales, entre las que se pueden mencionar a Darwin, Mineset, Clementine, INLEN, Explora, DBminer, DataMine, Quest o IBM Intelligent, como las de mayor difusión [47]. La selección de una herramienta de software en particular dependió de

requerimientos específicos. Más que una herramienta de aplicación, se requería de una suite de librerías que fueran accesibles desde un lenguaje de programación. Preferentemente, se buscaba de una herramienta de código abierto, ya que existía la posibilidad de modificar ó mejorar alguno de los algoritmos utilizados. Otro factor importante que influyó en la selección fue el precio, ya que uno de los propósitos era que la tecnología desarrollada fuera accesible a los distritos de riego (y al público en general).

De las herramientas revisadas, se determinó que el entorno para el análisis del conocimiento de Waikato (Weka[3], por sus iniciales en inglés) era la que satisfacía de mejor manera los requerimientos especificados. En palabras de Witten et al (1999) [48] "*Weka es una suite comprensiva de librerías de clases Java que implementa mucho del estado del arte del aprendizaje automático y algoritmos de minería de datos*". El propósito de Weka es *permitir a los usuarios acceder a una variedad de técnicas de aprendizaje automático para los propósitos de experimentación y comparación utilizando conjuntos de datos del mundo real* [49].

Otras herramientas que se utilizaron se han mencionado ya en la sección de recursos (punto 4.1.3).

4.2. Comprensión de los datos

Durante esta etapa, se realiza una primera inspección a los recursos de datos recabados en el inventario de recursos. El objetivo es listar los distintos datos que están disponibles para el proyecto, así como detectar los requerimientos para datos más detallados.

4.2.1. Recolección inicial de datos

Los datos fueron extraídos de los recursos mostrados en la tabla 3 (punto 4.1.3). La información que se consideró útil fue extraída en forma de series de datos, las cuales son expuestas en esta sección, describiendo el contenido a nivel de atributo. La serie de datos se muestra en la tabla 5.

Serie de datos #1: Estadísticas agrícolas SINHDR. Extraída del recurso de datos #1.
Período temporal: 1995-2003*
Descripción: Archivo con información estadística de índole agrícola. Los archivos están en formato Paradox[4].

Atributo	Información del atributo
Distrito	Clave oficial del distrito de riego, asignada por CONAGUA.
Módulo	Clave local del módulo de riego, asignada por el propio sistema SINHDR.
Ciclo	Clave que representa en un solo atributo al ciclo y al año agrícola. Así, P/V1997 indica que el registro representa un dato del año agrícola 1996-1997 en el ciclo "primavera-verano".
Cultivo	Cultivo de referencia
Superficie sembrada	Superficie sembrada con el cultivo de referencia. Se expresa en hectáreas.

[3] El "weka" es una curiosa ave nativa de Nueva Zelanda (lugar donde se localiza la Universidad de Waikato) de tamaño aproximado al de una gallina [49].
[4] Paradox es un formato de base de datos relacional desarrollada originalmente por ANSA Software, pasando posteriormente al control de Borland y desarrollada actualmente por COREL [50].

requerimientos específicos. Más que una herramienta de aplicación, se requería de una suite de librerías que fueran accesibles desde un lenguaje de programación. Preferentemente, se buscaba de una herramienta de código abierto, ya que existía la posibilidad de modificar ó mejorar alguno de los algoritmos utilizados. Otro factor importante que influyó en la selección fue el precio, ya que uno de los propósitos era que la tecnología desarrollada fuera accesible a los distritos de riego (y al público en general).

De las herramientas revisadas, se determinó que el entorno para el análisis del conocimiento de Waikato (Weka[3], por sus iniciales en inglés) era la que satisfacía de mejor manera los requerimientos especificados. En palabras de Witten et al (1999) [48] "*Weka es una suite comprensiva de librerías de clases Java que implementa mucho del estado del arte del aprendizaje automático y algoritmos de minería de datos*". El propósito de Weka es *permitir a los usuarios acceder a una variedad de técnicas de aprendizaje automático para los propósitos de experimentación y comparación utilizando conjuntos de datos del mundo real* [49].

Otras herramientas que se utilizaron se han mencionado ya en la sección de recursos (punto 4.1.3).

4.2. Comprensión de los datos

Durante esta etapa, se realiza una primera inspección a los recursos de datos recabados en el inventario de recursos. El objetivo es listar los distintos datos que están disponibles para el proyecto, así como detectar los requerimientos para datos más detallados.

4.2.1. Recolección inicial de datos

Los datos fueron extraídos de los recursos mostrados en la tabla 3 (punto 4.1.3). La información que se consideró útil fue extraída en forma de series de datos, las cuales son expuestas en esta sección, describiendo el contenido a nivel de atributo. La serie de datos se muestra en la tabla 5.

Serie de datos #1: Estadísticas agrícolas SINHDR. Extraída del recurso de datos #1.
Período temporal: 1995-2003*
Descripción: Archivo con información estadística de índole agrícola. Los archivos están en formato Paradox[4].

Atributo	Información del atributo
Distrito	Clave oficial del distrito de riego, asignada por CONAGUA.
Módulo	Clave local del módulo de riego, asignada por el propio sistema SINHDR.
Ciclo	Clave que representa en un solo atributo al ciclo y al año agrícola. Así, P/V1997 indica que el registro representa un dato del año agrícola 1996-1997 en el ciclo "primavera-verano".
Cultivo	Cultivo de referencia
Superficie sembrada	Superficie sembrada con el cultivo de referencia. Se expresa en hectáreas.

[3] El "weka" es una curiosa ave nativa de Nueva Zelanda (lugar donde se localiza la Universidad de Waikato) de tamaño aproximado al de una gallina [49].
[4] Paradox es un formato de base de datos relacional desarrollada originalmente por ANSA Software, pasando posteriormente al control de Borland y desarrollada actualmente por COREL [50].

Superficie cosechada	Superficie efectivamente cosechada. Se expresa en hectáreas.
Lámina de riego	Es una medida lineal que representa la cantidad de volumen aplicada por unidad de superficie[5]. Se expresa en centímetros (cm).
Costo de producción	Costo de producción por unidad de superficie. Se expresa en pesos por hectárea ($/ha).
Rendimiento	Cantidad de producto obtenido por unidad de superficie. Se expresa en toneladas por hectárea (ton/ha).
Precio	Precio al que será vendido el producto obtenido. Se expresa en pesos por tonelada ($/ton).
Industria	Industria destino del producto obtenido (por ej. harinera, mercado, forrajera, aceitera, etc.).
Destino	Indica si el producto será para consumo nacional o de exportación.

Tabla 5. Descripción de la serie de datos #1.

* Aunque el período temporal esté indicado de 1995 al año 2003, es de 1995 al año 2001 donde existe información para todos los distritos. En los años agrícolas 2002 y 2003 únicamente existe información del distrito de riego 038.

Serie de datos #2: Estadística de cultivos sembrados por distrito de riego. Información distribuida por CONAGUA (extraída por recurso de datos #2).
Período temporal: 1985-2005.
Descripción: Estadística reportada en un conjunto de documentos electrónicos que se emiten de forma anual en formato Acrobat PDF (Portable Document Format). Los documentos contienen varios conjuntos de estadísticas, pero se identificó únicamente la de cultivos como útil para el proyecto. La tabla 6 describe en términos de atributos la información obtenida en esta serie.

Atributo	Información del atributo
Distrito de riego	Identificador del distrito de riego.
Ciclo	Ciclo agrícola de referencia.
Cultivo	Cultivo de referencia.
Superficie sembrada	Superficie sembrada con el cultivo de referencia. Se expresa en hectáreas.
Superficie cosechada	Superficie efectivamente cosechada. Se expresa en hectáreas.
Rendimiento	Cantidad de producto obtenido por unidad de superficie. Se expresa en toneladas por hectárea (ton/ha).
Producción	Cantidad de producto obtenido en la superficie cosechada. Se expresa en toneladas (ton).
Precio medio rural	Precio al que será vendido el producto obtenido. Se expresa en pesos por tonelada ($/ton).
Valor de la cosecha	Ingreso total obtenido por concepto de venta del producto.

Tabla 6. Descripción de la serie de datos #2.

Para utilizar la información contenida en los documentos electrónicos fue necesario utilizar un mecanismo intermedio de captura y estructuración de la información, ya que el formato original no es el adecuado para aplicarle un algoritmo de minería de datos.

Serie de datos #3: Serie de datos de Información Climatológica. Extraída del recurso de datos #3.
Período temporal: 1960-2004.
Descripción: Base de datos de formato propietario[6] que se accede a través de una interfaz de consulta. Los atributos de la base de datos serían los siguientes [52]:

Atributo	Información del atributo
Año	Año al que corresponde la observación
Estación	Identificador de la estación climatológica origen de la observación.
Día	Periodo de tiempo de la observación (el día es la base).*

[5] La lámina de riego se calcula de la siguiente manera: Lra=VA/S, donde VA=Volumen de agua aplicado y S=Superficie regada [51].
[6] Base de datos con formato desconocido y accesible únicamente a través del sistema.

Temperatura observada	Grados centígrados observados a las 8:00 hrs.
Temperatura mínima	Temperatura mínima observada, en grados centígrados (ºC).
Temperatura máxima	Temperatura máxima observada, en grados centígrados (ºC).
Precipitación	Precipitación total acumulada en el día (de 8:00 am a 8:00 am), en milímetros (mm).
Evaporación	Evaporación total acumulada en el día (de 8:00 am a 8:00 am), en milímetros (mm).
Tormenta	Entero que indica la presencia de tormenta (1=Si hubo, 0=no hubo).
Granizo	Entero que indica la presencia de granizo (1=Si hubo, 0=no hubo).
Niebla	Entero que indica la presencia de niebla (1=Si hubo, 0=no hubo).
Cobertura del cielo	Entero que indica la modalidad de la cobertura del cielo (0=despejado, 1=medio nublado, 2=nublado).

Tabla 7. Descripción de la serie de datos #3.

* El sistema tiene la facilidad de consultar los promedios de las observaciones, esto a nivel mensual y anual.

Serie de datos #4: Estadística productiva del distrito de riego 038 para el año 2006.
Extraída del recurso de datos #4.
Período temporal: 2006.
Descripción: Series de datos de estadística de producción almacenada de origen en formato de hoja de cálculo. Cada serie representa la información de un módulo de riego. El formato de las series se muestra en la tabla 8.

Atributo	Información del atributo
Ciclo	Nombre del ciclo al que se refiere el bloque de información (primavera-verano, otoño-invierno, perennes).
Cultivo	Cultivo de referencia.
Superficie sembrada	Superficie sembrada con el cultivo de referencia. Se expresa en hectáreas.
Superficie cosechada	Superficie efectivamente cosechada. Se expresa en hectáreas.
Lámina de riego	Es una medida lineal que representa la cantidad de volumen aplicada por unidad de superficie. Se expresa en centímetros (cm).
Costo de producción	Costo de producción por unidad de superficie. Se expresa en pesos por hectárea ($/ha).
Rendimiento	Cantidad de producto obtenido por unidad de superficie. Se expresa en toneladas por hectárea (ton/ha).
Precio	Precio al que será vendido el producto obtenido. Se expresa en pesos por tonelada ($/ton).

Tabla 8. Descripción de la serie de datos #4.

Como podrá notarse, la serie de datos #4 es muy similar a la serie de datos #1, salvo por la omisión de los atributos distrito, módulo, año agrícola, industria y destino. El distrito es conocido, ya que se sabe que estas hojas pertenecen al distrito de riego 038. Lo mismo sucede con el año agrícola, ya que la información representa los datos del año agrícola 2005-2006. Cada serie de datos tiene asociado el identificador y nombre del módulo al que pertenecen, por lo que el módulo también es conocido. Como se verá, los atributos industria y destino serán desestimados en la siguiente etapa de la metodología CRISP-DM, por lo que se puede decir que la serie de datos #4 es una extensión de la serie de datos #1.

En resumen, se obtuvieron 4 series de datos de los recursos recopilados en la etapa de recolección inicial de datos. Tres de estas series están dedicadas a la información productiva de distritos de riego. Una serie almacena información climatológica de estaciones distribuidas a nivel nacional. En la siguiente etapa se realiza una descripción más detallada de los datos contenidos en cada serie.

4.2.2. Descripción de los datos

En esta sección se describen los datos adquiridos. Propiedades de los datos, como el formato, número de registros, tipos de campos y otras características propias de los datos, son identificadas en esta sección, donde se trata de identificar si los datos adquiridos son suficientes para satisfacer los requerimientos del proyecto.

La descripción de las series de datos recopiladas para el proyecto se proporciona en el anexo 2 de esta tesis, tablas A2.1-A2.4.

4.2.2.1. Análisis de los datos adquiridos

Se identificaron 4 series de datos en los recursos de datos disponibles para el proyecto. Tres de estas series contienen información de producción agrícola. Estas series almacenan información del mismo tipo y se puede visualizar que existe duplicidad de información, ya que los períodos de tiempo se traslapan.

El atributo rendimiento (objetivo principal del proyecto) fue localizado en las series de datos #1, #2 y #4, en ambos a nivel de cultivo. Un inconveniente con la serie #2 es su formato de origen, que, aunque digital, carece de la estructura requerida por un manejador de base de datos para ser utilizada de forma directa.

Otros atributos importantes y localizados de manera directa fueron la superficie sembrada, la superficie cosechada y el precio. Incluso, la superficie se pudo localizar a nivel de cultivo, régimen y distrito.

La duplicidad en los datos es menos conflictiva que la falta de los mismos. En la sección de "requerimientos" se planteó la necesidad de que el modelo contemplara factores esenciales del ciclo de producción del cultivo, como por ejemplo, el volumen de agua utilizado. En la sección anterior "descripción de los datos" se puede apreciar que ninguna de las series contempla a nivel de cultivo el atributo volumen. Sin embargo, existe en las series de datos #1 y #4 el atributo lámina de riego, que podría aproximarse al volumen a través de una operación utilizando el atributo superficie sembrada (también disponible en ambas series).

En la futura sección "preparación de datos", se verá como afectan estos detalles en la conformación de la serie de datos base del modelado.

4.2.3. Exploración de los datos

La exploración de datos aborda los aspectos del problema de minería que pueden ser manejados utilizando consultas, visualización gráfica y reportes [33]. La exploración de datos para minería de datos toma mucho de lo desarrollado en el área del análisis exploratorio de datos (EDA, por la siglas en inglés) y que fue introducida en 1977 por John Tuckey [53], como una medida para analizar los datos cuando hay un bajo nivel de conocimiento tanto sobre el sistema causante como de la información contextual.

El análisis exploratorio de datos puede ser descrito como una generación de hipótesis basada en los datos. Se examina la información en busca de estructuras que puedan indicar relaciones más profundas entre registros o atributos [27]. La exploración también ayuda en detectar errores técnicos antes de efectuar análisis costosos[7], evitando resultados indeseables causados por problemas ocultos presentes en los datos [54].

Diversas técnicas EDA fueron aplicadas en el proyecto de minería de datos de esta tesis. Éstas se agruparon en categorías, en función del tipo de tarea que estaban apoyando. Así, se utilizaron técnicas para: (a) obtener un resumen de datos, (b) visualizar variables únicas, (c) visualizar relaciones entre dos variables y (d) la visualización de más de dos variables.

4.2.3.1. Resumen de datos del proyecto

4.2.3.1.1. Técnicas para agrupar datos

En la sección 4.2.2.1 se observa que el atributo rendimiento se encuentra en tres series de datos distintas (#1, #2 y #4). Para la exploración de datos, se seleccionó la serie de datos #1 debido a que es la que contempla el periodo de tiempo más largo, es de origen una base de datos relacional y presenta atributos que pueden ser consultados a distintos niveles o dimensiones (distrito, modulo, ciclo, cultivo). Las series de datos #2 y #4 no fueron analizadas con estas técnicas debido a que poseían muy poca información relacionada con el problema.

Una forma básica de explorar el atributo rendimiento es obteniendo una vista resumida de los datos de manera agrupada a cualquiera de los distintos niveles que conforman la serie de datos.

Las consultas agrupadas en una base de datos se realizan por medio de la agregación de valores. La agregación en una base de datos se refiere a la combinación de varios valores en uno, esto por los operadores de suma o de maximización, por ejemplo. Un agregado es en general una cantidad computada desde una base de datos cuyos valores dependen de varios registros en la base de datos [27].

A continuación se exponen diversas funciones de agregación que se aplicaron para la exploración de los datos del proyecto.

- La **media** es un resumen simple del promedio de una colección de valores [27]. La media tiene la propiedad de ser una medida de tendencia central, esto en el sentido que minimiza la diferencia de suma de cuadrados entre los valores de datos. Pero se debe tener cuidado, ya que es sensible a datos atípicos en el conjunto de datos, y por lo tanto poco representativa en el caso de que la distribución sea asimétrica [55]. La media simple es definida como:

$$\overline{\mu} = \sum_{i=1}^{n} x(i) / n \qquad \textbf{(2.5)}$$

[7] En términos computacionales.

Donde:
$x(i)$.- Es el valor de la observación *i*.
n.- Es el número de observaciones.

La **mediana** estima el centro de distribución de un atributo. En datos de una dimensión, la mediana es el valor más central cuando los valores están ordenados [54]. En el caso de atributos categóricos (como especies o colores) la mediana no está definida.

$$Me = \begin{cases} x(\frac{n}{2}+\frac{1}{2}) & \text{si } n \text{ es impar} \\ \frac{x(\frac{n}{2})+x(\frac{n}{2}+1)}{2} & \text{si } n \text{ es par} \end{cases} \quad \textbf{(2.6)}$$

Donde:
$x(i)$ Es el valor de la observación número *i*. En el caso de la mediana, se asume que el conjunto de observaciones está ordenado.
n Es el número de observaciones.

La **varianza** es una medida de dispersión o variabilidad. Está definida como el promedio de las diferencias cuadradas entre la media y los valores de datos individuales [27]:

$$\bar{\sigma}^2 = \sum_{i=1}^{n} (x(i)-\mu)^2 / n \quad \textbf{(2.7)}$$

Donde:
$x(i)$.- Es el valor de la observación número *i*.
μ.- Es la media del conjunto.
n.- Es el número de observaciones.

La **desviación** mide lo lejos que están situados los datos respecto a su centro de gravedad, la media. La desviación típica es representativa de la dispersión del conjunto de datos sólo si la media es representativa de su centro [55].

$$\bar{\sigma} = \sqrt{\sum_{i} (x(i)-\mu)^2 / n} \quad \textbf{(2.8)}$$

Donde:
$x(i)$.- Es el valor de la observación número *i*.
μ.- Es la media del conjunto.
n.- Es el número de observaciones.

Una medida alternativa de dispersión que puede ser más representativa en el caso de que la distribución sea asimétrica o en presencia de datos atípicos, es el rango

intercuartílico [55]. Los cuartíles son los puntos que separan el conjunto de datos en cuatro partes del mismo tamaño. Se empieza con la mediana (Me), para obtener dos grupos de datos (a la izquierda y derecha de Me), posteriormente, para cada grupo se obtiene la mediana. El primer cuartíl es la mediana del grupo de datos que queda a la izquierda de Me, mientras que el tercer cuartíl es la mediana del grupo que queda a la derecha. El rango intercuartílico, común en algunas aplicaciones, es la diferencia entre el tercer y primer cuartíl [27][55].

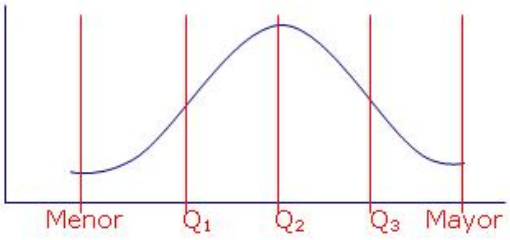


Figura 9. Representación gráfica de los cuartíles (Q_x) en una distribución normal.

$$RIC = Q_3 - Q_1 \qquad \textbf{(2.9)}$$

Donde:
Q1.- Es el primer cuartíl (mediana de la primera mitad de observaciones).
Q3.- Es el tercer cuartíl (mediana de la segunda mitad de observaciones).

Regla: Se consideran como atípicos los datos que son menores de $\boldsymbol{Q_1} - 1.5 \times$ RIC, o mayores de $\boldsymbol{Q_3} + 1.5 \times$ RIC [55].

4.2.3.1.2. Aplicación de las técnicas para agrupar datos en el proyecto de minería

La obtención de resúmenes de datos por medio del EDA se aplicó para identificar los tipos de error que se presentaron en los datos, sobre todo, aquellos relacionados con valores desproporcionados e irreales.

Detectando errores con el análisis exploratorio de datos

Para el caso de la serie de datos #1, se sabe que los valores reales de algunos de los atributos numéricos no deberían de variar de manera desproporcionada con respecto a una media histórica. Encontrar valores con esta característica podría ser indicativo de un probable error de captura, fallo de conversión o cálculo erróneo. Los atributos rendimiento, precio, lámina de riego y costo de producción entran en este tipo de atributos. Para ejemplificar lo anterior, la figura 10 muestra un conjunto de valores de rendimiento del cultivo maíz. De los 10 valores, 9 son correctos, y el valor del registro número 7, es incorrecto.

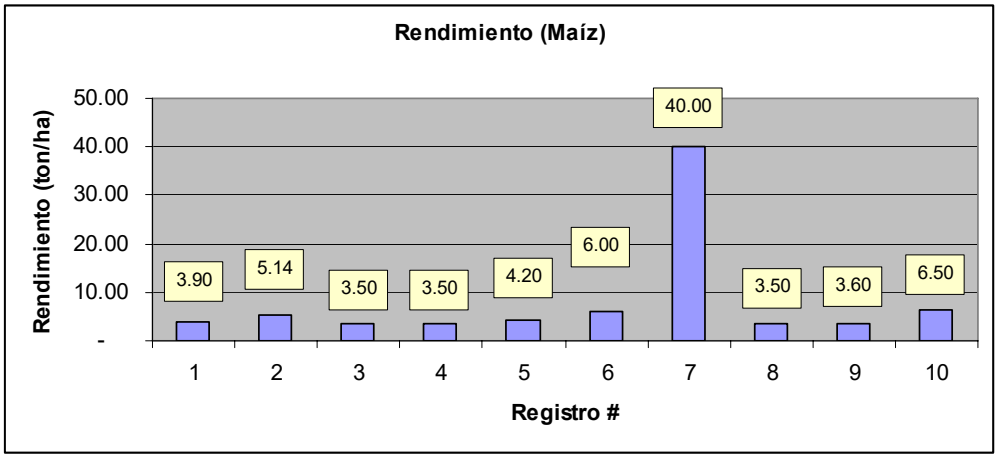


Figura 10. Valores de muestra del rendimiento para el cultivo Maíz (con errores).

La media del conjunto de valores de la figura 9 tiene un valor de 7.98. La desviación estándar se calculó en 11.30. El porcentaje de variación de la desviación estándar con respecto a la media equivale a un 142 %. Este porcentaje de variación podría indicarnos que existe una alta probabilidad de valores presentes en la muestra que se alejan de la media. Lo más probable es que el valor correspondiente al registro #7 sea 4.00 y no 40.00. La sección 4.3.2, ***limpieza de datos***, proporciona los mecanismos para lidiar con este tipo de fallos una vez detectados.

Las medidas propuestas para el análisis de la serie de datos #1 se muestran en la tabla 9.

#	Medida	Tipo de resumen	Nivel de agrupamiento	Objetivo
1	Rendimiento promedio por cultivo	Media	Cultivo	Obtener el rendimiento promedio de cada tipo de cultivo.
2	Desviación en el rendimiento promedio	Desviación estándar	Cultivo	Conocer la cantidad que se desvían los rendimientos de la media de cada tipo de cultivo.
3	Precio promedio por cultivo	Media	Cultivo	Obtener el precio promedio de cada tipo de cultivo.
4	Desviación estándar del rendimiento promedio por cultivo	Desviación estándar	Cultivo	Conocer la desviación con respecto a la media de los promedios de los cultivos.
5	Ingreso bruto[8] obtenido por cultivo	Suma	Cultivo	Detectar la influencia de las variaciones presentes en el rendimiento y el precio en el ingreso bruto del cultivo.
6	Mediana del ingreso neto[4] de los cultivos.	Mediana	Cultivo	Identificar los cultivos que se encuentran arriba de la mediana en cuanto a ingreso neto se refiere, esto como probables candidatos de futuros modelos ó análisis más profundos (serían los cultivos que más aportan a la producción).
7	Rango intercuartílico del rendimiento de los cultivos	Rango intercuartílico	Cultivo	Conocer que tan asimétrico es el rendimiento de los cultivos.
8	Ingreso neto obtenido por distrito de riego	Suma	Distrito	Identificar los distritos que más contribuyeron al ingreso total en el período de tiempo indicado.
9	Número de cultivos promedio por distrito de riego	Media	Distrito	Muestra el número de cultivos que en promedio se siembran en cada distrito.

Tabla 9. Medidas de interés para el proyecto (serie de datos del recurso #1).

[8] Los atributos "ingreso bruto" e "ingreso neto" no existen en la serie de datos original. Son atributos derivados a partir de los atributos rendimiento, superficie, precio y costo de producción. Para mayor información en la creación de atributos, consultar el punto 4.3.3, *construcción de datos*.

Un error que puede afectar en la forma en como los resúmenes son agrupados, es la incongruencia en los valores categóricos de los atributos nominales. Para ejemplificar esto, considerar el cálculo de la medida #5 de la tabla 9, el ingreso por cultivo. Una consulta efectuada con la agregación suma al atributo ingreso permite ver que existen hasta 14 nombres distintos relacionados con el mismo cultivo (tabla 10).

Nombre del cultivo	Ingreso (M$)	Cantidad
MAIZ	341,632,160.09	2,849
MAíZ	11,996.93	4
MAIZ (SEMILLA)	488.40	1
MAIZ BLANCO	936.57	2
MAIZ DULCE	3,664.08	2
MAIZ FORR. VERDE	14.25	1
MAIZ FORRAJE	3,598.04	2
MAIZ FORRAJE VERDE	3.43	1
MAIZ FORRAJERO	85,939.11	18
MAIZ GRANO	23,700.85	11
MAIZ P.	13,053.10	3
MAIZ TAR.	2,680.08	3
MAIZ TEM.	952.07	2
MAIZ V.	1,285.45	1

Tabla 10. Inconsistencias en el nombre del cultivo Maíz.

Algunas de las fallas en los valores categóricos del atributo "cultivo" podrían ser corregidas sin mayor conocimiento del contexto del problema (por ej., cambiar "MAíZ" por "MAIZ"). Sin embargo, la mayoría de las incongruencias mostradas en la tabla 10 requieren mayor información, información que en algunos casos podría solamente venir de una persona especializada en el área o en todo caso, con conocimientos más profundos del sistema origen de la serie de datos utilizada.

Derivado de este ejemplo, se puede ver que la etapa de *exploración de datos* está íntimamente ligada con las etapas *calidad de datos* y *limpieza de datos*. Muchas de las medidas de interés para el proyecto son afectadas por inconsistencias entre los datos, y a la vez, varias de las medidas implementadas ayudan a detectar inconsistencias en los datos (por ej. la desviación estándar). La razón de mencionar la calidad de los datos a este nivel, obedece a que durante este proyecto, la limpieza de los datos es necesaria para obtener buenos resultados en la fase de exploración. De manera similar, la *exploración de los datos* ayuda a establecer el nivel de calidad en los datos, para mejorarla en la etapa de *limpieza de datos*. Las tareas se vuelven recurrentes entre sí. Para fines prácticos, se publicarán en esta sección los resultados de la *exploración de datos*, aclarando que, de manera interactiva, se hicieron transiciones entre las etapas mencionadas con el fin de verificar y corregir los problemas presentes en los datos.

A continuación se exponen los resultados de los resúmenes de datos aplicados a los datos del proyecto.

Medidas relacionadas con rendimientos, precios e ingresos de los cultivos (Medidas #1, # 2, #3, #4 y #5).

La tabla 11 expone de manera parcial (sólo para 10 cultivos) los resultados del cálculo de las medidas #1, #2, #3, #4 y #5 indicadas en la tabla 9.

Cultivo	Rendimiento (Ton/ha)			Precio (M$/Ton)			Ingreso Bruto (M$)	Cantidad
	Promedio	Desviación estándar	% de desviación con respecto a la media	Promedio	Desviación estándar	% de desviación con respecto a la media		
AGUACATE	7.22	5.95	**82.35**	4.62	2.22	47.97	21,669.75	24
ALFALFA	32.73	29.04	**88.73**	5.29	58.18	**1,099.79**	166,993,084.69	933
ALFALFA (SECA)	12.90	15.70	**121.69**	1.70	0.14	8.32	416.40	2
ALGODÓN	2.90	0.93	32.18	23.44	268.49	**1,145.29**	20,961,394.18	776
AVENA	13.90	10.15	**73.02**	2.14	18.39	**859.74**	686,686.88	566
AVENA (SECA)	3.19	1.81	**56.82**	1.62	0.68	42.12	4,905.08	25
BERENJENA	32.96	4.94	14.98	2.52	0.67	26.67	191,375.18	17
BERMUDA	6.83	10.95	**160.34**	15.63	9.41	**60.22**	131,571.74	27
CACAHUATE	3.17	2.11	**66.74**	4.16	1.12	26.90	370,261.57	132
CALABAZA	14.53	7.11	48.90	6.96	76.69	**1,102.40**	3,828,761.56	327
CAMOTE	19.34	6.53	33.76	2.52	1.03	40.94	2,726.33	6

Tabla 11. Cálculo de diferentes métricas a nivel cultivo de la serie de datos #1.

La tabla 11 muestra también que existen porcentajes de desviación con respecto a la media para los atributos rendimiento y precio que exceden el 50 % (se muestran sombreados y en negrillas). Como se puede apreciar, algunas desviaciones resultan tan desproporcionadas, que con toda seguridad se puede afirmar que existe un fallo en la población de datos (atributo precio, cultivos "ALFALFA", "ALGODÓN", "AVENA" y "CALABAZA"). Más claramente, se puede observar que estas desviaciones altas están asociadas a un ingreso también muy alto, y que presuntamente, no corresponden con la realidad. La misma tabla, pero ya corregida se puede consultar en la sección de *limpieza de datos*, en el punto 4.3.2.4.

La técnica descrita es una definición para un dato atípico (outlier) definido en Knorr (1998), como un dato cuyas observaciones rebasan por más de tres desviaciones estándar a la media (obs > 3σ), para el caso de distribuciones normales. En la sección de *limpieza de datos* se proporciona más información al respecto.

Medida # 6: Mediana del ingreso neto de los cultivos

La mediana del ingreso neto acumulado de los cultivos para el período 1995-2001, fue calculada en 29,186.19 miles de pesos. Este valor obedece al cultivo maíz en su modalidad forraje, localizado en la posición 60 de un total de 119 cultivos presentes en la serie de datos #1 (una vez efectuado el proceso de limpieza que se describe en la sección 4.3.2.4.). Los 20 cultivos que aportan más a la producción desde el punto de vista económico en el período de tiempo indicado se muestran en la tabla 12.

Cultivo	Ingreso neto (M$)
TRIGO	10,798,870.56

MAIZ	10,790,012.71
CAÑA	9,149,318.37
TOMATE	8,308,341.89
ALFALFA	6,537,911.25
HORTALIZAS	5,788,274.43
CHILE	4,461,959.00
PAPA	4,105,963.32
VID	3,656,148.27
PASTO	2,496,608.35
SORGO	2,260,073.54
FRIJOL	2,210,084.39
ESPARRAGO	2,201,520.67
GARBANZO	1,852,849.95
CALABAZA	1,602,503.82
ALGODON	1,435,924.27
ARROZ	922,105.43
FORRAJE	830,218.74
PEPINO	827,648.25
CEBOLLA	816,260.34

Tabla 12. Los 20 cultivos que más contribuyeron al ingreso (período 1995-2001).

Ingreso neto y número promedio de cultivos por distrito de riego (medidas # 8 y #9)

La tabla 13 muestra el ingreso neto y el número de cultivos promedio de los distritos de riego en el período de 1995-2001 (medidas 8 y 9 de la tabla 9). Sólo se muestran los primeros 10 distritos, de un total de 84. A través de esta tabla se identifican los distritos que más ingreso han obtenido en el período de tiempo indicado, así como el número promedio de cultivos sembrados cada año.

No.	Distrito	Ingreso neto acumulado (M$)	Número de cultivos promedio por año
1	075	20,488,552.77	18
2	010	9,557,185.10	21
3	063	4,232,841.68	12
4	051	3,851,006.93	9
5	035	3,196,476.49	13
6	041	3,021,771.27	10
7	014	3,010,250.90	10
8	038	2,884,858.24	21
9	003	2,646,824.52	6
10	016	2,496,652.57	8

Tabla 13. Los 10 distritos que más contribuyeron al ingreso (período 1995-2001).

En general, se observa que hay cierta relación entre la variedad de cultivos y el ingreso neto que se obtiene. Aquellos distritos que perciben más ingresos netos tienen la característica que siembran una gran variedad de cultivos. Los distritos que perciben más ingresos (tabla

13) siembran en promedio 12 cultivos, mientras que los 20 distritos que percibieron menos siembran en promedio 5 cultivos al año.

4.2.3.2. Muestra de datos de una variable

El primer uso del histograma [27] en el proyecto, fue para probar que los valores de los atributos asociados a la producción agrícola siguen un comportamiento normal. De lo contrario, muchas de las métricas utilizadas para resumir datos del proyecto perderían efectividad (por ej. la media ó la desviación estándar).

La gráfica de la figura 11 muestra el número de registros que se agrupan por rangos específicos del atributo rendimiento. Los rangos fueron obtenidos utilizando la técnica de discretización que se describe en la sección 4.4.3.4.1.

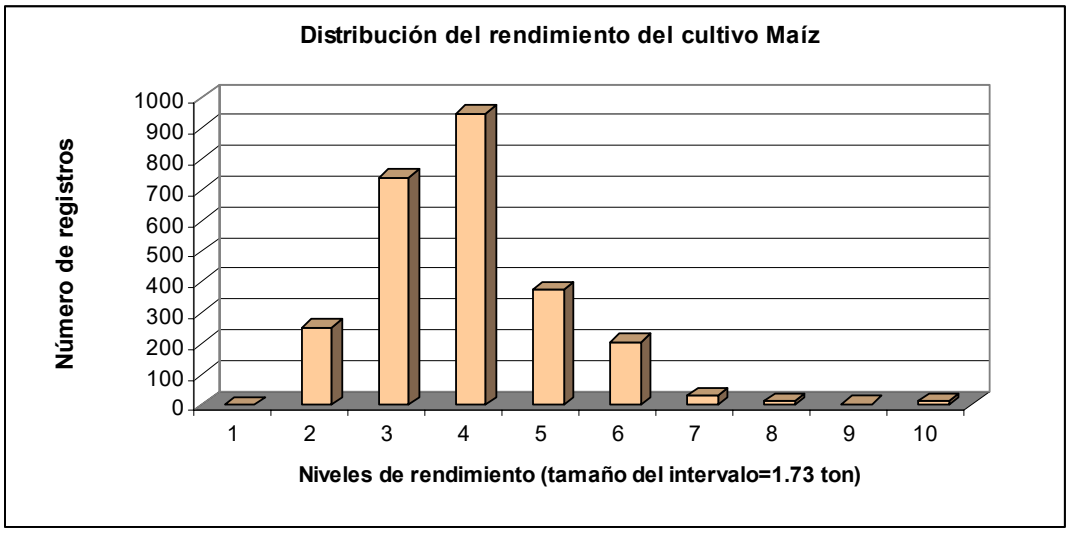


Figura 11. Distribución de datos para el atributo rendimiento.

La mayoría de los rendimientos de los cultivos presentaron un comportamiento similar al del cultivo maíz, que se muestra en la figura 11. Como se puede ver, las observaciones de los registros se acumulan alrededor de la media, denotando el comportamiento de una distribución normal.

Otro uso de los histogramas fue para obtener información de un aspecto desconocido del proyecto. Al inicio, se ignoraba que ciclos son los que concentran el mayor número de registros de siembra de cultivos. El histograma de la figura 12 muestra la concentración de registros por ciclo agrícola. El texto P-V corresponde a la abreviación de Primavera-Verano, para denotar a los cultivos que son sembrados entre estas dos estaciones del año. O-I es para otoño invierno y PER se utiliza para los perennes. Por medio de la visualización de la figura 12 es posible apreciar que en todos los años, el mayor número de siembras se da en el ciclo otoño-invierno.

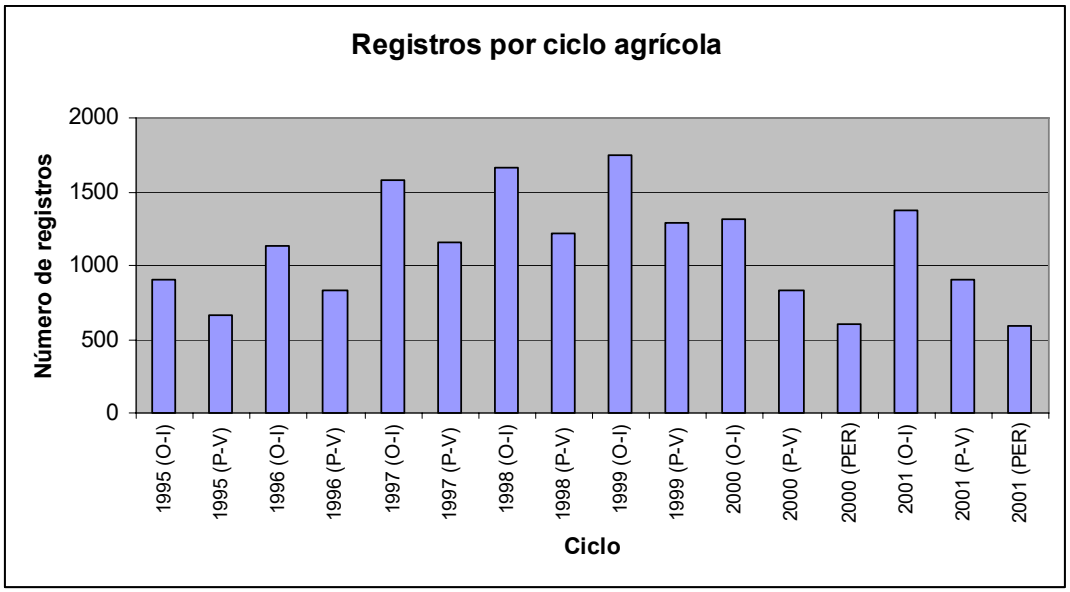


Figura 12. Concentración de registros por ciclo agrícola

También, por medio de la visualización del histograma, se puede denotar que de los distritos que más ingresos obtuvieron, son los distritos de riego 038 y 041 los que cuentan con un mayor número de registros en la serie de datos #1. Esto se puede apreciar en el histograma de la figura 13.

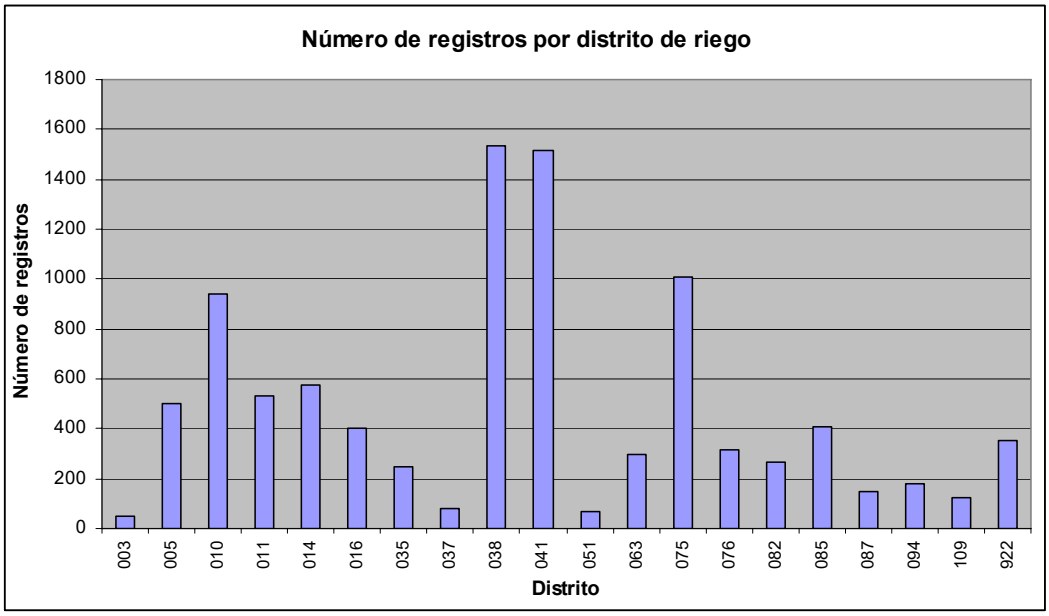


Figura 13. Concentración de registros por distrito de riego

A continuación se listarán las técnicas empleadas para mostrar relaciones entre dos variables.

4.2.3.3. Muestra de relaciones entre dos variables

Un gráfico de dispersión o scatterplot (Chambers, 1983) [56] revela relaciones o asociaciones entre dos variables [27]. La figura 13 muestra un gráfico de dispersión para los atributos lámina de riego y rendimiento para el cultivo maíz.

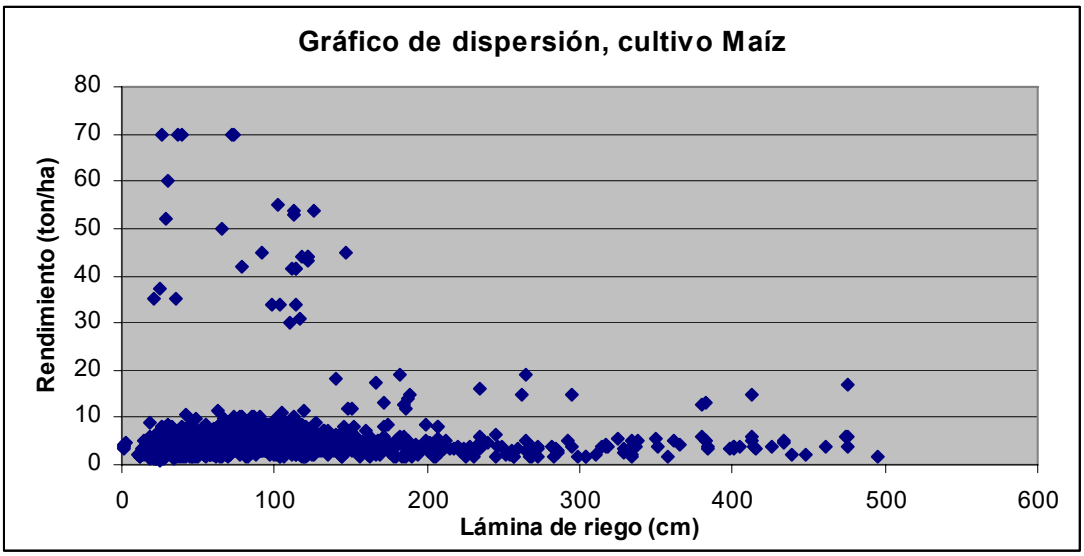


Figura 14. Gráfico de dispersión (lámina de riego/rendimiento) para el cultivo maíz

En la figura 14 se muestra que podría haber una relación entre la lámina de riego empleada y el rendimiento obtenido finalmente por el cultivo sembrado. Lamentablemente, esta relación no es apreciable a simple vista, debido a la densidad de puntos en el gráfico (2,612 registros). Esta problemática para el uso de los gráficos de dispersión ya había sido señalada por Hand et al, al mencionar que "*desafortunadamente, en minería de datos, los [scatterplots] no son siempre útiles. Si hay demasiados puntos de datos lo único que podremos observar será un rectángulo negro*" [27].

En el caso del proyecto, se utilizaron gráficas de este tipo para mostrar que el rendimiento de los cultivos variaba considerablemente de distrito a distrito. La figura 15 despliega los rendimientos promedio anuales del cultivo maíz para dos distritos de riego. Como se puede ver, en el mismo período de tiempo, el rendimiento para el mismo cultivo puede seguir un comportamiento totalmente distinto, siendo producto de factores particulares de cada distrito.

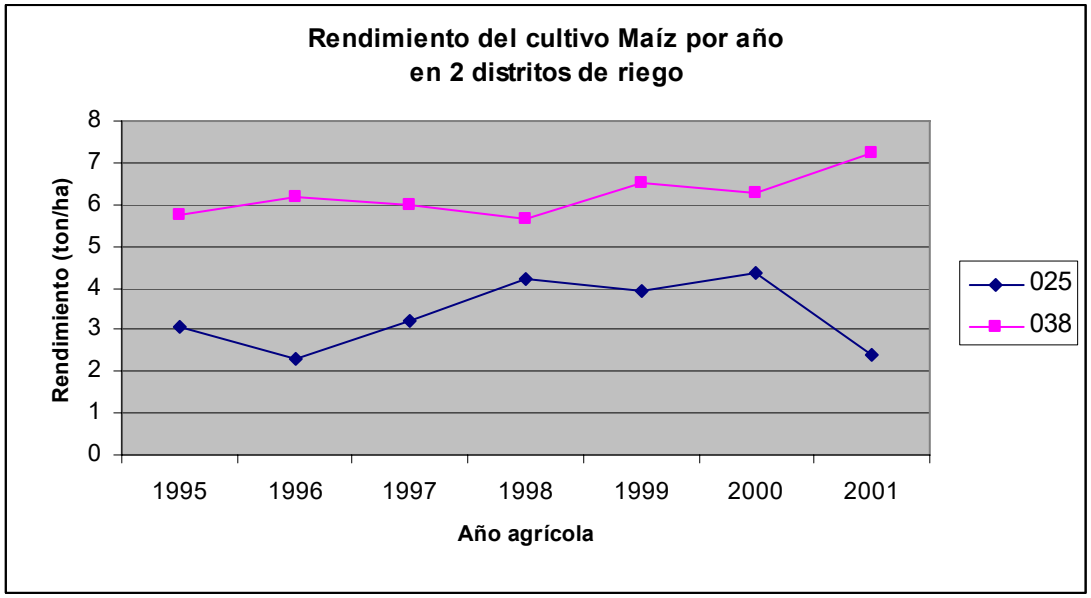


Figura 15. Gráfica del rendimiento promedio del maíz por distrito de riego por año.

Los datos de la serie de información climática (serie de datos #3) también pueden ser visualizados por medio de un gráfico XY con el atributo tiempo en el eje X. De hecho, el sistema de cómputo origen de dicha serie de datos proporciona estas gráficas como una opción de consulta de series de tiempo. Un ejemplo de este tipo de gráficas se da en la figura 16, donde se muestra el promedio mensual de las temperaturas observadas en el año 2000 en una estación climatológica ubicada en Cuernavaca, Morelos.

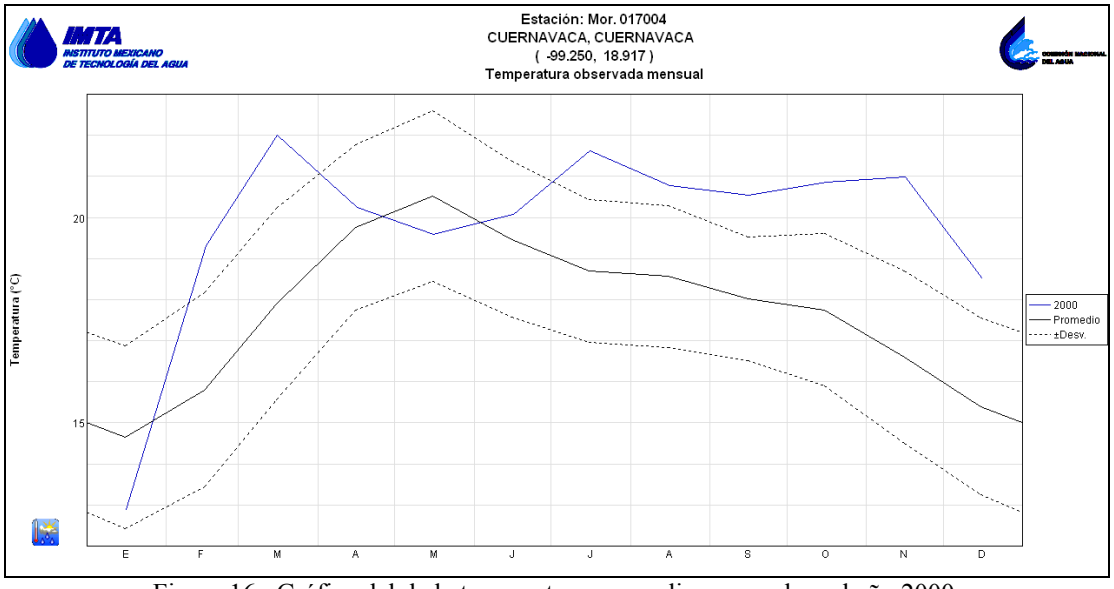


Figura 16. Gráfica del de la temperatura promedio mensual en el año 2000.

Los ejemplos aquí mostrados son sólo parte de los ejercicios realizados durante el desarrollo de esta tesis, los cuales son incluidos para ilustrar el tipo de procesamiento de la información que se da en las distintas etapas del método CRISP-DM.

El siguiente paso en la metodología consiste en la verificación de la calidad de los datos.

4.2.4. Verificación de la calidad de los datos

La etapa de *verificación de calidad de los datos* en CRISP-DM se dedica a describir los errores encontrados en la información recopilada para el proyecto. La calidad de datos es un tema muy complejo y no puede ser expuesto de manera detallada en esta tesis. Sin embargo, en el anexo 3, "Notas sobre la calidad de datos", se proporciona una introducción al tema que puede ser consultado para obtener más información al respecto.

Para describir los errores presentes en los datos se empleará la taxonomía propuesta en Kim et al (2001) [62]. Esta clasificación goza de una gran aceptación dentro de la terminología de *limpieza de datos*, y es usada con frecuencia para definir lo que es un "dato sucio". Una versión resumida de dicha taxonomía se muestra en la figura 17.

1 Dato faltante
1.1 Dato faltante donde no hay una condición de NULO-NO-PERMITIDO
1.2 Dato faltante donde una condición NULO-NO-PERMITIDO debe ser forzosa.
2 Dato existente, pero
2.1 Dato erróneo debido a
2.1.1 No cumplimiento de condiciones forzosas automáticas de integridad.
2.1.1.1 Condiciones de integridad soportadas en los sistemas de bases de datos relacionales actuales
2.1.1.1.1 Condiciones específicables por el usuario
2.1.1.1.1.1 Uso de un tipo de dato erróneo
2.1.1.1.1.2 Datos colgantes
2.1.1.1.1.3 Datos duplicados
2.1.1.1.2 Integridad garantizada a través del manejo de transacciones
2.1.1.1.2.1 Pérdida de actualización
2.1.1.1.2.2 Lectura sucia
2.1.1.1.2.3 Lectura irrepetible
2.1.1.1.2.4 Transacción perdida
2.1.2 Condiciones no forzosas de integridad
2.1.2.1 Error de entrada de datos involucrando un solo archivo o tabla
2.1.2.1.1 Error de entrada de datos que involucra a un solo campo
2.1.2.1.1.1 Entrada errónea
2.1.2.1.1.2 Falta de ortografía
2.1.2.1.1.3 Datos extraños
2.1.2.1.2 Error de entrada de datos que involucra a más de un campo
2.1.2.1.2.1 Entrada en el campo equivocado
2.1.2.1.2.2 Dato erróneo de un campo derivado
2.1.2.1 Inconsistencia a través de múltiples tablas/archivos
2.2 No error en el dato, pero es un dato inservible
2.2.1 Un dato diferente para la misma entidad a lo largo de múltiples bases de datos
2.2.2 Ambigüedad de datos, debido a
2.2.2.1 Uso de abreviaciones
2.2.2.2 Contexto incompleto
2.2.3 Conformación no estándar de los datos, debido a
2.2.3.1 Diferentes representaciones de un dato no compuesto
2.2.3.1.1 La transformación algorítmica no es posible
2.2.3.1.1.1 Abreviaciones
2.2.3.1.1.2 Alias/apodo (nick name)
2.2.3.1.2 La transformación algorítmica es posible
2.2.3.1.2.1 Formatos de codificación
2.2.3.1.2.2 Error de representación
2.2.3.1.2.3 Unidades de medición
2.2.3.2 Diferentes representaciones de un dato compuesto
2.2.3.2.1 Datos concatenados
2.2.3.2.1.1 Versión abreviada
2.2.3.2.1.2 Uso de caracteres especiales
2.2.3.2.1.3 Orden diferente
2.2.3.2.2 Datos jerárquicos

2.2.3.2.2.1	Versión abreviada
2.2.3.2.2.2	Uso de caracteres especiales
2.2.3.2.2.3	Orden diferente (ciudad-estado en lugar de estado-ciudad)

Figura 17. Taxonomía de los "datos sucios" por Kim et al (2001) [62]

Errores de datos presentes en la serie de datos #1. Estadísticas de producción agrícola SINHDR.

La fase de exploración de los datos aportó mucha información respecto al estado de los datos de este recurso. Las medidas empleadas para resumir datos y las técnicas de visualización revelaron que los atributos de este recurso se encontraban afectados por varios de los errores mostrados en la taxonomía de Kim (figura 17), de los cuales se muestran algunos ejemplos en la tabla 21

Atributo	Error	Cantidad de valores afectados	Comentarios
Cultivo	**2.1.2.1.1.2, 2.2.2.1**	2,084	Errores de mala ortografía en el nombre (por ej. ALGONON en lugar de ALGODON), abreviaciones para algunas palabras (AVENA FORR en lugar de AVENA FORRAJERA).
	2.1.2.1.1.3	9	Nombre de cultivo no identificable (por ej. AQUIL).
Superficie sembrada	**1.2**	385	Dato nulo.
	2.1.2.1.1.1	911	Datos con superficie sembrada en 0.
	2.1.2.1.1.1	995	Datos con superficie cosechada en 0.
Lámina de riego	**1.2**	1,786	Dato nulo.
	2.1.2.1.1.1	277	Lámina de riego en 0.
	2.2.3.1.2.3	375	Lámina de riego en metros. Detectadas por medio de desviaciones estándar.
	2.2.3.1.2.3	152	Lámina de riego en milímetros. Detectadas por medio de desviaciones estándar.
Costo de producción	**1.2**	3,171	Dato nulo.
	2.1.2.1.1.1	138	Costo de producción en 0.
	2.2.3.1.2.3	141	Costo de producción en pesos. Detectadas por medio de desviaciones estándar.
Rendimiento	**1.2**	2,028	Dato nulo.
	2.1.2.1.1.1	148	Rendimiento en 0.
	2.2.3.1.2.3	19	Rendimiento en kilogramos (por ej. 19000 en lugar de 19). Detectadas por medio de desviaciones estándar.
Precio	**1.2**	2155	Dato nulo
	2.1.2.1.1.1	135	Precio en 0
	2.2.3.1.2.3	80	Precio expresado en $/ton en lugar de M$/ton (por ej. 1500,1412, 1410, 1300,1200, 900 en lugar de 1.5, 1.412, 1.410, 1.300, 1.200, 0.90)

Tabla 14. Errores detectados en el recurso de datos #1.

La sección 4.3.2 describe la manera de como fueron manejados los errores mostrados en la tabla 14.

Errores presentes en la serie de datos #2: Información extraída de documentos electrónicos de estadísticas agrícolas CONAGUA.

- No se detectaron errores en este recurso de datos.

Errores presentes en la serie de datos #3: Información climatológica extraída del sistema ERIC III.

Antes de pasar por el proceso de limpieza, la información útil de esta serie de datos fue extraída y concentrada en hojas de cálculo (ver sección 4.3.1). El principal problema detectado en los datos extraídos fue la falta propia del dato, la que representa el sistema colocando una etiqueta "NO_D" en lugar del dato numérico. Tomando un valor "NO_D" como la representación de ausencia del valor, se diría que los errores detectados fueron del tipo **1.1** (según la taxonomía de la figura 17). La tabla 15 muestra el número de valores nulos detectados por mes y por variable. El número de observaciones (2,275) corresponde al número de variables (5), por el número de estaciones (13), por el número de años (35).

Variable	Obs. X mes	Valores nulos											
		Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Sep	Oct	Nov	Dic
Evaporación	455	341	345	343	344	348	344	342	351	348	347	343	344
Precipitación	455	306	306	310	308	309	309	307	317	310	309	310	310
Temperatura	455	326	328	328	330	330	328	327	336	329	325	325	329
Temperatura máxima	455	308	310	310	312	311	310	309	319	312	310	310	312
Temperatura mínima	455	308	310	310	312	312	310	309	319	312	310	310	312
Total	2,275	1,589	1,599	1,601	1,606	1,610	1,601	1,594	1,642	1,611	1,601	1,598	1,607
Valores no nulos		686	676	674	669	665	674	681	633	664	674	677	668

Tabla 15. Valores ausentes en la información extraída del sistema ERIC, recurso de datos #4.

El promedio de valores no nulos por mes es de 670, que representa el 29.45 % de las observaciones disponibles para todas las variables, presentándose una ausencia de datos del 70.54 %. Se espera que, a pesar de la gran falta de datos, la cantidad disponible sea suficiente para obtener un comportamiento definido de cada variable climática a lo largo del tiempo.

Errores de datos presentes en la serie de datos #4: Hojas de datos del distrito de riego 038.

El origen de la información juega un papel importante en la calidad de los datos. En el caso del recurso de datos #5, las hojas de datos habían formado parte de una metodología para la evaluación de desempeño de distritos de riego, para el cual, la veracidad de la información era crucial. Debido a esto, la información contenida en las hojas de datos manifestaba un nivel de calidad bastante alto, ya que no presentaba ninguno de los errores de la taxonomía mostrada en la figura 17.

4.3. Preparación de los datos

4.3.1. Selección de los datos

En esta fase seleccionan los datos serán utilizados para el análisis. El criterio de selección incluye la relevancia de los datos para las metas de la minería y las condiciones técnicas y de calidad (como restricciones en el tamaño ó tipos de datos) [33]. La selección de los datos es una tarea sumamente importante dentro del proceso de minería, y esto se debe a muchos factores. En Reinartz (2002)[63] se enfatiza que esta importancia se debe a que el tamaño de las bases de datos de hoy en día usualmente exceden el tamaño de los datos que los algoritmos de minería actuales pueden manejar. En John et al (1994) [64], se muestra como la selección de atributos irrelevantes afectan el desempeño y la precisión de los algoritmos de inducción del conocimiento.

Antes de pasar por el proceso de limpieza, la información útil de esta serie de datos fue extraída y concentrada en hojas de cálculo (ver sección 4.3.1). El principal problema detectado en los datos extraídos fue la falta propia del dato, la que representa el sistema colocando una etiqueta "NO_D" en lugar del dato numérico. Tomando un valor "NO_D" como la representación de ausencia del valor, se diría que los errores detectados fueron del tipo **1.1** (según la taxonomía de la figura 17). La tabla 15 muestra el número de valores nulos detectados por mes y por variable. El número de observaciones (2,275) corresponde al número de variables (5), por el número de estaciones (13), por el número de años (35).

Variable	Obs. X mes	Valores nulos											
		Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Sep	Oct	Nov	Dic
Evaporación	455	341	345	343	344	348	344	342	351	348	347	343	344
Precipitación	455	306	306	310	308	309	309	307	317	310	309	310	310
Temperatura	455	326	328	328	330	330	328	327	336	329	325	325	329
Temperatura máxima	455	308	310	310	312	311	310	309	319	312	310	310	312
Temperatura mínima	455	308	310	310	312	312	310	309	319	312	310	310	312
Total	2,275	1,589	1,599	1,601	1,606	1,610	1,601	1,594	1,642	1,611	1,601	1,598	1,607
Valores no nulos		686	676	674	669	665	674	681	633	664	674	677	668

Tabla 15. Valores ausentes en la información extraída del sistema ERIC, recurso de datos #4.

El promedio de valores no nulos por mes es de 670, que representa el 29.45 % de las observaciones disponibles para todas las variables, presentándose una ausencia de datos del 70.54 %. Se espera que, a pesar de la gran falta de datos, la cantidad disponible sea suficiente para obtener un comportamiento definido de cada variable climática a lo largo del tiempo.

Errores de datos presentes en la serie de datos #4: Hojas de datos del distrito de riego 038.

El origen de la información juega un papel importante en la calidad de los datos. En el caso del recurso de datos #5, las hojas de datos habían formado parte de una metodología para la evaluación de desempeño de distritos de riego, para el cual, la veracidad de la información era crucial. Debido a esto, la información contenida en las hojas de datos manifestaba un nivel de calidad bastante alto, ya que no presentaba ninguno de los errores de la taxonomía mostrada en la figura 17.

4.3. Preparación de los datos

4.3.1. Selección de los datos

En esta fase seleccionan los datos serán utilizados para el análisis. El criterio de selección incluye la relevancia de los datos para las metas de la minería y las condiciones técnicas y de calidad (como restricciones en el tamaño ó tipos de datos) [33]. La selección de los datos es una tarea sumamente importante dentro del proceso de minería, y esto se debe a muchos factores. En Reinartz (2002)[63] se enfatiza que esta importancia se debe a que el tamaño de las bases de datos de hoy en día usualmente exceden el tamaño de los datos que los algoritmos de minería actuales pueden manejar. En John et al (1994) [64], se muestra como la selección de atributos irrelevantes afectan el desempeño y la precisión de los algoritmos de inducción del conocimiento.

A un nivel abstracto, existen dos diferentes criterios para realizar la selección de datos: la relevancia y la representatividad [63]. En general, el criterio basado en la relevancia restringe la selección de los datos a la información que es apropiada para alcanzar el objetivo de la minería. En contraste, el criterio basado en la representatividad se asegura que la selección represente en su totalidad al conjunto de datos original (beneficie esto o no al desempeño del algoritmo). De manera más específica, En John et al (1994) [64] se indica que existen diferentes grados de relevancia, la relevancia débil y la relevancia fuerte. Una relevancia fuerte implica que un registro es indispensable en el sentido que no puede ser removido sin pérdida de exactitud en la predicción. Una relevancia débil implica que el registro puede contribuir algunas veces a la exactitud de la predicción.

En el caso del proyecto de predicción del rendimiento, el principal criterio utilizado para la selección de los datos fue el de relevancia, aunque en algunos casos (como la selección de información climática) se utilizó el criterio de representatividad. A continuación se describirá la manera en como se seleccionaron los datos para el proyecto. Esto de manera organizada por cada recurso de datos disponible en la información concentrada.

Selección de datos para la serie de datos #1: Estadísticas de producción agrícola SINHDR.

A diferencia de las otras series de datos seleccionadas para el proyecto, la serie de datos #1 fue sometida en su totalidad al proceso de *limpieza de datos*, esto de manera previa a la *selección de datos*. La razón fue que la *exploración de datos* reveló que esta serie se encontraba afectada por una gran cantidad de errores, por lo que fue necesario "limpiarla". Esta serie ha sido identificada como la más importante en el conjunto de información obtenida para el proyecto. Es la que aporta más información para el problema en cuestión, esto de manera horizontal (en número de atributos) y vertical (en número de registros).

El objetivo principal del proyecto de minería llevado por esta tesis es la obtención de los modelos que mejor se ajusten al problema de predicción del rendimiento de cultivos en distritos de riego. Se asumió que la metodología obtenida para generar los mejores modelos en un distrito de riego con seguridad funcionaría para cualquier otro. De esta manera, se decidió que la selección de datos se basara en la información de un solo distrito.

El distrito de riego seleccionado debe ser representativo desde el punto de vista económico. Un distrito de riego con ingresos altos proporciona más información del beneficio económico obtenido por sus cultivos que uno con bajos ingresos, esto debido a la información implícita en sus registros.

La variedad de cultivos es también un factor importante. Se ha visto como el número de cultivos está asociado hasta cierto punto con una mejoría en el ingreso neto. El contar con un modelo o conjunto de modelos que contemplen un número variado de cultivos abrirá las posibilidades de pruebas y de la obtención de mejores conclusiones.

En los distritos de riego, de un año a otro cambian las especificaciones para siembra, y cultivos que se han sembrado una vez podrían no volverse a sembrar en un período largo de tiempo. Con el fin de maximizar la utilidad y la efectividad de los modelos, se decidió

restringir éstos a los cultivos que presenten información en al menos 5 años de los 7 disponibles para todos los distritos en la serie de datos #1.

Otro factor importante a considerar, es el número de registros. El número de registros influye considerablemente en el desempeño de los algoritmos de minería de datos. El distrito a seleccionar debe contar con un número aceptable de registros.

Por las razones anteriormente expuestas, se decidió utilizar los registros pertenecientes al distrito de riego 038. Este distrito se encuentra entre los que más ingreso neto obtienen (posición #8 de la tabla 13). En esta misma serie de datos, el distrito 038 es el único que tiene información en los años 2002 y 2003. De la misma manera, el distrito tiene un número promedio de cultivos alto, sembrando en promedio 21 cultivos al año. Además, por el histograma de la figura 11, se puede apreciar que es el distrito para el cual se poseen más registros. Como un beneficio extra, se cuenta con información adicional de este distrito, aunque de origen no está en formato estructurado (recurso de datos hojas de datos #5). En su momento, esta información se utilizará para elaborar un conjunto de prueba y para validar los resultados de los modelos.

Se decidió también, utilizar únicamente los registros del ciclo otoño-invierno. La figura 12 indica que es en este ciclo donde se concentran la mayoría de los registros, y a fin de evitar agregarle complejidad al modelo, se decidió eliminar el atributo ciclo como un atributo de peso y acotar el modelo únicamente a los registros correspondientes de este ciclo. Desde luego, en las conclusiones generadas a partir de los modelos deberá considerarse que: a) son exclusivas del distrito de riego 038 y b) se atañen únicamente al ciclo otoño-invierno.

Para la selección de los cultivos se utilizó una consulta anidada que proporcionó el número de años en los que aparecía un cultivo específico en el distrito de riego 038, ciclo otoño-invierno. La tabla 16 muestra algunos cultivos del distrito y el número de años en los que aparecen, resaltando en gris a los cultivos que aparecen en 5 o más años.

Cultivo	Número de registros	Número de años en los que aparece
ALFALFA	**63**	**9**
ALGODON	**21**	**6**
AVENA	3	2
CARTAMO	**132**	**9**
CEBADA	1	1
CEBOLLA	10	2
CHICHARO	**21**	**9**
CHILE	18	2
EJOTE	7	2
FORRAJE	**11**	**5**
FRIJOL	**112**	**9**
FRUTALES	**23**	**5**
GARBANZO	**81**	**9**
GIRASOL	1	1
HORTALIZAS	**96**	**9**

HUERTA	1	1
LIRIO	2	2
MAIZ	**143**	**9**
PAPA	**89**	**9**
PAPAYA	1	1
SANDIA	5	2
SORGO	**45**	**8**
TOMATE	**51**	**5**
TRIGO	**138**	**9**
ZACATE	**9**	**6**
ZANAHORIA	2	2
ZEMPOALXOCHITL	7	4

Tabla 16. Cultivos sembrados en el distrito de riego 038 y número de años en los que el cultivo participó en la producción (1995-2003).

De la muestra de la tabla 16, se desprende que los cultivos seleccionados para el desarrollo de los modelos fueron: alfalfa, algodón, cártamo, chícharo, forraje, fríjol, frutales, garbanzo, hortalizas, maíz, papa, sorgo, tomate, trigo y zacate; 15 cultivos, haciendo un total de 1,035 registros, los cuales representan el 5.83 % del número de registros de la serie de datos #1 después del proceso de limpieza (ver punto 4.3.2), y el 4.99 % de la misma serie pero en su estado original (sin limpieza).

Por el lado de los atributos, de los atributos indicados en la etapa de *descripción de datos* (Anexo 2, tabla A2.1) se decidió eliminar al módulo con el fin de hacer más general el modelo. Otros atributos eliminados fueron industria y destino, que no influyen en el rendimiento (ni en ningún otro atributo de la serie). Puesto que se sabe que el modelado se realizaría para un solo distrito, se eliminó el atributo distrito. Al final, de los 12 atributos originales, se dejaron 8. Los atributos seleccionados fueron los siguientes: ciclo, cultivo, superficie sembrada, superficie cosechada, lámina de riego, costo de producción, rendimiento y precio. Pese a que se decidió evaluar los modelos para los cultivos sembrados en el ciclo otoño-invierno, el atributo ciclo fue dejado en la selección de datos ya que además de contener el ciclo contiene también el año agrícola, y este dato es importante para empatar la información de la serie con la información climática.

Selección de datos para la serie de datos #2: Documentos electrónicos de estadísticas agrícolas CONAGUA.

La selección de la información de la serie #1 abarca el período de tiempo de 1995 a 2003 para el distrito de riego 038. Se pretende robustecer este conjunto de datos hasta complementar el año 2006, con el fin de dedicar la información de los años 2004, 2005 y 2006 para probar los modelos desarrollados. De la serie de datos #2 se extrae la información para complementar los años 2004 y 2005, aunque existen dos atributos de la serie de datos #1 que no están presentes en la serie de datos #2: la lámina de riego y el costo de producción (ver sección de "descripción de los datos"). Los valores correspondientes a estos atributos serán derivados a través técnicas específicas para rellenar "huecos" en la información, tal como se describe en la sección 4.3.2.

De los archivos PDF fueron extraídos 51 registros. 22 correspondientes al año agrícola 2003-2004 y 29 del año agrícola 2004-2005. Los atributos fueron los mismos que para la serie #1, con la excepción de la lámina de riego, que no estaba presente. Después del filtro de los cultivos (tabla 16), los registros seleccionados quedaron en 21, 11 pertenecientes al año agrícola 2003-2004 y 10 para el año agrícola 2004-2005.

Selección de datos para la serie de datos #3: Sistema de consulta de información climatológica, ERIC III.

Hasta este punto, esta serie de datos no ha sido explorada en su totalidad. Esto debido principalmente a que el acceso a la base de datos del sistema de información climatológica no es directo, y se requiere llevar a cabo un proceso de exportación, depuración, formateo e importación para poder acceder a la información de manera estructurada. El costo de hacer lo anterior para toda la base de datos, un millón y medio de registros, no estaba justificado, debido a que posiblemente sólo se requeriría utilizar aquellos registros que empataran geográficamente con los distritos seleccionados en las otras series de datos.

Para extraer información del sistema ERIC III [52], es preciso indicar las estaciones climatológicas para las cuales se efectuará la consulta. Las estaciones climatológicas están distribuidas a lo largo de todo el país (ver figura 18), por lo que únicamente hace falta especificar cuáles están ubicadas en el área de interés. Para ello, el sistema permite seleccionar directamente la estación a través de una interfaz gráfica (utilizando el mapa de la figura 18), o bien, seleccionarla de manera directa a través de una lista de estaciones.

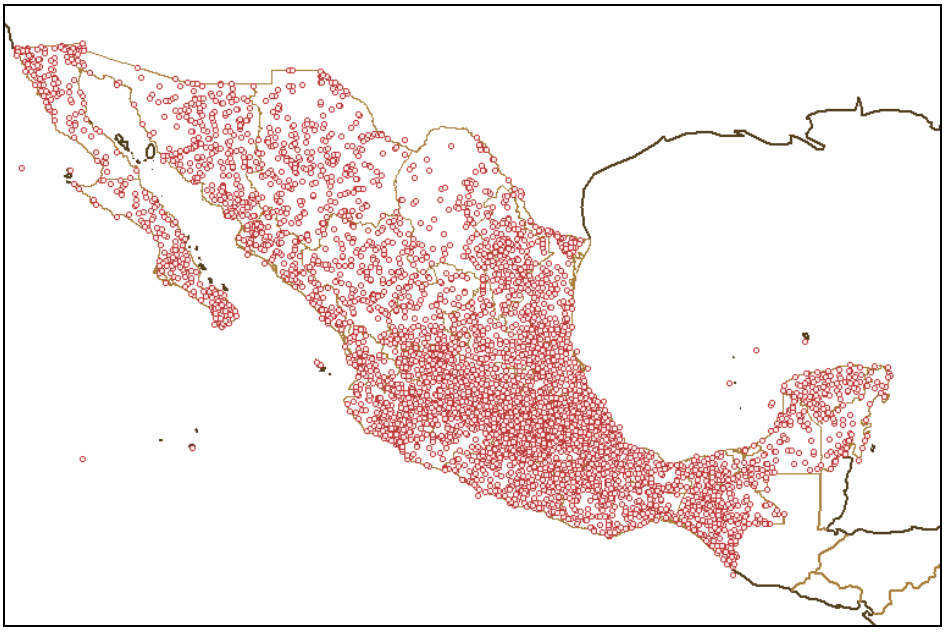


Figura 18. Mapa de estaciones climatológicas (sistema ERIC III).

Para obtener la lista de estaciones que se encontraban dentro de la zona del distrito de riego 038, se tuvo que obtener las coordenadas geográficas del distrito y comparar éstas con las coordenadas de las estaciones almacenadas en la base de datos del sistema. Para no realizar esta operación de manera manual, se exportaron las estaciones y sus coordenadas a una

hoja de cálculo en Microsoft Excel®, y por medio de una función de comparación se determinaron qué estaciones estaban dentro la zona de influencia del distrito de riego. El distrito de riego 038 se localiza entre las latitudes 25°45' y 27°15', y las longitudes 110°00' y 109°30' oeste. La lista de estaciones que se localizan dentro del área del distrito y sus coordenadas se muestran en la tabla 24.

ID	Localización	Estado	Longitud (W)	Latitud
26010	VILLA JUAREZ (A.BATEVITO)	Son.	-109.817	27.167
26034	ETCHOJOA, ETCHOJOA	Son.	-109.633	26.917
26044	HUATABAMPO, HUATABAMPO	Son.	-109.650	26.833
26051	EL LINDERO, NAVOJOA DGE	Son.	-109.750	26.900
26060	NACHUQUIS, NAVOJOA (DGE)	Son.	-109.533	27.067
26129	HUATABAMPO FF.CC.,	Son.	-109.617	26.833
26159	LUIS FF.CC., HUATABAMPO	Son.	-109.583	26.800
26194	SAN PEDRO, ETCHOJOA	Son.	-109.617	27.017
26211	YAVAROS, HUATABAMPO	Son.	-109.650	26.833
26282	JUPARE, HUATABAMPO	Son.	-109.750	26.767
26402	BASCONCOBE, ETCHOJOA	Son.	-109.567	26.921
26407	SEBAMPO, ETCHOJOA	Son.	-109.589	26.818
26419	HUATABAMPITO, HUATABAMPO	Son.	-109.700	26.767

Tabla 17. Estaciones climatológicas dentro del área del distrito

El siguiente paso fue seleccionar los atributos climáticos de interés para el proyecto. De los atributos descritos en la sección "descripción de los datos", se determinó que la temperatura promedio, la temperatura mínima, la temperatura máxima, la precipitación y la evaporación serían los atributos de influencia tomados en cuenta para el modelo. El granizo, la niebla y la cobertura del cielo fueron desestimados, ya que son fenómenos del clima aún más difíciles de predecir que la temperatura, la precipitación y la evaporación, por lo que su uso en el modelo se ve poco probable.

Otra cuestión a resolver, era el nivel de detalle de la información climática a integrar en el modelo. La información estaba disponible a nivel de detalle diario, pero un modelo que buscara predecir el rendimiento a futuro de los cultivos requiriendo para ello la estimación de la temperatura diaria (por ejemplo), resultaría bastante complicado de satisfacer por parte del usuario. Sin embargo, un pronóstico mensual de cualquiera de los atributos climáticos mencionados estaría al alcance de cualquier usuario, y esta fue la razón para incluir la información climática resumida a nivel mensual.

A pesar de que ya existía el antecedente para seleccionar únicamente el período de tiempo que comprende de 1995 a 2006[9] (esto dada la información de producción agrícola disponible), se decidió seleccionar de entrada un período más largo de tiempo con el fin de contar con información disponible en el caso de requerirse la construcción de algún modelo matemático para la estimación de valores faltantes en la serie. De esta manera, se seleccionó información del clima recopilada desde 1970 hasta el último año almacenado en

[9] Se puede ver en la selección de información de las series de datos #1 y #2.

el sistema ERIC III, el cual corresponde al año 2004. Lo anterior para las estaciones seleccionadas.

Las variables del clima de las estaciones fueron exportadas a archivos de formato texto (única opción ofrecida por el sistema). La información resumida mensual fue tomada de estos archivos y llevada a hojas de cálculo de Microsoft Excel con el fin de determinar su grado de completes y contabilizar los registros. En total, se obtuvieron 2,275 filas de información del clima, la cual incluía los datos temperatura, precipitación, evaporación, estación, año (1970-2004) y los valores para las observaciones en los meses del año de referencia.

La sección de *verificación de la calidad de los datos* describe el estado de la información extraída del sistema ERIC III. La etapa de *limpieza de datos* describe la forma en como fueron corregidos los errores detectados en la serie de datos climáticos y la fase *construcción de datos* la forma en como nuevos valores fueron agregados a las serie.

Nota: Previo a la integración de datos fue necesario hacer una nueva selección de información, pero esto únicamente para delimitar el período de tiempo de forma que empatara con el período correspondiente a la información climática.

Selección de datos para la serie de datos #4: Hojas de datos del distrito de riego 038.

Estas hojas de datos poseen un apartado de información de producción agrícola, en la cual se incluyen los mismos atributos que el recurso de datos #1, esto para los ciclos primavera-verano, otoño-invierno y perennes del año agrícola 2006. Cada hoja representa la información de un módulo de riego. De estas hojas se extrajeron 178 registros correspondientes a la información de producción del ciclo otoño-invierno. De éstos, 130 correspondían a los cultivos seleccionados del recurso de datos #1.

La información extraída de este recurso, junto la información del recurso #3, conforman el conjunto de prueba para los modelos desarrollados. Juntos, abarcan el período de información de los años 2004, 2005 y 2006.

4.3.2. Limpieza de datos

Un gran problema que apenas esta empezando a ser reconocido es el hecho de que los datos en las fuentes de datos origen están usualmente "sucios" [62]. De manera amplia, los datos "sucios" incluyen datos faltantes, datos erróneos y datos que no guardan una representación estándar entre ellos.

Una vez que un dato ha sido reconocido como "dato sucio", lo siguiente será emplear un método acorde al tipo de error que permita su corrección, o de lo contrario, optar por la eliminación del dato. De esta manera, (Maletic, 2005) [61] señala que en el proceso de limpieza de datos se define en tres fases:

- Definir y determinar los tipos de errores.
- Buscar e identificar instancias de error.

- Corregir los errores descubiertos

Las primeras dos fases ya han sido abordadas en secciones anteriores. Los tipos de errores se abordaron en la sección 4.2.4 correspondiente a la etapa de *verificación de la calidad de los datos*, mientras que la búsqueda e identificación de las instancias de error presentes en los datos se abordaron desde la etapa de *exploración de los datos*. Esta sección estará dedicada a la corrección de los errores detectados.

Muchas son las técnicas existentes para abordar los errores presentes en los datos. En el anexo 4 se proporciona una descripción general de los métodos utilizados con más frecuencia para llevar a cabo la limpieza de los datos. A continuación, se describe la manera en como se corrigieron los errores presentes en los recursos de datos del proyecto.

4.3.2.1. Tratamiento de los errores de datos detectados en el proyecto

En la sección de "calidad de los datos" se identificaron los tipos de errores que presentaban los recursos de datos obtenidos. De manera resumida, los tipos de errores detectados en los recursos se muestran en la tabla 18.

Tipo de error[10]	Descripción
1.1	Dato faltante donde no hay una condición de NULO-NO-PERMITIDO
1.2	Dato faltante donde una condición NULO-NO-PERMITIDO debe ser forzosa.
2.1.2.1.1.2	Falta de ortografía
2.2.2.1	Uso de abreviaciones
2.1.2.1.1.3	Datos extraños
2.1.2.1.1.1	Entrada errónea
2.2.3.1.2.3	Unidades de medición

Tabla 18. Tipos de errores presentes en los recursos de datos

Agrupando los errores de la tabla 18 en las técnicas de solución expuestas en el anexo 4, los errores tipo 1.2 y 1.3 se resolverían con las técnicas de "datos faltantes". Los errores tipo 2.1.2.1.1.2 y 2.2.2.1 con las técnicas de "registros duplicados" (para ser más precisos, con la "desambiguación de nombres"), y los errores 2.1.2.1.1.1, 2.1.2.1.1.3 y 2.2.3.1.2.3 se resuelven con las técnicas de "datos atípicos".

Errores de datos presentes en la serie datos #1. Estadísticas de producción agrícola SINHDR.

Atributo	Error	Cantidad de valores afectados	Comentarios	Descripción del método de limpieza de datos utilizado
Distrito	**1.2**	1	Dato nulo.	Eliminación del registro.
Cultivo	**2.1.2.1.1.2, 2.2.2.1**	2,084	Error de mala ortografía en el nombre (por ej. ALGONON en lugar de ALGODON), abreviaciones para algunas palabras (AVENA FORR en lugar de AVENA FORRAJERA).	Se empleó una técnica de agrupamiento (clustering) basado en la similaridad del nombre, con el nombre correcto como centroide. Posteriormente, se cambiaron los nombres de los cultivos agrupados por el del centroide correspondiente.

[10] De acuerdo a la clasificación de Kim et al (2001) [62] indicada en la sección 4.2.4

				Sin embargo, hubieron nombres de cultivos para los que fue necesario consultar a terceros con el fin de saber si existían nombres equivalentes en la serie de datos.
	2.1.2.1.1.3	828	Nombre de cultivo no identificable (por ej. AQUIL).	Eliminación de los registros.
Superficie sembrada	**1.2**	385	Dato nulo.	Eliminación de los registros.
	2.1.2.1.1.1	911	Datos con superficie sembrada en 0.	Eliminación de los registros.
Lámina de riego	**1.2**	1,786	Dato nulo.	Uso del dato promedio de acuerdo al cultivo.
	2.1.2.1.1.1	277	Lámina de riego en 0.	Uso del dato promedio de acuerdo al cultivo.
	2.2.3.1.2.3	375	Lámina de riego en metros. Detectadas por medio de desviaciones estándar.	Multiplicación por 100 del valor, para expresarlo de manera uniforme en cm.
	2.2.3.1.2.3	152	Lámina de riego en milímetros. Detectadas por medio de desviaciones estándar.	División por 10 del valor para expresarlo de manera uniforme en cm.
Costo de producción	**1.2**	3,171	Dato nulo.	Uso del dato promedio, de acuerdo al cultivo y al año.
	2.1.2.1.1.1	138	Costo de producción en 0.	Uso del dato promedio, de acuerdo al cultivo y al año.
	2.2.3.1.2.3	141	Costo de producción en pesos. Detectadas por medio de desviaciones estándar.	División del valor entre 1000, para expresarlo de manera uniforme en M$/ha.
Rendimiento	**1.2**	2,028	Dato nulo.	Eliminación del registro. Es el atributo a ser modelado, no es aconsejable usar un valor estimado.
	2.1.2.1.1.1	148	Rendimiento en 0.	Eliminación del registro. Es el atributo a ser modelado, no es aconsejable usar un valor estimado.
	2.2.3.1.2.3	19	Rendimiento en kilogramos ó en dimensiones desconocidas (por ej. 19000 en lugar de 19). Detectadas por medio de desviaciones estándar.	División del valor entre 1000, para expresarlo de manera uniforme en Ton/ha. Los registros con valor atípico al que no se le encontró ninguna correspondencia fueron eliminados.
Precio	**1.2**	2,155	Dato nulo	Uso del dato promedio, de acuerdo al cultivo y al año.
	2.1.2.1.1.1	135	Precio en 0	Uso del dato promedio, de acuerdo al cultivo y al año.
	2.2.3.1.2.3	80	Precio expresado en $/ton en lugar de M$/ton (por ej. 1500,1412, 1410, 1300,1200, 900 en lugar de 1.5, 1.412, 1.410, 1.300, 1.200, 0.90)	División del valor entre 1000, para expresarlo de manera uniforme en M$/Ton.

Tabla 19. Errores presentes en el recurso de datos #1 y la solución empleada.

En la tabla 19, se puede observar que se eliminaron 2,276 registros de cultivos que presentaban un rendimiento con valor 0, vacío o atípico. El rendimiento es un atributo crítico (a diferencia de la lámina de riego, el precio y el costo de producción), ya que será el atributo a predecir en la etapa de modelado. No es recomendable utilizar una técnica para reemplazar los valores faltantes en el atributo objetivo, ya que se influye (generalmente de manera negativa) en el desempeño final del modelo.

Después de las correcciones, la serie de datos #1 quedó con 17,962 registros, pero a diferencia de los registros originales, los atributos de estos registros presentaban menos ruido, y por lo tanto, una mejor calidad. Como ejemplo, la consulta anteriormente mostrada en la tabla 11 (rendimientos, precios promedio y desviaciones por cultivo), se muestra actualizada en la tabla 20. Como se podrá observar, las desviaciones respecto a la media son ahora mucho menores. Se siguen resaltando las desviaciones que exceden en 50% a la

media, pero el proceso de limpieza reveló que estas desviaciones hasta cierto grado eran normales. De hecho, para el caso de ambos atributos (rendimiento y precio) se consideraron normales las desviaciones del hasta un 150% (esto en base a la revisión hecha a los atributos que presentaban tales desviaciones). Anteriormente, en el atributo rendimiento se habían observado desviaciones de hasta un 160%, y en el atributo precio de 1,102.40 %, esto como producto de distintos errores de datos que afectaban a la serie.

Cultivo	Rendimiento (Ton/ha)			Precio (M$/Ton)			Ingreso Bruto (M$)	Cantidad
	Promedio	Desviación estándar	% de desviación con respecto a la media	Promedio	Desviación estándar	% de desviación con respecto a la media		
AGUACATE	7.22	5.95	**82.35**	4.62	2.22	47.97	21,669.75	24
ALFALFA	32.75	29.05	**88.70**	0.97	0.67	**69.49**	10,889,293.99	932
ALFALFA (SECA)	2.10	0.42	20.20	1.70	0.14	8.32	70.80	2
ALGODON	2.90	0.93	32.18	4.13	1.13	27.42	6,494,575.68	776
AVENA	13.90	10.15	**73.02**	0.82	1.17	**143.25**	365,558.33	566
AVENA (SECA)	3.19	1.81	**56.82**	1.62	0.68	42.12	4,905.08	25
BERENJENA	32.96	4.94	14.98	2.52	0.67	26.67	191,375.18	17
BERMUDA	0.41	0.13	31.15	24.61	6.55	26.63	131,571.74	27
BROCOLI	12.37	3.52	28.50	3.24	1.52	46.86	268,127.47	47
CACAHUATE	3.17	2.11	**66.74**	4.16	1.12	26.90	370,261.57	132
CALABAZA	14.53	7.11	48.90	2.71	2.23	**82.04**	2,267,699.19	327
CAMOTE	19.34	6.53	33.76	2.52	1.03	40.94	2,726.33	6

Tabla 20. Cálculo de diferentes métricas a nivel cultivo de la serie de datos #1 después del proceso de limpieza de datos.

Otro ejemplo de mejoría en la calidad de datos se puede ser la consulta del ingreso por cultivo para el cultivo Maíz. Una vez que la ambigüedad de nombres ha sido resuelta, los resultados son mucho más claros.

Nombre del cultivo	Ingreso (M$)	Cantidad
MAIZ	29,424,109.41	2,668
MAIZ (SEMILLA)	488.4	1
MAIZ FORRAJE	89,554.84	22

Tabla 21. Consulta del ingreso bruto del cultivo maíz.

<u>Serie de datos #3:</u> Información climatológica.

En la sección correspondiente a la calidad de los datos se había expuesto que el único problema detectado era la ausencia de datos que se presentaban en los registros correspondientes a las observaciones reportadas por cada estación climatológica.

Una vez que la información estuvo estructurada en la hoja de cálculo, ésta fue importada al manejador de base de datos Microsoft Access® con el fin de realizar una operación de resumen de datos agrupados utilizando la media aritmética, a manera de desaparecer el atributo estación. Los 2,275 registros fueron agrupados en 175, con 35 registros (uno por año) por cada una de las 5 variables disponibles. La tabla 22 muestra un extracto de los registros, mostrando datos de 1990 al año 2004 de la variable temperatura promedio, para

todos los meses del año. Como se puede apreciar, visto de manera agrupada, son pocos los años que quedan sin un valor promedio.

Año	Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Sep	Oct	Nov	Dic
1980	12.09	14.17	14.67	17.47	22.85	27.79	28.00	26.23	25.74	23.19	16.22	13.31
1981	14.34	13.26	14.56	19.47	22.96	26.51	28.37	28.91	27.16	22.71	16.91	12.10
1982	11.21	11.76	14.18	18.33	21.97	26.46	29.20	28.70	26.47	20.18	17.05	13.32
1983	13.57	12.99	15.10	16.90	21.02	25.37	28.47	26.63	27.31	22.58	16.30	12.97
1984	11.81	11.51	13.40	17.11	23.07	26.72	26.60	26.51	25.93	21.30	15.69	14.31
1985	12.00	12.17	14.27	18.53	20.59	26.24	27.37	27.30	26.41	21.44	15.27	10.97
1986	13.09	13.10	15.39	18.84	20.70	26.60	27.20	27.32	24.67	21.12	17.93	12.56
1987	9.35	11.10	12.28	16.95	19.66	24.56	27.53	25.82	25.56	22.99	16.57	11.79
1988	12.00	13.70	14.28	18.60	21.40	26.35	26.82	27.70	26.20	23.00	15.86	12.17
1989	10.00	13.43	15.23	19.03	21.60	24.85	28.25	26.65	27.05	23.10	17.85	12.10
1990	11.40	11.95	14.60	20.30	23.00	27.95	26.55	25.80	25.50	20.95	16.80	13.05
1991	10.10	16.30	14.70	16.35	22.10							
1992												
1993									23.90	20.50	16.15	12.70
1994	10.30	10.10	12.95	14.05	16.60	24.30	26.35	25.90	25.55	21.05	15.45	12.45
1995	9.60	13.15	12.95	12.30	16.70	20.50	26.05	27.45	25.45	20.25	15.95	10.10
1996	7.85	11.80	12.50	14.30	18.15	23.90	25.65	25.65	24.30	20.80	15.25	12.40
1997	12.35	9.20	13.10	16.05	21.05	23.85	26.65	26.45	25.50	20.80	16.50	9.50
1998	10.30	10.00	10.80	11.10	15.00	21.20	26.00	26.40	24.90	21.80		
1999												
2000						26.10	26.90	25.80	24.60	20.60	11.70	10.20
2001	9.50	9.60	10.70	13.20	19.20	23.50	25.80	25.70	25.10	19.50	15.40	8.20
2002	8.30	10.60	10.70	15.20	22.50	27.50	29.10	28.80	28.30	22.20	13.90	9.80
2003	11.20	12.50	10.60	13.70	19.50	22.40	27.40	27.40	26.80	22.80	14.80	8.50
2004	9.7	8.7	12.4	14	18.9							

Tabla 22. Promedio de temperaturas agrupados por año/mes de 1980 al año 2004.

Las restantes variables climatológicas (temperatura máxima, temperatura mínima, precipitación, evaporación) presentaban un estado similar al que se muestra en la tabla 18.

Una de las técnicas descritas como auxiliares para llenar datos vacíos es el análisis de regresión. En Kessler (2003) [55] indica que *hay dos utilidades principales al disponer de un modelo de regresión: podemos primero explicar la manera en la que cambios en los valores de una variable explicativa induce cambios en el valor de la variable respuesta. Por otra parte, si dispongo de un modelo para la evolución de la variable respuesta, me permite también realizar predicciones del valor que tomará para valores de las explicativas que no hemos observado.* Utilizando este último enfoque, un modelo de regresión permitirá inferir el valor de las observaciones que no se tienen disponibles en la tabla 22.

De la tabla 30 se deriva que la variable dependiente será el valor de la observación mensual, mientras que la variable independiente será el año. Lo siguiente será decidir que tipo de regresión se adapta mejor al comportamiento de las observaciones existentes. La figura 19 muestra la forma en cómo están distribuidos los valores de las observaciones a lo

largo de los años para el mes de enero (en la gráfica se incluyen observaciones desde 1970). En ella se pueden apreciar las interrupciones de información debido a los años que presentan ausencia de datos.

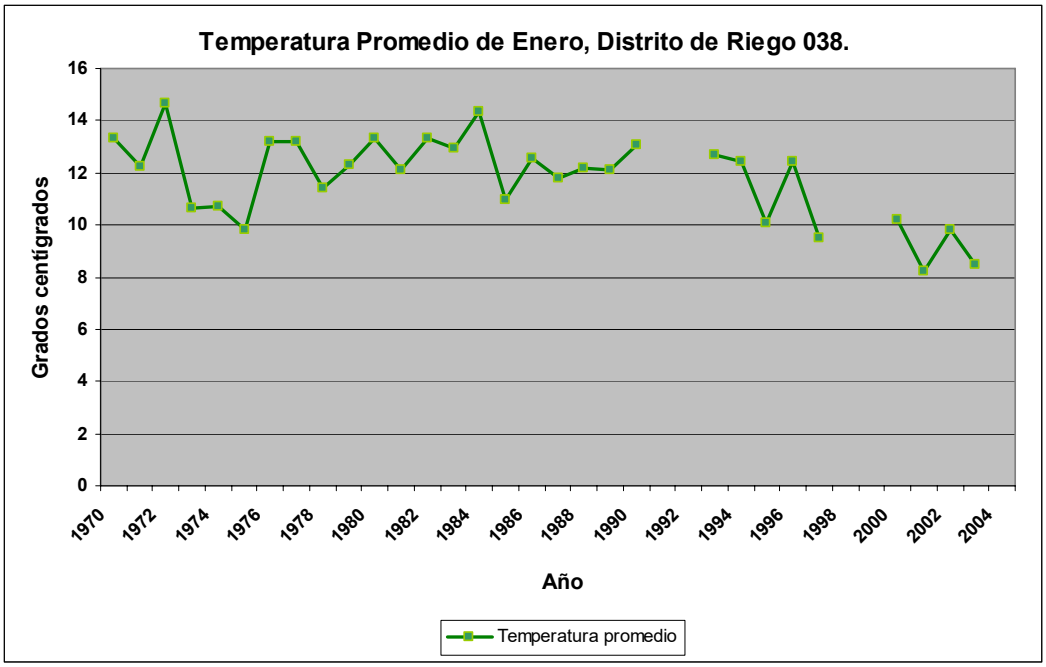


Figura 19. Observaciones de la temperatura promedio para el DR 038, 1970-2004

En la siguiente gráfica (Figura 20), hay tres líneas adicionales a los puntos de datos de las observaciones reales. Cada línea está implementada con un modelo de regresión distinto. En los tres casos, el modelo contempla una sola variable independiente (el año), pero en el primer caso, el modelo es una recta (con la variable independiente a la primera potencia), en el segundo, una curva cúbica (el modelo está en función de la variable independiente a la primera, segunda y tercera potencia) y finalmente, el tercer modelo, que representa una curva a la sexta potencia.

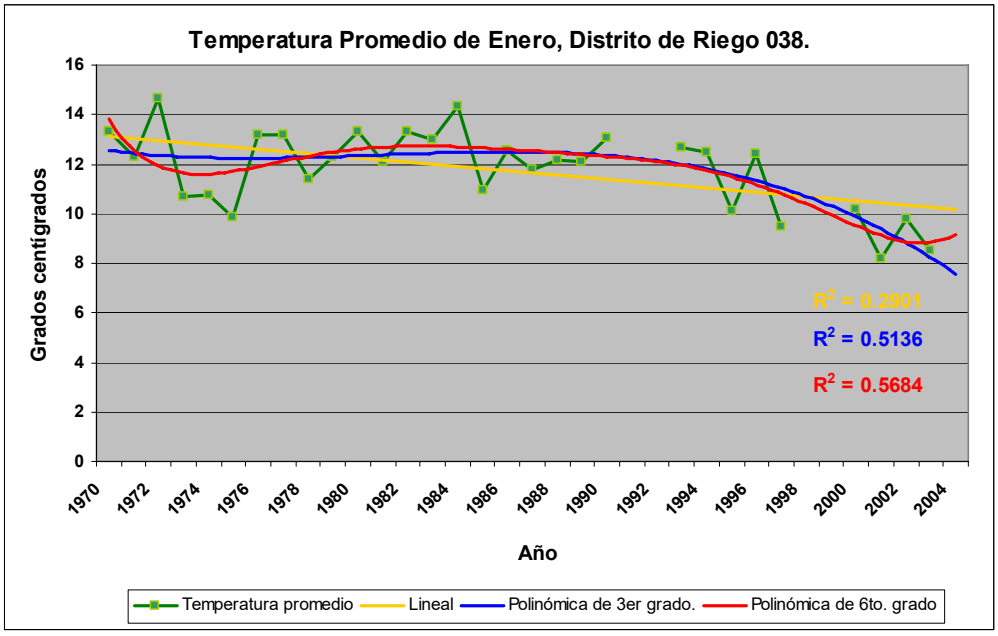


Figura 20. Observaciones de la temperatura promedio para el DR 038, 1970-2004

Las siguientes ecuaciones corresponden a cada una de las gráficas mostradas en la figura 20.

Tipo de curva	Función	Coeficiente de correlación
Recta	$Y = -0.0877x + 13.237$	0.2901
Curva de 3er grado.	$Y = -0.0005x^3 + 0.0161x^2 - 0.154x + 12.686$	0.5136
Curva de 6to grado	$y = 4E\text{-}07x^6 - 5E\text{-}05x^5 + 0.0021x^4 - 0.045x^3 + 0.4922x^2 - 2.3955x + 15.739$	0.5684

Tabla 23. Ecuaciones de los modelos de regresión para los datos de la temperatura promedio, mes de enero.

Por medio del coeficiente de correlación, se observa que el mejor ajuste se obtiene con la curva de sexto grado. Ejercicios similares se realizaron en los meses y variables restantes, obteniéndose una respuesta similar. La curva de regresión sexto grado fue la mejor aproximación encontrada para las series de datos del clima.

Para rellenar los datos faltantes de manera automatizada, se creó un programa de cómputo que permite leer un archivo en formato delimitado por comas, seleccionar dos atributos (uno como variable dependiente y otro como variable independiente) y calcular un modelo de regresión con la información de ambos atributos. El programa, creado en el lenguaje DELPHI 6, permite la selección de los atributos y el modelo de regresión a implementar. También incluye un mecanismo para el rellenado de datos faltantes con los modelos calculados. La figura 21 muestra la pantalla principal del programa, y la tabla 24, la misma serie de la tabla 22, pero complementada con los valores calculados con las regresiones generadas con el programa (resaltados con un asterisco).

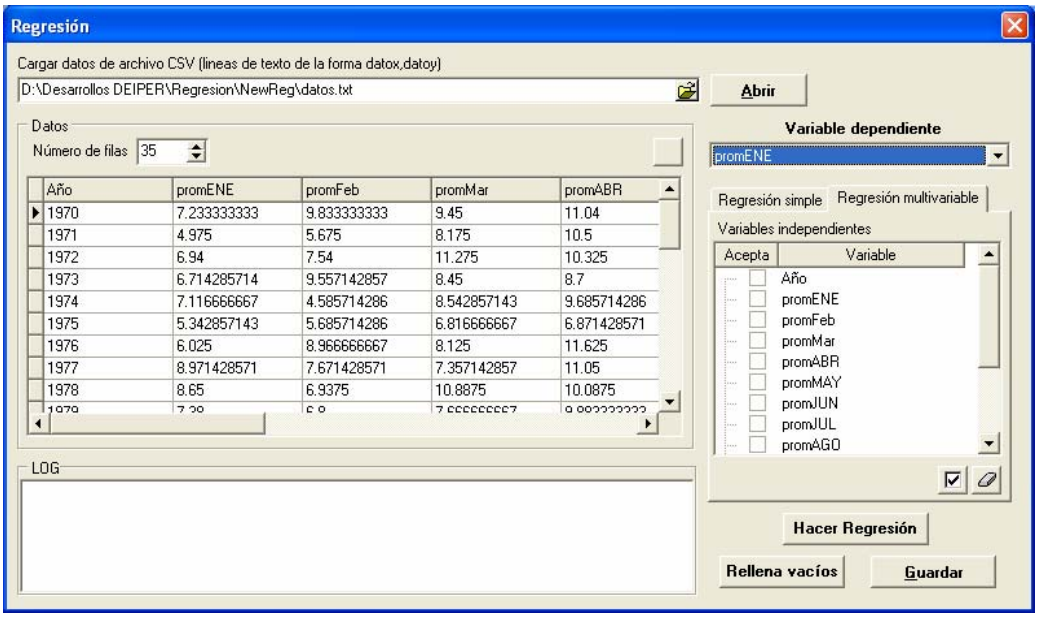


Figura 21. Pantalla del programa para el cálculo de modelos de regresión.

Año	Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Sep	Oct	Nov	Dic
1980	12.09	14.17	14.67	17.47	22.85	27.79	28.00	26.23	25.74	23.19	16.22	13.31
1981	14.34	13.26	14.56	19.47	22.96	26.51	28.37	28.91	27.16	22.71	16.91	12.10
1982	11.21	11.76	14.18	18.33	21.97	26.46	29.20	28.70	26.47	20.18	17.05	13.32
1983	13.57	12.99	15.10	16.90	21.02	25.37	28.47	26.63	27.31	22.58	16.30	12.97
1984	11.81	11.51	13.40	17.11	23.07	26.72	26.60	26.51	25.93	21.30	15.69	14.31
1985	12.00	12.17	14.27	18.53	20.59	26.24	27.37	27.30	26.41	21.44	15.27	10.97
1986	13.09	13.10	15.39	18.84	20.70	26.60	27.20	27.32	24.67	21.12	17.93	12.56
1987	9.35	11.10	12.28	16.95	19.66	24.56	27.53	25.82	25.56	22.99	16.57	11.79
1988	12.00	13.70	14.28	18.60	21.40	26.35	26.82	27.70	26.20	23.00	15.86	12.17
1989	10.00	13.43	15.23	19.03	21.60	24.85	28.25	26.65	27.05	23.10	17.85	12.10
1990	11.40	11.95	14.60	20.30	23.00	27.95	26.55	25.80	25.50	20.95	16.80	13.05
1991	10.10	16.30	14.70	16.35	22.10	*24.32	*26.52	*26.49	*25.39	*21.60	*16.72	*12.21
1992	*10.33	*12.49	*14.05	*16.21	*19.30	*23.88	*26.39	*26.45	*25.26	*21.47	*16.64	*12.09
1993	*10.10	*12.16	*13.73	*15.52	*18.80	*23.50	*26.29	*26.41	23.90	20.50	16.15	12.70
1994	10.30	10.10	12.95	14.05	16.60	24.30	26.35	25.90	25.55	21.05	15.45	12.45
1995	9.60	13.15	12.95	12.30	16.70	20.50	26.05	27.45	25.45	20.25	15.95	10.10
1996	7.85	11.80	12.50	14.30	18.15	23.90	25.65	25.65	24.30	20.80	15.25	12.40
1997	12.35	9.20	13.10	16.05	21.05	23.85	26.65	26.45	25.50	20.80	16.50	9.50
1998	10.30	10.00	10.80	11.10	15.00	21.20	26.00	26.40	24.90	21.80	*14.53	*10.39
1999	*9.83	*10.43	*11.09	*13.15	*18.65	*24.03	*26.40	*26.19	*25.04	*20.42	*14.08	*9.94
2000	*9.95	*10.32	*10.77	*13.29	*19.14	26.10	26.90	25.80	24.60	20.60	11.70	10.20
2001	9.50	9.60	10.70	13.20	19.20	23.50	25.80	25.70	25.10	19.50	15.40	8.20
2002	8.30	10.60	10.70	15.20	22.50	27.50	29.10	28.80	28.30	22.20	13.90	9.80
2003	11.20	12.50	10.60	13.70	19.50	22.40	27.40	27.40	26.80	22.80	14.80	8.50
2004	9.7	8.7	12.4	14	18.9	*22.62	*29.14	*29.42	*29.07	*24.50	*16.70	*9.18

Tabla 24. Promedio de temperaturas agrupados por año/mes de 1980 al año 2004, con los valores nulos reemplazados con estimaciones de regresión.

Al final de la etapa de ***limpieza de datos***, la serie de datos climáticos quedó con 174 registros, que representaban la información de 5 variables climáticas para los 12 meses de los años de 1970 a 2004.

4.3.3. Construcción de datos

Esta tarea involucra operaciones constructivas de preparación de datos, tal como la producción de atributos derivados, registros nuevos completos o la transformación de valores para atributos ya existentes [33].

Derivación de atributos para el proyecto

En Otero et al (2003) [85] se señala que el propósito de esta tarea es construir nuevos atributos fuera de los originales, transformando la representación original de los datos en una nueva donde las regularidades de los datos son detectadas más fácilmente por el algoritmo de clasificación, el cual tiende a mejorar la precisión de la predicción de este último. El mismo autor señala que existen dos enfoques para la construcción de atributos, el enfoque del pre-procesamiento y el enfoque intercalado:

- En el enfoque del pre-procesamiento, el proceso de construcción de atributos es independiente del algoritmo de aprendizaje inductivo que será utilizado para la extracción del conocimiento de los datos. En otras palabras, la calidad del candidato a nuevo atributo es evaluada directamente accediendo a los datos, sin la ejecución de ningún algoritmo de aprendizaje. En este enfoque, el método de construcción de atributos realiza un preprocesado de los datos, y los nuevos atributos construidos pueden ser dados a diferentes tipos de métodos de aprendizaje inductivo.

- En el enfoque intercalado, el proceso de construcción de atributos es intercalado con la aplicación del algoritmo de aprendizaje inductivo. La calidad de los candidatos a nuevos atributos es evaluado por el resultado de la ejecución del algoritmo de aprendizaje utilizado para la extracción de conocimiento, por lo tanto, la utilidad de los atributos construidos tiende a ser limitada por el algoritmo utilizado.

En CRISP-DM, la construcción de atributos está contemplada como parte del proceso de preparación de los datos y el enfoque de esta tarea es precisamente, el enfoque del pre-procesamiento.

El problema de la generación automática de atributos ha recibido una atención significativa durante la última década. Una variedad de algoritmos han sido desarrollados para mejorar el aprendizaje utilizando diferentes métodos de construcción de atributos. Estos algoritmos difieren en la forma de representar los atributos, en las técnicas de construcción y en los formatos de salida [86]. Básicamente, los algoritmos se pueden dividir en dos clases, aquellos que están orientados a problemas de dominio específico y los que son de propósito general.

Como ejemplo de algoritmos de construcción de atributos específicos del dominio, se tiene a Hirsh y Japkowicz (1994) [87], desarrollado en el entorno de la biología molecular. Otros

algoritmos más generales utilizan una forma de representación de atributo que puede ser empleada por diferentes dominios y problemas utilizando una conjunto fijo de operadores de construcción. Muchos algoritmos, como ***fringe*** [88], ***citre*** [89], ***ib3-ci*** [90], ***lfc*** [91] y ***gala*** [92], utlizan un conjunto mínimo de operadores lógicos como {¬,^} para expresar relaciones booleanas existentes entre atributos de datos. ***fringe***, ***lfc*** y ***gala*** operan en el marco del aprendizaje por árboles de decisión, el cual es utilizado para definir su contexto de construcción de atributos. A pesar de que estos algoritmos utilizan un lenguaje representativo idéntico y cuentan con la misma técnica de aprendizaje, sus enfoques de construcción pueden diferir [86].

En el caso del proyecto, no fue necesaria la intervención de ninguno de los algoritmos mencionados. Su mención es para propósitos de referencia únicamente. Aunque se empleó conocimiento del dominio, no hubo necesidad de emplear más allá de operaciones matemáticas básicas. A continuación, se describen los atributos creados para las series de datos del proyecto.

Atributos agregados a la serie de datos #1. Estadísticas de producción agrícola SINHDR.

Un aspecto importante no reflejado directamente por los atributos de esta serie de datos es el aspecto económico. Información como el ingreso bruto, los costos totales y el beneficio neto (utilidad) generado por la superficie sembrada por cultivo no figuraban en el conjunto de datos original. Tales datos eran importantes para la toma de decisiones, ya que permiten darse una mejor idea de la importancia de los cultivos desde el punto de vista del ingreso que permiten obtener.

Aunque los atributos ingreso, costo total y beneficio neto no existían en la serie original, los datos necesarios para calcular o inferir dicha información sí estaban presentes. La tabla 25 muestra la manera en como estos atributos fueron calculados.

Atributo	Forma de cálculo	Unidades	Descripción
Ingreso bruto	Ingreso bruto = Superficie cosechada x Rendimento x Precio	Miles de pesos (M$)	Representa la ganancia bruta obtenida por la venta del producto agrícola.
Costos totales	Costo total=Superficie sembrada x costo de producción.	Miles de Pesos (M$)	Representa el gasto total realizado en la siembra de la superficie indicad.
Ingreso neto	Ingreso neto = Ingreso bruto-Costo total	Miles de pesos (M$)	Representa la ganancia neta (utilidad) obtenido de la venta del producto agrícola.
Año agrícola	Separación del atributo ciclo	String (5)	Representa el año agrícola.

Tabla 25. Atributos agregados al recurso de datos #1 y forma de cálculo.

En la tabla 25 se puede apreciar que fue agregado un atributo que representa al año agrícola. El sistema de cómputo origen de la serie de datos #1 utilizaba una clave para representar al ciclo con el fin de almacenar, en el mismo campo, al año agrícola. De esta manera, un valor alfanumérico para el atributo ciclo de "O/I99-00" hacía referencia al ciclo otoño-invierno del año agrícola 1999-2000. Por conveniencia, se decidió separar la información del atributo ciclo en dos atributos, uno dedicado al almacenamiento del ciclo y otro dedicado a almacenar el año agrícola. Posteriormente, el atributo ciclo fue eliminado, ya que se sabe que la información pertenece al ciclo primavera-verano. Los 4 atributos de la tabla 33 fueron sumados a los 8 atributos resultantes de la selección de datos, dejando en total 12 atributos. La eliminación del atributo ciclo dejó la serie de datos #1 en 11 atributos.

Atributos agregados a la serie de datos #2. Estadísticas de producción CONAGUA.

Se plantea el uso de esta serie junto con la #4 para realizar pruebas. Por esta razón, ambas series (#2 y #4) deben contar con los mismos atributos que posee la serie #1. La etapa de ***selección de datos*** seleccionó 21 registros con casi los mismos atributos que contiene la serie #1. Los atributos faltantes, la lámina de riego y el costo de producción, deben ser agregados a la serie. Para agregar el atributo lámina de riego a la serie se utilizó el valor de la lámina promedio para cada uno de los cultivos bajo estudio. Al finaliza la etapa de construcción de datos, la serie #2 manifestaba 11 atributos, que representaban los mismos atributos que la serie de datos #1. Las series #3 y #4 se encontraban completas en cuestión de atributos. No fue necesario recurrir a la construcción de atributos para estas series.

Derivación de nuevos registros

Los registros generados deben ser completamente registros nuevos, los cuales agregan nuevo conocimiento o representan nuevos datos que de otra manera no estarían representados [33]. En el caso del proyecto que atañe a esta tesis, sólo fue necesaria la inclusión de nuevos registros para la serie de datos #3, relacionada con la información climatológica.

Registros generados para la serie de datos #3. Información climatológica.

A fin de completar la serie hasta el año 2006, fue necesaria la generación de los registros correspondientes a los años 2005 y 2006. Dado que para cada año se agregan cinco registros más (cada uno representa las observaciones de una variable en 12 meses del año), se agregaron en total 10 registros a los 175 registros resultantes del proceso de limpieza. En un principio se intentó utilizar la misma técnica usada en la etapa de limpieza de datos, la cual consistía en hacer una regresión polinomial de sexto grado para cada atributo de cada variable presente en la serie. Este método no resultó óptimo, ya que al evaluar las regresiones para los años 2005 y 2006, se observó que los valores calculados se alejaban por completo de la media como consecuencia de la curva polinomial (figura 22).

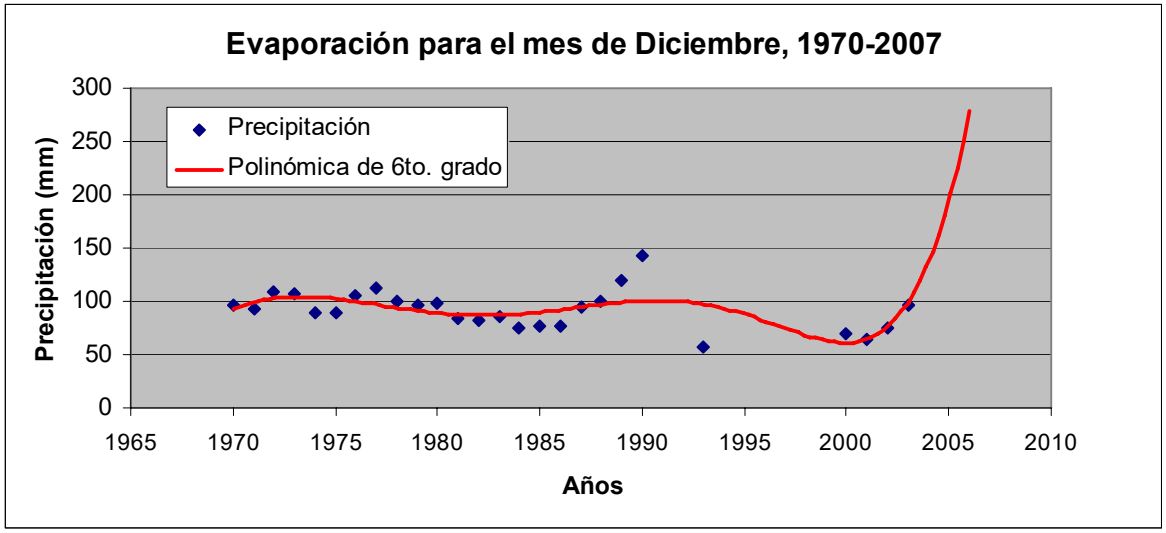


Figura 22. Aproximación por regresión polinomial para los datos de precipitación.

La figura 24 muestra que la aproximación polinómica de 6° grado perdía su efectividad después del año 2004. Haciendo diversas pruebas, se verificó que la aproximación lineal ofrecía muy poca ventaja frente a la media aritmética, por lo que se decidió utilizar la media para compensar los registros faltantes de la serie del clima. Después de la construcción de datos, la serie de datos climáticos quedó con 185 registros.

4.3.4. Integración de los datos

La integración de datos es un reto reciente en las aplicaciones que necesitan realizar consultas a través de múltiples y heterogéneas fuentes de datos autónomas [93]. La integración de datos es definida ampliamente como la combinación de datos desde diferentes fuentes sobre el mismo individuo o unidad [94]. Esta definición incluye encadenamientos entre encuestas y datos administrativos, así como desde dos o más fuentes administrativas. Una aplicación alternativa de la teoría de la integración de datos está en la identificación de registros en un solo archivo que pertenece al mismo individuo o unidad. Otros términos utilizados para describir el proceso de la integración de datos incluyen "encadenamiento de registros" y "correspondencia de datos" [94].

El propósito de la integración de datos en CRISP-DM es preparar una entidad o grupo de entidades cuyos atributos hagan referencia a un solo objeto común. Generalmente, las herramientas de minería aplican los algoritmos sobre una sola fuente de datos (que puede ser una consulta realizada a entidades distintas), por lo que es bastante frecuente que la ***integración de datos*** se dedique a producir una *entidad única organizada* que almacena la información proveniente de diversas fuentes. En el caso del proyecto de minería de esta tesis, este será el enfoque de la integración de datos, donde se producirá una entidad única que integre los datos recopilados a través de las series que han sido preparadas hasta este punto.

Como se puede ver, la integración de datos es un proceso con una fuerte influencia en la calidad de los datos. Al respecto, En Dasu et al (2003) [54] se indica que *la mayoría de los*

problemas intratables de la calidad de datos surgen durante la integración de datos, el proceso en el cual múltiples insumos de datos son puestos juntos para crear un conjunto de datos rico y completo. La etapa de ***limpieza de datos*** eliminó la mayoría de los problemas que usualmente pudiera causar la integración de datos, y como se verá, el verdadero trabajo durante esta fase consistió en diseñar una fuente de datos única y completa con miras al proceso de modelado.

4.3.4.1. Estado que guardan las series del proyecto previo a la integración de datos

Las series de datos del proyecto han sido mencionadas desde la etapa de ***recolección inicial de datos***. Estas fuentes han sufrido transformaciones a lo largo del proceso de minería. La reducción de atributos, la eliminación de registros y la limpieza de datos fueron algunas de las transformaciones que afectaron las series durante el proceso. En el siguiente resumen se describe el estado previo a la etapa de ***integración de datos*** que guardan las series

Serie de datos #1: Estadísticas agrícolas SINHDR. La fuente original, que consistía de 20,976 registros se redujo a 17,962 durante la fase de limpieza de datos. Además, estos registros guardaban una mejor calidad que los originales. Posteriormente, con la selección de datos, se decidió que únicamente se utilizaría la información correspondiente a un distrito de riego, el distrito de riego Río Mayo, con número 038, un grupo de 15 cultivos y un ciclo agrícola. Esta selección dejó 1035 registros de la serie original (ver selección de datos, punto 4.3.1). Por el lado de los atributos, de los 12 atributos originales se eliminaron dos en la temprana fase de descripción de los datos. La etapa de selección eliminó otros dos, el distrito de riego y el módulo. La fase de construcción de datos generó cuatro nuevos atributos, pero eliminó uno, con lo cual, la serie quedó con 11. Aunque estos atributos se han dejado hasta la fase de integración de datos, es probable que previo a la etapa de modelado se haga una nueva selección y algunos queden fuera de esa etapa. La razón para seguir manejando todos los atributos hasta esta etapa obedece a que su información será considerada en el desarrollo del algoritmo de optimización de la producción, el cual se aborda en el capítulo número 5 de esta tesis. Los 11 atributos y 1,035 registros generados representan la información de producción agrícola del distrito de riego 038 para los años 1995-2003. El diagrama de la figura 23 muestra las transformaciones que sufrió la serie de datos de estadísticas agrícolas.

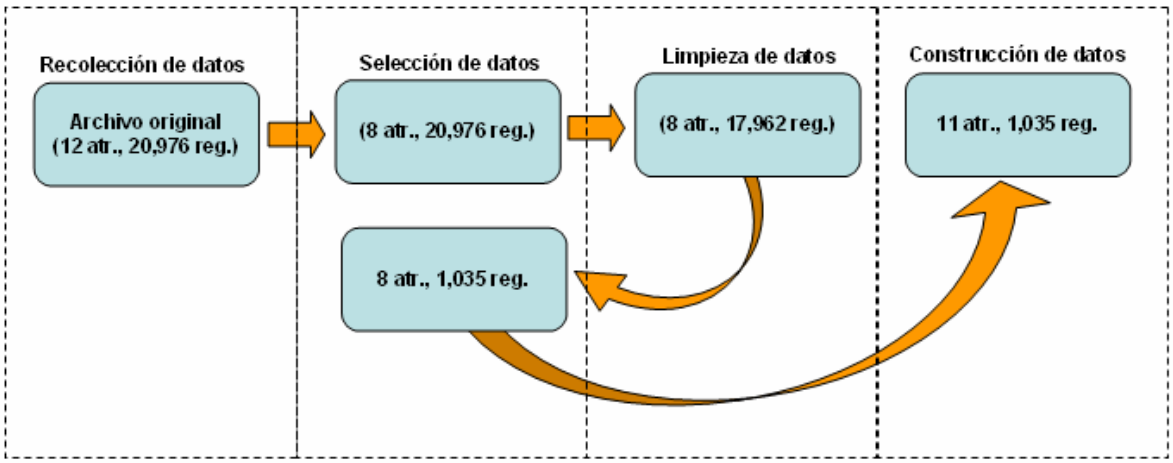


Figura 23. Cambios en número de atributos y registros para la serie de datos #1.

Serie de datos #2: Estadísticas de producción de documentos electrónicos CONAGUA. El recurso original de donde surge esta serie constaba de 24 documentos en formato PDF. La serie seleccionada se tomó de un solo archivo, el cual contenía la estadística del año 2006. Del documento fueron extraídos 51 registros, correspondientes al distrito de riego 038, ciclo primavera-verano. Posteriormente, al seleccionar únicamente los cultivos que habían sido determinados para el modelado, el número de registros se redujo a 21. El proceso de limpieza de datos no alteró la serie, pero el proceso de construcción de datos agregó 2 atributos más con el fin de dejar la serie de datos #2 con los mismos atributos de la serie #1. Después del proceso de construcción de datos la serie de datos #2 contaba con 11 atributos y 21 registros, que representaban la información de producción a nivel distrito de los años 2004 y 2005 del distrito de riego 038.

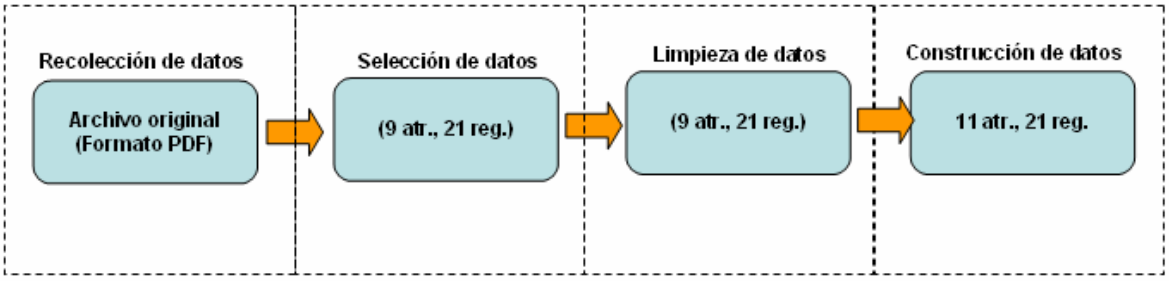


Figura 24. Cambios en número de atributos y registros para la serie de datos #2.

Serie de datos #3: Información climatológica. La información de esta serie sufrió grandes transformaciones durante el proceso de minería. El numero de registros potenciales en la base de datos propietaria era de 1,500,000. En la selección de datos, la forma estimada de los atributos originales fue alterada en forma, de manera tal que los atributos iniciales se convirtieron en el atributo variable y los meses (almacenados anteriormente como valores) pasaron a ser los atributos. La selección original quedó entonces con 2,275 registros y 15 atributos. Posteriormente, la etapa de ***limpieza de datos*** agrupó los registros y eliminó el atributo estación, por lo cual, la serie quedó en 175 registros y 14 atributos, abarcando la información desde 1970 al año 2004. La construcción de datos generó los registros faltantes hasta cubrir el período 1970-2006, con lo que dos años más de información fueron añadidos, quedando la serie en 185 registros y 14 atributos. Finalmente, previo a la etapa de integración de datos se hizo otra selección, esto para acotar el período de tiempo al manejado por las series agrícolas, quedando la serie del clima con 14 atributos y 65 registros.

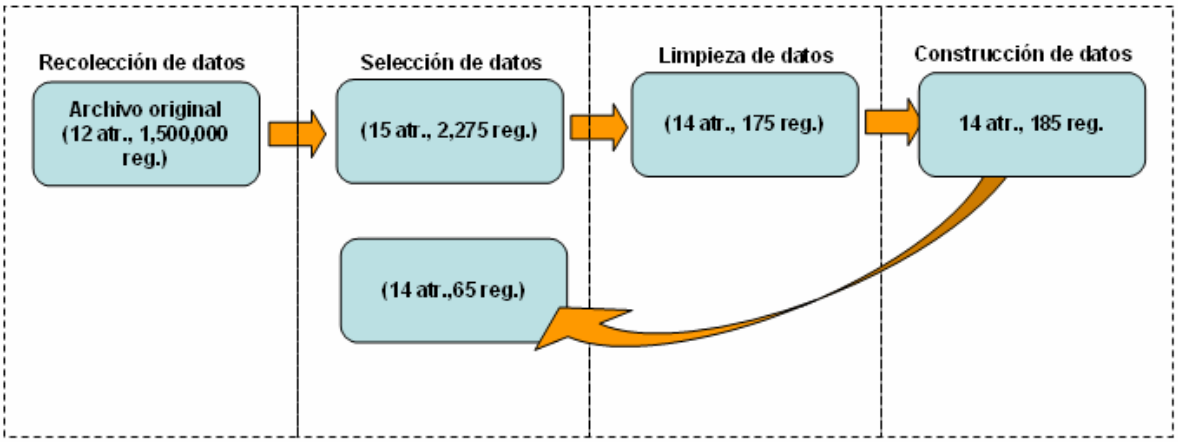


Figura 25. Cambios en número de atributos y registros para la serie de datos #3.

Serie de datos #4: Estadística de producción extraída de hojas de datos del distrito de riego 038. La información de esta fuente fue muy concisa desde el origen. Todos los registros extraídos de las hojas de datos sumaban originalmente 770 registros. En la selección de datos, se desestimaron 4 atributos que no eran necesarios, distrito, módulo, ciclo y producción total, además de filtrar la serie original para los cultivos seleccionados de la serie #1. Esto dejó la serie #4 en 130 registros y 9 atributos. La limpieza de datos dejó intacto el número de registros y de atributos. La construcción de datos agregó 2 atributos a la serie, esto con el fin de que la serie #4 quedara en el mismo formato que la serie #1 y la serie #2. Después de la construcción de datos, la serie #4 quedó con 11 atributos y 130 registros, que representaban la información de producción agrícola del distrito de riego 038 del año 2006.

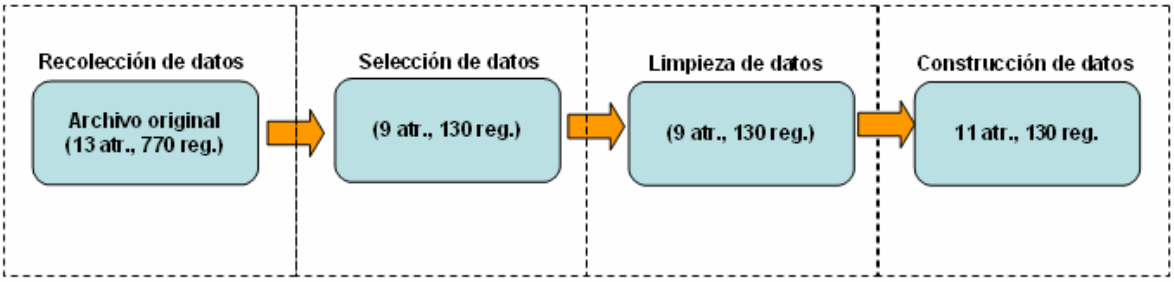


Figura 26. Cambios en número de atributos y registros para la serie de datos #4.

4.3.4.2. Integración de datos del proyecto

La salida de la etapa de integración de datos es una entidad única resultado de la combinación ordenada de las series de datos del proyecto. Para hacer tal combinación, es necesario definir la forma o estructura que tendrán estos datos combinados (ver figura 27).

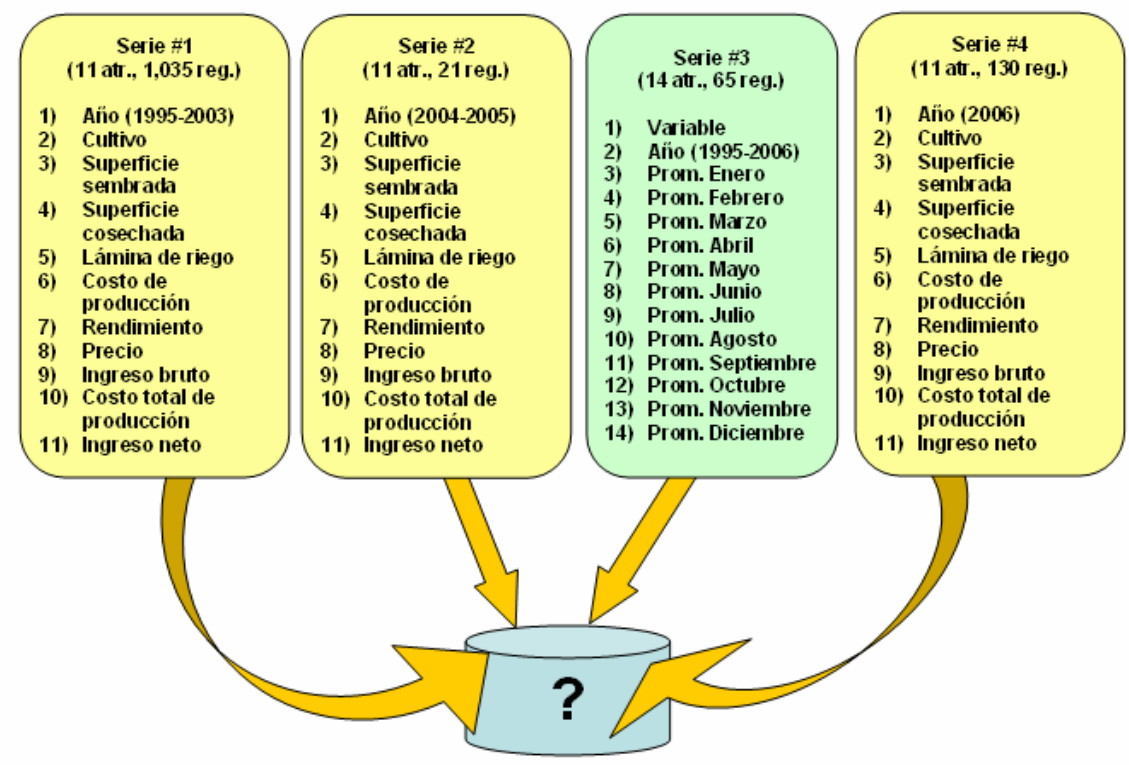


Figura 27. Definiendo la forma que tendrán los datos integrados.

Las series #1, #2 y #4 almacenan el mismo tipo de información, pero en períodos de tiempo distintos. No existe ningún traslape entre los registros, por lo contrario, se complementan entre sí. La integración de estas tres series se pudo hacer simplemente uniéndolas de forma vertical en una sola, insertando los registros de forma secuencial. Finalmente, se tuvieron dos tipos de series, la serie de datos de información agrícola y la serie de datos de información climática (figura 28).

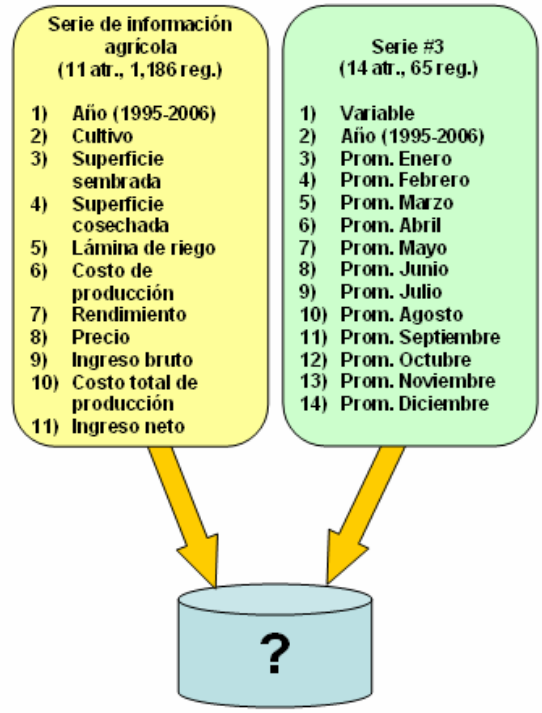


Figura 28. Definiendo la integración de series de datos distintas.

El principal problema para la integración de datos es que las series de datos que almacenan tipos distintos de información tienen atributos diferentes. Una opción para la integración de dichas series es la localización de un atributo único que permita ligar los registros en ambas series de datos. Este enfoque es conocido como el empate exacto (*exact matching*).

Como lo señala la guía para integración de datos [94], *no hay incertidumbre en el empate exacto. Los registros coinciden a través del identificador único o no lo hacen. El problema es cuando la calidad de las variables no es lo suficientemente buena como para garantizar que el valor del identificador único está disponible, es correcto y único.* La misma guía recomienda que si el empate perfecto por si solo no puede producir un conjunto de datos lo suficientemente robusto e integrado, entonces se utilice el encadenamiento probabilístico.

En las series de datos del proyecto se identificó como atributo único al *año*. Aparentemente, ambos tipos de series coinciden en este atributo, por lo que lo que el único trabajo requerido sería unir los atributos de ambas series para producir una sola. Pero no fue tan sencillo, dado que se presentaron dos problemas:

1) El atributo *variable* de la serie de datos del clima puede tomar a lo largo de la serie 5 valores distintos, cada uno representando una variable climática (precipitación, temperatura, evaporación, etc.). Para integrar este atributo, se tendría que repetir 5 veces la información de un registro de la serie de datos agrícolas para empatar cada registro con una variable climática distinta. La otra opción, es convertir cada variable en 12 atributos extras para la serie de datos agrícolas, agregando en total 60 atributos a la serie.

2) El atributo *año* de la serie de datos agrícolas tiene diferente significado que el atributo *año* de la serie de datos del clima. En el primer caso, se hace referencia a un año agrícola, el cual abarca un intervalo de tiempo situado entre dos años "normales", más concretamente, desde los meses de octubre del año anterior hasta el mes de septiembre del año referenciado en la serie de datos. Por otro lado, la serie de datos del clima maneja los años de forma normal, con los meses de enero a diciembre. Si se empataran los datos del clima con los de producción agrícola de manera directa, las observaciones mensuales de las variables climáticas estarían desfasadas, y serían incongruentes con la realidad.

Para el primer caso, no existía razón para mantener el atributo variable como atributo, ya no puede ser tomado ni como atributo dependiente y tampoco como atributo independiente. Es un identificador del registro. Esto decantó la solución hacia la segunda opción, transformar el atributo variable y transponerlo en la serie, agregando 60 atributos más a la serie de datos de producción.

En el segundo problema, los valores de los atributos de la serie del clima fueron tomados de dos años distintos, para que empataran de manera correcta con el año de referencia en la serie de datos agrícolas. Así, el orden de los atributos del clima fue cambiado, para iniciar los años en el mes de octubre y terminarlos en septiembre.

Para realizar la integración de datos, la información de todas las series fue llevada al programa de hoja de cálculo Microsoft Excel. Los registros de las series #1, #2 y #4 fueron colocados de forma consecutiva, de manera ordenada de acuerdo al valor del atributo año. Para integrar los datos del clima, se utilizó una función de búsqueda proporcionada por el programa, adecuada para buscar los valores del clima tomando como parámetros el año, la variable y el mes, variando el año de acuerdo a la posición vertical de la celda y variando la variable y el mes de acuerdo a la posición horizontal de la celda destino.

La serie de datos resultante del proceso de integración tiene la estructura mostrada en la tabla A5.1 del anexo 5, con un total de 1,186 registros y 71 atributos. Esta serie se identifica como la *serie de datos objetivo*.

4.3.5. Formateo de datos

El formateo de datos se refiere principalmente a modificaciones sintácticas hechas a los datos que no alteran su significado, pero que deben ser realizadas por ser requeridas por la herramienta de modelado [33]. Estas modificaciones pueden ser el orden de los atributos, la construcción de un identificador único, una estructura de almacenamiento específica (ej., base de datos relacional, archivo de texto, hoja de cálculo, etc.).

En la etapa de ***selección de herramientas*** se mencionó que para la labor de modelado de datos se utilizaría el software para de minería denominado Weka. Weka es un tanto flexible con el formato de la serie de datos a modelar, la cual puede provenir de a) un archivo de

texto en formato delimitado por comas, b) un archivo de formato ARFF[11] y c) una base de datos accesible por JDBC[12].

Para el proyecto de minería de la tesis, se decidió utilizar una base de datos en MySQL para almacenar la serie de datos objetivo. MySQL es una de las bases de datos relaciones soportadas por el manejador JDBC, por lo que Weka tiene un acceso transparente a la base de datos. Weka no tiene restricciones en el orden de los atributos, aunque por definición, asume que el último atributo de la serie es el atributo a modelar.

4.4. Modelado

Un modelo es una representación abstracta de un proceso del mundo real [27]. Se dice que la construcción de modelos en minería de datos es una actividad conducida por los datos (*data-driven*), esto en referencia a que trata de capturar las relaciones manifestadas entre los datos.

4.4.1. Selección de las técnicas de modelado

Como primer paso en el modelado, se debe seleccionar la técnica de modelado que será utilizada inicialmente. Si se van a aplicar múltiples técnicas, entonces este paso se debe ejecutar para cada técnica de forma individual [33].

Como se vio en la sección 4.1.11.1, no todas las técnicas son aplicables para todo tipo de problemas. Entre las técnicas disponibles y su selección hay conjunto de requerimientos admistrativos y condiciones que limitan las opciones disponibles (ver figura 29). Pudiera ser, incluso, que la herramienta seleccionada no sea la más adecuada técnicamente para el problema a resolver, pero que, sin embargo, satisface los requerimientos administrativos y las condiciones impuestas.

[11] ARFF (Attribute-Relation File Format). Formato de archive atributo-relación. Archivo de texto ASCII que describe una lista de instancias que comparten un conjunto de atributos [95].
[12] JDBC (Java Database Connectivity). Estándar para la conectividad independiente entre el lenguaje de programación Java y un amplio rango de bases de datos [96]. Weka esta desarrollado completamente en Java.

texto en formato delimitado por comas, b) un archivo de formato ARFF[11] y c) una base de datos accesible por JDBC[12].

Para el proyecto de minería de la tesis, se decidió utilizar una base de datos en MySQL para almacenar la serie de datos objetivo. MySQL es una de las bases de datos relaciones soportadas por el manejador JDBC, por lo que Weka tiene un acceso transparente a la base de datos. Weka no tiene restricciones en el orden de los atributos, aunque por definición, asume que el último atributo de la serie es el atributo a modelar.

4.4. Modelado

Un modelo es una representación abstracta de un proceso del mundo real [27]. Se dice que la construcción de modelos en minería de datos es una actividad conducida por los datos (*data-driven*), esto en referencia a que trata de capturar las relaciones manifestadas entre los datos.

4.4.1. Selección de las técnicas de modelado

Como primer paso en el modelado, se debe seleccionar la técnica de modelado que será utilizada inicialmente. Si se van a aplicar múltiples técnicas, entonces este paso se debe ejecutar para cada técnica de forma individual [33].

Como se vio en la sección 4.1.11.1, no todas las técnicas son aplicables para todo tipo de problemas. Entre las técnicas disponibles y su selección hay conjunto de requerimientos admistrativos y condiciones que limitan las opciones disponibles (ver figura 29). Pudiera ser, incluso, que la herramienta seleccionada no sea la más adecuada técnicamente para el problema a resolver, pero que, sin embargo, satisface los requerimientos administrativos y las condiciones impuestas.

[11] ARFF (Attribute-Relation File Format). Formato de archive atributo-relación. Archivo de texto ASCII que describe una lista de instancias que comparten un conjunto de atributos [95].
[12] JDBC (Java Database Connectivity). Estándar para la conectividad independiente entre el lenguaje de programación Java y un amplio rango de bases de datos [96]. Weka esta desarrollado completamente en Java.

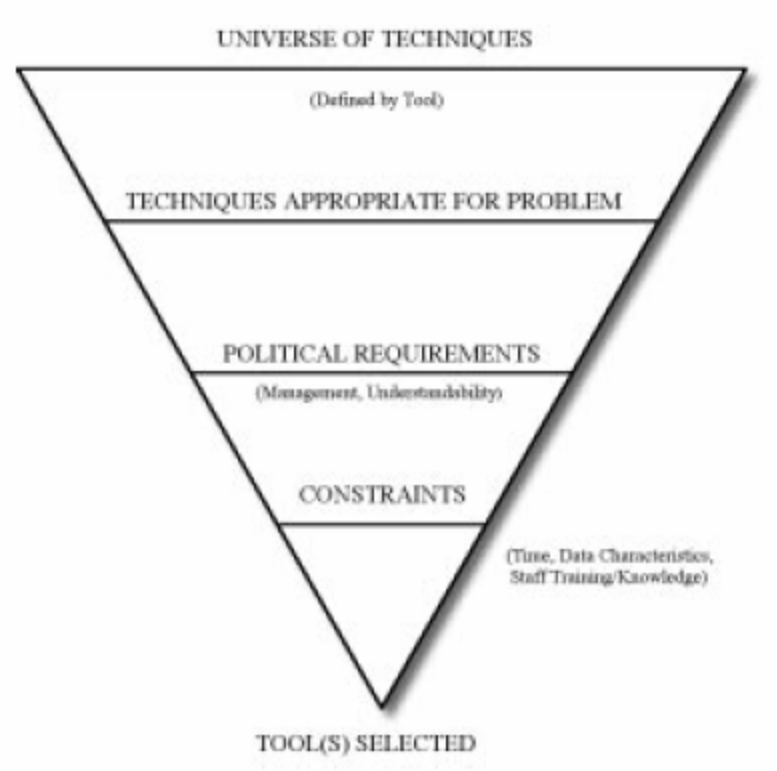


Figura 29. Universo de técnicas (CRISP-DM) [33]

Las técnicas seleccionadas se mencionaron ya en la sección 4.1.11.2. Como se comentó entonces, la selección obedeció a que son técnicas probadas con anterioridad para el problema de predicción del rendimiento de cultivos, y además, son técnicas sugeridas por la guía CRISP-DM para abordar un problema de predicción. Las técnicas seleccionadas fueron las siguientes:

a) Regresión multivariable.
b) Redes neuronales.
c) Inducción de reglas.
d) Modelos de árbol para regresión

La mayoría de estas técnicas están clasificadas como técnicas para problemas de predicción. Tal es el caso de la regresión multivariable, las redes neuronales y los modelos de árbol. Los árboles de decisión están más bien ubicados como una herramienta de clasificación [33], pero un problema de predicción puede también ser transformado en un problema de clasificación (como se verá más adelante), por lo que no se puede considerar que su uso esté fuera de lugar.

Cada una de las técnicas mencionadas (con excepción de la regresión multivariable) representa a un grupo de algoritmos o modelos específicos que presentan una estructura definida la cual caracteriza al grupo. Así, las redes neuronales se pueden implementar en un modelo Perceptrón, Adaline, Hopfield o mapa de Kohonen (por mencionar a algunos). Los árboles de decisión se generan mediante un algoritmo ID3, un C4.5, RandomTree, etc. Para los árboles de regresión existen la metodología CART, el modelo de árbol M5 y otras.

La descripción de cada una de las técnicas seleccionadas se proporciona en el anexo 6 de esta tesis. La selección específica del algoritmo de red neuronal, el algoritmo para la inferencia de reglas y el algoritmo de árbol de regresión se realiza en las siguientes secciones.

4.4.1.1. Selección del modelo de red neuronal

En el punto 2 del anexo 6 se proporcionan las bases teóricas de las redes neuronales, junto con una descripción de las topologías de red más utilizadas.

Revisando la información al respecto, se decidió que la red neuronal a evaluar para el problema de predicción del rendimiento fuera una red tipo Perceptrón Multicapa con el algoritmo de aprendizaje de retropropagación. La selección estuvo basada en los siguientes factores:

a) **Referencias.** Problemas parecidos al de esta tesis han sido abordados en el pasado con la implementación redes neuronales tipo perceptrón multicapa y el algoritmo de retropropagación. Como ejemplo, se puede citar Liu et al (1999) [39] , Drummond et al (2000) [38], Shearer et al (1999) [124], entre otros.
b) **Facilidad de uso y documentación.** Por el análisis realizado, se llegó a la conclusión de que el algoritmo de perceptrón multicapa tiene una mayor difusión, por lo que existe una mayor documentación del mismo. Esto a su vez, facilita la implementación del algoritmo, ya que es común encontrar ejemplos y ejercicios para llevar a la práctica dichas redes.
c) **Precisión.** Se ha encontrado que las arquitecturas basadas en el algoritmo de retropropagación son más precisas a la hora de clasificar, que, como por ejemplo, los mapas autoorganizativos de Kohonen [125][126]. Según propone Lee (2002) [126], las redes backpropagation proveen una mejor probabilidad de acierto y aprendizaje posterior cuando se dispone de suficiente información para ser aprendida en su fase de entrenamiento.
d) **Flexibilidad.** Estudios previos demuestran que, con el número suficiente y adecuado de unidades de proceso internas, la arquitectura de red neuronal backpropagation es capaz de aproximar bien cualquier función arbitraria (Funahashi, 1989; Hornik et al., 1989) [127].
e) **Eficiencia.** Para casos predictivos, se propone más eficiente el uso de arquitectura BP frente a los SOM de Kohonen, aún cuando se considera esta última arquitectura como una herramienta de clasificación alternativa, dependiente del grado de tolerancia en la precisión que se desee obtener [127].

Los parámetros para implementar la red perceptrón multicapa, como el tiempo de entrenamiento, la taza de aprendizaje y los atributos de entrada, se indicarán en la etapa de construcción de los modelos (sección 4.4.3.3.1).

4.4.1.2. Selección del algoritmo para la generación de reglas

Existen varios tipos de algoritmos utilizados para la obtención de reglas. En el punto 3 del anexo 6 se proporciona una breve descripción de los algoritmos más comunes para la generación de reglas Para la selección de un algoritmo en particular en el proyecto de minería de esta tesis, se optó por realizar una pre-evaluación de los algoritmos para la obtención de reglas más comunes, obtener una "vista preliminar" de su eficiencia en un conjunto de prueba y posteriormente implementar en la etapa de construcción del modelo aquel del que se hayan obtenido mejores resultados.

Materiales y métodos de la evaluación

Para hacer la evaluación se seleccionó de la serie de datos objetivo un conjunto de prueba con los registros de producción agrícola que comprenden los años agrícolas de 1994-1995 al 2000-2001.

Como se describió en la sección de selección de herramientas (sección 4.1.11.2), como herramienta de apoyo en la aplicación de algoritmos de minería de datos (dentro de los cuales se encuentran los de generación de reglas) se utiliza Weka. El propósito de Weka es permitir a los expertos acceder a una gran variedad de técnicas de aprendizaje automático para propósitos de investigación utilizando conjuntos de datos reales (McQueen et al, 1995)[49].

Para comparar los algoritmos se utilizarán las siguientes métricas básicas: número de reglas, efectividad (instancias correctamente clasificadas/ número total de instancias), error relativo absoluto y error cuadrático relativo (estas últimas métricas se describen en la etapa de diseño de pruebas, sección .4.4.2.1).

La mecánica de la selección es la siguiente: primero se hará una el cálculo de las métricas utilizadas en la evaluación. El siguiente paso, será comparar los algoritmos en base a los resultados de las métricas. Finalmente, se concluye con la selección del algoritmo.

Algoritmos bajo análisis

Los algoritmos sometidos a la evaluación fueron los siguientes:

- **Id3** [138].
- **J48** [68].
- **Random Tree (árbol aleatorio)**.
- **REPTree** [139].
- **Conjunctive Rule** [139].
- **JRip** [141]..
- **NNge** [142].
- **OneR** [143].
- **PART**. [139] [144].
- **Ridor** [139].

En el punto 3 del anexo 6 se proporciona una descripción general de los algoritmos para inducción de reglas aquí utilizados.

Evaluación

El conjunto de entrenamiento utilizado consistió en una serie de 810 registros históricos de productividad, cada uno de la forma:

Año agrícola
Cultivo
Superficie (ha)
Volumen (Mm3)
Temperatura promedio mensual (°C) [1..12]
Temperatura máxima del mes (°C) [1..12]
Temperatura mínima del mes (°C) [1..12]
Precipitación (mm) [1..12]
Evaporación (mm)
Rendimiento (ton/ha)

Tabla 26. Atributos originales del conjunto de entrenamiento.

Varios de los modelos clasificadores previamente citados comparten el problema de lidiar con atributos continuos, o cuando ellos son capaces de hacerlo, la presencia de este tipo de atributos incrementa la complejidad computacional del método de aprendizaje, como se señala en Molina y Béjar (1999) [145]. Para fines de la aplicación de los clasificadores, fue necesario convertir el atributo *rendimiento*, que originalmente un valor continuo, a un valor discreto, para ello se utilizó una de las técnicas más simples de discretización, denominada discretización por intervalos de igual anchura (Equal Width Interval, EWI). La técnica consiste en asignar a cada registro un valor discreto que corresponde con la posición que guarde el valor del atributo continuo en una tabla ordenada de rangos que son elaborados en base al valor mayor y menor del atributo continuo y al número de rangos solicitado[13].

Ciertos algoritmos (por ejemplo Id3) dependen de que todos los atributos sean discretos. En estos casos se utilizó la misma técnica descrita en el párrafo anterior. En ambos casos se utilizó un número original de 30 rangos, pero que Weka optimizó para dejarlo en el número que se adecuaba más a cada atributo.

Otra modificación al conjunto original fue la eliminación del atributo año agrícola antes de aplicar el clasificador. Esto es porque no se buscaba clasificar los ejemplos en función del año.

Una vez preparado el conjunto de entrenamiento, se aplicaron cada uno de los algoritmos de clasificación mencionados con los siguientes resultados:

Clasificador	Reglas generadas	Instancias clasificadas		Efectividad (%)	Error relativo absoluto	Error cuadrático
		Bien	Mal			

[13] Las técnicas de discretización se abordan con más detalle en la sección de construcción de los modelos (sección 4.4.3.2).

						relativo
Id3*	13	581	229	71.73	39.59	63.04
J48	67	540	270	66.67	52.12	72.33
Random Tree	923	810	0	100.00	0.00	0.00
REPTree	27	484	326	59.75	62.73	79.49
Conjunctive Rule	1	286	524	35.31	91.94	96.04
JRip	12	429	381	52.96	75.32	86.95
NNge	394	810	0	100.00	0.00	0.00
OneR	13	435	0	53.70	52.90	103.05
PART	113	582	228	71.85	43.07	65.75
PRISM*	444	450	360	55.56	50.78	100.97
Ridor	819	428	382	52.84	53.89	104.01

* Estos clasificadores no pueden trabajar con atributos de valores continuos.

Tabla 27. Resultado de la aplicación de los clasificadores.

La tabla 27 muestra el resultado de la ejecución de los algoritmos sobre el conjunto de entrenamiento. De un análisis de la tabla se desprenden las siguientes observaciones:

a) Los algoritmos *Id3* y *PRISM* sólo pueden trabajar con atributos discretos. Para aplicarlos, se hizo necesario discretizar todos los atributos. Esto los hace poco deseables para una aplicación real, ya que para cada nueva instancia que se presente será necesaria la discretización de todos sus atributos, e incluso cada inclusión de un nuevo ejemplo podría cambiar los rangos originales, haciéndose necesaria la reconstrucción por entero del clasificador.
b) Lamentablemente, los algoritmos que mostraron una eficiencia del 100 % manifiestan la tendencia de generar reglas en demasía. Tal es el caso del *Random Tree*, con 923 reglas (superior incluso al número de instancias), y el *NNge*, con 394 reglas. Este aspecto es importante, ya que el tamaño del conjunto de reglas tiene una fuerte influencia en la comprensión y utilización de las mismas [146].
c) Se observa la tendencia en todos los algoritmos a aumentar su eficiencia a medida de que aumenta el número de reglas (ver *NNge* y *Random Tree* en la tabla 35). Lamentablemente, siempre se busca obtener el menor número de reglas posibles, con el fin de tener un clasificador robusto. Ejemplos de clasificadores robustos (para el problema bajo análisis) los tenemos en la tabla 27 con *PART* y *J48*. *Id3* se descarta por las razones expuestas en el inciso a).
d) Muy malos clasificadores resultan *PRISM* y *Ridor* para el problema en cuestión. Con un número alto de reglas, ofrecen una eficiencia menor a la de otros clasificadores que presentan un número menor de reglas.

Cabe señalar, que los clasificadores están fuertemente parametrizados, por lo que su eficacia está condicionada a que se aplique en condiciones sumamente parecidas a las que originaron el conjunto de entrenamiento utilizado. En el caso del problema en cuestión, un clasificador desarrollado específicamente para un distrito de riego seguramente sufrirá un gran deterioro en su eficacia para un distrito distinto.

Las pruebas realizadas favorecieron a J48 como la elección para el algoritmo de inducción de reglas a implementar para el proyecto. Su facilidad para lidiar con atributos continuos, la

precisión y el número de reglas fueron los factores para la selección de este algoritmo. En el punto 3 del anexo 6 se proporciona una descripción detallada del algoritmo J48.

4.4.1.3. Selección del algoritmo de árbol de regresión

Un árbol de regresión es una pieza constante o una pieza de estimación lineal de una función de regresión, construido por el particionamiento recursivo de los datos y del espacio muestral. Su nombre se deriva de la práctica de visualizar las particiones como un árbol de decisión, desde el cual los papeles de las variables predictoras pueden ser inferidas [150]. La sección 4 del anexo 3 proporciona una descripción general de los tipos de modelos de árbol de regresión.

Los modelos basados en árboles de regresión son una herramienta de uso reciente en el contexto de la estimación del rendimiento de cultivos agrícolas. En Roel y Plant (2004) [173] se muestra una aplicación de CART [155] para detectar factores subyacentes que afectan el rendimiento en dos campos de arroz de California. En Lobell et al (2005) [44] se utilizan árboles de regresión CART para "entender" las variaciones en el rendimiento del trigo en un panorama bajo riego. En Paul et al (2004) [174] se utiliza CART en la elaboración de un modelo la evaluación de riesgo por la siembra de un tipo particular de maíz. Muy pocos trabajos que utilicen el algoritmo M5 [163] para la estimación del rendimiento han sido encontrados, esto quizá debido a que su aparición es más reciente.

Para efectos de esta tesis, se propone elaborar modelos de regresión para el rendimiento de cultivos basándose en árboles generados por el algoritmo M5. La ventaja de M5 sobre CART, es que los árboles generados son mucho más pequeños [175]. Esto reduce el tiempo de análisis y favorecen el aprovechamiento del árbol generado. Además, dado que no existen muchas aplicaciones de M5 para la estimación del rendimiento, este ejercicio puede aportar información valiosa sobre el desempeño de esta técnica de modelado aplicada a la información de cultivos agrícolas.

En la sección 4 del anexo 3 puede verse la descripción del el algoritmo M5' (M5 prime), que es el utilizado para la implementación en Weka de los árboles de regresión. La diferencia de M5' con M5, es que este último deja valores constantes en las "hojas" del árbol, mientras que M5' deja ecuaciones lineales.

4.4.2. Generación del diseño de pruebas

El diseño de pruebas es la construcción de un procedimiento o mecanismo para validar el modelo y probar su calidad. Por ejemplo, en un problema de clasificación es común utilizar la proporción de errores como una medida de la calidad de los modelos obtenidos por la minería de datos. Normalmente se separa la serie de datos en un conjunto de entrenamiento y otro de prueba, se construye el modelo con el conjunto de entrenamiento y se estima su efectividad en un conjunto separado de prueba [33].

En Hand et al (2001) [27] se señala la importancia de contar con un conjunto de prueba al indicar que *para obtener estimaciones de la desviación en el desempeño futuro de un modelo, es preciso medir su desempeño utilizando un conjunto de datos, el cual debe ser*

independiente del conjunto utilizado para construir y seleccionar el modelo. De la misma forma, se hace la advertencia que estas propiedades teóricas obtenidas de la evaluación al conjunto de prueba no son siempre una guía efectiva para medir el desempeño futuro del modelo.

Revisando la serie de datos objetivo y agrupando los registros por cultivo, se obtiene el número de registros para cada uno de los cultivos seleccionados (ver tabla 28).

Cultivo	Número de registros
Alfalfa	63
Algodón	23
Cártamo	150
Chícharo	27
Forraje	11
Fríjol	129
Frutales	23
Garbanzo	96
Hortalizas	102
Maíz	161
Papa	105
Sorgo	51
Tomate	77
Trigo	156
Zacate	11
Total	**1,185**

Tabla 28. Número de registros por cultivo de la serie de datos objetivo.

De acuerdo a lo indicado por la guía, cada uno de los conjuntos de los cultivos mostrados en la tabla 28 debe ser separado en un subconjunto de prueba y en un subconjunto de entrenamiento[14]. El modelo se construye la información del subconjunto de entrenamiento y entonces es validado con la información del subconjunto de prueba. Este tipo de validación es denominada validación simple. La primera decisión que se debe tomar en un diseño de pruebas es precisamente el tamaño de ambos conjuntos. Aquí se depende enormemente del número de registros en la serie de datos objetivo. En conjuntos pequeños no existen muchas posibilidades para decidir sobre el tamaño, pero en conjuntos grandes sí. Conjuntos de entrenamiento grandes resultan en modelos más estables, pero deben reservarse los registros suficientes para en el conjunto de prueba para evaluar la calidad del modelo. Normalmente, se reserva un 20 % de la muestra para pruebas [97]. En el caso del proyecto de esta tesis, se decidió reservar un 30% de los registros pertenecientes a cada cultivo para prueba. Los grupos de registros para los cultivos quedaron divididos como se muestra en la tabla 29.

Cultivo	Registros para entrenamiento	Registros para prueba	Número total de registros
Alfalfa	50	13	63
Algodón	18	5	23

[14] Otros nombres para *conjunto de prueba* es *datos de prueba* ó *datos de evaluación*.

Cártamo	120	30	150
Chícharo	21	6	27
Forraje	8	3	11
Fríjol	103	26	129
Frutales	18	5	23
Garbanzo	76	20	96
Hortalizas	81	21	102
Maíz	128	33	161
Papa	84	21	105
Sorgo	40	11	51
Tomate	61	16	77
Trigo	124	32	156
Zacate	8	3	11
Total	**940**	**245**	**1,185**

Tabla 29. División de la serie de datos para entrenamiento y prueba.

Otro tipo de evaluación es la denominada *validación cruzada* (*cross validation*) [187][188]. La validación cruzada es el diseño experimental más utilizado entre los investigadores en aprendizaje automático [189]. Es una extensión de la validación simple, donde la división de dos conjuntos independientes es repetida de manera aleatoria en múltiples ocasiones, con cada vez estimándose un nuevo modelo del conjunto destinado al entrenamiento y una desviación por parte de la aplicación del modelo al conjunto de prueba [27]. Estas desviaciones son promediadas para obtener una desviación total.

Existen numerosas variantes de la validación cruzada. La que será utilizada para el proyecto es una de las más comunes, denominada validación cruzada de k-dobleces (k-fold cross validation), la cual consiste en dividir el conjunto de ejemplos de que se dispone en k conjuntos disjuntos de igual tamaño, $T_1,..,T_k$. Se realizan *k* experimentos, usando como conjunto de entrenamiento en la iteración i-ésima $\cup_{j\neq i} T_j$ y como conjunto de prueba T_i. Cada algoritmo da lugar a una muestra de k estimaciones del error, y las diferencias entre dos algoritmos se juzgan mediante un contraste acerca de las diferencias entre las medias o las medianas del error muestral [189].

La validación cruzada es popular en la práctica, esto porque es simple y razonablemente robusta. Sin embargo, si el particionamiento es repetido *m* veces atrae un costo de *m* veces la complejidad del método basado en la validación simple [27]. Esto debe ser tomado en cuenta, ya que incrementa el tiempo de la etapa de pruebas.

Ambos métodos, la validación simple y la validación cruzada, son utilizados en la evaluación de los modelos de esta tesis. Para evaluar los modelos, ambas técnicas de validación van acompañadas de un conjunto de métricas que permiten cuantificar el desempeño de cada modelo en aspectos específicos, como la efectividad de predicción, la desviación del valor real o la ventaja con respecto a otros mecanismos de predicción, como la media. Las métricas de evaluación se describen en la siguiente sección de esta tesis.

4.4.2.1. Métricas utilizadas para evaluar los modelos de regresión

Simbología

En esta sección se empleará la simbología que se describe en la tabla 30.

Símbolo	Descripción
n	Número total de observaciones.
y(i)	Valor real de la observación *i*.
$\hat{y}(i)$	Valor estimado por el modelo de regresión para la observación *i*.
$\bar{y}$	Media del conjunto de observaciones.

Tabla 30. Simbología empleada para la definición de métricas.

El enfoque tradicional para evaluar el desempeño de los modelos de regresión está basado en métricas aditivas de los errores [177], que dependen de los residuos indicados en (2.10).

$$r_i = y(i) - \hat{y}(i) \qquad \textbf{(2.10)}$$

Utilizando estos residuos se define un conjunto de métricas que permiten evaluar el desempeño cuantitativo de los modelos. Las métricas empleadas se describen en la tabla 31.

No.	Nombre	Unidades	Cálculo	Descripción
1	Eficiencia promedio.	%	$\frac{\sum_{i=1}^{n}\left(1-\frac{\lvert r_i\rvert}{y(i)}\right)}{n}\times 100$	Promedio de eficiencia porcentual en la predicción del valor estimado con respecto al valor real.
2	Error absoluto medio (Mean Absolute Error)	Ton/ha	$\frac{\sum_{i=1}^{n}\lvert r_i\rvert}{n}$	Promedio de desviaciones absolutas del valor estimado con respecto al valor real.
3	Raíz del error cuadrático medio (root mean squared error)	Ton/ha	$\sqrt{\frac{\sum_{i=1}^{n} r_i^2}{n}}$	Es una de las medidas más utilizadas para medir el éxito de una predicción numérica. Proporciona el valor del error en la misma dimensión que el valor real y el valor estimado.
4	Error absoluto relativo (relative absolute error)	%	$\frac{\sum_{i=1}^{n}\lvert r_i\rvert}{\sum_{i=1}^{n}\lvert y_i-\bar{y}\rvert}x100$	Error relativo con respecto a la media. Un valor bajo denota un modelo más preciso, un valor de 0 denota un modelo estadístico perfecto.
5	Raíz del error cuadrático relativo (root relative squared error)	%	$\sqrt{\frac{\sum_{i=1}^{n} r_i^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}}x100$	Es el error relativo total con respecto a la media. La raíz cuadrada es para dar la misma dimensión que los valores estimados. Esto exagera los casos en donde el error de predicción fue significativamente mayor que la media.

Tabla 31. Métricas utilizadas para la evaluación de los modelos de regresión

En general, las métricas expuestas en la tabla 31 permitirán evaluar todo modelo que proporcione un valor numérico como resultado de la aplicación de un modelo para predicción. Por lo tanto, utilizando tales métricas se puede evaluar el desempeño de la regresión lineal múltiple, las redes neuronales y los modelos de árbol de regresión. Los modelos basados en reglas, cuyos resultados son valores categóricos, requieren de métricas específicas que permitan evaluar aspectos relacionados con la clasificación nominal y no con valores continuos. Sin embargo, si se define un mecanismo para inferir un valor numérico de un valor categórico, las métricas descritas en la tabla 31 también pueden ser

aplicadas a dicho valor. En esta tesis utilizamos un método para obtener un valor numérico representativo de una clasificación hecha a base de un árbol J48, por lo que las métricas de la tabla 31 también son aplicadas a los resultados de los árboles J48.

4.4.2.2. Métricas utilizadas para la evaluación de reglas

La evaluación de reglas es un tema abierto e importante en el área de la investigación en minería de datos. El problema de la evaluación de reglas es generalmente manejado como la detección de "reglas interesantes" utilizando "métricas de interés" [178][179][180]. Estas técnicas típicamente emplean el conjunto de datos como un todo para minar las reglas, y entonces filtran o priorizan las reglas descubiertas de diversas maneras [178].

Para fines del proyecto, se seleccionaron un conjunto de técnicas para evaluar las reglas generadas con el algoritmo J48. Este conjunto de técnicas se toman de la clasificación realizada por Geng y Hamilton (2006) [179]. Estos mismos autores proponen una clasificación para las métricas, la cual se basa en la división por *métricas objetivas* (que se basan únicamente en la información proporcionada por los datos, sin conocimiento del dominio ó del usuario), *métricas subjetivas* (toman en cuenta los datos y al usuario) y *métricas semánticas* (considera la semántica y las explicaciones surgidas de los patrones). Estas tres divisiones abarcan nueve aspectos que describen el grado o nivel de interés aportado por las reglas (figura 30).

Medidas objetivas
- Basadas en probabilidad (generalidad y fiabilidad).
- Basadas en la forma de las reglas
  - Peculiaridad
  - Sorpresa
  - Concisión
    - Reglas no redundantes
    - Descripción de longitud de mínima (Minimum Description Lenght, MDL).

Medidas subjetivas
- Sopresa
- Novedad

Medidas semánticas
- Utilidad
- Funcionalidad

Figura 30. Clasificación de los aspectos para determinar el interés de una regla [179]

Los nueve aspectos en los que se puede medir el interés de una regla son los siguientes (Geng et al, 2006) [179]:

1. Concisión. Un patrón es conciso si contiene relativamente pocos pares atributo-valor. Un modelo es conciso si contiene pocos patrones.
2. Generalidad/cobertura. Un patrón es general si cubre un subconjunto relativamente largo del conjunto de datos original. La generalidad o cobertura mide la comprensión de un patrón, que es la fracción de todos los registros que concuerdan con el patrón.

3. Confiabilidad. Un patrón es confiable si la relación descrita por el el patrón ocurre en un alto porcentaje de casos.
4. Peculiaridad. Un patrón es peculiar si está lejos de otros patrones descubiertos de acuerdo a alguna métrica de distancia. Los patrones peculiares son generados de datos peculiares (outliers), los cuales son relativamente pocos en número y significativamente diferentes del resto de los datos.
5. Diversidad. Un patrón es diverso si sus elementos difieren significativamente unos de otros. Un conjunto de patrones es diverso si si los patrones en el conjunto difieren significativamente unos de otros.
6. Novedad. Un patrón es novedoso para una persona si el o ella no lo conocían de manera previa y no es capaz de inferirlo a partir de otros patrones conocidos. Ningún sistema de minería de datos puede representar todo lo que el usuario sabe, y entonces, la novedad no puede ser medida de manera explícita con referencia al conocimiento del usuario.
7. Sorpresa. Un patrón es sorprendente (o inesperado) si contradice el conocimiento existente de un usuario o sus expectativas [181][182][183][184]. Los patrones sorprendentes son interesantes porque identifican fallas en el conocimiento previo existente y pueden sugerir un aspecto de los datos que necesita más estudio.
8. Utilidad. Un patrón es útil si su uso por parte de una persona contribuye a alcanzar una meta. Diferentes personas pueden tener metas divergentes respecto al conocimiento que pueda ser extraído del conjunto de datos.
9. Funcionalidad/aplicación.. Un patrón es aplicable si en algún dominio provoca la toma de decisiones sobre futuras acciones en su dominio [185][186]. La funcionalidad algunas veces está asociada con una estrategia de selección de patrones.

Las métricas objetivas basadas en probabilidad que evalúan los aspectos de generalidad y confiabilidad han sido estudiadas ampliamente por los investigadores [179]. Para fines del proyecto, se utilizará un subconjunto de este tipo de métricas para evaluar los modelos generados por el algoritmo de inducción de reglas J48.

Para definir las métricas se utilizará la simbología mostrada en la tabla 32.

Símbolo	Descripción
A	Denota el antecedente de una regla.
B	Denota el consecuente de una regla.
P(A)	Denota la probabilidad de A, calculada como se ve en (33.a).
P(A|B)	Denota la probabilidad condicional de A dado B. Calculada como se ve en (33.b).
P(B|A)	Denota la probabilidad condicional de B dado A. Calculada como se ve en (33.c).
P(AB)	Denota la probabilidad conjunta de que A y B ocurran. Calculada como se ve en (33.d).
n(A)	Número de registros que satisfacen A
n(AB)	Número de registros que satisfacen ambos, A y B
N	Número total de registros.

Tabla 32. Simbología para definir las métricas utilizadas en la inducción de reglas.

El cálculo de P(A), P(A|B), P(B|A) y P(AB) es como sigue:

$$P(A) = \frac{n(A)}{N} \qquad P(A \mid B) = \frac{n(AB)}{n(B)}$$

(a) **(b)**

$$P(B \mid A) = P(A/B)\left(\frac{P(B)}{P(A)}\right) \qquad P(AB) = P(B \mid A) \cdot P(A)$$

(c) **(d)**

(2.11)

Las métricas se describen en la tabla 33.

Métrica	Aspecto que cubre	Cálculo	Descripción
Soporte	Generalidad	*P(AB)*	Es la probabilidad de que una transacción contenga ambos, A y B. Porcentaje de instancias que la regla predice correctamente.
Cobertura	Generalidad	*P(A)*	Porcentaje de casos que abarca la regla (porcentaje de veces que el antecedente concurre).
Confianza/ Precisión	Fiabilidad	*P(B|A)*	Porcentaje de veces que la regla se cumple cuando se aplica.
Valor agregado	Fiabilidad	*P(B|A)-P(B)*	Porcentaje de veces que la regla se cumple de más (por sobre su valor de fiabilidad).
IS	Generalidad + Fiabilidad	$\sqrt{\frac{P(AB)}{P(A)P(B)}}\, xP(AB$	
Piatetsky-Shapiro	Generalidad+ Fiabilidad	*P(AB)-P(A)P(B)*	

Tabla 33. Métricas utilizadas en la inducción de reglas.

Además de las métricas de la tabla 33, se utiliza la métrica más común para la evaluación de reglas, denominada precisión de predicción, que se define como:

$$Precision = \frac{Número\ de\ ejemplos\ correctamente\ clasificados}{Número\ total\ de\ ejemplos\ de\ prueba} \qquad \textbf{(2.12)}$$

En esta tesis se propone un método para comparar de forma numérica un valor categórico. En tales condiciones, es posible utilizar también las métricas definidas en la tabla 32. Los resultados de la aplicación de todas las métricas se describen en la sección 4.4.4, correspondiente a la evaluación de los modelos.

4.4.3. Construcción de los modelos

Durante la fase de construcción de modelos se hace uso de la herramienta de modelado y la serie de datos resultante de la fase de *preparación de datos*, con el propósito de generar uno o más modelos [33].

La herramienta de Minería de Datos Weka es utilizada para generar los modelos de predicción del rendimiento de cultivos. Por esta razón, la fase de construcción de modelos está enmarcada por una fuerte participación de esta herramienta, la cual interviene a lo largo de todo el proceso de modelado.

La mecánica utilizada para la construcción de modelos fue la siguiente:

1. Se realiza una última selección de atributos. Algunos atributos fueron considerados en la fase de preparación de datos debido a su posible uso en la etapa de optimización, pero no son tomados en cuenta en la etapa de modelado.
2. Se utiliza la herramienta Explorer[15] de la suite Weka para la construcción de los modelos. La generación de los modelos con un determinado algoritmo lleva la siguiente secuencia de comandos o instrucciones:

   Selección de los registros. Esta acción se realiza por medio de la opción "*Open DB"* del diálogo *Preprocess* del Weka Explorer. Este diálogo recibe como entrada los parámetros para ejecutar una consulta a un manejador de base de datos soportado por el controlador ODBC. En el caso de la tesis, una consulta típica es del tipo "*SELECT [lista_campos] FROM [tabla_hechos] WHERE Cultivo=[nombre_cultivo]*". Donde [lista_campos] es la lista de los identificadores para cada uno de los atributos mostrados en la tabla 41, [tabla_hechos] es el nombre de la tabla que almacena la información obtenida durante el pre-procesamiento de los datos. [nombre_cultivo] es el nombre del cultivo para el cual se realiza la consulta. Si se obtiene un modelo global de todos los cultivos, entonces la condición sobre el atributo/campo *Cultivo* se omite.

   Una vez que los registros son atraídos al *Explorer*, puede ser necesario algún tipo de preprocesado antes de integrarlos al modelo. El caso más común, es requerir de la discretización del atributo (muy usual en la construcción de los árboles J48). En tal caso, se selecciona el filtro *Unsupervised/Discretize* en el diálogo *Preprocess*.

   El siguiente paso viene con la selección del tipo de modelo. En el diálogo *Classify* se selecciona el tipo de algoritmo.

   En general, los algoritmos requieren de la correcta especificación de sus parámetros para proporcionar una salida óptima. Los parámetros se especifican a través del diálogo *GenericObjectEditor* (dando clic en el cuadro con el nombre del clasificador). Para esta tesis, los parámetros de construcción de cada modelo se describen en su correspondiente sección.

   Se establecen las opciones de prueba en el cuadro *test options* del diálogo *Classify*.

   Se especifica el atributo objetivo de la clasificación, que en el proyecto usualmente es el campo *Rendimiento*, o bien, su versión discretizada.

   Se construye el modelo.

[15] El Weka Explorer es uno de los componentes de la interfaz gráfica de la suite Weka. Como su nombre lo dice, es una herramienta para explorar datos, la cual ofrece las técnicas más comunes para colaborar las distintas tareas que pueden integrar un proceso de minería de datos.

3. El modelo generado se valida contra el conjunto de entrenamiento. Se aplican las técnicas de evaluación descritas en la sección 4.4.2.
4. El modelo generado se valida contra el conjunto de prueba. Se aplican las técnicas descritas en la sección 4.4.2.

4.4.3.1. Una última selección de atributos.

Como se había mencionado con anterioridad, muchos de los atributos "arrastrados" durante el proceso de preparación de los datos fueron conservados debido a que proporcionan información en la fase de la elaboración del modelo de optimización de esta tesis. Tal es el caso de los atributos con información económica, como el precio, el ingreso bruto, el costo de producción, etc. Estos atributos no son incluidos en la fase de construcción de los modelos de la etapa de minería de datos, por lo que son omitidos como entradas para la herramienta de minería Weka. Otros atributos omitidos son el año agrícola (que desaparece porque no se desea discriminar sobre el año), la superficie cosechada (que casi siempre es igual a la superficie sembrada), y el cultivo. Este último atributo desaparece de la selección porque se realiza un modelo para cada uno de los 15 cultivos seleccionados. De manera que los atributos que sirven de entrada para todos los algoritmos de modelado utilizados se exponen en la tabla A7.1 del anexo 7.

4.4.3.2. Construcción de los modelos de regresión multivariable

4.4.3.2.1. Parámetros utilizados en la construcción de los modelos de regresión multivariable

Para la construcción de los modelos de regresión se utilizó la opción *classifiers/functions/linearregresion* del diálogo classify. Los parámetros requeridos por Weka para la regresión lineal múltiple se muestran en la figura 31.

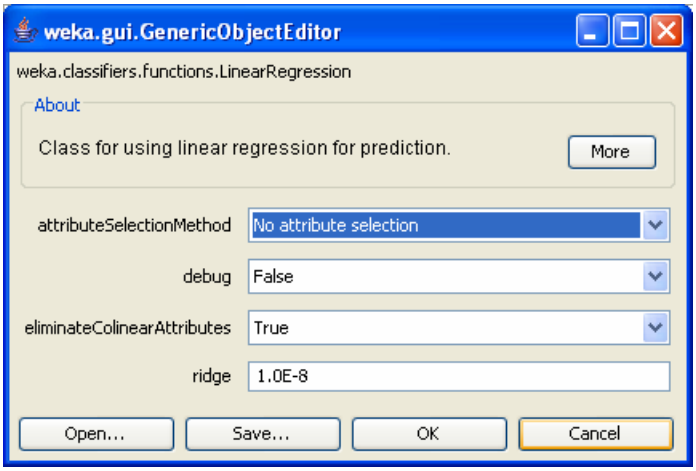


Figura 31. Parámetros para el modelo de regresión

El modelo de regresión múltiple tiene los siguientes parámetros:

a) *attributeSelectionMethod*. Permite indicar el tipo de algoritmo o mecanismo para realizar la selección de atributos que participan en el modelo de regresión. Algunas técnicas son el método del algoritmo M5[16] [190] o la selección avara utilizada por la métrica de información Akaike.
b) *debug*. Produce información de salida hacia la consola.
c) *eliminateColinearAttributes*. La colinealidad se refiere en un estricto sentido a la presencia de relaciones lineares exactas dentro de un conjunto de variables, típicamente un conjunto de variables predictoras utilizadas en un modelo de regresión lineal [191].
d) *ridge*. El valor del parámetro ridge (ver descripción de modelos).

Los valores de los parámetros utilizados para la construcción de todos los modelos de regresión lineal del proyecto aparecen en la tabla 34.

Parámetro Weka	Valor establecido para el proyecto	Comentario
attributeSelectionMethod	No attribute selection	Se desea que todos los atributos aparezcan en el modelo.
debug	False	No se requiere información para depurar.
eliminateColinearAttributes	False	Se desea que todos los atributos aparezcan en el modelo. La colinearidad se detecta y señala en la fase de evaluación.
Ridge	1.0E-8	Valor de facto en Weka. Acerca el modelo a la regresión multivariable tradicional.

Tabla 34. Parámetros para la regresión lineal multivariable.

4.4.3.2.2. Descripción de los modelos de regresión multivariable generados

Recordando, el modelo de regresión lineal multivariable se describe como:

$$Y_i = \sum_{j=1}^{k} B_j X_{ij} + \in_i \qquad \textbf{(2.13)}$$

Los valores de los atributos/coeficientes $\boldsymbol{B_j}$ correspondientes a cada modelo de cada uno de los cultivos seleccionados se muestran en la tabla A8.1 del anexo 8.

Los resultados de las métricas para la evaluación se describen en la siguiente sección.

4.4.3.2.3. Resultados de los modelos de regresión multivariable

La tabla 35 muestra los resultados de la aplicación de las métricas propuestas en la sección de *diseño de pruebas* a los modelos de regresión multivariable de los cultivos. La tabla 35 muestra la evaluación por medio de validación cruzada, mientras que los resultados obtenidos de la validación simple pueden ser consultados en la tabla A8.2 del anexo 8.

[16] Moverse a través de los atributos removiendo aquel con el valor del coeficiente estandarizado más pequeño hasta que no se observa mejora en la estimación del error dado por el criterio de información Akaike [139].

Cultivo	Eficiencia promedio (%)	Error absoluto medio (ton/ha)	Raíz del error cuadrático medio (ton/ha)	Error relativo absoluto (%)	Raíz del error cuadrático relativo (%)	Número de instancias
Alfalfa	87.77	1.71	2.19	99.53	100.91	63
Algodón	75.19	0.67	0.82	112.53	121.73	23
Cártamo	78.79	0.44	0.82	102.41	100.49	150
Chícharo	43.15	3.14	3.98	148.21	129.74	27
Forraje	0.00	10.17	13.83	295.14	344.68	11
Fríjol	76.12	0.42	0.60	109.57	121.50	129
Frutales	69.93	4.41	5.60	93.86	94.58	23
Garbanzo	78.93	0.38	0.50	105.40	105.30	96
Hortalizas	69.12	3.69	4.70	88.17	92.74	102
Maíz	81.69	1.07	1.66	92.30	95.82	161
Papa	61.45	6.02	8.37	111.25	112.68	105
Sorgo	76.58	0.93	1.17	113.24	101.46	51
Tomate	40.33	7.36	10.19	105.83	107.22	77
Trigo	91.97	0.43	0.57	92.81	94.77	156
Zacate	82.17	2.24	2.58	79.42	70.04	11
Promedios	**67.55**	**2.87**	**3.84**	**116.64**	**119.58**	**1,185**

Tabla 35. Métricas calculadas para la regresión multivariable (validación cruzada).

Como se puede ver en la tabla, la regresión lineal multivariable tiene un error absoluto promedio de **2.87** ton/ha y una eficiencia promedio del **67.55** %. Estos datos deben verse sólo con carácter informativo, ya que, como se ha mencionado, la media es afectada en la presencia de desviaciones, y la mejor forma de evaluar los modelos de regresión lineal es de manera independiente, cultivo a cultivo. El hecho de que la regresión lineal proporcione resultados negativos para el conjunto de datos de un cultivo no debe ser razón para desestimar los modelos de los cultivos restantes. En la sección 4.5.1 (evaluación de resultados) se realiza una comparación de modelos a nivel de cultivo para cada una de las técnicas de modelado seleccionadas.

Un hecho a resaltar, es que mientras la validación simple proporcionó en apariencia un coeficiente de correlación más alto que el de la validación cruzada, la precisión de la predicción utilizando validación simple es afectada con un error mayor que el arrojado por la validación cruzada. Esto se debe a que los factores que contribuyeron a la información reflejada en los registros del conjunto de prueba variaron significativamente con los factores que participaron en la generación de la información de los registros de entrenamiento. Aquí, "factores" debe entenderse como aquellos mecanismos del mundo "real" que contribuyeron a la generación de información, como cuestiones climáticas, económicas e incluso sociales involucrados en el proceso de producción agrícola. La validación cruzada realiza *m* muestreos aleatorios para fabricar los conjuntos de entrenamiento y prueba, por lo que dichos conjuntos estarán un poco más equilibrados, conformados de información reciente y no tan reciente.

En la tabla 35 se muestran sombreados algunos valores de las columnas "Error absoluto relativo" y "Raíz del error cuadrático relativo". Tales valores corresponden a predicciones numéricas que tuvieron mayor certeza que el utilizar la media como estimador. Esta

también es una buena forma de evaluar los modelos de predicción, ya que se comparan contra el "peor" estimador disponible. Para la regresión lineal, se observa que de los errores absolutos relativos de los modelos de seis cultivos están por debajo del 100%, lo que significa que los modelos de regresión lineal de 6 cultivos predicen mejores resultados que la media.

4.4.3.3. Construcción de los modelos perceptrón multicapa

4.4.3.3.1. Parámetros utilizados en la construcción de los modelos perceptrón multicapa

La generación de un modelo perceptrón multicapa entrenado por retropropagacion en Weka requiere los parámetros mostrados en la figura 32.

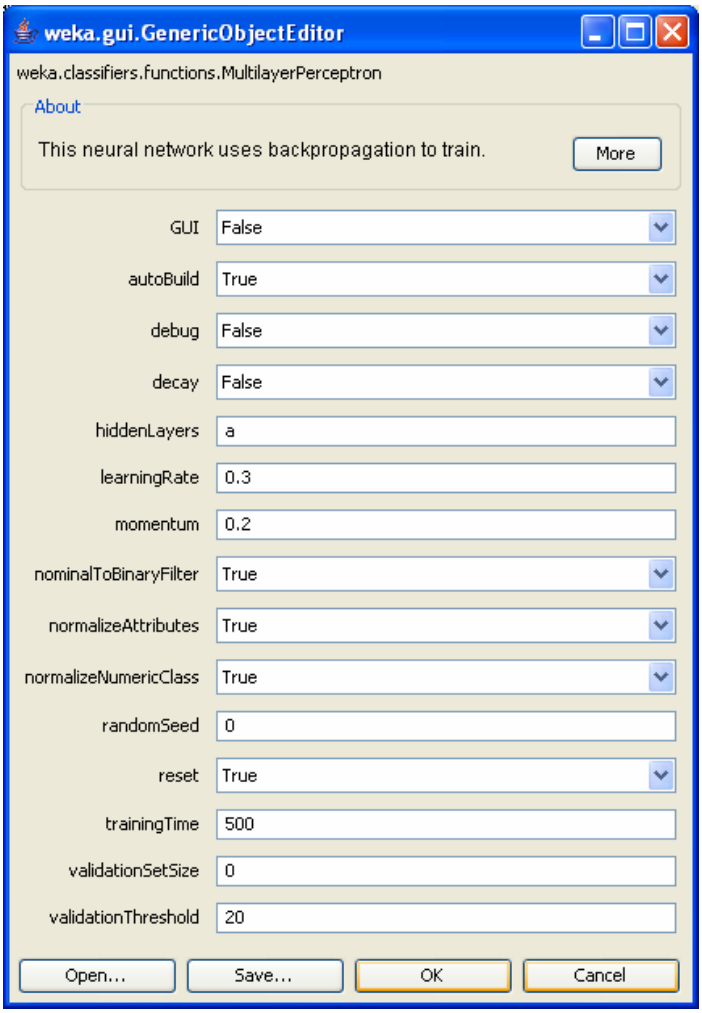


Figura 32. Parámetros para el modelado de redes perceptrón multicapa.

La descripción de los parámetros es la siguiente [139]:

a) *GUI*. Ofrece una interfaz de usuario. Esto permite pausar y alterar la red neuronal durante el entrenamiento.

b) *autoBuild*. Agrega y conecta las capas ocultas en la red.
c) *debug*. Agrega información adicional a la consola para depuración.
d) *decay*. Provoca que la taza de aprendizaje disminuya conforme de acuerdo al número de época (epoch).
e) *hiddenLayers*. Define el número de capas ocultas de la red neuronal. Es una lista de valores positivos enteros, separados por comas. Existen comodines: 'a'=(núm. de atributos + número de clases)/2,'i'=número de atributos, 'o'=número de clases, 't'=número de atributos+número de clases.
f) *learningRate*. Tasa de aprendizaje. La cantidad que afecta los pesos cuando éstos son actualizados.
g) *momentum*. Momentum aplicado a los pesos cuando éstos son actualizados.
h) *nominalToBinaryFilter*. Preprocesa las instancias con un filtro que convierte los valores nominales en binarios.
i) *normalizeAttributes*. Normaliza los atributos.
j) *normalizeNumericClass*. Normaliza la clase si ésta es numérica.
k) *randomSeed*. Semilla utilizada para inicializar el generador de números aleatorios que son utilizados para la incialización de los pesos que son utilizados en las conexiones entre los nodos.
l) *reset*. Permite que se reinicie la red cuando se ha alcanzado una tasa de aprendizaje muy baja (sólo si la GUI está activada).
m) *trainingTime*. Número de ciclos o épocas (epochs) de entrenamiento.
n) *validationSetSize*. El tamaño porcentual del conjunto de validación (el entrenamiento continuará hasta que se observe que el error en el conjunto de entrenamiento se ha vuelto peor o se alcanza el número especificado de épocas).
o) *validationThreshold*. Utilizado para terminar la prueba de validación. Este valor indica cuantas veces en un registro el error de validación puede ser peor antes de que el entrenamiento sea detenido.

Los valores de los parámetros utilizados para la construcción de las redes neuronales del proyecto de minería de esta tesis se proporcionan en la tabla 36.

Parámetro Weka	Valor establecido para el proyecto	Comentario
GUI	False	
autoBuild	True	
debug	False	
decay	False	La tasa de aprendizaje es constante durante todo el entrenamiento.
hiddenLayers	(Variable)	Se determinó según el cultivo (ver explicación).
learningRate	0.3	Valor de facto Weka.
momentum	0.2	Valor de facto Weka.
nominalToBinaryFilter	True	Valor de facto Weka. No afecta a los modelos de los cultivos, ya que no hay atributos nominales.
normalizeAttributes	True	Valor de facto Weka. Es lo más óptimo.
normalizeNumericClass	True	Valor de facto Weka. Es lo más óptimo.
randomSeed	0	Valor de facto Weka.
reset	True	Valor de facto Weka.
trainingTime	(Variable)	Se determinó según el cultivo (ver explicación).
validationSetSize	0	Valor de facto Weka. La realización se hace por medio de la validación cruzada.
validationThreshold	20	Valor de facto Weka. No afecta, dado que el validationSetSize es 0.

Tabla 36. Parámetros para la construcción de los modelos de redes perceptrón multicapa

Como se puede ver en la tabla 36, los únicos parámetros que fueron cambiados con respecto a la configuración de facto ofrecida por Weka para la construcción de los modelos fueron los parámetros *hiddenLayers* y *trainingTime*. El parámetro *hiddenLayers* prácticamente dicta la topología de la red, describiendo el número de capas ocultas y el número de nodos por capa. *trainingTime* indica el número de ciclos dedicados al entrenamiento. Ambos parámetros influyen de manera directa en el desempeño de la red.

Establecer ambos parámetros es en sí un problema de diseño de la red neuronal. Pese a que existen trabajos ([192][193]) que proponen métodos para encontrar de forma automática el diseño óptimo de una red neuronal artificial, esta sigue siendo por mucho un proceso de prueba y error que depende grandemente de la experiencia con aplicaciones similares [194]. Bajo esta perspectiva, se propuso un método para encontrar los parámetros óptimos de los modelos de red neuronal de los cultivos, basándose en la prueba sistemática de valores para los parámetros *hiddenLayers* y *trainingTime*.

La prueba consistió en buscar entre diferentes configuraciones de red y por distintos tiempos de entrenamiento. Esta búsqueda se dividió en dos partes, una que buscaba de forma secuencial entre tiempos de entrenamiento relativamente pequeños y de la cual se infería un comportamiento del desempeño de la red (por ej., si se incrementa la precisión conforme aumentan los tiempos de entrenamiento, la precisión es oscilatoria, los beneficios son mínimos, etc.). La segunda parte consistía en mejorar el tiempo obtenido en la etapa anterior, utilizando la información inferida de la primera etapa. El proceso anterior se repite para cada topología planteada (número de capas y número de nodos propuestas).

La métrica utilizada para la comparación fue el coeficiente de correlación de a) valores generados por las redes neuronales entre b) los valores reales. El coeficiente de correlación denota que tanta aproximación existe entre ambas variables basándose en la diferencia de sus valores.

Para efectos del proyecto se utilizó una topología de red de una sola capa. Trabajos relacionados [39][38] habían basado sus resultados en topologías de una sola capa, por lo que se decidió usar una configuración similar. Para los nodos, se probaron redes de 1, 3, 10 y 15 nodos en la capa oculta (la única excepción fue para elaborar la red del cultivo forraje, ya que fue necesario utilizar 40 nodos ocultos). La métrica registrada para comparar los resultados fue la el coeficiente de correlación entre los valores reales y los valores calculados. En el anexo 9 se muestran los resultados para cada una de las topologías probadas junto con los tiempos de entrenamiento utilizados. En la tabla 37 se señalan las topologías con los tiempos que generaron los mejores resultados.

Cultivos	Coeficiente de correlación	Nodos en la capa oculta	Tiempo de entrenamiento (epochs)
Alfalfa	0.241	10	100
Algodón	0.365	10	600
Cártamo	0.244	1	20
Chícharo	0.183	1	20
Forraje	0.157	40	100

Fríjol	0.109	10	3300
Frutales	0.467	15	1000
Garbanzo	0.235	1	7000
Hortalizas	0.380	10	125
Maíz	0.398	15	150
Papa	0.157	15	1000
Sorgo	0.217	15	200
Tomate	0.285	10	200
Trigo	0.364	1	225
Zacate	0.687	1	200

Tabla 37. Topologías que generaron mejores resultados para las redes perceptrón multicapa para cada uno de los cultivos evaluados.

Los números de nodos y tiempos indicados en la tabla 37 fueron los utilizados para la construcción de los modelos de redes neuronales del proyecto. En la siguiente sección se describen los modelos obtenidos con estas configuraciones.

4.4.3.3.2. Descripción de los modelos de red neuronal perceptrón multicapa generados

La forma general de una RNA es la de un modelo de "caja negra" en el sentido de que se conoce lo que entra al modelo, pero no se sabe por qué o como es que la red produce una determinada salida [195]. La figura 42 muestra una representación gráfica general del tipo de red producida para el proyecto. Para hacer particular la red de la figura 33, lo único que falta es especificar los pesos N_iS_j que van de las neuronas en la capa de entrada a las neuronas de la capa oculta.

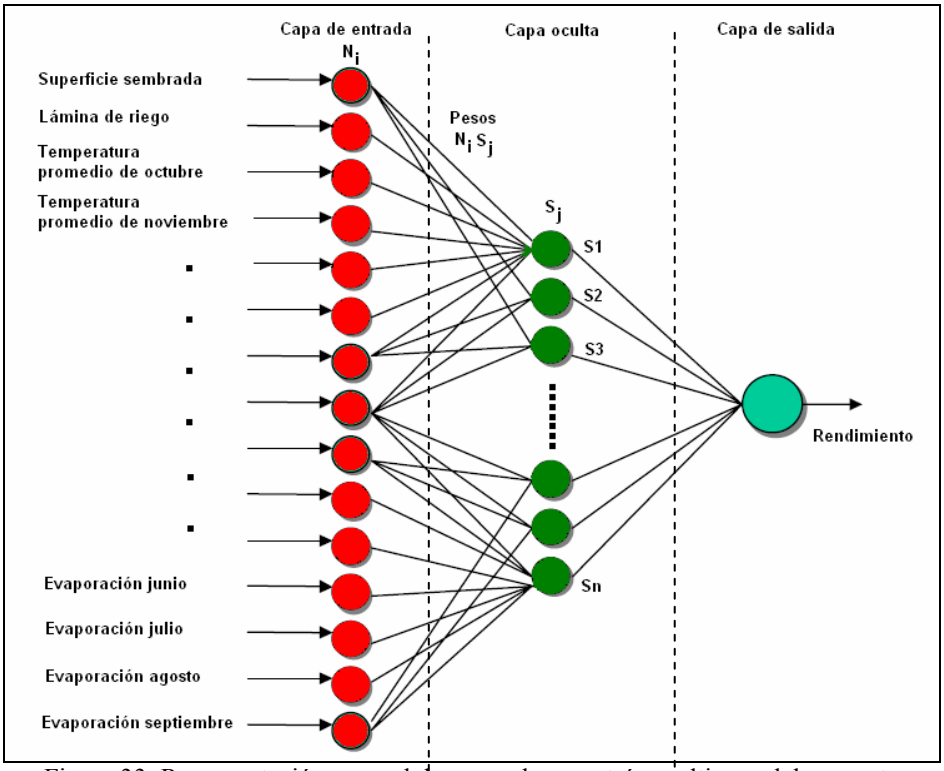


Figura 33. Representación general de una red perceptrón multicapa del proyecto

Los valores de las métricas de evaluación para los modelos perceptrón multicapa de los cultivos se presentan en la siguiente sección.

4.4.3.3.3. Resultados de los modelos de red neuronal perceptrón multicapa

La tabla 38 muestra los resultados de aplicar las métricas propuestas para evaluación en los modelos de redes neuronales perceptrón multicapa para el esquema de validación cruzada. Los resultados de la validación simple se muestran en la tabla A9.5 del anexo 9.

Cultivo	Eficiencia promedio (%)	Error absoluto medio (ton/ha)	Raíz del error cuadrático medio (ton/ha)	Error relativo absoluto (%)	Raíz del error cuadrático relativo (%)	Número de instancias
Alfalfa	85.45	2.08	2.56	121.03	117.59	63
Algodón	64.38	1.00	1.24	167.55	184.10	23
Cártamo	75.64	0.50	0.83	115.33	101.84	150
Chícharo	50.28	2.63	3.38	123.82	110.33	27
Forraje	0.00	1.20	3.99	3.49	9.93	11
Fríjol	63.50	0.61	1.29	158.01	260.64	129
Frutales	54.87	5.98	7.37	127.30	124.55	23
Garbanzo	78.15	0.37	0.49	104.11	103.88	96
Hortalizas	66.59	4.09	5.31	97.77	104.82	102
Maíz	82.13	1.08	1.64	92.37	94.86	161
Papa	48.91	9.24	13.88	170.63	186.93	105
Sorgo	75.52	0.96	1.27	117.61	110.15	51
Tomate	49.00	7.72	10.18	111.08	107.13	77
Trigo	91.49	0.45	0.58	97.96	96.91	156
Zacate	76.67	3.00	3.40	106.54	92.37	11
Promedios	**64.17**	**2.73**	**3.83**	**114.31**	**120.40**	**1,185**

Tabla 38. Métricas de error calculadas para las redes perceptrón multicapa por el método de validación cruzada.

Comparando contra los cálculos por validación simple, los resultados por validación cruzada son más optimistas, y sus correlaciones guardan todos sentidos positivos. El error absoluto medio es en promedio **2.73**, y el error absoluto relativo es de **114.31**. Esto debido a la uniformidad presente en los conjuntos evaluados por el método de validación cruzada. Más adelante en esta tesis, estas desviaciones se compararán contra las producidas por otros métodos de aprendizaje automático y las técnicas estadísticas descritas.

4.4.3.4. Construcción de los modelos de árbol por el algoritmo J48

4.4.3.4.1. Un paso previo: la discretización de atributos para utilizar el algoritmo J48.

El algoritmo J48 es la versión del C4.5 implementado en Weka. En el punto 3 del anexo 6 se describió de manera detallada el funcionamiento de este algoritmo, y se mencionó que el atributo a predecir (la variable dependiente) debe ser un valor discreto. Para fines del

proyecto, la discretización afecta al atributo rendimiento, el cual representa el producto obtenido por hectárea de un cultivo determinado. La discretización de los valores del rendimiento consistirá en representar por medio de un valor discreto un segmento de valores continuos dentro del cual cae el valor de la variable dependiente.

Un proceso típico de discretización se puede ver en la figura 34. Éste consiste de cuatro pasos: (1) ordenamiento de los valores continuos, (2) evaluación de los puntos de corte o la determinación de los intervalos adyacentes para combinación. (3) de acuerdo con algún criterio, dividir o combinar los intervalos de valores continuos, y (4), detener la discretización en algún punto [198].

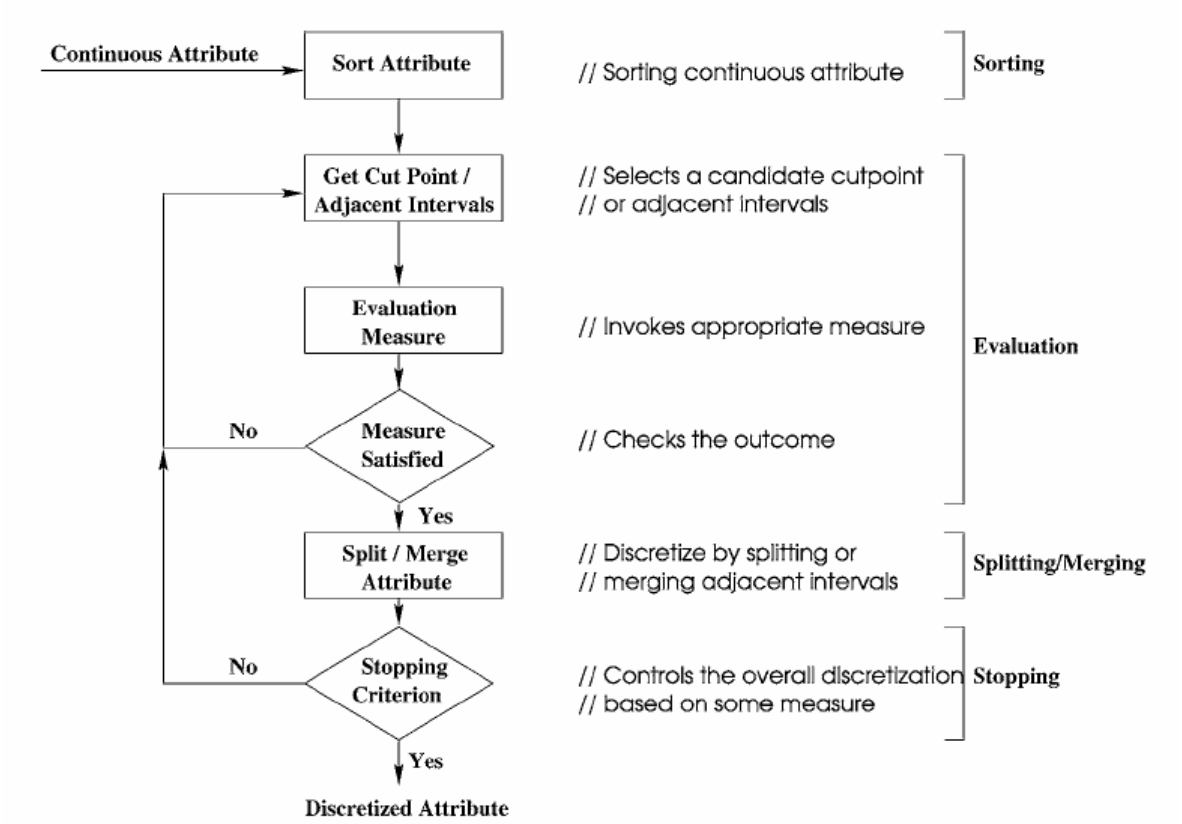


Figura 34. Proceso de discretización por Liu et al (2000) [198]

Existen muchas formas de discretización, pero principalmente, éstas se dividen en formas de discretización supervisada y no supervisada [196]. Básicamente, la discretización supervisada considera la información existente de clases previamente establecidas sobre algún atributo. La discretización no supervisada utiliza únicamente la información numérica existente en el conjunto.

Como se indica en Liu et al (2000) [198], si no existe información de clase disponible, la discretización no supervisada es la única elección. Esta es la condición del atributo rendimiento de la serie de datos objetivo del proyecto, donde no existe ninguna información de clase disponible que pueda ser utilizada para realizar discretización supervisada.

Existen relativamente pocos métodos para la discretización no supervisada. Los métodos más simples son la construcción de intervalos de igual anchura (Equal Width Intervals, EWI) y la construcción de intervalos de igual frecuencia (Equal Frecuency Intervals, EFI). EWI divide el rango de valores observados en para una variable en ***k*** divisiones de igual tamaño, donde ***k*** es un parámetro proporcionado por el usuario [196]. Este método es vulnerable a los datos atípicos presentes en el atributo, ya que pueden influir fuertemente en los rangos generados [197]. EFI es bastante similar a EWI, ya que divide una variable continua en ***k*** intervalos donde dadas ***m*** instancias, los grupos estarán conformados por ***m/k*** (posiblemente repetidos) valores adyacentes [196].

En el proyecto de minería que atañe a esta tesis, se decidió utilizar la técnica EWI por su sencillez y facilidad de implementación. Se consideró que los problemas que pudieran derivarse por posibles datos atípicos fueron superados por el tratamiento de la información de la etapa de *limpieza de datos*.

La discretización de un atributo para un conjunto de datos dado por el método EWI sería el siguiente:

a) *Ordenamiento*. Se ordenan los registros de acuerdo a los valores del atributo a discretizar. Lo normal es que este ordenamiento se haga en forma ascendente, que es lo más natural.

b) *Determinación de los puntos de corte*. Para determinar los puntos de corte, es necesario calcular primero el tamaño del intervalo. Para ello, se utiliza la siguente ecuación [196]:

$$\delta = \frac{x_{max} - x_{min}}{k} \qquad \textbf{(2.14)}$$

Donde:

δ es la amplitud del intervalo.

x_{max} es el valor máximo en el atributo.

x_{min} es el valor mínimo en el atributo.

k es el número de intervalos.

El primer punto de corte se situará de forma de incluir los registros cuyos valores del atributo a discretizar estén entre x_{min} y $x_{min}+\delta$. Dado que los registros están ordenados, estos deben ser los primeros en orden ascendente (hasta que se localice un registro cuyo valor sea mayor que $x_{min}+\delta$). El siguiente punto de corte abarcará los registros cuyos valores en el atributo a discretizar se localizen entre $x_{min}+\delta$ y $x_{min}+2\delta$. Este procedimiento se sigue de manera sucesiva, hasta el último punto de corte, que estará situado justo antes del registro cuyo valor es mayor a $x_{min}+ (k-1)\delta$.

c) *Asignar el valor discreto*: Los valores discretos son asignados a un nuevo atributo o a una versión transformada del atributo original. Generalmente, se almacena el índice que indica el número del intervalo correspondiente, pero también puede

asignarse una literal (‘a’..’z’) ó un string (por ejemplo, uns cadena que indique el límite inferior y superior del intervalo, como ‘0.34..0.88’).

La discretización tiene un gran impacto en el desempeño de los algoritmos de inducción de reglas. La selección de un ***k*** demasiado pequeño generalmente produce reglas más precisas, pero puede generar intervalos tan grandes que estos carezcan de utilidad práctica. Por el contrario, un valor de ***k*** demasiado grande frecuentemente incrementa el número de reglas. Lamentablemente, la mayoría de los algoritmos que buscan el valor óptimo de ***k*** recaen precisamente en las técnicas de discretización supervisada, y generalmente, encontrar el mejor valor de ***k*** en la discretización no supervisada es un proceso de prueba y error que no puede ser aplicada universalmente a cada atributo [199].

Para observar como afecta la discretización del atributo rendimiento a la construcción de los árboles J48 de la serie de datos objetivo, se optó por realizar una prueba con el conjunto de entrenamiento y la discretización del atributo con distintos valores para ***k***. Así, se generaron 20 árboles J48 para el atributo rendimiento discretizado de 10 hasta 200 intervalos, con incrementos de tamaño 10 (10, 20, 30,…, 200). De cada arbol generado, se registraron a) el tamaño del intervalo (en unidades del atributo rendimiento), b) número de reglas generadas, c) número de instancias correctamente clasificados, d) error de predicción promedio, e) raíz de la media del error cuadrático, f) error relativo absoluto. La tabla A10.1 del anexo 10 muestra la estadística completa de la prueba. La figura 35 muestra el desempeño de los árboles generados (en número de reglas y eficiencia) para cada uno de los valores de k

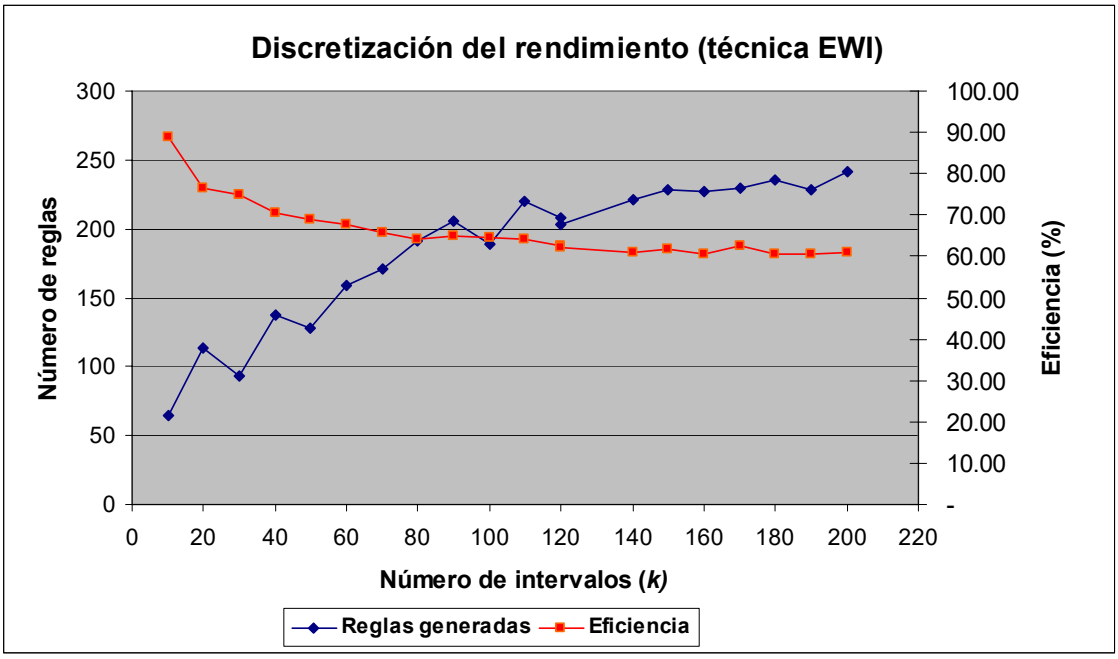


Figura 35. Discretización del rendimiento y su efecto en la construcción de árboles J48.

Como se puede ver en la figura 35, cuando el número de intervalos aumenta, la precisión como clasificador del árbol J48 disminuye. El beneficio en utilizar un mayor número de intervalos es que su precisión con respecto al rendimiento predicho aumenta, ya que la

clase predicha corresponde a un intervalo que es más pequeño conforme el número de intervalos aumenta. Para clarificar esto, obsérvese la tabla 36. Cuando se selecciona discretizar el rendimiento en 10 intervalos, cada intervalo corresponde a un rango de 4.15 toneladas. Si se le pregunta a un productor agrícola, seguramente hará la observación de que predecir el rendimiento de un cultivo con un margen de error de 4.15 ton/ha (en el peor caso) no le es demasiado útil. Su opinión seguramente cambiará al ofrecer una precisión de 0.5 toneladas (correspondiente a un número de intervalos cercano a 80). Con 80 intervalos, el número de reglas aumenta a 191 (para todo el conjunto de prueba), que se hace complicado de leer, lo cual es también un factor a tomar en cuenta al seleccionar el número de intervalos.

Intervalos	Tamaño del intervalo (ton/ha)	Hojas (número de reglas)
10	4.15	65
20	2.08	114
30	1.38	93
40	1.04	138
50	0.83	128
60	0.69	159
70	0.59	171
80	0.52	191
90	0.46	205
100	0.42	189
110	0.38	220
120	0.35	208
120	0.35	203
140	0.30	221
150	0.28	228
160	0.26	227
170	0.24	230
180	0.23	235
190	0.22	228
200	0.21	242

Tabla 39. Número de intervalos sometidos a prueba y su efecto en el tamaño del intervalo y el número de reglas generadas para todo el conjunto de prueba.

La decisión final para el valor de ***k*** fue un número personalizado por cultivo que favoreciera tanto la clasificación como a la precisión del intervalo del rendimiento. El número de reglas se dejó en segundo término, pero aún se siguió contemplando como un factor de importancia. Se propuso buscar intervalos que midieran por lo menos el 10% del tamaño del intervalo global del rendimiento de cada cultivo (lo cual significaba en un principio un valor de ***k***=10). El hacer este cálculo de manera directa implicaba aceptar los valores atípicos presentes en los datos del atributo rendimiento, por lo que se buscó una manera de penalizar aquellos intervalos globales de los cultivos que presentaran más desviaciones en sus datos, con el fin de hacer intervalos más precisos. El mecanismo utilizado consistió en afectar de manera negativa el tamaño del intervalo global, al restar de éste el porcentaje que corresponde a la división de la desviación estándar entre la media. Lo anterior generó intervalos más pequeños para los cultivos que presentaran una mayor

desviación en sus datos. La tabla 40 muestra la forma en como se calcularon los números de intervalos para cada uno de los cultivos presentes en la serie de datos objetivo.

Cultivo	Valor mínimo *(ValMin)*	Valor máximo *(ValMax)*	Promedio *(m)*	Desviación estándar *(DE)*	% Desviación respecto a la media *PDE=(DE/m)* 100*	Rango global *RG=ValMax - ValMin*	Rango normalizado *RN=RG* (1-PDE/100)*	Tamaño del intervalo *δ=RN* 10%*	Número de intervalos *k=RG/δ*
Alfalfa	8.50	20.00	14.65	2.15	14.70	11.50	9.81	0.98	12
Algodón	1.80	4.00	2.86	0.63	22.12	2.20	1.71	0.17	13
Cártamo	0.80	9.90	2.13	0.81	38.05	9.10	5.64	0.56	16
Chícharo	2.00	15.00	5.71	2.93	51.35	13.00	6.32	0.63	21
Forraje	2.50	15.00	10.22	3.67	35.92	12.50	8.01	0.80	16
Fríjol	0.50	3.60	2.11	0.49	23.31	3.10	2.38	0.24	13
Frutales	7.00	26.00	20.04	5.56	27.72	19.00	13.73	1.37	14
Garbanzo	1.00	3.00	1.93	0.46	24.00	2.00	1.52	0.15	13
Hortalizas	2.04	35.00	17.44	4.96	28.44	32.96	23.58	2.36	14
Maíz	1.20	16.22	6.76	1.72	25.45	15.02	11.19	1.12	13
Papa	3.00	42.00	29.74	7.37	24.78	39.00	29.34	2.93	13
Sorgo	1.50	8.50	4.89	1.13	23.03	7.00	5.39	0.54	13
Tomate	4.40	60.00	19.81	9.26	46.74	55.60	29.61	2.96	19
Trigo	3.50	7.00	5.55	0.60	10.73	3.50	3.12	0.31	11
Zacate	9.00	20.00	13.36	3.44	25.78	11.00	8.16	0.82	13

Tabla 40. Cálculo del número de intervalos para cada cultivo.

Los cálculos o variables utilizados en el cálculo de los intervalos se muestran en los encabezados de las columnas de la tabla 40. El número de intervalos que aparece en la última columna de la tabla fue el utilizado para construcción de los modelos de árbol J48 del proyecto. Los parámetros utilizados y la descripción de los modelos se proporcionan en las siguientes secciones.

4.4.3.4.2. Parámetros utilizados en la construcción de los modelos de árbol del algoritmo J48

La figura 36 muestra la interfaz de usuario utilizada por Weka para la introducción de los valores de los parámetros que intervienen en la construcción de los modelos de árbol J48.

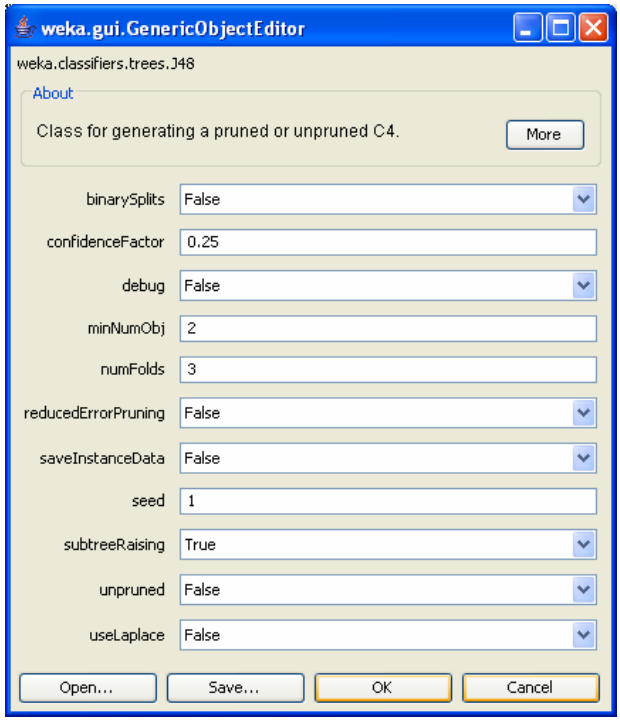


Figura 36. Parámetros para generar un árbol J48 en Weka.

Los parámetros son los siguientes:

a) *binarySplits*.- Indica cuando utilizar cortes binarios en atributos nominales durante la construcción de los árboles.
b) *confidenceFactor*. Factor de confianza utilizado para la poda (valores más pequeños incrementan la poda).
c) *debug*. Si el valor del parámetro es verdadero, el clasificador envía información adicional a la consola.
d) *minNumObj*. El número mínimo de instancias por hoja.
e) *numFolds*. Determina la cantidad de datos utilizados para reducir el error de la poda. Una división es utilizada para la poda, el resto para crecimiento del árbol.
f) *reducedErrorPruning*. Utiliza el error reducido para la poda en lugar del mecanismo del C4.5.
g) *saveInstanceData*. Cuando guardar los datos para visualización.
h) *seed*. La semilla utilizada para la toma aleatoria de datos cuando se produce un error de poda.
i) *subtreeRaising*. Considerar el ascenso de un subárbol durante la poda.
j) *unpruned*. Indica si efectuar la poda o no.
k) *useLaplace*. Indica si debe utilizar Laplace para el suavizado de la cuenta de hojas.

Los valores de los parámetros utilizados para la construcción de los árboles J48 se da en la tabla 41.

Parámetro Weka	Valor establecido para el proyecto	Comentario
binarySplits	False	Valor de facto.
confidenceFactor	0.25	Valor de facto.

debug	False	Valor de facto.
minNumObj	2	Valor de facto.
numFolds	3	Valor de facto.
reducedErrorPruning	True	Valor de facto.
saveInstanceData	False	Valor de facto.
seed	1	Valor de facto.
subtreeRaising	True	Valor de facto.
unpruned	False	Valor de facto.
useLaplace	False	Valor de facto.

Tabla 41. Parámetros para la construcción de los árboles J48.

Como se aprecia en la tabla 41, no fue necesario alterar la configuración sugerida por Weka para la construcción de los árboles. En la siguiente sección se describen los modelos de árbol obtenidos, y en la sección 4.4.3.4.4 se proporciona información relacionada con la precisión de dichos modelos.

4.4.3.4.3. Descripción de los modelos generados basados en reglas del algoritmo J48

Los modelos de árbol J48 se representan como un árbol invertido que va realizando pruebas sobre atributos y tomando distintos caminos hasta concluir con una decisión (o una asignación de clase). En la figura 46 se muestra una representación de un modelo de árbol generado por Weka para el caso del cultivo zacate.

```
J48 pruned tree
------------------

TempDic <= 10.2
|   SuperficieSembrada <= 58.37: '(11.538462-12.384615]' (4.0/1.0)
|   SuperficieSembrada > 58.37: '(19.153846-inf)' (2.0)
TempDic > 10.2: '(9.846154-10.692308]' (2.0/1.0)

Number of Leaves  :   3
```

Figura 37. Árbol J48 generado en Weka para la información del cultivo zacate.

En el árbol de la figura 37, la primera prueba se realiza sobre el atributo *TempDic* (que representa la temperatura promedio del mes de diciembre). Si el valor de la temperatura de diciembre es menor a 10.2 grados, entonces la siguiente prueba recae en la comparación del valor del atributo *SuperficieSembrada*. Esta tiene también dos caminos, una para la superficie mayor a 58.37 ha, la cual conduce a que se obtiene un rendimiento aparente de entre 11.5 y 12.38 ton/ha; la otra para superficie menor a 58.37 ha, la cual produce un rendimiento mayor a 19.15 ton/ha. Si la temperatura de diciembre es mayor a 10.2 grados, entonces se produce un rendimiento entre 9.8 y 10.69 ton/ha.

La conversión de un árbol de decisión a reglas de producción es realmente sencilla. Únicamente se debe hacer el seguimiento desde la raíz hasta cada una de las hojas, modelando cada bifuración como una condición que forma parte del antecedente de la regla. Cada condición adicional que se une al antecedente es precedida por un operador Y (AND). La hoja se modela como el consecuente de dicha regla, y se repite esta secuencia por cada hoja del árbol. De esta manera, el árbol de la figura 37 quedaría representado por el siguiente conjunto de reglas:

SI *(TempDic <= 10.2)* ***Y*** *(SuperficieSembrada <= 58.37)* ***ENTONCES***
RendimientoCategorico = '(11.538462-12.384615]'

SI *((TempDic <= 10.2)* ***Y*** *(SuperficieSembrada > 58.37)* ***ENTONCES***
RendimientoCategorico = '(19.153846-inf)'

SI *(TempDic > 10.2)* ***ENTONCES***
RendimientoCategorico = '(9.846154-10.692308]'

En total, se generaron 15 modelos de árbol J48, uno para cada cultivo seleccionado. En la figura A10.1 del Anexo 10, se muestra la representación del árbol J48 del cultivo alfalfa. La siguiente sección presenta los resultados de la evaluación realizada a los modelos de árbol J48.

4.4.3.4.4. Resultados de los modelos de árbol J48

Los resultados de los árboles J48 son evaluados en dos aspectos distintos. El primer aspecto consiste en evaluar los árboles J48 como un conjunto de reglas generadas para clasificación de los registros de producción agrícola. Para ello, se utilizan las métricas para evaluación de reglas definidas en la sección 4.4.2. (diseño de pruebas). El segundo aspecto que se evalúa es la precisión numérica que pudiera ser obtenida de tales reglas, generando un valor continuo a partir de la clase, de modo que este valor pueda ser comparado con el valor real.

En la tabla 42 se muestran los resultados de la métrica precisión de la predicción para la clasificación realizada por los árboles J48 desarrollados. La tabla incluye los resultados de la validación cruzada y la validación simple.

	Validación simple				Validación cruzada			
Cultivo	Número de instancias	Número de reglas	Instancias correctamente clasificadas (%)	Estancias clasificadas incorrectamente (%)	Número de instancias	Número de reglas	Instancias correctamente clasificadas (%)	Estancias clasificadas incorrectamente (%)
Alfalfa	13	21	30.77	69.23	63	16	23.81	76.19
Algodón	5	5	0	100	23	4	4.35	95.65
Cártamo	30	10	43.33	56.67	150	13	58.67	41.33
Chícharo	6	7	0	100	27	6	25.93	74.07
Forraje	3	4	33.33	66.67	11	3	9.09	90.91
Fríjol	26	32	7.69	92.31	129	23	11.63	88.37
Frutales	5	8	40	60	23	2	13.04	86.96
Garbanzo	20	16	20	80	96	13	23.96	76.04
Hortalizas	21	5	9.52	90.48	102	3	35.29	64.71
Maíz	33	20	21.21	78.79	161	19	47.83	52.17
Papa	21	26	9.52	90.48	105	17	39.05	60.95
Sorgo	11	11	18.18	81.82	51	12	17.65	82.35
Tomate	16	19	6.25	93.75	77	15	19.48	80.52
Trigo	33	39	18.18	81.82	156	27	28.21	71.79

Zacate	3	5	0	100	11	3	0	100
Total	**246**	**228**	**17.2**	**82.8**	**1185**	**176**	**23.86**	**76.14**

Tabla 42. Evaluación de los árboles J48 como clasificadores de los registros de producción agrícola.

En la tabla 42 se puede observar que no hay grandes diferencias entre los resultados de la validación simple y la validación cruzada para la métrica de precisión de predicción. De manera global, se puede decir que alrededor de un 20% de las instancias presentes en la serie de datos objetivo son clasificadas correctamente, mientras el 80% restante no lo son. El panorama cambia cuando se observa la precisión de forma independiente por cultivo. Por ejemplo, para el cártamo se alcanzan a clasificar correctamente hasta un 58.67 % de las instancias por la prueba de validación cruzada (y un 43.33 % en la validación simple). Para el zacate, en ambas pruebas de validación se obtiene una precisión de 0. Resumiendo, se observa que los niveles de precisión en la predicción para los dos tipos de validación son bajos, ya que la gran mayoría no alcanza ni un 50% de valor de precisión.

Además de la precisión de predicción, en el *diseño de pruebas* se establecieron métricas que permiten medir cada una de las reglas generadas a partir de los árboles J48. Esto se realiza con el fin de seleccionar las reglas que manifiesten un mejor desempeño con el fin de obtener conjuntos robustos y relativamente pequeños de las mismas. La mejor aportación de las reglas como clasificadores es su enorme poder descriptivo, el cual aumenta cuando el número de reglas disminuye pero la calidad de soporte y cobertura se mantiene. Para el caso de esta tesis, las métricas de evaluación de reglas fueron aplicadas con el propósito de obtener una calificación cuantitativa de la calidad impresa en las reglas generadas y no como un criterio para la selección de reglas.

La tabla 43 muestra el valor promedio de las métricas para cada uno de los conjuntos de reglas generados. Realmente, el verdadero valor de la aplicación de las métricas se tiene en la calificación a nivel de regla, lo que proporciona un parámetro para aceptar o rechazar la regla. Los promedios mostrados en la última fila de la tabla 43 son valores que proporcionan una vista general de la calidad de las reglas en conjunto, pero para aceptar o rechazar una regla específica los valores de las métricas deben analizarse a nivel individual.

Cultivo	Soporte promedio	Cobertura promedio x regla	Confianza/ precisión promedio	Valor agregado promedio	IS	Piatetsky-Shapiro
Alfalfa	0.64	0.06	0.63	0.50	0.47	0.03
Algodón	0.44	0.25	0.51	0.40	0.71	0.07
Cártamo	0.76	0.08	0.69	0.38	0.42	0.02
Chícharo	0.71	0.17	0.72	0.57	0.77	0.09
Forraje	0.63	0.33	0.58	0.38	0.76	0.13
Fríjol	0.55	0.04	0.62	0.52	0.44	0.02
Frutales	0.50	0.50	0.59	0.34	0.76	0.10
Garbanzo	0.58	0.08	0.72	0.57	0.57	0.03
Hortalizas	0.56	0.33	0.60	0.41	0.77	0.05
Maíz	0.75	0.05	0.74	0.54	0.47	0.02
Papa	0.68	0.06	0.71	0.61	0.57	0.03
Sorgo	0.63	0.08	0.68	0.55	0.56	0.04

Tomate	0.62	0.07	0.61	0.52	0.55	0.03
Trigo	0.02	0.04	0.69	0.56	0.40	0.02
Zacate	0.75	0.33	0.75	0.50	0.86	0.16
Promedio	**0.59**	**0.16**	**0.66**	**0.49**	**0.60**	**0.06**

Tabla 43. Promedios de las métricas aplicadas para la evaluación de reglas por cultivo.

Además de evaluar la precisión de los modelos J48 como clasificadores, también se propuso comparar de manera numérica la clasificación realizada por los árboles. Para ello, se tomó el valor medio del intervalo representado por la clase asignada a cada registro durante la discretización. Este valor se asumió como la predicción numérica del modelo de árbol, y fue comparado contra el valor real. Para ejemplificar esto, supóngase que el atributo rendimiento de los registros del cultivo alfalfa fueron discretizados en las 12 clases que se muestran en la tabla 44.

Intervalo #	Inicio	Fin	Medio
1	8.50	9.46	8.98
2	9.46	10.42	9.94
3	10.42	11.38	10.90
4	11.38	12.33	11.85
5	12.33	13.29	12.81
6	13.29	14.25	13.77
7	14.25	15.21	14.73
8	15.21	16.17	15.69
9	16.17	17.13	16.65
10	17.13	18.08	17.60
11	18.08	19.04	18.56
12	19.04	20.00	19.52

Tabla 44. Discretización del rendimiento para el cultivo alfalfa.

Ahora, asúmase que se desea clasificar una instancia del cultivo alfalfa con los siguientes valores en sus atributos:

- Superficie sembrada (SupSembrada):126 ha.
- Precipitación de mayo (PrecipMay): 0.00 mm.
- Temperatura promedio de julio (TempJul): 26.9 °C.
- Rendimiento:16 ton/ha.

Dado el árbol J48 para el cultivo alfalfa (figura A10.1 del anexo 10), la instancia con la información proporcionada quedaría clasificada con un rendimiento de clase '(11.375-12.333333]' (intervalo #4 de la tabla 44). Esta clase, que realmente representa un intervalo de valores continuos, puede ser representada por el valor medio de 11.85 ton/ha, al sumar al valor inicial del intervalo la mitad de la distancia entre éste valor y el extremo final. Usando este valor, la diferencia con el rendimiento real se calcula como:

$$\boldsymbol{r} = 16.00\text{-}11.85$$
$$\boldsymbol{r} = 4.15.$$

Siguiendo este procedimiento se calcularon todas las diferencias para los conjuntos de prueba fabricados en la etapa de *diseño de pruebas*. Con estas diferencias se calcularon las métricas propuestas para comparaciones de modelos de regresión presentadas en la sección 4.4.2.1. Los resultados para la validación cruzada son mostrados en la tabla 45. La tabla A10.2 del anexo 10 muestra los resultados para la validación simple.

Cultivo	Eficiencia promedio (%)	Error absoluto medio (ton/ha)	Raíz del error cuadrático medio (ton/ha)	Error relativo absoluto (%)	Raíz del error cuadrático relativo (%)	Número de instancias
Alfalfa	86.79	1.871	2.454	114.427	119.409	63
Algodón	71.49	0.839	1.040	149.487	163.721	23
Cartamo	84.87	0.341	0.785	91.389	101.967	150
Chicharo	74.90	1.765	3.054	87.280	106.006	27
Forraje	33.10	4.332	5.479	137.500	153.126	11
Frijol	72.11	0.542	0.700	145.721	145.389	129
Frutales	57.60	6.432	8.069	152.680	148.917	23
Garbanzo	76.33	0.441	0.607	127.499	134.778	96
Hortalizas	69.14	4.039	5.570	104.864	118.158	102
Maiz	82.94	1.019	1.751	92.290	107.604	161
Papa	69.17	5.314	8.291	99.847	114.156	105
Sorgo	74.84	1.066	1.388	127.880	123.009	51
Tomate	46.79	8.019	10.698	115.064	115.852	77
Trigo	91.59	0.453	0.651	104.774	112.747	156
Zacate	78.46	3.154	4.137	129.598	128.722	11
Total	**71.34**	**2.642**	**3.645**	**118.687**	**126.237**	**1,185**

Tabla 45. Resultados de la métricas aplicadas para la evaluación de precisión numérica de las reglas (validación cruzada).

La eficiencia de la estimación numérica a partir de las clasificaciones fue estimada en un **71.34** %. La desviación absoluta promedio que se presentó en la validación cruzada fue de **2.642** ton/ha, mientras que la de la validación simple se dio en **3.626** ton/ha.

Una desventaja que se presenta en la validación del método del valor medio es la necesidad de conocer el valor real del atributo rendimiento en una instancia que ya está discretizada. La comprobación de la validación simple fue relativamente fácil de hacer, ya que los registros eran pocos y fue posible identificar el valor real del atributo rendimiento para un registro por su posición en el conjunto de datos de prueba. Para la validación cruzada, donde ambos conjuntos (entrenamiento y prueba) son elaborados al azar de manera iterativa, fue necesario mantener un registro del valor continuo del rendimiento de las instancias que se toman para prueba, a fin de que pueda ser utilizado para compararse con el valor medio de la clase asignada.

En la tabla 45 también se puede ver que los errores absolutos relativos calculados para cada cultivo favorecen en su mayoría a la media como estimador, cuando ésta es comparada contra la estimación numérica obtenida de los árboles J48.

4.4.3.5. Construcción de los árboles de regresión M5

4.4.3.5.1. Parámetros utilizados en la construcción de los árboles regresión M5

El algoritmo M5 (M5Prime en Weka) [165] permite generar un árbol de modelos de regresión a partir de un conjunto de datos. Los parámetros en Weka para la ejecución del algoritmo se muestran en la figura 38.

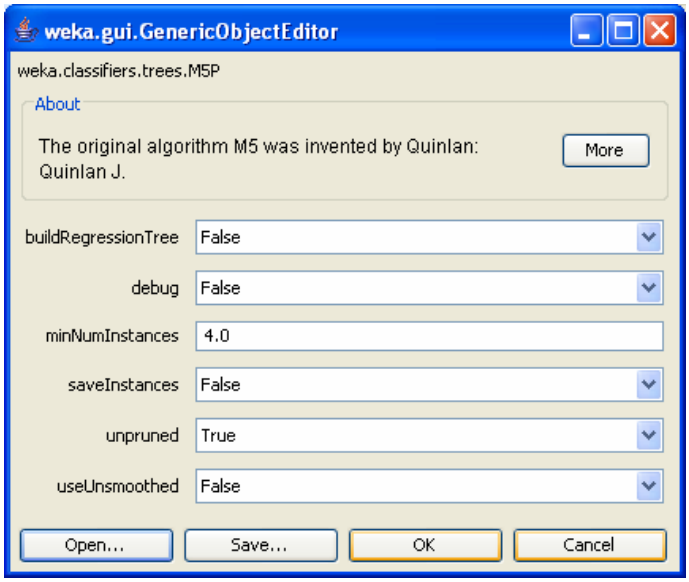


Figura 38. Parámetros para generar un modelo de árbol de regresión M5 en Weka.

Los parámetros son los siguientes:

a) *buildRegressionTree*. Si el valor de este parámetro es *True*, entonces Weka construye un árbol de regresión con valores reales en sus hojas. En caso de que el valor del parámetro sea *False* (por defecto), produce un árbol con modelos lineales en las hojas.
b) *debug*. Proporciona información adicional a la consola para depuración. Valor por de facto: *False*.
c) *minNumInstances*. Número mínimo de instancias para el criterio de paro (ver descripción del algoritmo M5, sección 4.4.1.4). Valor por de facto: *4*.
d) *saveInstances*. Almacena las instancias automáticamente. Valor por de facto: 4.
e) *unpruned*. Si está activado, entonces el árbol es podado utilizando la medida de error de los modelos de las hojas. Valor por de facto: *False*.
f) *useUnsmoothed*. Indica si se realizará el proceso de suavizado (*False*) o si no se realizará (*True*).

En el caso del proyecto, los valor que cambiaron con respecto a la configuración de facto fueron los parámetros *buildRegressionTree* y *unpruned*. *buildRegressionTree* fue cambiado a true para producir árboles de regresión (algoritmo M5 original). *unpruned* fue puesto en *False*, ya que diversas pruebas realizadas con los conjuntos de entrenamiento probaron que los árboles de modelos sin poda proporcionaban un resultado más certero, pero generaban

un número de reglas más alto. Nuevamente, se dio prioridad a la precisión de la predicción sobre el número de reglas, y la poda fue deshabilitada.

En la siguiente sección se describe la estructura de los árboles de regresión generados.

4.4.3.5.2. Descripción de los árboles de regresión M5

Un árbol de regresión M5 tiene un aspecto muy parecido al de un árbol J48. La excepción, es que en lugar de asignaciones de clase, en las hojas se valores constantes. La figura 39 muestra el árbol de regresión generado para el cultivo zacate con la información del conjunto de entrenamiento. Como se puede observar, cada hoja tiene asociada una variable *LMi* la cual describe un valor de rendimiento (el atributo clasificado).

```
SuperficieSembrada <= 73.185 :
|   TempJul <= 26.025 : LM1 (3/0%)
|   TempJul >  26.025 : LM2 (3/39.911%)
SuperficieSembrada >  73.185 : LM3 (2/0%)

LM num: 1
Rendimiento = 13.1706

LM num: 2
Rendimiento = 13.1389

LM num: 3
Rendimiento = 14.4853
```

Figura 39. Árbol de modelos generado por M5 para el cultivo zacate.

El árbol de la figura 39 correspondiente a la información del cultivo zacate, que es uno de los más pequeños. El árbol con más grande fue el del maíz, con un total de 56 reglas. Como ejemplo, en el anexo 11 (figura A11.1) se proporciona el árbol de regresión del cultivo alfalfa.

La siguiente sección proporciona información sobre la eficiencia de los árboles de modelos generados.

4.4.3.5.3. Resultados de los árboles de regresión M5

Los resultados de la evaluación a los árboles de modelos M5 por el método de validación cruzada se presentan en la tabla 46. Los resultados de la validación simple pueden ser consultados en el anexo 11, tabla A11.1.

Cultivo	Número de reglas	Eficiencia promedio (%)	Error absoluto medio (ton/ha)	Raíz del error cuadrático medio	Error absoluto relativo (%)	Raíz del error cuadrático relativo (%)	Número de instancias
Alfalfa	26	88.36	1.63	2.14	94.92	98.26	63.00
Algodón	11	79.56	0.53	0.62	88.17	92.54	23.00
Cártamo	62	79.51	0.41	0.78	93.56	96.31	150.00
Chícharo	11	59.69	1.97	2.99	93.00	97.37	27.00
Forraje	5	35.84	3.79	4.32	109.83	107.62	11.00
Fríjol	56	78.41	0.38	0.50	98.58	99.90	129.00

Frutales	10	67.02	4.76	5.95	101.28	100.47	23.00
Garbanzo	38	80.19	0.35	0.46	97.92	97.67	96.00
Hortalizas	39	65.77	4.05	5.07	96.80	100.18	102.00
Maíz	66	81.16	1.07	1.67	92.15	96.76	161.00
Papa	43	64.61	5.48	7.46	101.23	100.48	105.00
Sorgo	22	78.43	0.80	1.13	98.50	97.35	51.00
Tomate	31	51.71	6.80	9.18	97.74	96.59	77.00
Trigo	65	91.93	0.43	0.55	92.03	91.97	156.00
Zacate	5	79.02	2.90	3.74	102.77	101.51	11.00
Total	**490**	**72.08**	**2.36**	**3.10**	**97.23**	**98.33**	**79.00**

Tabla 46. Resultados de las métricas para la evaluación de los árboles de regresión (validación cruzada).

Como se puede observar en la tabla 46, la eficiencia promedio para los modelos generados por el algoritmo M5 fue de **72.08** %. El error absoluto medio mostró una desviación promedio en el rendimiento de **2.36** ton/ha. Muy buenos resultados con respecto a la media mostró este algoritmo, ya que la mayoría de sus errores relativos se mostraron abajo del 100%, favoreciendo a los resultados de los árboles de modelos.

La siguiente sección proporciona una evaluación global de los indicadores para todos los tipos de modelados realizados.

4.5. Evaluación

4.5.1. Evaluación de resultados

Pasos previos a la evaluación trataron con factores como la precisión y generalidad de cada modelo. Este paso determina el grado en el cual el modelo satisface los objetivos del negocio y busca determinar si existe alguna razón por la cual el modelo o modelos desarrollados sean deficientes. Este compara los resultados con el criterio de evaluación definido al inicio del proyecto [33].

El criterio de éxito al inicio del proyecto fue el obtener modelos que lograran una eficiencia mayor en la predicción del rendimiento con respecto la técnica estadística clásica de la regresión multivariable. Por esta razón, la regresión multivariable fue la primera técnica de modelado explorada. Las otras tres técnicas sometidas a la evaluación pertenecen a la comunidad de aprendizaje automático y fueron tomadas en cuenta por ser técnicas que pueden inferir comportamientos no lineales de los datos.

Para fines de la evaluación, esta tesis se centra en el análisis de tres de las métricas propuestas para la comparación de resultados de modelos de regresión: efectividad promedio, error absoluto medio y el error absoluto relativo.

4.5.1.1. Efectividad promedio

La efectividad promedio expresa el porcentaje en el que los resultados estimados se aproximaban a los reales. La tabla 47 muestra el concentrado de resultados de las métricas para cada uno de los algoritmos empleados. Como se puede apreciar, se resaltan los

Frutales	10	67.02	4.76	5.95	101.28	100.47	23.00
Garbanzo	38	80.19	0.35	0.46	97.92	97.67	96.00
Hortalizas	39	65.77	4.05	5.07	96.80	100.18	102.00
Maíz	66	81.16	1.07	1.67	92.15	96.76	161.00
Papa	43	64.61	5.48	7.46	101.23	100.48	105.00
Sorgo	22	78.43	0.80	1.13	98.50	97.35	51.00
Tomate	31	51.71	6.80	9.18	97.74	96.59	77.00
Trigo	65	91.93	0.43	0.55	92.03	91.97	156.00
Zacate	5	79.02	2.90	3.74	102.77	101.51	11.00
Total	**490**	**72.08**	**2.36**	**3.10**	**97.23**	**98.33**	**79.00**

Tabla 46. Resultados de las métricas para la evaluación de los árboles de regresión (validación cruzada).

Como se puede observar en la tabla 46, la eficiencia promedio para los modelos generados por el algoritmo M5 fue de **72.08** %. El error absoluto medio mostró una desviación promedio en el rendimiento de **2.36** ton/ha. Muy buenos resultados con respecto a la media mostró este algoritmo, ya que la mayoría de sus errores relativos se mostraron abajo del 100%, favoreciendo a los resultados de los árboles de modelos.

La siguiente sección proporciona una evaluación global de los indicadores para todos los tipos de modelados realizados.

4.5. Evaluación

4.5.1. Evaluación de resultados

Pasos previos a la evaluación trataron con factores como la precisión y generalidad de cada modelo. Este paso determina el grado en el cual el modelo satisface los objetivos del negocio y busca determinar si existe alguna razón por la cual el modelo o modelos desarrollados sean deficientes. Este compara los resultados con el criterio de evaluación definido al inicio del proyecto [33].

El criterio de éxito al inicio del proyecto fue el obtener modelos que lograran una eficiencia mayor en la predicción del rendimiento con respecto la técnica estadística clásica de la regresión multivariable. Por esta razón, la regresión multivariable fue la primera técnica de modelado explorada. Las otras tres técnicas sometidas a la evaluación pertenecen a la comunidad de aprendizaje automático y fueron tomadas en cuenta por ser técnicas que pueden inferir comportamientos no lineales de los datos.

Para fines de la evaluación, esta tesis se centra en el análisis de tres de las métricas propuestas para la comparación de resultados de modelos de regresión: efectividad promedio, error absoluto medio y el error absoluto relativo.

4.5.1.1. Efectividad promedio

La efectividad promedio expresa el porcentaje en el que los resultados estimados se aproximaban a los reales. La tabla 47 muestra el concentrado de resultados de las métricas para cada uno de los algoritmos empleados. Como se puede apreciar, se resaltan los

resultados de aquellas técnicas que obtienen un valor mayor de efectividad a nivel modelo de cultivo. Así, en la fila correspondiente al cultivo Alfalfa, se resalta el resultado del algoritmo M5 ya que fue el que obtuvo el valor más alto para la métrica de efectividad promedio. En el caso del cultivo del cártamo (fila 3), el método para extraer un valor numérico del árbol J48 proporciona mayor certeza de predicción que cualquier otra técnica.

Cultivo	Regresión multivariable	Perceptrón multicapa	J48	M5
Alfalfa	87.77	85.45	86.79	88.36
Algodón	75.19	64.38	71.49	79.56
Cártamo	78.79	75.64	84.87	79.51
Chícharo	43.15	50.28	74.90	59.69
Forraje	0.00	0.00	33.10	35.84
Fríjol	76.12	63.50	72.11	78.41
Frutales	69.93	54.87	57.60	67.02
Garbanzo	78.93	78.15	76.33	80.19
Hortalizas	69.12	66.59	69.14	65.77
Maíz	81.69	82.13	82.94	81.16
Papa	61.45	48.91	69.17	64.61
Sorgo	76.58	75.52	74.84	78.43
Tomate	40.33	49.00	46.79	51.71
Trigo	91.97	91.49	91.59	91.93
Zacate	82.17	76.67	78.46	79.02
Promedios	**67.55**	**64.17**	**71.34**	**72.08**

Tabla 47. Resultados de la métrica de efectividad promedio para cada una de las técnicas de modelado (validación cruzada).

El algoritmo que registró la efectividad promedio más alta un mayor número de veces fue el árbol de regresión M5, con 7 resultados superiores, y una efectividad promedio de **72.08**%. La diferencia es realmente poca con respecto los modelos generados por el algoritmo J48, que registró en 5 ocasiones los resultados más altos, con un resultado promedio de **71.34** %. El siguiente en la lista sería la técnica de regresión lineal multivariable, con 3 resultados favorables y un promedio de efectividad del **67.55** %. Los modelos de red neuronal perceptrón multicapa no registraron ningún resultado que sobresaliera a los demás, y en promedio obtuvo una efectividad del **64.17** %.

Los resultados cambian en la validación simple (tabla A12.1 del anexo 12). En este caso, el algoritmo M5 obtiene en 7 ocasiones las efectividades promedio más altas, seguido por la regresión multivariable (con 4 resultados favorables). Los modelos de red neuronal obtienen 3 y el algoritmo J48, únicamente un resultado superior. El promedio de efectividad fue de **71.93**, **63.31**, **51.40** y **62.03** %, respectivamente. Al igual que en la validación cruzada, los resultados de la validación simple favorecen al algoritmo M5, con una ventaja mucho más marcada que la que le guarda el algoritmo J48 en la validación cruzada. Nótese que en la validación simple todos los modelos pierden efectividad, pero M5 casi mantiene la efectividad que mostró en la validación cruzada.

La diferencia en la validación cruzada es muy poca para los algoritmos J48 y M5. En la validación simple, la ventaja de M5 es más visible. En ambos tipos de validaciones, la red

neuronal mostró resultados más bajos que la regresión multivariable. Utilizando estos argumentos, una jerarquía sugerida para la métrica de efectividad promedio sería la siguiente: M5, J48, regresión multivariable y perceptrón multicapa.

4.5.1.2. Error absoluto medio

El error absoluto medio muestra el promedio de la desviación en la estimación del rendimiento con respecto al valo real, esto en las mismas unidades que la variable objetivo. La tabla 48 muestra el error absoluto medio para cada una de las técnicas de modelado seleccionadas.

Cultivo	Regresión multivariable	Perceptrón multicapa	J48	M5
Alfalfa	1.71	2.08	1.87	1.63
Algodón	0.67	1.00	0.84	0.53
Cártamo	0.44	0.50	0.34	0.41
Chícharo	3.14	2.63	1.77	1.97
Forraje	10.17	1.20	4.33	3.79
Fríjol	0.42	0.61	0.54	0.38
Frutales	4.41	5.98	6.43	4.76
Garbanzo	0.38	0.37	0.44	0.35
Hortalizas	3.69	4.09	4.04	4.05
Maíz	1.07	1.08	1.02	1.07
Papa	6.02	9.24	5.31	5.48
Sorgo	0.93	0.96	1.07	0.80
Tomate	7.36	7.72	8.02	6.80
Trigo	0.43	0.45	0.45	0.43
Zacate	2.24	3.00	3.15	2.90
Promedios	**2.87**	**2.73**	**2.64**	**2.36**

Tabla 48. Resultados de la métrica de error absoluto medio para cada una de las técnicas de modelado (validación cruzada).

En la tabla 48 se muestran sombreados los mejores resultados para cada cultivo. A diferencia de la efectividad promedio, en el error absoluto medio se considera un mejor resultado a aquel que arroje el valor más bajo. Así, por ejemplo, en el caso de la alfalfa, el algoritmo M5 obtiene el mejor resultado, ya que su promedio de desviación es de **1.63** ton/ha, que es mejor que los resultados de **1.71**, **1.87** y **2.08** de la regresión múltiple, el árbol J48 y el perceptrón multicapa, respectivamente. Contabilizando resultados, el algoritmo M5 obtiene en global 5 resultados que están sobre otros algoritmos, y en promedio, obtiene un valor de **2.36** ton/ha de error absoluto. Le sigue en número de mejores resultados el algoritmo J48, con cuatro modelos mejores y un promedio de error de **2.64** ton/ha. Son curiosos los casos de la regresión multivariable y el perceptrón multicapa, ya que aunque la regresión manifiesta 3 resultados que sobresalen a la demás técnicas, su valor de error absoluto promedio es mayor que el presentado por la red neuronal (pese a que esta última técnica sólo presenta un modelo con el mejor resultado). Esto quiere decir que, aunque tiene muy buenos resultados, cuando la regresión multivariable presenta un error de desviación, este suele ser mayor que los errores generados por las técnicas de aprendizaje automático. Como ejemplo, se pueden observar los resultados del cultivo forraje, donde la

regresión produce un error que es extremadamente alto comparado con los resultados de otras técnicas (10.17 ton/ha).

Los modelos basados en perceptrón multicapa alcanzan una magnitud de error promedio de **2.73** ton/ha con una sola estimación favorable, aunque es de mencionar, que es precisamente en esta desviación donde la regresión multivariable manifestó un error mayor de desviación.

En la validación simple (tabla A12.2 del anexo 12), el algoritmo M5 muestra una ventaja sobre las estimaciones realizadas por otras técnicas al obtener en 10 resultados el valor más bajo de de error, manifestando un error promedio de **2.83** ton/ha. Este error global es el más bajo del conjunto. Le sigue el algoritmo J48 (**3.63** ton/ha), con un resultado con el valor de error más bajo. El perceptrón multicapa tiene un error promedio de **5.19** ton/ha, pero sólo uno de sus modelos obtuvo un resultado por encima de las otras técnicas. La regresión múltiple manifiesta en 3 ocasiones un resultado de error menor que las otras técnicas, pero en promedio, manifiesta un error mayor que cualquiera de las otras técnicas al presentar una magnitud de **6.32** ton/ha, 3 toneladas más que el mejor resultado.

Por los resultados descritos, una jerarquía propuesta para clasificar las técnicas para la métrica de error absoluto medio es la siguiente: M5 (primero), J48, perceptrón multicapa y regresión multivariable (último).

4.5.1.3. Error absoluto relativo

El error absoluto relativo indica en porcentaje la magnitud de la desviación, pero toma como base de comparación de la diferencia producida entre la media y el valor real. Una magnitud menor al 100 % significa que la diferencia absoluta del valor estimado es mejor que la diferencia absoluta respecto a la media. Los resultados de esta métrica se muestran en la tabla 48. Se resaltan en sombreado los resultados más bajos a nivel de cultivo, y con un (*) todos aquellos valores que resultaron menores al 100 %.

Cultivo	Regresión multivariable	Perceptrón multicapa	J48	M5
Alfalfa	99.53*	121.03	114.43	94.92*
Algodón	112.53	167.55	149.49	88.17*
Cártamo	102.41	115.33	91.39*	93.56*
Chícharo	148.21	123.82	87.28*	93.00*
Forraje	295.14	3.49*	137.50	109.83
Fríjol	109.57	158.01	145.72	98.58*
Frutales	93.86*	127.30	152.68	101.28
Garbanzo	105.40	104.11	127.50	97.92*
Hortalizas	88.17*	97.77*	104.86	96.80*
Maíz	92.30*	92.37*	92.29*	92.15*
Papa	111.25	170.63	99.85*	101.23
Sorgo	113.24	117.61	127.88	98.50*
Tomate	105.83	111.08	115.06	97.74*
Trigo	92.81*	97.96*	104.77	92.03*

Zacate	79.42*	106.54	129.60	102.77
Promedios	**116.64**	**114.31**	**118.69**	**97.23**

Tabla 49. Resultados de error absoluto relativo para cada una de las técnicas de modelado (validación cruzada).

Como se puede apreciar en la tabla 49, los modelos M5 produjeron 8 valores más bajos del error absoluto relativo con respecto al resto de las técnicas bajo análisis. También, es el algoritmo M5 el que obtuvo el promedio más bajo de todos, al registrar un **97.23**% de error en promedio. Fue la única técnica de modelado cuyo promedio descendió debajo del 100%. En número de resultados le sigue la regresión múltiple y los árboles J48, con 3 resultados favorables y error absolutos relativos de **116.64**% y **118.69**%, respectivamente. Nuevamente, se nota que cuando ocurren desviaciones en los modelos de regresión, éstas suelen ser más grandes que las que ocurren con los otros métodos. Las redes perceptrón multicapa generaron el siguiente error promedio más bajo: **114.31**%, pero sus modelos únicamente demostraron en una ocasión ser mejores estimadores que la media, aunque esto de una forma muy marcada, ya que dicho modelo logró un error de tan solo el **3.5**% (cultivo forraje), el valor de error más pequeño del conjunto entero de resultados.

Por otro lado, la validación simple confirmó a M5 como el algoritmo que obtiene los mejores resultados frente a la media, al obtener nuevamente en 8 ocasiones el error más bajo alcanzado con un promedio del **98.41**%, muy similar al obtenido durante la evaluación por validación cruzada. El algoritmo J48 registró un error del **134.71**%, obteniendo en 3 ocasiones los errores más bajos a nivel cultivo. La regresión lineal multivariable obtuvo un promedio de error del **188.79**%, y obtuvo resultados favorables en 3 de sus modelos a nivel cultivo. Las redes perceptrón multicapa fueron las que manifestaron los errores absolutos relativos más grandes, al obtener en promedio un **279.90**%. Aún así, en 2 de sus modelos obtuvieron los mejores resultados. La desviación tan anormal de esta última técnica, se debe en gran parte al modelo resultante para el fríjol, el cual produce error frente a la media del **1,512.41** %. Sin dicho valor, el promedio del error de la red neuronal alcanza un promedio del **191.86** %, mostrando una medida similar a la obtenida por la regresión lineal multivariable.

La jerarquía propuesta para la métrica del error absoluto relativo es: M5 (primer), J48, regresión múltiple y perceptrón multicapa (último).

4.5.1.4. Selección de la mejor técnica de modelado

La tabla 50 muestra una clasificación jerárquica de las técnicas de modelado dada la evaluación de las métricas expuestas en los puntos anteriores (4.5.1.1, 4.5.1.2 y 4.5.1.3). Se han incluido las métricas de "raíz del error cuadrático medio" y "raíz del error cuadrático relativo" con el fin de proporcionar un mayor soporte a la selección. La clasificación se basa exclusivamente en los resultados de las métricas, las que finalmente miden la habilidad del modelo para predecir un valor numérico. Cuestiones relacionadas con las desventajas (o ventajas) que se presentan para construir o entrenar cada uno de los modelos son analizadas y comentadas en la sección de conclusiones de este capítulo.

Como se puede ver, la mayoría de las métricas favorecen al algoritmo M5 como la herramienta de modelado que proporciona los mejores resultados. En segunda posición (si se puede decir de este modo), se encuentra la inducción numérica realizada a los árboles generados por el algoritmo J48.

Posición	Eficiencia promedio	Error absoluto medio	Raíz del error cuadrático medio	Error absoluto relativo	Raíz del error cuadrático relativo	Selección
1	M5	M5	M5	M5	M5	**M5**
2	J48	J48	J48	J48	RM	J48
3	RM	PM	PM	RM	PM	RM/PM
4	PM	RM	RM	PM	J48	RM/PM

Tabla 50. Clasificación jerárquica por su desempeño en las métricas para cada una de las técnicas de modelado.

El perceptrón multicapa fue ubicado en 3ra y 4ta posición, porque no se encontraron argumentos suficientes para indicar que se desempeñaba mejor que la regresión multivariable. La eficiencia promedio de la regresión multivariable fue en su mayoría superior a los modelos de red perceptrón multicapa. Las métricas de "error absoluto medio" y la "raíz del error cuadrático medio" mostraron mejores resultados en el perceptrón multicapa, pero en la validación simple, el perceptrón multicapa manifestó mayores desviaciones con respecto a la media que cualquier otra técnica. Otras desventajas de la redes, como la dificultad para el entrenamiento, se discuten en la sección de conclusiones.

En las tablas 51 y 52 se pueden observar algunos de los argumentos utilizados para la clasificación de la tabla 50. En la tabla se muestran los resultados promedios obtenidos de las métricas para cada una de las técnicas de modelado aplicadas. J48 y M5 mantienen resultados muy cerrados en las métricas de "eficiencia promedio", "error absoluto promedio" y "raíz del error cuadrático medio". Lo que finalmente hace resaltar a M5 como un mejor estimador para el caso del rendimiento promedio, es su comparación con respecto a la media como estimador. Esto se puede observar en los valores de los indicadores "error absoluto relativo" y "raíz del error cuadrático relativo", donde M5 mantiene una marcada diferencia con respecto al resto de técnicas, siendo consistentes en ambos tipos de validaciones (simple y cruzada).

Validación simple				
	RM	**PM**	**J48**	**M5**
Eficiencia promedio (%)	63.31	51.40	62.03	71.93
Error absoluto medio (ton/ha)	6.32	5.19	3.63	2.83
Raíz del error cuadrático medio	7.71	6.46	4.56	3.61
Error absoluto relativo (%)	188.51	279.90	134.71	98.41
Raíz del error cuadrático relativo (%)	163.92	273.53	137.54	97.57

Tabla 51. Resultados promedio de cada métrica y técnica de modelado aplicada (validación simple).

Validación cruzada				
	RM	**PM**	**J48**	**M5**
Eficiencia promedio (%)	67.55	64.17	71.34	72.08
Error absoluto medio (ton/ha)	2.87	2.73	2.64	2.36
Raíz del error cuadrático medio	3.84	3.83	3.64	3.10

Error absoluto relativo (%)	116.64	114.31	118.69	97.23
Raíz del error cuadrático relativo (%)	119.58	120.40	126.24	98.33

Tabla 52. Resultados promedio de cada métrica y técnica de modelado aplicada (validación cruzada).

Otra razón que hace parecer mejor a M5 que el resto de técnicas seleccionadas, es su robustez. Comparando los resultados, se puede observar que los valores de las métricas de validación simple y validación cruzada casi no sufren cambios en lo que al algoritmo M5 se refiere. Esto habla de un modelo que mantiene su eficiencia pese a que el conjunto de prueba es cambiado. En el caso de las otras técnicas, existen diferencias importantes en los resultados por validación simple y validación cruzada.

En lo que respecta al objetivo de la tesis, se puede decir que, finalmente, se encontró que las técnicas J48 y M5 obtienen una mejor precisión en la predicción del rendimiento de los cultivos. De estas dos, M5 tiene un mejor resultado con respecto a la media, por lo cual, es seleccionado como el mejor algoritmo para la estimación del rendimiento. Los resultados en la aplicación de los modelos perceptrón multicapa no mostraron grandes diferencias respecto a la regresión multivariable, por lo que no se puede afirmar que proporcionen una eficiencia mayor en la predicción del rendimiento.

Los resultados aquí mostrados no deben ser interpretados como una calificación general del desempeño de los algoritmos. El desempeño de un algoritmo de minería está fuertemente ligado a las propiedades intrínsecas de los datos para los cuales es aplicado. También debe tomarse en cuenta que los resultados expresados en esta tesis fueron obtenidos bajo los parámetros aquí especificados (con valores seleccionados de forma general). Distintos valores para los parámetros, o incluso, algunas técnicas de optimización aplicadas a los mismos algoritmos aquí evaluados, podrían hacer variar los resultados e inclinar la balanza hacia cualquier otro algoritmo.

4.5.2. Revisión del proceso

CRISP-DM recomienda hacer una revisión entera del proceso de minería de datos con el fin de determinar si existe un factor importante o tarea que haya pasado inadvertida. A este punto del ejercicio de minería de datos, una revisión del proceso toma la forma de una revisión de aseguramiento de calidad [33].

Durante esta etapa se hicieron múltiples revisiones y ajustes. Los datos y conclusiones aquí mostradas son el resultado final del proceso entero, incluyendo la etapa de revisión.

4.5.3. Determinando los próximos pasos

De acuerdo a los resultados de la evaluación y a la revisión del proceso, el proyecto decide como preceder en esta etapa. El proyecto necesita decidir cuando finalizar el proceso y moverse hacia la implantación de los resultados, o cuando iniciar nuevas iteraciones o configurar un proyecto completamente nuevo de minería de datos [33].

Para fines de esta tesis, se consideró que con la selección de la técnica de minería de datos que predice mejor el rendimiento, y la comparación de ésta con la técnica de regresión

multivarable, es como concluye el proyecto de minería de datos. Los siguientes pasos consistirán en la implementación de un algoritmo de optimización utilizando los modelos resultantes del proceso de minería de datos (capítulo 5).

5. Desarrollo de un algoritmo para la optimización de la producción agrícola

El capítulo 4 de la tesis describe como se utiliza la minería de datos para encontrar los modelos que predicen con mayor eficiencia el rendimiento generado por los cultivos seleccionados en el distrito de riego bajo estudio. Durante este capítulo se utilizan dichos modelos en un esquema de optimización como elementos clave en el cálculo del ingreso a obtener por la siembra de los cultivos en el distrito. Es importante señalar por lo tanto la importancia de la labor de minería de datos para esta tesis, ya que es el trabajo que finalmente proporciona las herramientas para realizar el algoritmo de optimización que se presenta en este capítulo.

La optimización trata el problema de encontrar la mejor de un conjunto posible de opciones para un determinado objetivo, esto sin violar un cierto número de restricciones. En términos matemáticos, la optimización es el problema de minimizar (o maximizar) una función prescrita, la función objetivo, mientras se obedecen un número de restricciones de igualdad y de desigualdad [200].

Esta sección de la tesis se dedica tratar el problema de la producción agrícola como un problema de optimización, el cual tiene como objetivo la maximización del ingreso obtenido de la venta de cultivos agrícolas. En primera instancia, se proporciona una noción general de lo que es un modelo de optimización, para describir en las mismas condiciones al problema de la maximización del ingreso. Como se verá, no hay solución trivial al problema, por lo que, en segunda instancia, se propone un algoritmo para realizar dicha maximización.

5.1. Formulación de la producción agrícola como un problema de optimización

La formulación de un problema de optimización involucra la toma de instrucciones, la definición general de metas y requerimientos de una actividad dada, y la trascripción de éstos en una serie de declaraciones matemáticas bien definidas. De manera más precisa, la formulación involucra [201]:

1. La selección de un o más variables de optimización.
2. La selección de una función objetivo.
3. La identificación de una serie de condiciones.

En Linares et al (2001) [202] se proporciona la siguiente descripción para los elementos anteriormente citados:

- *Función objetivo*. La función objetivo es una medida cuantitativa del funcionamiento del sistema que se desea optimizar (maximizar o minimizar). Como ejemplo de funciones objetivo se pueden mencionar: la minimización de los costes variables de operación de un sistema eléctrico, la maximización de los beneficios netos de venta de ciertos productos, la minimización del cuadrado de las

desviaciones con respecto a unos valores observados, la minimización del material utilizado en la fabricación de un producto, etc.

- *Variables.* Las variables representan las decisiones que se pueden tomar para afectar el valor de la función objetivo. La función objetivo y las restricciones ó condiciones deben ser todas funciones de una o más variables de optimización. Desde un punto de vista funcional, se pueden clasificar en variables independientes o principales o de control, y variables dependientes o auxiliares o de estado, aunque matemáticamente son todas iguales. En el caso de un sistema eléctrico serán los valores de producción de los grupos de generación o los flujos por líneas. En el caso de la venta, la cantidad de cada producto fabricado y vendido. En el caso de la fabricación de un producto, sus dimensiones físicas.

- *Restricciones.* Las restricciones representan el conjunto de relaciones (expresadas mediante ecuaciones e inecuaciones) que ciertas variables están obligadas a satisfacer. Por ejemplo, las potencias máxima y mínima de operación de un grupo de generación, la capacidad de producción de la fábrica para los diferentes productos, las dimensiones del material bruto del producto, etc.

En términos matemáticos, en una gran cantidad de problemas de optimización pueden ser expresados en la siguiente forma [201]:

Encontrar un vector de variables de optimización $\boldsymbol{x}=(x_1, x_2, \ldots, x_n)^T$, tal que se minimice una función objetivo o de costo $\boldsymbol{f(x)}$.		
Sujeto a:		
$\boldsymbol{g_i(x)} \leq 0$	i =1,2,….,m	Restricciones de desigualdad del tipo "menor que". m denota el número de restricciones.
$\boldsymbol{h_i(x)} = 0$	i =1,2,….,p	Restricciones de igualdad. p denota el número de restricciones.
$\boldsymbol{LI_i \leq x_i \leq LU_i}$	i =1,2,….,n	Límites de variables de optimización. n denota el número de restricciones, LI_i y LU_i son las cotas entre las cuales x_i puede variar, no pudiendo ésta ser menor a LI_i ni superior a LU_i.

Figura 40. Forma estándar de un problema de optimización [201].

En estos términos, la maximización del beneficio económico neto obtenido por la cosecha de cultivos agrícolas en un distrito de riego ya había sido planteado en la sección 2.1.1 como:

$$BN = (IB_1 - C_1)S_1 + (IB_2 - C_2)S_2 + \ldots + (IB_n - C_n)S_n \qquad \textbf{(5.1)}$$

Donde:
i = 1, 2, 3,…,n. Donde n = número de cultivos presentes en la optimización.
BN.- Beneficio económico neto. Expresado en unidad monetaria ($).

IB_i.- Ingreso bruto por *ha* del cultivo *i*. Éste es obtenido de multiplicar el rendimiento por unidad de superficie (ton/ha) del cultivo *i* por el precio unitario del producto cosechado ($/ha).
C_i.- Costo de producción por unidad de superficie del cultivo *i* ($/ha).
S_i.- Área cosechada del cultivo *i* (ha).

Haciendo una analogía con lo descrito en la figura 40, se puede distinguir que:

Variables de optimización: $S=(S_1, S_2, S_3, ...,, S_n)$. Superficie utilizada para la siembra de los cultivos 1, 2, 3,…, *n*.
La función objetivo: $f(S) = BN = (IB_1 - C_1)S_1 + (IB_2 - C_2)S_2 + ... + (IB_n - C_n)S_n$, con el propósito de maximizar *BN*.

Con las restricciones siguientes:

a) La superficie total disponible para siembra. La suma de las superficies de los cultivos no debe exceder la superficie total que el distrito tiene disponible para sembrar. A la superficie está asociado el volumen, ya que este depende de la cantidad de hectáreas que sean sembradas. Esta restricción se modela de la siguiente manera:

$$S_1 + S_2 + ... + S_n \leq ST \qquad \text{Para } i = 1, 2, 3, \ldots, n \qquad \textbf{(5.2)}$$

S_i se define en la expresión 5.1. *ST* es la superficie total disponible para siembra del distrito de riego.

b) La superficie disponible por cultivo. Por algunas razones (como la preparación especial del terreno, determinadas propiedades del suelo, disposición de sistemas de riego especializados, étc), las áreas disponibles para siembra de cada cultivo están delimitadas. Es decir, no se puede disponer libremente de toda la superficie para cualquier cultivo. Por esta razón, es importante señalar el rango de superficie disponible para sembrar cada cultivo. Estas restricciones se modelan de la siguiente manera:

$$LI_i \leq S_i \leq LU_i \qquad \text{Para } i = 1, 2, 3, \ldots, n \qquad \textbf{(5.3)}$$

Estas restricciones (única para cada cultivo) puede descomponerse en dos desigualdades de la forma $S_i \geq LI_i$ y $S_i \leq LU_i$. LI_i y LU_i representan las cotas de la superficie del cultivo *i*, denotando el rango entre el cual es posible variar el valor de S_i.

Como se podrá observar, se han dejado fuera de este planteamiento inicial a las variables IB_i y C_i, presentes ambas en la expresión 5.1. Para convertir la expresión 5.1 en un problema estándar de programación lineal, la operación (IB_i - C_i) debe poder expresarse en términos cuantitativos, como un coeficiente constante dentro del modelo. C_i representa el costo de producción ($/ha), y es un dato que depende de condiciones económicas de mercado fuera del alcance de este proyecto, por lo que se considera que es una cantidad

cuantitativa que debe ser ingresada al modelo. La variable IB_i representa el ingreso bruto que se puede obtener por cada unidad de producto agrícola vendida ($/ton). El ingreso bruto está en función del precio de venta del cultivo i (P_i) y del rendimiento del cultivo (R_i), de manera que:

$$IB_i = P_i \times R_i \quad \textbf{(5.4)}$$

Donde:
IB_i es el ingreso bruto obtenido por la venta del cultivo *i* ($).
P_i es el precio del cultivo *i* ($/ton).
R_i es el rendimiento del cultivo *i* (ton/ha).

El precio de venta (P_i) es, al igual que el costo, un valor cuantitativo dependiente de condiciones del mercado, por lo que se considera que su valor es un dato de entrada que debe ser proporcionado al modelo. La variable R_i, sin embargo, representa la cantidad de producto de un determinado cultivo que se puede obtener por cada unidad de superficie sembrada, y es, por lo tanto, una variable extremadamente compleja que depende de muchos factores, por lo que debe de ser considerada como una función que se encuentra dentro de la propia función de maximización del ingreso de la producción agrícola.

Ya en Palacios et al (1986) [5], se había hablado sobre la dificultad de plantear la maximización de la producción agrícola como un problema que pudiera ser resuelto con técnicas de programación lineal, esto debido a la numerosa cantidad de condiciones involucradas en el problema. La propuesta de esta tesis recarga el comportamiento impredecible del modelo de optimización en la variable rendimiento, para la cual se seleccionó ya una técnica de modelado no lineal que se obtuvo como resultado del proceso de minería de datos. Se determinó el uso de la minería de datos, ya que ésta es precisamente una disciplina que nos permite descubrir aspectos desconocidos de los datos a partir de la aplicación de algoritmos especializados. Como resultado, se obtuvieron una serie de modelos que representan una función del rendimiento en los términos mostrados en la siguiente expresión (5.5).

$$R_i = f_i(S_i, L_i, Clima) \quad \textbf{(5.5)}$$

Donde:
f_i .- Representa la función de estimación del rendimiento para el cultivo *i*.
R_i.- Rendimiento del cultivo *i* (ton/ha).
S_i.- Área cosechada del cultivo *i* (ha).
L_i.- Lámina de riego aplicada al cultivo *i* (mm).
$Clima_{m,v}$. Es la distribución mensual de variables climáticas (60 variables en total, 12 por cada tipo de variable: temperatura promedio, temperatura máxima, temperatura mínima, precipitación y evaporación). Las temperaturas se miden en °C, mientras que la precipitación y la evaporación se miden en milímetros (mm). La variable $Clima_{m,v}$ es una matriz *m* x *v* de 12 x 5 elementos, donde cada fila v_i representa un tipo de variable (temperatura, precipitación, etc.) y cada columna m_i representa un mes del año. El clima es independiente del cultivo, lo cual se representa por la ausencia del índice *i* en la variable. Es

importante señalar en esta descripción que en ocasiones se hará referencia a la información de las filas (que representan las variables) de manera independiente y únicamente por el nombre de la variable que representan, indicando por ejemplo *Temperatura* para hacer referencia a la fila de $Clima_{m,v}$ que representa la variable.

De esta manera, el modelo planteado inicialmente en la expresión 5.1 queda como se muestra en la siguiente expresión (5.6).

$$BN = (P_1 \times f_1(S_1, L_1, Clima) - C_1)S_1 + (P_2 \times f_2(S_2, L_2, Clima) - C_2)S_2 + \cdots + (P_n \times f_n(S_n, L_n, Clima) - C_n)S_n \quad \textbf{(5.6)}$$

Con las siguientes restricciones:

$S_1 + S_2 + \ldots\ldots\ldots\ldots + S_n \leq ST$

$LI_i \leq S_i \leq LU_i$ Donde $i = 1..n$, siendo n el número de cultivos presente en el problema.

La forma que adopta la función f_i depende de la técnica empleada para desarrollarla. En el capítulo 4 de esta tesis se exploraron 4 técnicas de modelado para la función de predicción del rendimiento: regresión lineal multivariable, redes neuronales perceptrón multicapa, árboles J48 y árboles de regresión M5. Una forma común de resolver la expresión 5.6 es reemplazar f_i por el rendimiento promedio del cultivo *i*, con lo cual la expresión 5.6 se convierte en lineal y puede ser resuelta por técnicas como el método Simplex (Danzing, 1947) [203]. Lamentablemente, el uso del promedio como un estimador es poco recomendable, ya que es vulnerable a desviaciones presentes en los datos. En la sección de resultados de este capítulo, se presenta una comparación de la maximización de la producción agrícola utilizando el algoritmo propuesto en esta tesis, y la maximización utilizando como estimador al rendimiento promedio.

5.2. Uso de los árboles de regresión M5 en el modelo de optimización

Durante el capítulo 4 de esta tesis se expuso la técnica de minería de datos como una herramienta para la búsqueda del modelo que arrojara los mejores resultados para predecir el rendimiento de los cultivos agrícolas. La herramienta seleccionada como resultado del proceso fue el algoritmo de árbol de regresión M5. Un ejemplo de estos modelos se presenta en la figura 41.

```
SuperficieSembrada <= 52.185 :
|  LaminaRiego <= 120.95 :
|  |  TempEne <= 9.55 : + 12.7852
|  |  TempEne >  9.55 :
|  |  |  TempOct <= 20.925 : + 12.7698
|  |  |  TempOct >  20.925 : + 12.7637
|  LaminaRiego >  120.95 : + 12.9328
SuperficieSembrada >  52.185 : + 14.0253
```

Figura 41. Forma de un modelo de árbol de regresión M5 para un cultivo específico.

En esta tesis se adoptan este tipo de modelos como la forma de las funciones $f_i(S_i, L_i, Clima)$ presentes en el modelo de optimización descrito en la expresión 5.6. En la figura 41, el atributo *SuperficieSembrada* representa a la variable S_i, el atributo LaminaRiego a la variable L_i y los atributos *TempEne* y *TempOct* hacen referencia al valor *Clima*[*Enero*, *Temperatura*] y *Clima*[*Octubre*, *Temperatura*], donde las variables *Octubre* y *Temperatura* deben ser cambiados por sus correspondientes identificadores numéricos. La figura 42 describe el árbol de la figura 41 con la notación tradicional de función. Por comodidad, se ha abreviado los índices de la matriz *Enero*, *Octubre*, y *Temperatura* de la matriz *Clima* con las letras *E*, *O* y *T*.

$$f_i(S_i, L_i, Clima) = \begin{cases} \text{Rendimiento (ton/ha)} & \text{Para:} \\ 12.7852 & -\infty \le S_i \le 52.185 \\ & -\infty \le L_i \le 120.95 \\ & -\infty \le Clima[E, T] \le 9.55 \\ 12.7698 & -\infty \le S_i \le 52.185 \\ & -\infty \le L_i \le 120.95 \\ & 9.55 < Clima[E, T] \le \infty \\ & 0 \le Clima[O, T] \le 20.925 \\ 12.7637 & -\infty \le S_i \le 52.185 \\ & -\infty \le L_i \le 120.95 \\ & 9.55 < Clima[E, T] \le \infty \\ & 20.925 > Clima[O, T] \le \infty \\ 12.9328 & -\infty \le S_i \le 52.185 \\ & 120.95 > L_i \le \infty \\ 14.0253 & 52.185 \le S_i \le \infty \end{cases}$$

Figura 42. Definición del árbol de la figura 41 con notación tradicional de función.

Nótese que los modelos de árbol M5 son particulares de cada cultivo. Para que estas funciones regresen un valor numérico, es preciso especificar los valores de las variables SuperficieSembrada (S_i), lámina de riego (L_i) y la matriz *Clima*. Las variables L_i y *Clima* son independientes del modelo de optimización, pero la superficie S_i no, por lo que la selección de un determinado rendimiento de los árboles de regresión M5 puede provocar un cambio en las restricciones de superficie del modelo de optimización.

Para ejemplificar lo anterior, supóngase que se tiene la maximización de la función objetivo representada en la expresión 5.6. Supóngase también, que la restricción de superficie del cultivo 1 es $30 \le S_1 \le 100$, lo que significa que el cultivo puede sembrarse en una extensión de 30 a 100 ha. Se sabe que la lámina de riego debido al volumen disponible para el cultivo andará entre 100 y 150 mm, y que las temperaturas promedio de

enero y octubre serán de 10 y 18 °C, respectivamente. Asumiendo que la figura 42 representa la función $f_1(S_1, L_1, Clima)$ para el cultivo 1, entonces se tiene que:

Para una superficie $30 \le S_1 \le 100$, una lámina $100 \le L_1 \le 150$ y los valores *Clima*[*Enero*,*Temperatura*] = 10, *Clima*[*Octubre*,*Temperatura*] = 18, la función f_1 está definida como:

$$f_1(S_1, L_1, Clima) = \begin{cases} 12.7698 & \text{Para } 30 \le S_1 \le 52.185 \text{ (a)} \\ 14.0253 & \text{Para } 52.185 < S_1 \le 100 \text{ (b)} \end{cases}$$

Figura 43. Consulta al modelo de árbol de regresión M5 de la figura 41.

Como se habrá notado, con la información proporcionada de la lámina y del clima se deja a la función f_1 en términos de una sola variable (la superficie). Esto quiere decir que el valor para f_1 de (a) representa un rendimiento que está definido para una superficie de entre 30 y 52.185 ha, esto "cruzando" la restricción global, la restricción de la lámina del riego y la información del clima con la información provista por el árbol de regresión. El rendimiento en (b), es producto únicamente de la superficie, ya que el modelo de árbol indica que no influye otro atributo cuando la superficie excede las 52.185 ha (esto como producto de la información de la base de datos a la que fue sometido el algoritmo M5). El seleccionar el rendimiento(a) ó (b) para colocar en la expresión 5.6 atrae la consecuencia de tener que ajustar la restricción de la superficie, según los límites proporcionados por el árbol de regresión. En los casos donde el rango indicado por la restricción de superficie original se ubique totalmente por encima ó por debajo del valor de prueba del atributo superficie de un árbol de regresión no será necesario ningún ajuste.

El ejemplo anterior también muestra otra problemática del uso de modelos de árbol M5 con modelos de optimización: ¿qué valor de salida seleccionar?. Es decir, como se muestra en la figura 43, cada f_i en la expresión 5.6 puede regresar más de un valor para el rendimiento (junto con sus cambios en las restricciones $LI_i \le S_i \le LU_i$), así que se debe seleccionar un valor único para el modelo de optimización. En el ejemplo, parece obvia la selección del rendimiento (b), ya que indica un mayor rendimiento en una extensión mayor de superficie. Sin embargo, existe la posibilidad de que la extensión de superficie invertida en este rendimiento sea más útil con otro cultivo, o bien, que la extensión de superficie sea demasiado pequeña o demasiado grande como para empatarla en el modelo global, al considerar otros cultivos. De cualquier manera, cuando exista más de un rendimiento en la salida de una consulta a los modelos de árbol M5 (las funciones $f_i(S_i, L_i, Clima)$ indicadas en la expresión 5.6), será necesario hacer una búsqueda por la mejor combinación de rendimientos y superficies que maximicen el ingreso total obtenido de la ecuación 5.6.

La búsqueda de la combinación ideal de rendimientos no es una tarea trivial que pudiera ser resuelta de forma determinista. Cada cultivo agregado al modelo en la expresión 5.6 multiplica el espacio de búsqueda en función de los rendimientos probables del cultivo que satisfacen las restricciones impuestas por el modelo de optimización. Así, si dos cultivos

tienen tres rendimientos probables cada uno, entonces el espacio de búsqueda es de tamaño 3^2. El espacio de búsqueda crece de manera exponencial a medida que se incrementa el número de cultivos. Tomando como ejemplo a los cultivos seleccionados en la fase de minería de datos y utilizando el número de reglas generadas de los árboles M5 como el número máximo de rendimientos a obtener (en el peor de los casos), el número de combinaciones posible de rendimientos para los 15 cultivos de prueba es de:

$$26 \times 11 \times 62 \times 11 \times 5 \times 56 \times 10 \times 38 \times 39 \times 66 \times 43 \times 22 \times 31 \times 65 \times 5 = 5.09139E+20 \quad \textbf{(5.7)}$$

Como se ve, realizar una búsqueda exhaustiva a través de las múltiples combinaciones de rendimientos se visualiza como una tarea desgastante y consumidora de grandes recursos de cómputo y tiempo. Esto lleva a pensar en proponer una heurística de búsqueda que permita localizar la combinación óptima de rendimientos en un tiempo accesible y con un alto grado de confiabilidad. Es dentro de este contexto donde se propone el uso de un algoritmo genético como el mecanismo de búsqueda para la combinación óptima de rendimientos.

5.3. Algoritmos genéticos

Un algoritmo genético [204] (o AG para abreviar) es una técnica de programación inspirada en la evolución biológica como estrategia para resolver problemas. Dado un problema específico a resolver, la entrada del AG es un conjunto de soluciones potenciales a ese problema, codificadas de alguna manera, y una métrica llamada función de aptitud que permite evaluar cuantitativamente a cada candidata. Estas candidatas pueden ser soluciones que ya se sabe que funcionan, con el objetivo de que el AG las mejore, pero se suelen generar aleatoriamente [205].

Algunas de las diferencias que presentan los AG frente a otros métodos de optimización y búsqueda son [206][207]:

- Trabajan con una codificación del conjunto de parámetros y no los parámetros en sí.
- Buscan para un conjunto de puntos, no para uno solo (realizan una búsqueda poblacional).
- Utilizan información de una función la cual es la que optimizan, en lugar de usar derivadas u otro conocimiento adicional.
- Usan reglas de transición probabilísticas y no determinísticas.

5.3.1. Consideraciones y funcionamiento de un algoritmo genético

Para describir el funcionamiento de un algoritmo genético es preciso definir antes los elementos involucrados en el método de solución propuesto por este algoritmo:

Cromosoma. Los algoritmos genéticos utilizan una estructura de datos conocida como cromosoma [207]. Un cromosoma agrupa una serie de parámetros denominados genes, que en conjunto integran una posible solución al problema tratado por el AG. Existen diferentes representaciones para los genes, desde la codificación binaria (0's y 1's), las

cadenas alfanuméricas, u otras representaciones numéricas (como enteros y números reales).

Población. Una población es un subconjunto de todos los cromosomas, tomados como posibles soluciones individuales del problema a resolver. El tamaño de la población es un punto importante que se debe de considerar, ya que una población pequeña implica una convergencia muy rápida a un óptimo local, por otro lado, una población muy grande tiene como consecuencia el consumo de grandes recursos computacionales, por lo que se recomienda un balance en el tamaño de la población [207].

Función de aptitud. La función de adaptación (o aptitud) debe ser diseñada para cada problema de manera específica. Dado un cromosoma particular, la función de adaptación le asigna un número real, que se supone refleja el nivel de adaptación al problema del individuo representado por el cromosoma [208].

Operadores genéticos. Los operadores genéticos permiten generar nuevas soluciones (cromosomas) a partir de las ya existentes, ya sea por medio de intercambio o alteración de genes. Habitualmente, se definen dos operadores genéticos básicos: mutación y cruce [208].

- *Operador de cruce*. El operador de cruce toma dos padres seleccionados y corta sus ristras de cromosomas en una posición escogida al azar, para producir dos subristras iniciales y dos subristras finales. Después se intercambian las subristras finales, produciéndose dos nuevos cromosomas completos (véase la figura 44). Ambos descendientes heredan genes de cada uno de los padres. Este operador se conoce como operador de cruce basado en un punto.

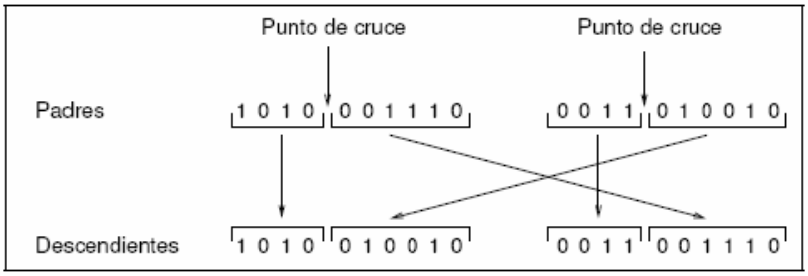


Figura 44. Operador de cruce basado en un punto [208]

- *Operador de mutación*. El operador de mutación se aplica a cada hijo de manera individual. Consiste en la alteración aleatoria (normalmente con probabilidad pequeña) de cada gen componente del cromosoma. La figura 45 muestra la mutación del quinto gen del cromosoma. Si bien puede en principio puede pensarse que el operador de cruce es más importante que el operador de mutación, ya que proporciona una exploración rápida del espacio de búsqueda, éste último se asegura de que ningún punto del espacio de búsqueda tenga probabilidad cero de ser examinado, y es de capital importancia para asegurar la convergencia de los AG [208].

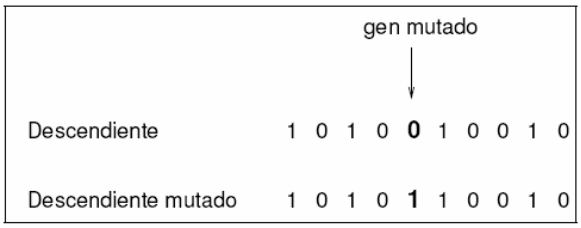


Figura 45. Operador de mutación [208]

Una vez descritos estos elementos, se procede a la descripción del funcionamiento de un algoritmo genético. Para ello, la figura 46 muestra la estructura algorítmica de un algoritmo genético simple.

```
BEGIN /* Algoritmo Genetico Simple */
    Generar una poblacion inicial.
    Computar la funcion de evaluacion de cada individuo.
    WHILE NOT Terminado DO
    BEGIN /* Producir nueva generacion */
        FOR Tamaño poblacion/2 DO
        BEGIN /*Ciclo Reproductivo */
            Seleccionar dos individuos de la anterior generacion,
            para el cruce (probabilidad de seleccion proporcional
            a la funcion de evaluacion del individuo).
            Cruzar con cierta probabilidad los dos
            individuos obteniendo dos descendientes.
            Mutar los dos descendientes con cierta probabilidad.
            Computar la funcion de evaluacion de los dos
            descendientes mutados.
            Insertar los dos descendientes mutados en la nueva generacion.
        END
        IF  la poblacion ha convergido THEN
            Terminado := TRUE
    END
END
```

Figura 46. Algoritmo genético simple (por Larrañaga, Inza y Moujahid, 2005 [208])

Se pueden distinguir varias operaciones importantes en el algoritmo de la figura 46. Como puede verse, el algoritmo inicia con la generación de una población inicial de individuos o cromosomas, que finalmente representan posibles soluciones al problema (pero codificadas). La función de evaluación mide la aptitud de cada solución, con el fin de proporcionar un parámetro que permita hacer una selección en la población con el fin de determinar individuos candidatos a los operadores de cruce y mutación. Estos operadores producen descendientes, los cuales son insertados en una nueva generación. El proceso se repite hasta que no se detectan cambios en las generaciones (convergencia), o bien, un determinado número de generaciones ha transcurrido.

Como se verá en la siguiente sección, el algoritmo propuesto por esta tesis para la maximización del ingreso económico de la producción agrícola, es, en sí, un algoritmo genético básico con un planteamiento particular de la función de aptitud. Este

planteamiento reduce el espacio de búsqueda original, acotando el tiempo de convergencia hacia una solución óptima, la cual está dada por la combinación correcta de rendimientos y superficies generadas a partir de la información disponible en árboles de regresión M5.

5.4. El algoritmo GenSM5 (Genético+Simplex+M5)

5.4.1. Consideraciones iniciales

El algoritmo propuesto ha sido llamado GenSM5 (GENético+Simplex+M5). La razón del nombre, es que utiliza un algoritmo genético como mecanismo de búsqueda de soluciones óptimas, las cuales son probadas por una función de aptitud basada en el algoritmo Simplex para maximización, el cual es alimentado a su vez por la información proveniente de árboles de regresión M5.

Las consideraciones más importantes a tomar en cuenta para cualquier algoritmo son los datos de entrada y los datos de salida. Como datos de entrada, el algoritmo GenSM5 necesita:

a) Los modelos de árbol de regresión de cada cultivo involucrado en el proceso de producción agrícola.
b) Las restricciones de superficie y lámina de riego por cultivo.
c) La restricción de superficie total disponible para siembra en el distrito.
d) La información de precios y costos de producción de cada uno de los cultivos.
e) La información del clima en el distrito de riego

Como salidas, el algoritmo GenSM5 proporciona:

a) La cantidad exacta de superficie a sembrar por cultivo.
b) El rendimiento esperado por cultivo.
c) El ingreso neto óptimo a obtener por la venta de la producción agrícola.
d) Las láminas de riego necesarias por cultivo. Esto se traduce al volumen de agua total requerido para la siembra de los cultivos.

En general, el tamaño del problema a resolver será de n variables, donde n depende del número de cultivos seleccionados. Como cada cultivo aporta dos restricciones de superficie (el límite inferior y el límite superior), entonces el número de restricciones es de nx2=2n. A este número se debe añadir una restricción más, la cual representa la restricción de la superficie total disponible, por lo que el número de restricciones queda en $2n$+1, donde n es el número de cultivos introducidos al algoritmo.

Dado que GenSM5 utiliza un algoritmo genético como mecanismo de búsqueda, es necesario definir también a los elementos que tradicionalmente componen un algoritmo genético en términos del problema y del modelo de solución propuesto. Estos elementos son: el cromosoma, la población, la función de aptitud y los operadores genéticos.

Cromosoma. Para la propuesta de solución, los cromosomas se codifican como una serie de valores enteros. Cada valor entero representa el índice en una lista de rendimientos (con superficie y lámina de riego asociadas) particular de cada cultivo. Los rendimientos y

superficies son extraídos de los árboles de regresión M5 y ajustados con las restricciones de superficie indicadas al modelo de optimización (tal como se describe en el punto 5.2 de este capítulo). Por lo tanto, el tamaño de un cromosoma depende del número de cultivos presentes en el problema, y representa una combinación única de los rendimientos de todos los cultivos. Para clarificar lo anterior, obsérvese la figura 47, que muestra un cromosoma de tamaño n (para un problema de n cultivos). Los índices I_1, $I_{1,\ldots,}I_n$ representan el índice en las listas de rendimientos LR_1, LR_2,…,LR_n. Cada lista almacena la información de los rendimientos y sus superficies asociadas. Al acceder a un gen del cromosoma, se accede a la información del rendimiento localizado en la posición indicada por el gen en la lista de rendimientos correspondiente.

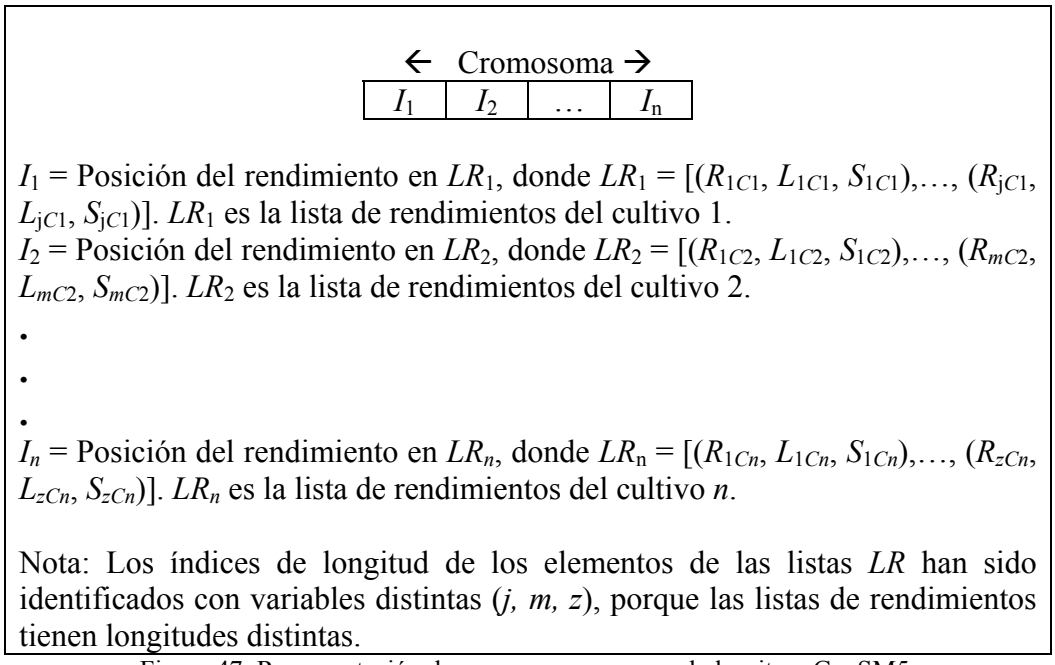


Figura 47. Representación de un cromosoma para el algoritmo GenSM5.

Población. Una población es un conjunto de cromosomas (combinaciones de rendimientos de cultivos). Cabe mencionar, sin embargo, que el universo de cromosomas disponibles para la generación de la población no son todas las combinaciones indicadas en (5.7), si no, únicamente aquellas que resulten de la integración de las restricciones iniciales de la superficie disponible para los cultivos, la lámina de riego y las condiciones climáticas con la información de los árboles de regresión M5.

Función de aptitud. La función de aptitud es la maximización del modelo indicado en (5.6), una vez que se ha hecho el reemplazo de las funciones $f(S_i, L_i, Clima)$ por los rendimientos indicados en el cromosoma a evaluar. Para tomar en cuenta la restricción de la superficie total, la maximización del ingreso es penalizada si se excede la superficie total. También, se consideran los ajustes en las restricciones de superficie generadas por la selección del rendimiento indicado en el cromosoma. Una vez realizados estos cambios, las variables precio y costo de producción son reemplazadas también por los valores

correspondientes (que deben ser proporcionados al modelo), dejando el modelo indicado en 5.6 como sigue:

$$BN = Coef_1 S_1 + Coef_2 S_2 + \cdots + Coef_n S_n \qquad \textbf{(5.8)}$$

Donde:
$Coef_i$ representa el resultado de la operación $(R_i \times P_i - C_i)$. Este coeficiente es el ingreso neto que se obtiene por cada unidad de superficie sembrada del cultivo *i*. Se expresa en pesos por hectárea ($/ha).

Con las siguientes restricciones:

$S_1 + S_2 + \ldots\ldots\ldots\ldots + S_n \leq ST$

$S_i L' \leq S_i \leq S_i U'$ Para $i = 1..n$, donde n es el número total de cultivos.

Los límites en las restricciones de las superficies de cada cultivo (S_i) han sido cambiados por $S_i L'$ *y* $S_i U'$ para indicar que dichos límites han cambiado como consecuencia de la información proporcionada por los modelos M5 (ver punto 5.2 en este capítulo).
La maximización del modelo presentado en (5.8) es trivial por Simplex una vez que la maximización del ingreso se reduce a términos de una sola variable: la superficie.
El valor regresado por la maximización es tomado como el indicador que mide el desempeño del cromosoma en la función de evaluación. Puede suceder, sin embargo, que una maximización produzca un gran ingreso, pero viole la restricción de superficie total. Esto debido a que la maximización se realiza con la información del cromosoma en turno, y los rendimientos indicados en el cromosoma pueden estar ligados a restricciones de superficies que en el mejor de los casos exceden a la superficie total. En tal caso, la función de aptitud penaliza el ingreso calculado por la maximización, en función de la desviación que se produzca entre la superficie calculada y la superficie total disponible. A un exceso mayor de la superficie calculada sobre la disponible corresponde un mayor castigo al ingreso, con el fin de advertir al algoritmo genético que el cromosoma no es viable.

Otra consideración, es que el algoritmo GenSM5 funciona de manera correcta únicamente cuando la superficie total disponible es mayor la cantidad mínima disponible indicada por las restricciones (al sumar el límite inferior del rango restrictivo de superficie por cultivo). Esto se debe a que el ciclo genético nunca encontrará la combinación que maximice el ingreso, ya que las restricciones indicadas por los rendimientos de los cromosomas están acotadas a las restricciones del problema.

Una vez proporcionados los elementos base para el planteamiento del problema de optimización del ingreso como un algoritmo genético, se procede a la descripción formal del algoritmo.

5.4.2. Descripción del algoritmo

La figura 48 describe a grandes rasgos el funcionamiento del algoritmo GenSM5. El algoritmo inicia con la generación u obtención de los modelos de árbol de regresión M5. Para esta tesis, este paso fue cubierto por la actividad de minería de datos (capítulo 4), cuyo uno de los resultados fue la generación de modelos de árbol de regresión M5 para cada uno de los cultivos seleccionados.

El siguiente paso consiste en la lectura de restricciones de superficie y láminas de riego, junto con los valores de precios y costos de producción asociados a cada cultivo presente en el problema. Los precios y costos de producción son valores numéricos reales, mientas que las restricciones de superficie y de lámina de riego son rangos de valores especificados por un límite inferior y un límite superior.

Otras restricciones que deben ser proporcionadas por el usuario son la superficie total disponible para siembra en el distrito y la información del clima. La superficie es un valor real que indica la cantidad total de superficie que puede ser dispuesta para la siembra de los cultivos. El clima es la información de 60 variables climáticas que representan el pronóstico de temperatura promedio, temperatura máxima, temperatura mínima, precipitación y evaporación para cada uno de los meses presentes en el año agrícola. De hecho, no se necesitan los valores de las 60 variables climáticas, si no únicamente, los valores de aquellas variables que intervienen en los modelos de árbol introducidos al algoritmo.

Los pasos anteriores abarcan lo que es la introducción de datos al algoritmo de optimización. Los siguientes pasos van encaminados al procesamiento de los datos de entrada en el marco de búsqueda de una solución óptima al problema planteado.

La siguiente etapa “cruza” las restricciones de superficie y lámina de riego por cultivo que fueron introducidas al algoritmo con la información de los árboles M5 tal como se indicó en la sección 5.2 de este capítulo. Este paso “poda” los árboles de regresión M5 para dejar únicamente los rendimientos (ó reglas) que satisfacen las restricciones de entrada de los cultivos. Los rendimientos, superficies y láminas de riego que satisfacen los requerimientos son concentrados de manera ordenada (con respecto al rendimiento) en una lista independiente por cultivo.

GenSM5 verifica entonces si el uso de los rendimientos que se encuentran al inicio de las listas (los rendimientos más altos) produce una salida satisfactoria. Para ello, se maximiza la función objetivo (indicada en (5.6)) con el método Simplex y con la información de los rendimientos presentes en el índice 0 de las listas. A este nivel, una salida satisfactoria es aquella que utiliza casi toda la superficie disponible (especificado por un parámetro, como por ejemplo, el 90%). Dado que de antemano se sabe que se están utilizando los mejores rendimientos, entonces puede asumirse que se ha encontrado una muy buena solución al problema y el algoritmo terminará antes de iniciar el ciclo genético. Esta primera maximización producirá los datos de salida esperados (descritos en la sección 5.4.1 de este capítulo).

Si las superficies por cultivo resultantes de la maximización no suman un mínimo de superficie requerida, entonces el algoritmo inicia el ciclo genético. Básicamente, el ciclo genético en GenSM5 es el mismo que se mostró en la sección 5.3.1 (algoritmo genético básico), pero con las consideraciones indicadas en la sección 5.4.1.

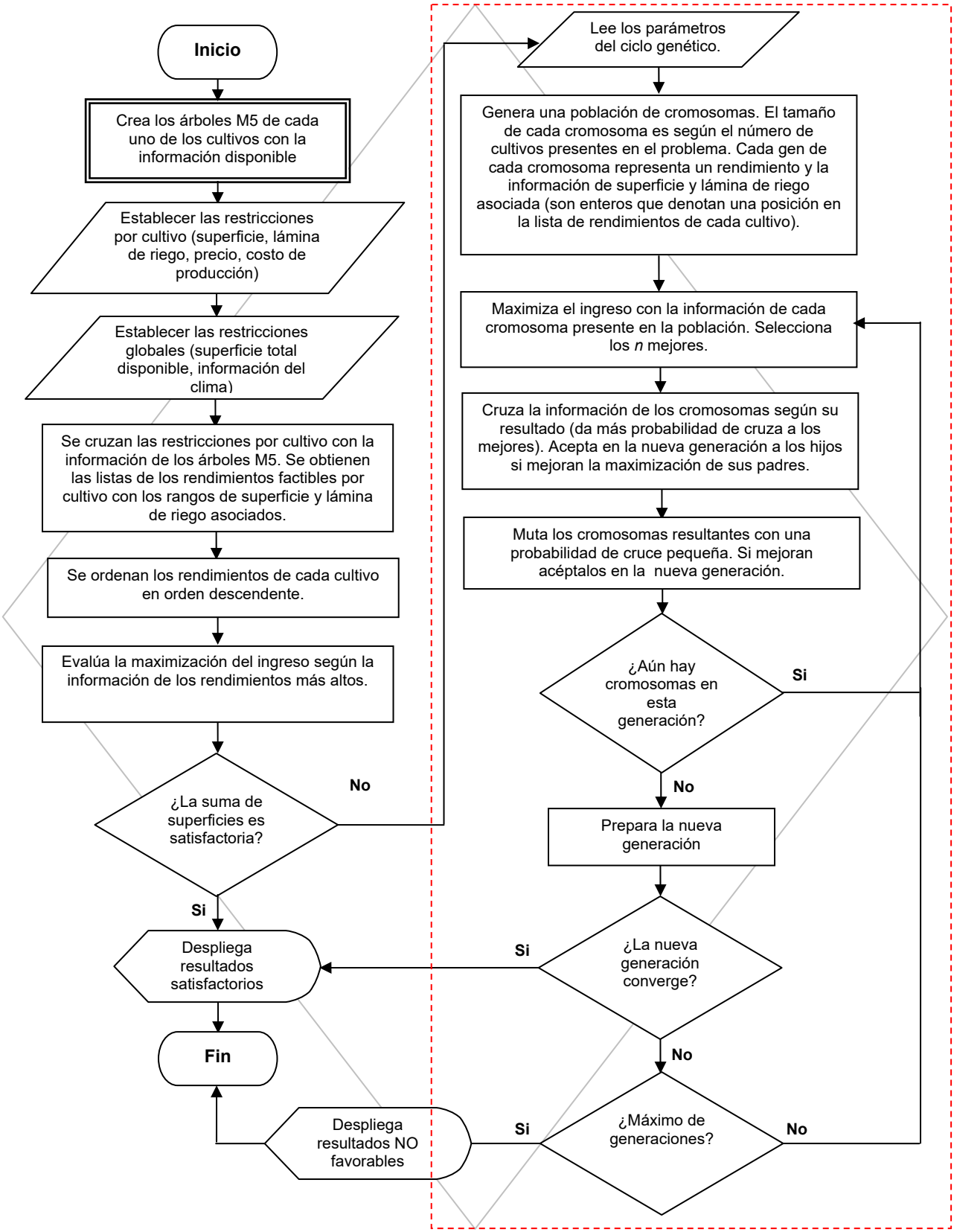


Figura 48. Funcionamiento del algoritmo GenSM5.

Para su ejecución, la parte genética del algoritmo GenSM5 requiere de los valores de los parámetros *tamaño de la población*, *probabilidad de cruce*, *número máximo de generaciones y probabilidad de mutación*. La búsqueda del algoritmo GenSM5 es guiada por el resultado de la evaluación de la maximización del ingreso indicado por el cromosoma en turno. Al final del ciclo genético, el algoritmo arroja la mejor combinación de rendimientos encontrada (aquella que produce un ingreso más alto que cualquier otra combinación probada). Las salidas del algoritmo GenSM5 son: la cantidad exacta de superficie a sembrar por cultivo, el rendimiento esperado por cultivo, el ingreso neto óptimo a obtener por la venta de la producción agrícola y las láminas de riego necesarias por cultivo.

El algoritmo GenSM5 fue implementado en una herramienta de software y aplicado a la base de datos que se preparó durante la etapa de minería de datos de esta tesis. Un ejemplo de la ejecución del algoritmo se muestra en la siguiente sección. Los resultados de las pruebas realizadas al algoritmo se presentan en la siguiente sección. Los detalles del desarrollo de la aplicación se proporcionan en el capítulo 6 de esta tesis.

5.5. Ejemplo de funcionamiento

Para ejemplificar el funcionamiento del algoritmo GenSM5 se presenta en esta sección un ejercicio realizado con los cultivos algodón y frutales sembrados ambos en el distrito de riego 038. Se utilizarán las etapas mostradas en la figura 48 para describir el desarrollo del ejercicio.

Paso 1.- Creación de los árboles M5 de cada uno de los cultivos incluidos en la optimización.

Los árboles M5 de los cultivos algodón y frutales son presentados en la figura 49. Se recuerda que la metodología que llevo a la obtención de estos modelos se describe en el capítulo 4 de esta tesis.

```
LaminaRiego <= 91.2 :
|  SuperficieSembrada <= 262.5 :
|  |  SuperficieSembrada <= 103.93 :
|  |  |  LaminaRiego <= 83.709 :
|  |  |  |  SuperficieSembrada <= 58 :
|  |  |  |  |  LaminaRiego <= 67.375 : 3.0348
|  |  |  |  |  LaminaRiego >  67.375 : 3.0344
|  |  |  |  SuperficieSembrada >  58 : 3.035
|  |  |  LaminaRiego >  83.709 : 3.0389
|  |  SuperficieSembrada >  103.93 : 3.0532
|  SuperficieSembrada >  262.5 :
|  |  SuperficieSembrada <= 379.86 : 2.9566
|  |  SuperficieSembrada >  379.86 : 2.9411
LaminaRiego >  91.2 :
|  SuperficieSembrada <= 463.99 :
|  |  TempOct <= 20.425 : 2.6609
|  |  TempOct >  20.425 :
|  |  |  TempEne <= 9.95 : 2.6626
|  |  |  TempEne >  9.95 : 2.6636
```

```
LaminaRiego <= 103.4 :
|  TempMar <= 11.795 :
|  |  SuperficieSembrada <= 32.5 :
|  |  |  SuperficieSembrada <= 21 :
|  |  |  |  SuperficieSembrada <= 7.645 : 18.7741
|  |  |  |  SuperficieSembrada >  7.645 : 18.7849
|  |  |  SuperficieSembrada >  21 : 18.6877
|  |  SuperficieSembrada >  32.5 : 18.7861
|  TempMar >  11.795 :
|  |  SuperficieSembrada <= 9.836 : 19.8463
|  |  SuperficieSembrada >  9.836 :
|  |  |  SuperficieSembrada <= 29 : 20.0024
|  |  |  SuperficieSembrada >  29 :
|  |  |  |  TempNov <= 15.35 : 19.9527
|  |  |  |  TempNov >  15.35 : 19.9674
LaminaRiego >  103.4 :
|  LaminaRiego <= 128.29 : 21.1405
|  LaminaRiego >  128.29 : 21.1691
```

| SuperficieSembrada > 463.99 : 2.6898	

Figura 49. Árboles de regresión M5 para los cultivos algodón (izquierda) y frutales (derecha).

Paso 2.- Establecer las restricciones por cultivo.

En este paso se establecen las restricciones que indican los rangos de superficie y lámina de riego que pueden ser utilizadas. Las restricciones impuestas para el ejercicio se muestran en la tabla 53. Como se puede observar, la superficie del cultivo algodón está acotada entre 227.06 ha y 340.60 ha. Esto quiere decir que durante la optimización, la superficie calculada no podrá ser menor a 227.06 ha y no podrá exceder las 340.60 ha. El rango de la lámina de riego del mismo cultivo está acotado entre 68.11 mm y 102.17 mm. El dato promedio de ambas variables se proporciona para fines de referencia.

	Superficie (ha)			Lámina de riego (mm)			Costo de producción (M$/ha)	Precio de venta (M$/ton)
Cultivo	Límite inferior	Límite superior	Promedio	Límite inferior	Límite superior	Promedio		
Algodón	227.06	340.60	283.83	68.11	102.17	85.14	10.00	4.08
Frutales	17.26	25.89	21.57	66.46	99.69	83.07	10.00	1.21

Tabla 53. Restricciones de cultivos para prueba del algoritmo de optimización.

La información del costo de producción y precio de venta también debe ser proporcionada al algoritmo, ya que es utilizada para calcular el beneficio económico a obtener por cada unidad de superficie utilizada en la optimización. Estos datos se muestran en las últimas dos columnas de la tabla 53.

Paso 3. Establecer las restricciones globales.

Además de las restricciones de superficie y lámina de riego por cultivo, debe establecerse la cantidad total de superficie con la que cuenta el distrito para la siembra de los cultivos y las condiciones estimadas del clima para el año agrícola (se asume que la planeación se está haciendo antes del inicio del año agrícola). Para el ejercicio, se asumió una superficie total disponible para siembra de 350 ha. Este límite fue estimado tomando como base la suma de los límites inferiores y superiores de las restricciones de superficie, donde se puede ver que como mínimo, en el distrito se pueden sembrar 244.06 ha (contando ambos cultivos) y como máximo, 366.48 ha. Se seleccionó 350 ha como una cantidad que se localiza entre ambas sumas. Para el clima, se observa por la figura 49 que las únicas variables climáticas que intervienen son las temperaturas promedio de los meses de octubre, noviembre, enero y marzo (tomando en cuenta ambos cultivos). Resumiendo, las restricciones globales fueron las siguientes:

- Superficie total disponible para siembra: 350 ha.
- Temperatura promedio estimada para octubre: 20.95°.
- Temperatura promedio estimada para noviembre: 14.90°.
- Temperatura promedio estimada para enero: 9.98°.
- Temperatura promedio estimada para marzo: 11.78°.

Paso 4. Cruza de las restricciones por cultivo con la información de los árboles M5

De la cruza de las restricciones impuestas a nivel de cultivo y la información de los árboles M5 surgen los rendimientos asociados para cada uno de los cultivos, junto con posibles ajustes en las restricciones de superficie. A continuación se describe la forma de obtener los rendimientos y las restricciones ajustadas para cada uno de los cultivos presentes en el ejercicio.

Algodón

Revisando el árbol de regresión de este cultivo (izquierda, figura 49), se observa que el primer atributo a prueba es la lámina de riego (*LaminaRiego <= 91.2*). La restricción para la lámina del algodón expresada en la tabla 53 indica que la lámina de este cultivo debe estar entre 68.11 y 102.17 mm. Tomando esto en cuenta, se deduce que el rango de la lámina de riego debe ser ajustado a 68.11 - 91.2 para los rendimientos que se localicen bajo la condición *LaminaRiego <= 91.2*. La linea 2 contempla la prueba *SuperficieSembrada <= 262.5*. Dado que la restricción de superficie para el algodón indica que la superficie aceptada debe estar entre las 227.06 y las 340.60 ha, los casos contemplados bajo la condición de *SuperficieSembrada <= 262.5* deben ajustar el rango de superficie a 227.06 – 262.50 ha, considerando el límite inferior impuesto por la restricción y el límite superior indicado por la condición. La condición de la línea 3 (*SuperficieSembrada <= 103.93*) está fuera del rango "filtrado" por las líneas anteriores (227.06 – 262.50 ha), por lo que todos los casos bajo la condición de la línea 3 quedan descartados. Así, la siguiente línea que influye en la selección de rendimientos es la 10, donde está la conclusión *SuperficieSembrada > 103.93 : 3.053*, que establece que se obtiene un rendimiento de 3.053 ton/ha en caso de que la superficie exceda los 103.93 ha. Ya que el rango de superficie obtenido hasta este punto es de 227.06 - 262.50 ha, y éste se localiza sobre las 103.93 ha indicadas por la condición de la línea 10, el rendimiento es aceptado. Como se habrá notado, la secuencia seguida hasta el punto de aceptar el rendimiento modificó los rangos de superficie sembrada y lámina de riego. La tabla 54 muestra las restricciones ajustadas para la obtención del rendimiento indicado.

Cultivo	Superficie (ha)		Lámina de riego (mm)		Rendimiento (ton/ha)
	Límite inferior	Límite superior	Límite inferior	Límite superior	
Algodón	227.06	262.50	68.11	91.20	3.05

Tabla 54. Restricciones de cultivos para prueba del algoritmo de optimización.

Siguiendo un proceso similar al indicado en el párrafo anterior, se obtienen todos los rendimientos del árbol M5 del algodón (izquierda, figura 49) que son accesibles dadas las restricciones indicadas en la tabla 53. Estos rendimientos se muestran en la tabla 55.

Rendimiento ID	Cultivo	Superficie (ha)		Lámina de riego (mm)		Rendimiento (ton/ha)
		Límite inferior	Límite superior	Límite inferior	Límite superior	
0	Algodón	227.06	262.50	68.11	91.20	3.05
1	Algodón	262.50	340.60	68.11	91.20	2.96

2	Algodón	227.06	340.60	91.20	102.17	2.66

Tabla 55. Rangos de superficie y rendimientos asociados para el cultivo algodón dadas las restricciones del ejercicio.

Se observa que tres diferentes rendimientos con tres tipos de restricciones diferentes pueden ser alcahazados dadas las restricciones iniciales y la información del árbol de regresión M5 del cultivo algodón.

Frutales

El proceso seguido para la obtención de los rendimientos de este cultivo (que realmente, es un grupo de cultivos con un comportamiento productivo similar) es el mismo que se describió para el cultivo algodón. Al final del análisis, la siguiente lista de rendimientos y restricciones asociadas son obtenidas para el cultivo frutales:

Rendimiento ID	Cultivo	Superficie (ha)		Lámina de riego (mm)		Rendimiento (ton/ha)
		Límite inferior	Límite superior	Límite inferior	Límite superior	
0	Frutales	17.26	21.00	66.46	99.69	18.78
1	Frutales	21.00	25.89	66.46	99.69	18.69

Tabla 56. Rangos de superficie y rendimientos asociados para el cultivo frutales dadas las restricciones del ejercicio.

Como se observa, dos diferentes rendimientos pueden ser obtenidos con las restricciones iniciales impuestas sobre el árbol de regresión M5 del cultivo frutales.

Al final, se observa que existen 3 tipos de rendimiento probables para el algodón y dos para el cultivo frutales. El objetivo del algoritmo GenSM5 será seleccionar la combinación de los dos rendimientos que permitan obtener el mayor ingreso posible, respetando completamente las restricciones ligadas a cada rendimiento. En total, se tienen 3 x 2 = 6 combinaciones distintas a probar en este ejercicio.

Paso 6. Se ordenan los rendimientos en orden ascendente

Los rendimientos obtenidos para los cultivos se pueden ver ordenados de mayor a menor en las tablas 55 y 56.

Paso 7. Evalúa la maximización del ingreso según la información de los rendimientos más altos.

Los rendimientos más altos y sus correspondientes restricciones son aquellos que se localizan en el índice 0 de las tablas 55 y 56. El planteamiento de la ecuación 5.6 establece la maximización de la producción de la siguiente manera (5.9):

$$MAX \quad BN = (P_1 \times f_1(S_1, L_1, Clima) - C_1)S_1 + (P_2 \times f_2(S_2, L_2, Clima) - C_2)S_2 + \cdots + (P_n \times f_n(S_n, L_n, Clima) - C_n)S_n \qquad \textbf{(5.9)}$$

Con las siguientes restricciones:

$S_1 + S_2 + ... + S_n \leq ST$

$S_i L \leq S_i \leq S_i U$ Donde $i = 1..n$, siendo n el número de cultivos presente en el problema.

La información de los costos (C_i) y los precios (P_i) se proporcionan en la tabla 53. Las funciones $f_i(S_i, L_i, Clima)$ serán reemplazadas con los rendimientos localizados en los índices 0 de las tablas 55 y 56, de donde serán tomadas también las restricciones del tipo $S_i L \leq S_i \leq S_i U$. El valor de *ST* corresponde a la superficie total disponible, indicada para el problema en 350 ha (paso número 3 de este ejemplo). Realizando las substituciones correspondientes, se tiene que el beneficio neto se maximiza por (5.10):

$$\text{MAX } BN = (4.08 \times 3.05 - 10.0)S_1 + (1.21 \times 18.78 - 10.0)S_2 \quad \textbf{(5.10)}$$
$$\text{MAX } BN = 2.444S_1 + 12.724S_2$$

Sujeto a:

$S_1 + S_2 \leq 350$

$227.06 \leq S_1 \leq 262.50$

$17.26 \leq S_2 \leq 21.0$

La restricción $227.06 \leq S_1 \leq 262.50$ es convertida en dos restricciones: $S_1 \geq 227.06$ y $S_1 \leq 262.50$. De la misma manera, la restricción $17.26 \leq S_2 \leq 21.0$ es convertida a $S_2 \geq 17.26$ y $S_2 \leq 21.0$.

La demostración del método Simplex para este problema está fuera del alcance de esta tesis. El problema descrito es sencillo y puede ser resuelto fácilmente por cualquier programa de investigación de operaciones disponible. En el caso de esta tesis, el algoritmo Simplex fue programado en la aplicación que implementa el algoritmo GenSM5, y es la misma librería la utilizada para resolver el problema planteado en la expresión 5.10.

La solución generada por el método Simplex para el problema de la expresión 5.10 es: S_1 = 262.5 (ha), S_2 = 21.0 (ha), con un valor de BN = 908.74 (miles de pesos).

La superficie total empleada dada la configuración de los rendimientos más altos arroja un total de 283.5 ha utilizadas de las 350 disponibles, dejando sin utilizar 66.50 ha, que representa un 19 % de la superficie total disponible. Como una medida preventiva a la ejecución de la búsqueda genética, GenSM5 contempla que el porcentaje de superficie alcanzado con los mejores rendimientos sea comparado con un porcentaje de alcance definido por el usuario como aceptable para el problema. Éste puede ser, por ejemplo, el 95 % de la superficie, o bien -si es aceptable para el usuario- un 90%. De cualquier modo, esto debe ser impuesto bajo el criterio de un productor. En el caso del ejercicio de demostración, se utilizó una suma de superficie deseable del 90 %. Dado que $S_1 + S_2$ alcanza sólo un 81 %

de la superficie total disponible, GenSM5 procedió con la búsqueda genética (paso número 8).

Paso 8. Búsqueda genética.

La búsqueda genética inicia con la especificación de los parámetros. En el caso del ejercicio, se utilizaron los siguientes parámetros:

- Tamaño del cromosoma (igual al número de cultivos): 2.
- Tamaño de la población: 4.
- Probabilidad de cruce: 0.7.
- Máximo de generaciones: 50.
- Probabilidad de mutación: 0.003.

Como se expuso en la sección 5.4.1 (consideraciones iniciales para GenSM5), un cromosoma representa una combinación de rendimientos de cultivos (en el caso de este ejercicio, tomada de las tablas 55 y 56). Así, el cromosoma [0,1] indica que se tomará el rendimiento identificado por el número 0 de la tabla del cultivo algodón, y el rendimiento identificado por el número 1 del cultivo frutales. El cromosoma [2,0] indican que se utilizarán el rendimiento 2 del cultivo algodón y el rendimiento 0 del cultivo frutales. Así, por ejemplo, una población inicial de individuos podría incluir a los cromosomas [0,0],[0,1],[1,1],[2,1], sometiendo esta población a los operadores de selección, cruzamiento y mutación hasta obtener al mejor individuo.

La prueba para determinar "al mejor" individuo corresponde a la maximización del ingreso de acuerdo a la información de los rendimientos indicados por medio del valor de los genes del cromosoma en turno. Esto se expone en detalle en la explicación de la función de aptitud descrita en la sección 5.4.1. Un ejemplo de la maximización para los mejores rendimientos (cromosoma [0,0]) se realizó en el paso 7 de este ejercicio.

Para el ejercicio, la búsqueda genética convergió rápidamente hacia el cromosoma [1,1], que representa la maximización con los rendimientos y restricciones de superficies localizadas en los índices 1 de las tablas 55 y 56. Esta maximización produce un beneficio neto de 996.05 miles de pesos, con una superficie utilizada de 350 ha (S_1 = 324.11 y S_2 = 25.89). Como se puede observar, aunque los rendimientos empleados no sean los más altos, se alcanza un beneficio superior al esperado por la utilización de aquellos, que en el paso 7 se determinó proporcionaban un ingreso de 908.754 miles de pesos.

Observaciones finales

El ejercicio pretende demostrar el funcionamiento del algoritmo tomando como ejemplo la información de dos cultivos del distrito de riego bajo estudio. La siguiente sección expone los resultados de la aplicación del algoritmo GenSM5 a los 15 cultivos seleccionados presentes en la información del distrito 038.

5.6. Resultados de la aplicación del algoritmo

Para las pruebas, el algoritmo GenSM5 fue aplicado a la información procesada durante la etapa de minería de datos de esta tesis. Esto incluye la ejecución del algoritmo para la información de los 15 cultivos agrícolas seleccionados en la sección 4.3.1 de esta tesis, en el período de tiempo de 1995-2006. Para las restricciones de superficie y lámina de riego por cultivo, se utilizaron rangos que variaban en un 20% abajo y 20 % arriba del promedio de cada dato. Así, por ejemplo, si la superficie promedio sembrada de un cultivo *i* es de 100 ha, entonces la restricción de superficie para el cultivo tiene la forma de $80 \leq SM_i \leq 120$, donde SM_i representa la superficie sembrada promedio del cultivo *i*. Los promedios de precios y costos de producción también fueron utilizados. La tabla 57 muestra las restricciones iniciales de superficie, lámina de riego, costo de producción y precio de venta utilizados.

Cultivo	Superficie (ha)		Lámina de riego (mm)		Costo de producción (M$/ha)	Precio de venta (M$/ton)
	Límite inferior	Límite superior	Límite inferior	Límite superior		
Alfalfa	84.09	126.14	82.64	123.96	5.00	1.53
Algodón	227.06	340.60	68.11	102.17	10.00	4.08
Cártamo	513.58	770.36	29.19	43.79	5.00	2.25
Chícharo	67.78	101.66	61.99	92.98	9.00	2.53
Forraje	109.70	164.55	76.83	115.25	4.00	1.56
Fríjol	146.12	219.18	61.69	92.53	7.00	5.55
Frutales	17.26	25.89	66.46	99.69	10.00	1.21
Garbanzo	218.87	328.31	32.84	49.26	7.00	5.65
Hortalizas	240.94	361.41	66.55	99.83	16.00	2.07
Maíz	638.04	957.06	65.77	98.66	7.00	1.37
Papa	348.37	522.56	63.32	94.98	37.00	3.24
Sorgo	44.32	66.47	63.48	95.22	6.00	1.18
Tomate	81.43	122.15	82.48	123.72	20.00	4.36
Trigo	2,740.06	4,110.10	61.88	92.83	7.00	1.50
Zacate	32.19	48.28	86.54	129.81	8.00	2.05

Tabla 57. Restricciones de cultivos para prueba del algoritmo de optimización.

Para la información de entrada del clima, se tomaron los promedios de las 60 variables involucradas con la información de la serie de datos climática para los años seleccionados en la etapa de minería de datos. La tabla 58 muestra los valores utilizados para cada una de las variables.

Variables climáticas						
Meses del año agrícola		Temperatura (°C)	Temperatura mínima (°C)	Temperatura máxima (°C)	Precipitación (mm)	Evaporación (mm)
	Oct	20.95	18.78	34.55	23.44	150.93
	Nov	14.9	12.66	29.5	17.16	99.78
	Dic	10.52	8.35	25.21	19.22	78.35
	Ene	9.98	7.63	25.68	4.26	83.12
	Feb	10.91	8.84	26.18	7.46	95.81
	Mar	11.78	9.6	28.54	0.9	144.63

	Abr	13.91	11.33	30.89	1.41	187.06
	May	19	14.85	35.32	0.41	234.9
	Jun	23.82	20.69	37.92	2.72	280.63
	Jul	26.68	24.18	38.44	62.02	268.67
	Ago	26.62	24.45	37.87	71.08	254.92
	Sep	25.55	23.54	37.02	75.16	204.25

Tabla 58. Restricciones del clima para prueba del algoritmo de optimización.

Las restricciones de las tablas 57 y 58, junto con la información almacenada en los árboles de regresión M5 de cada cultivo, generan los rendimientos mostrados en la tabla A13.1 (esto de acuerdo a la explicación dada en el punto 5.2 de esta tesis).

En la restricción de superficie total, se tomaron en cuenta dos escenarios. El primero corresponde a un estado de holgura, donde se dispone de más superficie de la que permiten manejar las restricciones de superficie para cada cultivo. El segundo escenario contempla un estado de austeridad, donde la superficie es reducida y su correcta distribución cobra mayor importancia. Sumando los límites superiores de las restricciones de superficie por cultivo, se nota que la superficie máxima con posibilidades de sembrar es de 8,264.72 ha. Tomando esto como base, se propone para el primer escenario una superficie total disponible de 10,000 ha, y para el segundo escenario, una superficie total de 6,000 ha.

En general, se tiene que la optimización se realiza para ambos escenarios con los siguientes parámetros:

- *Función objetivo*: Una función objetivo de 15 variables, cada una representando la superficie a sembrar de un determinado cultivo.
- *Restricciones*: De manera inicial, se tienen las 15 restricciones de superficie para cada uno de los cultivos. Éstas son indicadas en la tabla 57 bajo el encabezado de superficie (columnas *límite inicial* y *límite final*). Cada una de estas 15 restricciones es separada en dos, por lo que el número de restricciones aumenta a 30. A este número hay que agregar la restricción de la suma tota de superficie, la cual se indica que no debe de pasar de 10,000 ha (escenario 1), ó de 6,000 ha (escenario 2). En total, 31 restricciones son introducidas al algoritmo para los problemas representados en los dos escenarios mencionados.

Los parámetros para la parte genética del algoritmo fueron los siguientes:

- *Tamaño del cromosoma*: 15 (igual al número de cultivos).
- *Tamaño de la población*: 10.
- *Probabilidad de cruce*: 0.7.
- *Máximo de generaciones*: 50.
- Probabilidad de mutación: 0.003.

Para comparar los resultados de GenSM5, se utilizó el método Simplex para maximizar el modelo indicado en (5.6), pero utilizando los rendimientos promedio de cada cultivo para reemplazar las funciones $f(S_i, L_i, C\lim a)$. Las restricciones de las superficies de los cultivos fueron dejadas con sus valores originales, ignorando la información proporcionada

por los modelos de árbol de regresión generados en la etapa de minería de datos. En esta maximización, la información del clima y de la lámina de riego es ignorada completamente.

A continuación se exponen los resultados de la comparación.

Escenario no. 1: Superficie máxima disponible de 10,000 ha.

Los resultados de la ejecución de GenSM5 para una superficie total disponible de 10,000 ha se muestran en la tabla 65. Al cabo de 50 generaciones, GenSM5 encontró que el mejor cromosoma era [2,0,3,0,0,1,1,1,0,0,0,2,4,1,0] (en el anexo 13 se pueden consultar los rendimientos asociados a los identificadores mostrados en el cromosoma). Como se ve, aunque existen en mayor número los rendimientos localizados en la posición 0 (mejor rendimiento), se observa que en varios casos el uso de un rendimiento menor proporcionó mayor beneficio, esto debido a que la superficie relacionada con dicho rendimiento se ajustó de mejor manera al problema global. El cromosoma representa la mejor combinación de rendimientos y superficies encontrada para los cultivos seleccionados, los cuales a su vez se traducen en un ingreso neto potencial a obtener por la venta del producto. La tabla 59 muestra los rendimientos, superficies y láminas de riego estimados a partir de la optimización, junto con los totales de producción, ingreso bruto, costo total e ingreso neto calculados a partir de la superficie y rendimiento estimados.

Cultivo	Superficie (ha)	Lámina de riego (mm)	Rendimiento estimado (ton/ha)	Producción estimada (ton)	Ingreso bruto (M$)	Costo total (M$)	Ingreso neto (M$)
Alfalfa	126.14	103.30	13.96	1,760.65	2,693.80	630.71	2,063.09
Algodón	262.50	79.66	3.05	801.47	3,270.78	2,625.00	645.78
Cártamo	770.36	33.01	2.07	1,598.35	3,601.09	2,311.09	1,289.99
Chícharo	101.66	77.48	5.40	548.55	1,389.47	914.98	474.50
Forraje	164.55	82.25	10.55	1,736.44	2,715.80	658.19	2,057.60
Fríjol	156.10	91.99	2.10	327.13	1,815.56	1,092.67	722.90
Frutales	25.89	83.07	18.69	483.73	585.31	258.85	326.46
Garbanzo	328.31	41.05	1.91	628.68	3,554.56	2,298.17	1,256.39
Hortalizas	361.41	80.15	15.99	5,778.17	11,966.59	5,782.54	6,184.05
Maíz	957.06	85.29	6.32	6,045.87	8,282.84	6,699.45	1,583.39
Papa	522.56	71.72	32.91	17,196.69	55,665.69	19,334.57	36,331.12
Sorgo	66.09	85.02	4.64	306.43	362.82	264.35	98.47
Trigo	4,110.10	77.35	5.50	22,625.67	33,983.76	28,770.68	5,213.08
Zacate	48.28	103.75	12.76	616.19	1,261.35	386.22	875.13
Tomate	122.15	87.54	18.20	2,222.86	9,693.90	2,442.92	7,250.98
Totales	**8,123.16**	**78.84**			**140,843.32**	**74,470.39**	**66,372.93**

Tabla 59. Resultados del algoritmo GenSM5 para una superficie disponible de 10,000 ha.

La tabla 60, por otro lado, muestra los resultados de la maximización utilizando Simplex y la información de los rendimientos promedio de los cultivos. Las restricciones para la superficie de cada cultivo son las mismas utilizadas para GenSM5, mostradas con anterioridad en la tabla 57.

Cultivo	Superficie (ha)	Lámina de riego (mm)	Rendimiento esperado (ton/ha)	Producción esperada (ton)	Ingreso bruto (M$)	Costo total (M$)	Ingreso neto (M$)
Alfalfa	126.14	103.3	14.65	1,847.73	2,827.02	630.71	2,196.31
Algodón	340.6	85.14	2.86	974.44	3,976.70	3,405.95	570.75
Cártamo	770.36	36.49	2.13	1,637.02	3,688.21	2,311.09	1,377.12
Chícharo	101.66	77.48	5.72	581.01	1,471.70	914.98	556.72
Forraje	164.55	96.04	10.22	1,681.85	2,630.41	658.19	1,972.21
Fríjol	219.18	77.11	2.11	462.91	2,569.15	1,534.27	1,034.89
Frutales	25.89	83.07	20.04	518.71	627.64	258.85	368.79
Garbanzo	328.31	41.05	1.93	633.31	3,580.73	2,298.17	1,282.56
Hortalizas	361.41	83.19	17.44	6,303.70	13,054.95	5,782.54	7,272.41
Maíz	957.06	82.22	6.76	6,471.67	8,866.18	6,699.45	2,166.74
Papa	522.56	79.15	29.74	15,542.91	50,312.39	19,334.57	30,977.81
Sorgo	66.47	79.35	4.89	325.33	385.19	265.9	119.29
Trigo	4,110.10	77.35	5.55	22,823.37	34,280.70	28,770.68	5,510.02
Zacate	48.28	108.17	13.36	645.17	1,320.67	386.22	934.45
Tomate	122.15	103.1	19.81	2,419.22	10,550.23	2,442.92	8,107.31
Totales	**8,264.72**	**80.81**			**140,141.87**	**75,694.49**	**64,447.38**

Tabla 60. Resultados del método Simplex con rendimientos promedio y superficie total de 10,000 ha.

Comparando los totales de la tabla 59 y 60, se observa que la distribución de superficie y los rendimientos estimados por el algoritmo GenSM5 producen un ingreso neto de 66,372.93 M$, mientras el uso de Simplex con promedios indica que, con su distribución de superficie, se puede obtener un ingreso de 64,447.38 M$. La diferencia es de 1,925.55 M$. Además, la superficie indicada por GenSM5 es 146.56 ha menor, lo que habla de la efectividad del método, ya que en una superficie menor produce un ingreso mayor.

Algo muy importante, es que el rendimiento promedio es más inexacto que el rendimiento proporcionado por los árboles de regresión M5. Esto ya se demostró en la sección de resultados de los árboles M5 en la sección de minería de datos de esta tesis (punto 4.4.3.5.3). La tabla 61 muestra el rendimiento promedio comparado con el rendimiento predicho por los árboles de regresión M5, los cuales sí toman en cuenta la información de superficie, lámina de riego y clima provista como datos de entrada al modelo.

Cultivo	Rendimiento promedio	Rendimiento estimado por los modelos de árbol M5	Diferencia	Error (%)
Alfalfa	14.65	13.96	-0.69	4.71
Algodón	2.86	3.05	0.19	6.64
Cártamo	2.13	2.07	-0.06	2.82
Chícharo	5.72	5.40	-0.32	5.59
Forraje	10.22	10.55	0.33	3.23
Fríjol	2.11	2.10	-0.01	0.47
Frutales	20.04	18.69	-1.35	6.74
Garbanzo	1.93	1.91	-0.02	1.04
Hortalizas	17.44	15.99	-1.45	8.31
Maíz	6.76	6.32	-0.44	6.51

Papa	29.74	32.91	3.17	10.66
Sorgo	4.89	4.64	-0.25	5.11
Trigo	5.55	5.50	-0.05	0.90
Zacate	13.36	12.76	-0.60	4.49
Tomate	19.81	18.20	-1.61	8.13
Promedio				5.02

Tabla 61. Comparación del rendimiento promedio y las estimaciones realizadas por los árboles de regresión M5.

En la tabla 61 se puede apreciar el porcentaje del error entre la estimación de los árboles M5 y el rendimiento promedio. En la columna diferencia se puede ver que la mayoría de los rendimientos promedio son sobreestimaciones, y aún con dichas sobreestimaciones, la maximización del algoritmo GenSM5 supera a la maximización con rendimientos promedio.

Escenario no. 2: Superficie máxima disponible de 6,000 ha.

La tabla 62 muestra los resultados de la ejecución del algoritmo GenSM5 cuando la superficie disponible está restringida a 6,000 ha, el cromosoma encontrado corresponde a la combinación [2,0,3,0,0,0,1,1,0,1,0,1,0,1,0] (en el anexo 13 se pueden consultar los rendimientos asociados a los identificadores mostrados en el cromosoma). La tabla 63 muestra los resultados de la ejecución del método Simplex con la información de rendimientos promedios y la misma restricción de superficie. En ambos casos, la superficie es de 6,000 ha (el máximo posible a utilizar), pero se puede observar, que los resultados del algoritmo GenSM5 indican un beneficio económico neto superior en 2,512.89 M$ con respecto a la estimación del método Simplex.

Cultivo	Superficie (ha)	Lámina de riego (mm)	Rendimiento esperado (ton/ha)	Producción esperada (ton)	Ingreso bruto (M$)	Costo total (M$)	Ingreso neto (M$)
Alfalfa	126.14	103.3	13.96	1,760.65	2,693.80	630.71	2,063.09
Algodón	262.5	79.66	3.05	801.47	3,270.78	2,625.00	645.78
Cártamo	770.36	33.01	2.07	1,598.35	3,601.09	2,311.09	1,289.99
Chícharo	101.66	77.48	5.4	548.55	1,389.47	914.98	474.5
Forraje	164.55	82.25	10.55	1,736.44	2,715.80	658.19	2,057.60
Fríjol	219.18	91.99	2.1	459.53	2,550.42	1,534.27	1,016.15
Frutales	25.89	83.07	18.69	483.73	585.31	258.85	326.46
Garbanzo	328.31	41.05	1.91	628.68	3,554.56	2,298.17	1,256.39
Hortalizas	361.41	80.15	15.99	5,778.17	11,966.59	5,782.54	6,184.05
Maíz	957.06	68.84	6.31	6,041.56	8,276.94	6,699.45	1,577.49
Papa	522.56	71.72	32.91	17,196.69	55,665.69	19,334.57	36,331.12
Sorgo	66.47	72.03	4.73	314.62	372.51	265.9	106.61
Trigo	1,923.48	67.74	5.52	10,607.99	15,933.20	13,464.35	2,468.84
Zacate	48.28	103.75	12.76	616.19	1,261.35	386.22	875.13
Tomate	122.15	87.54	18.2	2,222.86	9,693.90	2,442.92	7,250.98
Totales	**6,000.00**	**76.24**			**123,531.41**	**59,607.21**	**63,924.18**

Tabla 62. Resultados del algoritmo GenSM5 para una superficie disponible de 6,000 ha.

Cultivo	Superficie (ha)	Lámina de riego (mm)	Rendimiento esperado (ton/ha)	Producción esperada (ton)	Ingreso bruto (M$)	Costo total (M$)	Ingreso neto (M$)
Alfalfa	126.14	103.3	14.65	1,847.73	2,827.02	630.71	2,196.31
Algodón	340.6	85.14	2.86	974.44	3,976.70	3,405.95	570.75
Cártamo	770.36	36.49	2.13	1,637.02	3,688.21	2,311.09	1,377.12
Chícharo	101.66	77.48	5.72	581.01	1,471.70	914.98	556.72
Forraje	164.55	96.04	10.22	1,681.85	2,630.41	658.19	1,972.21
Fríjol	219.18	77.11	2.11	462.91	2,569.15	1,534.27	1,034.89
Frutales	25.89	83.07	20.04	518.71	627.64	258.85	368.79
Garbanzo	328.31	41.05	1.93	633.31	3,580.73	2,298.17	1,282.56
Hortalizas	361.41	83.19	17.44	6,303.70	13,054.95	5,782.54	7,272.41
Maíz	957.06	82.22	6.76	6,471.67	8,866.18	6,699.45	2,166.74
Papa	522.56	79.15	29.74	15,542.91	50,312.39	19,334.57	30,977.81
Sorgo	66.47	79.35	4.89	325.33	385.19	265.9	119.29
Trigo	1,845.38	77.35	5.55	10,247.42	15,391.62	12,917.69	2,473.93
Zacate	48.28	108.17	13.36	645.17	1,320.67	386.22	934.45
Tomate	122.15	103.1	19.81	2,419.22	10,550.23	2,442.92	8,107.31
Totales	6,000.00	80.81			121,252.79	59,841.50	61,411.29

Tabla 63. Resultados del método Simplex con rendimientos promedio y superficie total de 6,000 ha.

Observaciones finales

Una desventaja del algoritmo GenSM5 (y de los genéticos en general) es el tiempo consumido en la búsqueda de la solución óptima. En los ejercios realizados, se notó que el tiempo promedio invertido en la ejecución del algoritmo GenSM5 con los parámetros especificados fue de 8 segundos. Una sola maximización Simplex se lleva alrededor de 0.0026 segundos. Aún así, estos 8 segundos es poco tiempo comparado con el tiempo invertido en realizar una búsqueda exhaustiva a través de todas las combinaciones de rendimientos posibles. La tabla 64 muestra el número de combinaciones posibles dados los rendimientos expuestos en la tabla A13.1 del anexo 3. Como se observa, el número de combinaciones posibles llega a 20,160,000 maximizaciones. Multiplicando esta cantidad por el tiempo que lleva una maximización Simplex se calcula que el tiempo que tomaría una búsqueda exhaustiva a través de todas las maximizaciones posibles es de 52,416 segundos, que representa un tiempo de 14.56 horas.

Cultivo	Número de posibles rendimientos	Combinaciones
ALFALFA	3	3
ALGODON	3	9
CARTAMO	7	63
CHICHARO	1	63
FORRAJE	2	126
FRIJOL	5	630
FRUTALES	2	1260
GARBANZO	2	2520
HORTALIZAS	2	5040
MAIZ	2	10080
PAPA	5	50400
SORGO	5	252000
TOMATE	5	1260000
TRIGO	8	10080000

ZACATE	2	**20,160,000**

Tabla 64. Número de combinaciones necesarias en una búsqueda exhaustiva.

Si bien el tiempo estimado de la búsqueda exhaustiva para este ejercicio podría no justificar el uso de una heurística para realizar dicha búsqueda, se invita a considerar el caso de contar con 20 cultivos a sembrar en lugar de 15. Con cultivos similares a los utilizados por este ejercicio, se considera que el número de combinaciones necesarias en una búsqueda exhaustiva con 20 cultivos aumentaría a 2,540,160,000, con un tiempo total consumido de 1,834.56 horas, que representa a su vez 76.44 días. Aumentando el número de cultivos a 25, se calcula que el tiempo total necesario sería de 6,115 días (16.75 años). Como se ve, se trata de un problema extremadamente complejo cuyo tiempo de solución aumenta de manera exponencial con respecto al número de cultivos involucrados, con lo cual el uso de una heurística de búsqueda no está solamente justificada, si no que es necesaria.

También debe tomarse en cuenta que son las restricciones las que influyen en el número de rendimientos seleccionables para el problema. Cerrando mucho los rangos de superficie y lámina de riego se accede a rendimientos demasiado específicos, con lo cual el número de combinaciones se reduce (aunque también el rango de variación de la superficie). Abriendo los rangos se provoca que más rendimientos sean tomados en cuenta para el problema, lo que aumenta el espacio de búsqueda. Suponiendo que se aceptaran todos los rendimientos de los cultivos bajo análisis, el número máximo de combinaciones calculadas sería de 5.09139E+20 (expresión 5.7), cuya búsqueda exhaustiva tomaría más de más de 109 millones de años en llevarse a cabo.

6. Implementación del modelo predictivo y del algoritmo de optimización en una herramienta de Software

El algoritmo GenSM5 fue implementado en una aplicación de software con el objetivo de proporcionar una herramienta de administración que permitiera al usuario construir distintos escenarios de la producción agrícola en un distrito de riego y observar la distribución de recursos sugerida por el algoritmo. La pantalla principal de la aplicación se muestra en la figura 50.

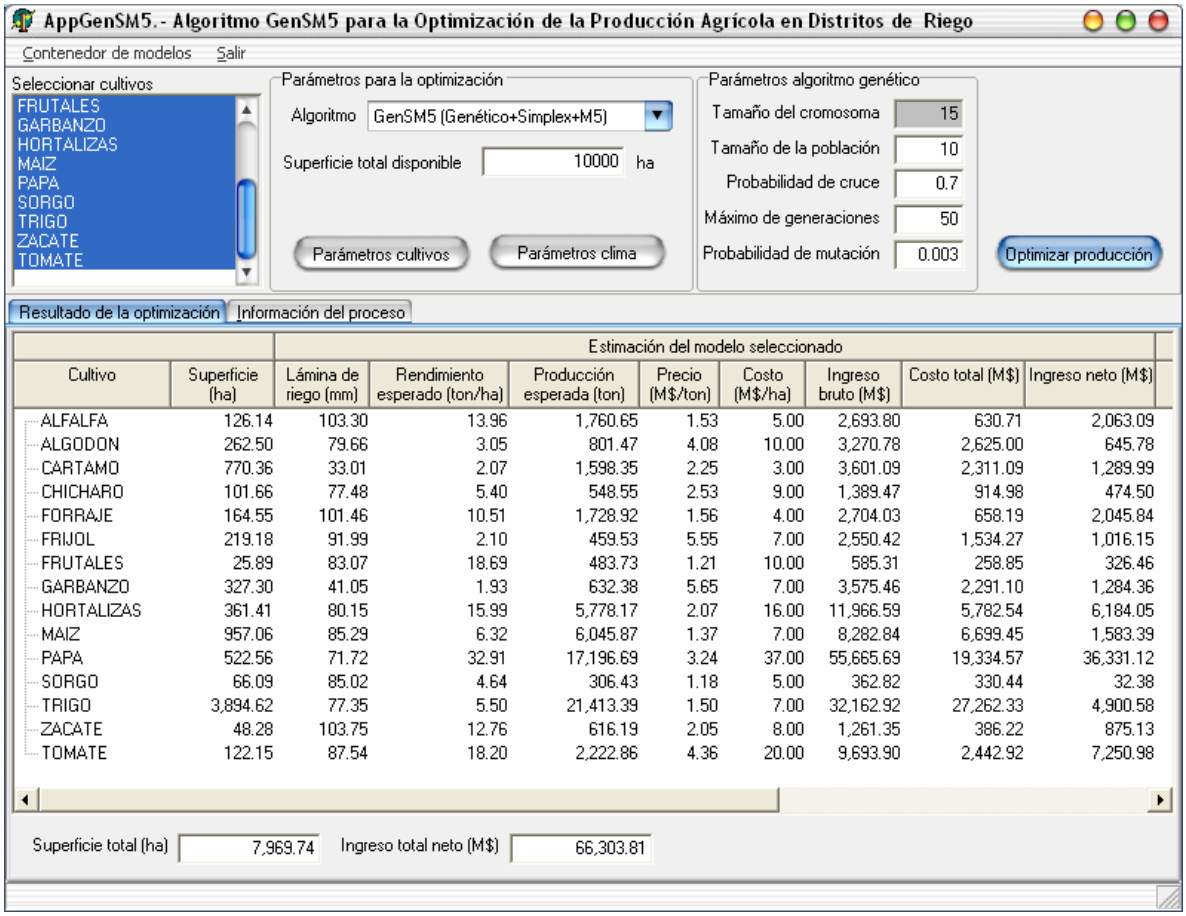
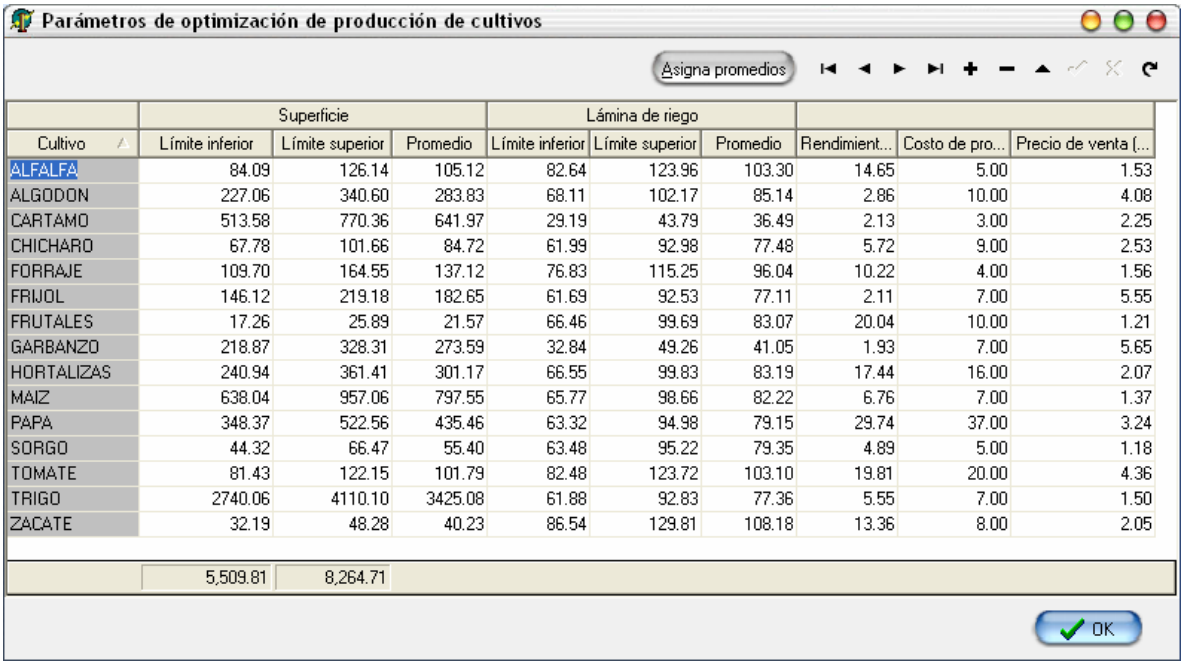


Figura 50. Pantalla principal de la aplicación AppGenSM5

La primera acción que el usuario debe realizar es la selección de cultivos a incluir en el modelo. Para ello, se utiliza la lista localizada en la esquina superior derecha de la ventana. Una vez seleccionados los cultivos, debe indicarse a la aplicación el método utilizado para llevar a cabo la optimización. Se encuentran disponibles los algoritmos *GenSM5* y *Simplex*. GenSM5 hace uso de la información de los modelos de árbol de regresión (si está disponible), mientras que Simplex solo utiliza los rendimientos promedio. El algoritmo Simplex fue introducido en la aplicación con el fin de contar con un método para contrastar los resultados generados por el algoritmo GenSM5 (ver capítulo anterior de esta tesis). En cualquiera de los dos casos, se requiere que el usuario especifique la superficie total

disponible, esto en el cuadro de edición localizado debajo de la lista para la selección del algoritmo.

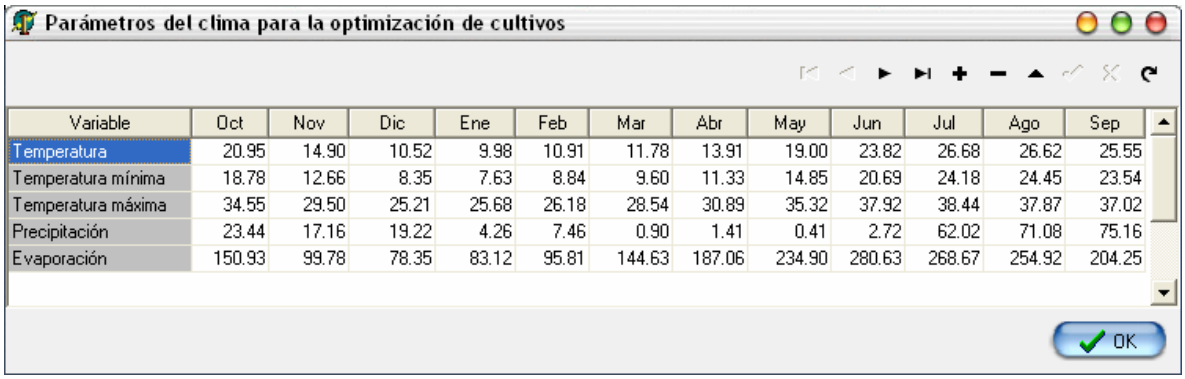
Es necesario que se introduzcan las restricciones por cultivo requeridas por el algoritmo de optimización. Para ello, se utiliza el diálogo "*Parámetros de optimización de producción de cultivos*", que se abre al dar clic en el botón "*parámetros cultivos*". La figura 51 muestra el diálogo utilizado para la introducción de las restricciones de los cultivos seleccionados.



	Superficie			Lámina de riego					
Cultivo	Límite inferior	Límite superior	Promedio	Límite inferior	Límite superior	Promedio	Rendimient...	Costo de pro...	Precio de venta [...
ALFALFA	84.09	126.14	105.12	82.64	123.96	103.30	14.65	5.00	1.53
ALGODON	227.06	340.60	283.83	68.11	102.17	85.14	2.86	10.00	4.08
CARTAMO	513.58	770.36	641.97	29.19	43.79	36.49	2.13	3.00	2.25
CHICHARO	67.78	101.66	84.72	61.99	92.98	77.48	5.72	9.00	2.53
FORRAJE	109.70	164.55	137.12	76.83	115.25	96.04	10.22	4.00	1.56
FRIJOL	146.12	219.18	182.65	61.69	92.53	77.11	2.11	7.00	5.55
FRUTALES	17.26	25.89	21.57	66.46	99.69	83.07	20.04	10.00	1.21
GARBANZO	218.87	328.31	273.59	32.84	49.26	41.05	1.93	7.00	5.65
HORTALIZAS	240.94	361.41	301.17	66.55	99.83	83.19	17.44	16.00	2.07
MAIZ	638.04	957.06	797.55	65.77	98.66	82.22	6.76	7.00	1.37
PAPA	348.37	522.56	435.46	63.32	94.98	79.15	29.74	37.00	3.24
SORGO	44.32	66.47	55.40	63.48	95.22	79.35	4.89	5.00	1.18
TOMATE	81.43	122.15	101.79	82.48	123.72	103.10	19.81	20.00	4.36
TRIGO	2740.06	4110.10	3425.08	61.88	92.83	77.36	5.55	7.00	1.50
ZACATE	32.19	48.28	40.23	86.54	129.81	108.18	13.36	8.00	2.05
	5,509.81	8,264.71							

Figura 51. Diálogo para la introducción de restricciones de superficie, lámina de riego, costo de producción y precio de venta para los cultivos.

El botón "*parámetros clima*" en la pantalla principal permite registrar los valores de las variables climáticas con el fin de establecer las condiciones del año agrícola de referencia. La figura 52 muestra el diálogo para la introducción de las condiciones climáticas.



Variable	Oct	Nov	Dic	Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Sep
Temperatura	20.95	14.90	10.52	9.98	10.91	11.78	13.91	19.00	23.82	26.68	26.62	25.55
Temperatura mínima	18.78	12.66	8.35	7.63	8.84	9.60	11.33	14.85	20.69	24.18	24.45	23.54
Temperatura máxima	34.55	29.50	25.21	25.68	26.18	28.54	30.89	35.32	37.92	38.44	37.87	37.02
Precipitación	23.44	17.16	19.22	4.26	7.46	0.90	1.41	0.41	2.72	62.02	71.08	75.16
Evaporación	150.93	99.78	78.35	83.12	95.81	144.63	187.06	234.90	280.63	268.67	254.92	204.25

Figura 52. Pantalla principal de la aplicación AppGenSM5

En el caso de que se esté utilizando el algoritmo GenSM5 para la optimización del ingreso, es necesario indicar los parámetros requeridos para el ciclo genético presente en el algoritmo. La aplicación ofrece para ello el diálogo que se muestra en la figura 53.

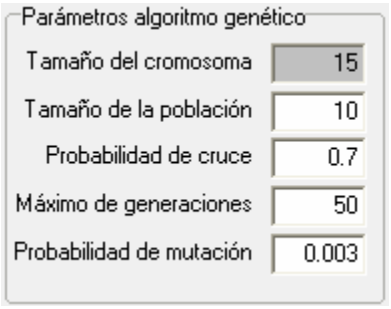


Figura 53. Diálogo para la introducción de los parámetros del ciclo genético.

Para proceder con la optimización (una vez seleccionado el algoritmo, las restricciones y los parámetros), el usuario deberá dar clic en el botón "*optimizar producción*". En el caso de utilizar el algoritmo Simplex la salida se obtiene casi de inmediato. En el caso del algoritmo GenSM5, el programa demorará un poco más. La salida de la optimización se visualiza en una tabla que se encuentra al centro del diálogo mostrado en la figura 50. En esta tabla se pueden apreciar los siguientes datos:

- Cultivo.
- **Superficie (ha).**
- **Lámina de riego (mm).**
- **Rendimiento estimado (ton/ha).**
- **Producción esperada (ton).**
- Precio (M$/ton).
- Costo de producción (M$/ha).
- **Ingreso bruto (M$)**
- **Costo total (M$)**
- **Ingreso neto (M$)**

Los datos en negrillas son estimaciones producto de la optimización. En el caso del uso del algoritmo Simplex, el rendimiento indicado es el dato promedio de los cultivos (no es producto de la una estimación del algoritmo). La figura 54 proporciona una vista completa de la tabla de salida generada por el programa.

		Estimación del modelo seleccionado							
Cultivo	Superficie (ha)	Lámina de riego (mm)	Rendimiento esperado (ton/ha)	Producción esperada (ton)	Precio (M$/ton)	Costo (M$/ha)	Ingreso bruto (M$)	Costo total (M$)	Ingreso neto (M$)
ALFALFA	126.14	103.30	13.96	1,760.65	1.53	5.00	2,693.80	630.71	2,063.09
ALGODON	262.50	79.66	3.05	801.47	4.08	10.00	3,270.78	2,625.00	645.78
CARTAMO	770.36	33.01	2.07	1,598.35	2.25	3.00	3,601.09	2,311.09	1,289.99
CHICHARO	101.66	77.48	5.40	548.55	2.53	9.00	1,389.47	914.98	474.50
FORRAJE	164.55	101.46	10.51	1,728.92	1.56	4.00	2,704.03	658.19	2,045.84
FRIJOL	219.18	91.99	2.10	459.53	5.55	7.00	2,550.42	1,534.27	1,016.15
FRUTALES	25.89	83.07	18.69	483.73	1.21	10.00	585.31	258.85	326.46
GARBANZO	327.30	41.05	1.93	632.38	5.65	7.00	3,575.46	2,291.10	1,284.36
HORTALIZAS	361.41	80.15	15.99	5,778.17	2.07	16.00	11,966.59	5,782.54	6,184.05
MAIZ	957.06	85.29	6.32	6,045.87	1.37	7.00	8,282.84	6,699.45	1,583.39
PAPA	522.56	71.72	32.91	17,196.69	3.24	37.00	55,665.69	19,334.57	36,331.12
SORGO	66.09	85.02	4.64	306.43	1.18	5.00	362.82	330.44	32.38
TRIGO	3,894.62	77.35	5.50	21,413.39	1.50	7.00	32,162.92	27,262.33	4,900.58
ZACATE	48.28	103.75	12.76	616.19	2.05	8.00	1,261.35	386.22	875.13
TOMATE	122.15	87.54	18.20	2,222.86	4.36	20.00	9,693.90	2,442.92	7,250.98

Superficie total (ha) 7,969.74 Ingreso total neto (M$) 66,303.81

Figura 54. Tabla de salida de la ejecución del sistema.

Además de los datos mencionados, se muestran cuatro columnas que permiten comparar los resultados del algoritmo Simplex y el rendimiento estimado por los árboles M5. Éstas no son de utilidad cuando se utiliza el algoritmo GenSM5 para la optimización.

La aplicación permite exportar los resultados generados por medio de la opción "*exportar a memoria*" del menú auxiliar que aparece al dar clic derecho sobre la tabla de resultados.

La superficie total y el ingreso neto calculado por el algoritmo se muestran en cuadros de texto ubicados debajo de la tabla de resultados.

La aplicación también permite consultar la información generada durante el procesamiento realizado por el algoritmo. Para ello, el usuario debe seleccionar el "tab" con el nombre "información del proceso", que se localiza encima de la tabla de resultados (ver figura 55).

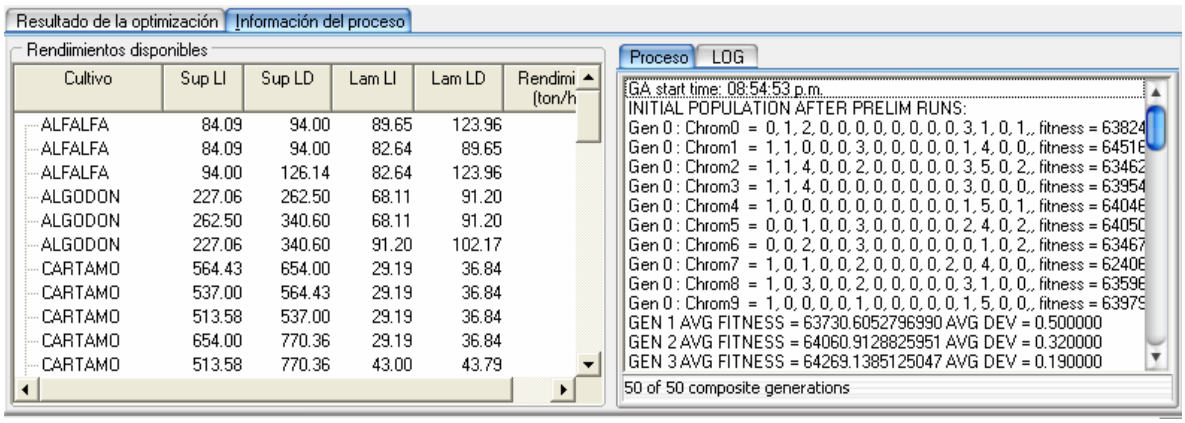


Figura 55. Diálogo que muestra información del proceso realizado por el algoritmo GenSM5

En el diálogo de la figura 55 se pueden distinguir dos secciones: información de rendimientos disponibles e información del proceso. La tabla de "*rendimientos disponibles*"

muestra los rendimientos (con los rangos de superficie y lámina de riego) extraídos de los árboles de regresión M5 y que han sido contrastados con la información de las restricciones de los cultivos. La sección de *proceso* muestra los cromosomas generados y el proceso evolutivo seguido por la parte genética de GenSM5.

Para introducir los modelos de árbol de regresión M5 que fueron originalmente generados por la herramienta de minería de datos Weka, fue necesaria la inclusión de un módulo en la aplicación que leyera los árboles de regresión y almacenara su información en un formato tratable por la aplicación. Esta es la tarea del contenedor de modelos, el cual se accede desde la pantalla principal por medio de la opción "*información de árboles M5*" del menú "*contenedor de modelos*". La figura 56 muestra la interfaz del módulo.

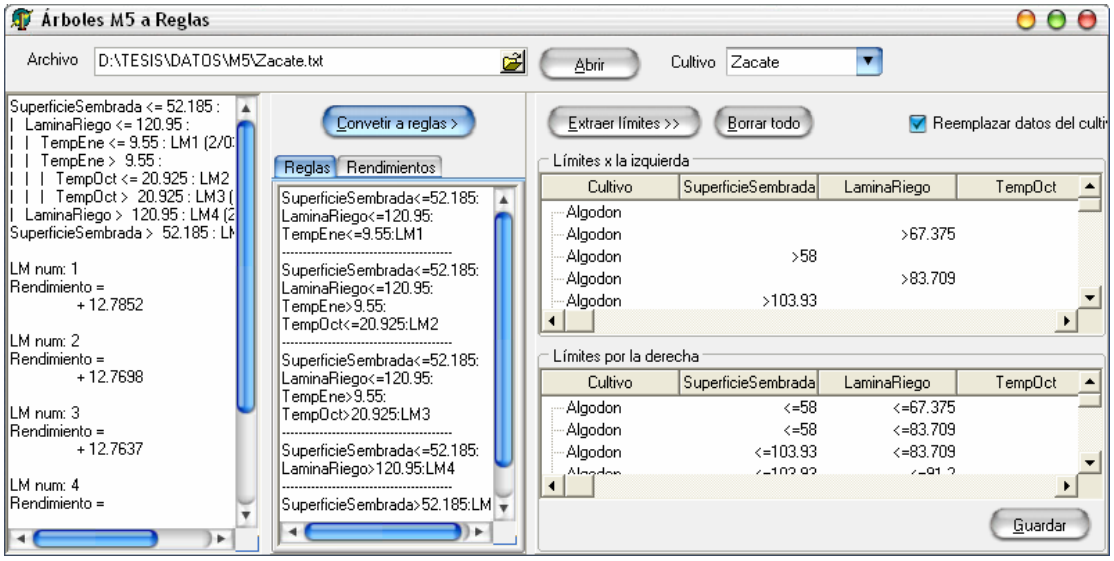


Figura 56. Módulo para la extracción de información de los árboles de regresión M5.

El diálogo de la figura 56 permite abrir un archivo de texto que almacena la representación de Weka para un árbol de regresión y lo procesa con el fin de extraer y almacenar las reglas expresadas en el árbol. Para ello, separa los valores de prueba en dos tablas, una dedicada a almacenar las pruebas sobre atributos del tipo "<" ó "<=", y otra que almacena las pruebas del tipo ">"y ">=". Las pruebas del tipo "=" se almacenan en ambas tablas. Además de las pruebas, estas tablas almacenan también el valor del rendimiento (que también es extraído del archivo de texto que almacena la representación del árbol). Las tablas son consultadas por la aplicación para extraer la información de los rendimientos que aplican bajo ciertos valores de superficie, lámina de riego y las variables climáticas.

6.1. Consideraciones técnicas

La aplicación fue desarrollada en el lenguaje de programación Delphi 6. La información utilizada por la aplicación se almacena en una base de datos tipo MySQL. Las pruebas de ejecución del algoritmo (implementado en la aplicación) se realizaron en una computadora con procesador Intel Centrino Duo a 1.66GHz, con 1 GB de memoria RAM y un disco duro de 80 GB. El sistema operativo para el cual se desarrolló la aplicación fue Windows XP.

7. Conclusiones

En la presente tesis se aplicó el proceso de descubrimiento de conocimiento en bases de datos a un conjunto de información de producción agrícola generada en un distrito de riego. La práctica generó modelos de predicción del rendimiento para un grupo de cultivos con fuerte influencia en el distrito. Durante la práctica, se propuso comparar la efectividad de predicción de modelos generados por medio de técnicas de aprendizaje automático contra la efectividad de predicción de técnicas estadísticas tradicionales. Se llegó a la conclusión de que la técnica de árbol de regresión M5 (perteneciente al primer grupo de técnicas) obtiene una certeza de predicción superior del 4.53% frente a la técnica de la regresión multivariable. Una vez que los mejores modelos fueron seleccionados, éstos fueron utilizados en un algoritmo genético para obtener la combinación de rendimientos que maximizan el ingreso del distrito al sembrar los cultivos seleccionados. Con estos resultados como base, se concluye que el objetivo principal de esta tesis planteado como "*el aplicar el proceso de descubrimiento de conocimiento en bases de datos para desarrollar un modelo de predicción que permitiera optimizar el beneficio económico obtenido por la producción agrícola generada en un distrito de riego*" fue plenamente alcanzado. El algoritmo propuesto para la optimización del ingreso incrementa el beneficio económico alcanzado en cerca de un 2.99 %, mientras ahorra un 1.77 % de la superficie (para las pruebas realizadas). Cabe mencionar que el algoritmo desarrollado utiliza como estimadores a los modelos de árbol de regresión M5, los cuales consideran la influencia que los factores cantidad de superficie, lámina de riego y clima tienen sobre la producción agrícola. El rendimiento estimado por estos modelos tienen una certeza superior frente al promedio de al menos 2.77 %, beneficio que debe ser sumado al algoritmo de optimización propuesto.

A continuación se exponen una serie de conclusiones sobre aspectos específicos del proceso desarrollado durante la tesis.

7.1. Sobre el proceso de minería de datos

En esta tesis se llevó a cabo un proyecto de minería de datos utilizando la metodología CRISP-DM. El propósito de dicha práctica fue doble: obtener un modelo de predicción del rendimiento de los cultivos y evaluar la utilidad de la metodología CRISP-DM [33] como una guía exitosa en los proyectos de minería de datos. Las conclusiones derivadas de las actividades realizadas durante el proceso de minería de datos son expuestos en esta sección. Las conclusiones específicas sobre la utilidad de CRISP-DM se exponen en el siguiente punto.

Las conclusiones se agrupan por etapa:

c) **Comprensión del negocio.** La "comprensión del negocio" fue una parte vital para la conclusión exitosa del proyecto de minería. Muchos conceptos necesarios para la asimilación del problema de la maximización de la producción agrícola tuvieron que ser aprendidos durante esta etapa, pero contribuyó a plantear de manera correcta el problema y los objetivos de la práctica de minería de datos. Como resultado de esta

etapa, se concluyó que la variable más importante dentro del proceso de producción de los cultivos agrícolas es el **rendimiento**, lo cual conllevó a dos cosas: a) plantear dicha variable como el objetivo de la labor de minería de datos y b) decidir que los modelos desarrollados serían del tipo predictivo.

d) Comprensión de los datos. El origen de la información influye en el tiempo invertido en la preparación de los datos para la aplicación de los algoritmos de minería. El contexto/sistema origen en el que se generan y almacenan los datos proporciona tanta o más información sobre el estado de los datos que los datos mismos. Las fuentes de datos a las que se tuvo acceso para establecer la serie de datos objetivo tuvieron orígenes distintos. Aquellas que provenían de sistemas donde el aseguramiento de la calidad de la información había sido prioritaria, no requirieron prácticamente tiempo de preparación, mientras que aquellas que provenían de sistemas con pocos o nulos mecanismos de validación se llevaron la mayor parte del tiempo de pre-procesamiento. Algunos de los errores presentes en los datos no hubieran podido ser resueltos de no conocer la forma en como fueron almacenados. Por lo tanto, se resalta la importancia de llevar un registro confiable de los sistemas o medios que dan origen a la información utilizada durante el proyecto, y de ser posible, acceder a dichos sistemas para obtener información del contexto en el que se generaron los datos.
Las etapas de *comprensión del negocio* y la *comprensión de los datos* son una particularidad de la metodología CRISP-DM. La mayoría de los autores que describen el proceso de minería de datos se centran en tres etapas: la *preparación de datos*, el *modelado* y la *evaluación*.

e) Preparación de los datos. Muchos autores señalan que la preparación de los datos es con frecuencia la actividad que consume la mayor parte del tiempo invertido en un proyecto de minería de datos [209][61][65]. Incluso se habla de porcentajes de tiempo, como del 40% ó incluso del 80 al 90%. La presente tesis coincide totalmente con dichos autores, al constatar que dicha actividad requirió aproximadamente del 50% del tiempo total del proyecto. La tarea específica que consumió más tiempo fue la de limpieza de datos, esto por encima de tareas como la selección, la construcción o la integración de los datos.
Durante la fase de limpieza de datos se aplicó la taxonomía mostrada en Kim et al (2001) [62] para identificar los datos erróneos presentes en la serie de datos objetivo. Los errores detectados fueron corregidos con técnicas estadísticas que se describen en esta tesis.
El resultado final de la etapa de preparación de los datos fue una serie de datos integrada de información de producción agrícola y climática, la cual presentaba 64 atributos (ver anexo 7) y 1,186 registros.

f) Modelado y evaluación de los modelos. Durante estas etapas se evaluaron diferentes técnicas de modelado para el problema de la predicción del rendimiento de cultivos. Se compararon técnicas estadísticas (regresión multivariable y la media) y técnicas de aprendizaje automático (red neuronal perceptrón multicapa [210], árbol de clasificación J48 [68] y árbol de regresión M5 [165]). La selección de algoritmos se basó principalmente en la bibliografía consultada, ya que se utilizaron técnicas que habían sido utilizadas en problemas parecidos por otros autores [37][38][39][40][41]. Para

evaluar el desempeño de los algoritmos, se generó un diseño de pruebas que consistió en la aplicación de cinco métricas en un conjunto de datos seleccionado por medio de los métodos de validación simple y validación cruzada. Las métricas seleccionadas permitieron medir la eficiencia de predicción, la cantidad promedio de error y la certeza comparada contra la media.

Para el modelado, la serie de datos objetivo fue dividida en 15 conjuntos de registros, cada uno representando la información de un cultivo. Las métricas y los métodos de validación fueron aplicados a cada uno de los conjuntos, y estos fueron comparados para cada técnica de modelado seleccionada.

Como conclusión, se obtuvo que la técnica que representaba más ventajas para la predicción del rendimiento en la serie de datos seleccionada fue el algoritmo de árbol de regresión M5. Este algoritmo presentó una efectividad promedio del 72 .08 % con una desviación promedio de error absoluto de 2.36 ton/ha, frente al 71.45 % y 2.64 ton/ha de su competidor más cercano, el árbol de clasificación J48. La efectividad promedio de la regresión lineal múltiple fue estimada en 67.55 %, con un error promedio absoluto de 2.87 ton/ha.

Las redes neuronales perceptrón multicapa no presentaron grandes avances frente a la regresión multivariable, pero merecen una mención aparte de los otros métodos de aprendizaje utilizados. Las redes neuronales dependen enormemente de dos factores: el tiempo de aprendizaje y la topología de la red. La forma de seleccionar los parámetros para los modelos desarrollados se expone en la sección 4.4.3.3.1, que básicamente, fue un proceso de prueba y error guiado por los resultados obtenidos en cada prueba. Esta actividad consumió mucho del tiempo empleado en la generación de los modelos, y aún así, es posible que las topologías y los tiempos de aprendizaje óptimos para cada conjunto de registros no hayan sido del todo "descubiertos". El tiempo invertido en esta búsqueda y los resultados obtenidos hacen ver aún mejor a las otras técnicas evaluadas, cuyo costo en tiempo para la generación de los modelos es mucho menor que el consumido por las redes neuronales. Sin embargo, no se exploraron todas las topologías y tiempos de aprendizaje posibles, por lo que la existencia de otras configuraciones de red que proporcionen mejores resultados que los obtenidos en esta tesis no está descartada.

En contraste, la expresividad de los árboles de clasificación y regresión es una ventaja comparada con el modelo de caja negra que es una red neuronal. En la sección de conclusiones dedicadas al modelo de optimización de la producción agrícola se describe como esta expresividad es aprovechada dentro de un contexto de optimización.

Las pruebas realizadas permiten concluir también, que las técnicas de aprendizaje automático de minería de datos sí ofrecen ventajas sobre las técnicas estadísticas tradicionales, y en el caso particular de esta tesis, ofrecen mayor certeza en la predicción del rendimiento de cultivos. Estos resultados están fuertemente ligados a la serie de datos preparada e integrada durante este proyecto, por lo que no deben ser tomadas como verdades absolutas que pueden ser aplicadas a cualquier otro conjunto de datos.

El uso de los modelos generados se puede consultar en la sección dedicada a las conclusiones del algoritmo de optimización (sección 7.2 de esta tesis)

7.1.1. Sobre CRISP-DM

Las siguientes son algunas observaciones referentes a la metodología CRISP-DM, y se basan en el documento CRISP-DM versión 1.0 (guía paso a paso para minería de datos), publicada por Chapman, Clinton, Kerber y otros en 1999 [33].

- CRISP-DM está totalmente orientado a los negocios (si no totalmente, sí en su mayor parte). Esto provoca que algunos de los conceptos manejados por la metodología sean difíciles de empalmar en otro ámbito (por ejemplo, el científico), al no contar con una explicación más amplia o general de su significado. Estos inconvenientes se presentan sobre todo en la fase de comprensión del negocio (Bussines understanding), la cual está inundada de conceptos relacionados con los negocios, pero con muy poca información al respecto. Por ejemplo, resulta complicado distinguir exactamente entre los "objetivos del negocio" y los "objetivos de la minería de datos", o bien, distinguir entre el "criterio de éxito del negocio" y el "criterio de éxito de la minería de datos", sobre todo cuando la guía se está aplicando en un contexto que no es propiamente el de negocios.

- La secuencia entre las fases de un proyecto de minería no es rígida, pero la guía no logra reflejarlo. En la página 13, en la descripción del modelo de referencia, la guía CRISP-DM señala que el ir y venir entre las fases es siempre necesario. Pero ni el modelo de referencia ni la guía de usuario proporcionan información específica sobre cuándo volver hacia alguna de las fases en particular.

- Demasiada generalidad. En el afán de hacer CRISP-DM un guía de propósito abierto, algunos de los conceptos se describen en términos tan generales, que caen en la ambigüedad. Hace falta descripciones más precisas de los conceptos manejados en la guía.

- Falta de ejemplos. Pese a que se buscaron, no fue posible encontrar ejemplos prácticos de proyectos completos de minería de datos que hayan sido desarrollados con la metodología.

Cabe señalar, que estas apreciaciones corresponden únicamente al punto de vista del autor de esta tesis.

7.2. Sobre el algoritmo para la optimización de la producción agrícola

Para lograr el objetivo de optimización, en esta tesis se propone un algoritmo basado en un algoritmo genético, cuya función de aptitud es la maximización del ingreso total producto de la siembra de un grupo de cultivos. Esta maximización utiliza los modelos de predicción del rendimiento generados durante la etapa de minería de datos.

El algoritmo fue probado en dos escenarios distintos de superficie disponible. El primero corresponde a un escenario de holgura, donde la superficie disponible excede la cantidad

máxima impuesta por las restricciones de los cultivos. El segundo escenario restringe la superficie y obliga al algoritmo a buscar la mejor distribución de superficie.

Las pruebas realizadas permiten concluir que, en un escenario de superficie ilimitada, el algoritmo proporciona beneficios como un incremento del 2.98 % en el ingreso (lo que equivale a un beneficio neto de $ 1,925,550), junto con un ahorro del 1.77 % en superficie sembrada (que representa un ahorro de 146.56 ha y un volumen de agua asociado no cuantificado). En un escenario de austeridad, GenSM5 proporciona una ganancia del 4.09% en el ingreso (que representa un beneficio neto de $ 2,512,890), con la máxima superficie disponible. Estos porcentajes se establecen en comparación con la maximización con el método Simplex, utilizando el rendimiento promedio de los cultivos.

7.3. Trabajo futuro

7.3.1. Sobre CRISP-DM

- Proponer una ampliación de la metodología CRISP-DM que clarifique muchos de los conceptos planteados en la versión 1.0. Se sabe que actualmente, un grupo abierto de investigadores colaboran en el desarrollo de la versión 2.0 de esta metodología, y es probable que esta observación ya sea obsoleta para este momento.

- Relacionado con lo anterior, se propone que en esta metodología se incorporen métricas de evaluación que proporcionen al usuario un punto de referencia específico para dirigir la secuencia del proceso de minería. Estas métricas deben ser aplicadas a las salidas obtenidas en cada fase, produciendo un parámetro que indica de manera explícita cuándo repetir ó avanzar entre las fases. Así, por ejemplo, una métrica de la desviación de los datos podría indicar cuándo repetir la fase de limpieza, o, en su defecto, cuándo avanzar hacia la construcción de los modelos.

7.3.2. Sobre minería de datos y cultivos agrícolas

- Realizar un análisis más profundo de los modelos de árbol de regresión M5 del rendimiento de los cultivos. Los árboles M5 tienen un enorme poder descriptivo, pero por cuestión de alcance, durante esta tesis no fue posible realizar un análisis más profundo de la información contenida en dichos modelos (por ejemplo, las relaciones entre los atributos).

- Probar con otras técnicas de modelado del rendimiento, basadas sobre todo en funciones probabilísticas (árboles bayesianos, cadenas de Markov, modelos ocultos de Markov, etc.).

- Incrementar los atributos de las series para incluir otros factores importantes involucrados en el proceso de producción (como propiedades del suelo, prácticas de manejo de la tierra, técnicas de riego, etc). Esto trae consigo la dificultad inherente de la obtención de la información necesaria para dicho propósito.

- Extender la aplicación del ejercicio a otros distritos de riego. Esto permitirá realizar un cruzamiento de la información generada por los modelos de varios distritos y descubrir factores que infieren de manera común en el rendimiento de los cultivos agrícolas.

7.3.3. Sobre la optimización del ingreso de la producción agrícola.

- Se propone como trabajo futuro una modificación a GenSM5 que puede acelerar la búsqueda de la mejor combinación de rendimientos. Esta consiste en dar preferencia en un principio a los rendimientos que se localizan al inicio de las listas, y profundizar en ellas a medida que transcurre el tiempo. Dicho de otra forma, la probabilidad de selección de rendimientos que se encuentran hacia el final de las listas aumenta conforme al tiempo, siendo en un inicio 0.

- Probar con técnicas basadas en otras heurísticas. La selección del algoritmo de optimización de esta tesis fue muy natural. La manera de organizar los rendimientos en una lista dio pie a la creación de un arreglo de posiciones, al cual se le encontró analogía con un cromosoma. Sin embargo, nada asegura que el algoritmo desarrollado con técnicas evolutivas sea mejor que otro basado en una heurística diferente. Se reserva como trabajo futuro la prueba con otro tipo de heurística (por ejemplo, recocido simulado, búsqueda tabú, colonias de hormigas, etc.) y comparar los resultados con el algoritmo desarrollado en esta tesis.

- Realizar una especificación genérica de GenSM5 para que pueda ser utilizado en otras áreas de aplicación. En esta tesis, el algoritmo se describe en el contexto de la producción agrícola, sin embargo, el algoritmo puede ser aplicado con cualquier otro tipo de problema, por lo que se requiere una especificación general del mismo.

8. Referencias

[1] Playán, E.; Mateos, L. *Modernization and optimization of irrigation systems to increase water productivity*. Proceedings of the 4th International Crop Science Congress, 26 Sep – 1 Oct 2004, Brisbane, Australia. 2004.

[2] CONAGUA. *Información de distritos de riego, programa de infraestructura hidroagrícola* [en línea]. México, D. F.. 2006. <http://www.cna.gob.mx/eCNA/Espaniol/Organismos/Central/Publicaciones/DistritoRiego_CNA.htm>.

[3] SEMARNAT. *Estadísticas de agricultura y ganadería* [en línea]. México, D. F. . 2006. <http://www.semarnat.gob.mx/estadisticas_ambientales/compendio/03actividades_humanas/agricultura.shtml.>

[4] CONAGUA. *Información de distritos de riego, programa de infraestructura hidroagrícola* [en línea]. México, D. F.. 2006. <http://www.cna.gob.mx/eCNA/Espaniol/Organismos/Central/Publicaciones/DistritoRiego_CNA.htm>.

[5] Palacios, E. ; García, A. E. *Introducción a la teoría de la operación de distritos y sistemas de riego*. Texcoco: Colegio de Posgraduados, Centro de Hidrociencias, 1986.

[6] SIAP. *Metodologías para la integración y análisis de indicadores y modelos del sector agropecuario* [en línea]. México D. F.: SEMARNAT, 2003. < www.siap.sagarpa.gob.mx/modelos/metodologias/**MetIndMod03**.pdf >

[7] Fayyad, U. M.; Piatetsky-Shaphiro, G.; Smith, P. *From Data Mining to Knowledge Discovery in Databases*. AAAI, Otoño 1996, p-37, 1996.

[8] CIENTEC. *Data Mining: Buscando un Tesoro Oculto Entre Sus Datos* [en línea]. Santiago de Chile. 2006. <http://www.cientec.com/Management/Management10.asp>.

[9] Drummond, S. T.; Sudduth, K. A.; Joshi, A.; Birrel, S. J,; Kitchen, N. R. *Statical and Neural Methods For Site-Specfied Yield Prediction*. Transactions of the ASAE, Vol. 46 (1). 2003.

[10] Mathews, R.; Blackmore S. *Using crop simulation models to determine optimum management practices in precision agriculture*. Precision Agriculture '97, 413–420. J. V. Stafford, ed. Oxford, U.K.: BIOS Scientific Publishers. 1997.

[11] Palacios J.E.; Sierrra A.; Bichir Y.M. *Estimación de la Superficie Sembrada, en el Distrito de Riego del Río Mayo, mediante Percepción Remota*. Artículo del XIII Congreso Nacional de Irrigación, ANEI 2005.

[12] Sifuentes E.; Ojeda W.; Gómez H. *Calibración de parámetros de cultivo en papa para la calendarización del riego en tiempo real bajo dos sistemas de riego en el distrito de riego 075, río fuerte, Sinaloa*. Artículo del IX Congreso Nacional de Irrigación, ANEI 1999.

[13] Sudduth, K. A.; Drummond, S. T.; Birrelli, S. J.; Kitchen N. R. *Analysis of Spatial Factors Influencing Crop Yield*. Proceedings of 3rd Int. Conf. On Precision Agriculture, pp. 129-140, 1996.

[14] FairIsaac. *A Discussion of Data Analysis, Prediction and Decision Techniques.* A Fair Issac White Paper, May 2003.

[15] Hache, C. *Site-specific Crop Response to Soil Variability in an Upland Field*. Master Thesis, University of Agriculture and Technology, Graduate School of Agriculture. Tokyo: Feb. 2003.

[16] Lobell, D.; Nicholas, K.; Field, C. *Weather-based yield forecasts developed for 12 California crops*. Californa Agriculture, Volume 60, No. 4. October-December 2006.

[17] Catherine, N. *Wheat Yield Prediction Modeling For Localized Optimization of Fertilizer and Herbicide Application*. A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, Montana State University. Bozeman, Montana: July 2004.

[18] Universidad Autónoma de Ciudad Juárez. *Función de producción para el algodón en el Valle de Juárez, caso San Agustín*. Universidad Autónoma de Ciudad Juárez, Programa de Licenciatura en Economía, 2004.

[19] Azofeifa, A. G.; Villanueva M. *Estimación de una función de producción: caso de Costa Rica*. Departamento de Investigaciones Económicas, Banco Central de Costa Rica, 1996.

[20] García, F. *Funciones de Producción y Programación Lineal*. Universidad de los Andes, Facultad de Ciencias Económicas y Sociales, Escuela de Administración y Contaduría Pública, Departamento de Empresas, Cátedra de Producción y Análisis de la Inversión. Mérida, Yucatán: Febrero de 2004.

[21] Cisneros E., O. X.; González E., J.: Cázares H. *El análisis retrospectivo como base de la planeación agrícola del módulo de riego 05 del Distrito de Riego 038, Río Mayo, Sonora*. Artículo del XIII Congreso Nacional de Irrigación, ANEI 2005.

[22] González C., A. *Aplicación del Benchmarking en los 16 módulos del Distrito de Riego 038, Río Mayo, Son*. Informe de Proyecto, IMTA-CNA, 2003.

[23] Red española de Minería de Datos y Aprendizaje. *Presentación* [en línea]. Fecha de consulta: 12 de Abril del 2006. <http://www.lsi.us.es/redmidas/>

[24] GAO, Z. *Decision-Making Support System for Irrigation Water management of Jingtai Chuan Pumping Irrigation Scheme at the Upper Reaches of Yellow River*. Watsave Workshop Paper Presented at 51st IEC, Cape Town, South Africa, 2000.

[25] Han, J.; Lamber, M. *Data Mining, Concepts and Techniques*. Morgan Kaufmann Publishers, 2001.

[26] Fayyad, U. M.; Piatetsky-Shaphiro, G.; Smith, P.; Uthurusamy, R. *Advances in Knowledge discovery and Data Mining*. AAAI Press, 1996.

[27] Hand, D.; Mannila, H.; Smyth, P. *Principles of Data Mining*. MIT Press, 2001.

[28] Hernández O., J.; Ramírez, M.J.; Ferri, C. *Introducción a la Minería de Datos*. Pearson, 2004.

[29] Cheeseman, P. *On Finding the Most Probable Model*. Chapt. 3 in Computational Models of Scientific Discovery and Theory Formation, pp. 73-95, Morgan Kaufmann, 1990.

[30] Fernández, E.J. *Asistente para la Gestión de Documentos de Proyectos de Explotación de Datos*. ITBA, Tesis de magíster, 2006.

[31] Montequín R., M. T.; Álvarez C., J. V.; Fernández M., J. M.; González V., A. *Metodologías Para la Realización de Proyectos de Data Mining*.

[32] SAS *Sitio oficial* [en línea]. Fecha de consulta: 12 de abril de 2006. <http://www.sas.com/>

[33] Chapman, P.; Clinton, J.; Kerber, R.; Khabaza, T.; Reinartz, T.; Shearer, C.; Wirth R. *CRISP-DM 1.0 Step-by-step data mining guide*. SPSS. 2000.

[34] Rodríguez Montequín, M. T.; Álvarez Cabal, J. V.; Mesa Fernández, J. M.; González Valdés, A. *Metodologías Para la Realización de Proyectos de Data Mining*. Actas del VII Congreso Internacional de Ingeniería de Proyectos. Pamplona, España, Octubre 2003.

[35] KDNuggets. *Data Mining Methodology (Apr 2004)* [en línea]. Polls. Fecha de consulta: 13 de Marzo de 2007. <http://www.kdnuggets.com/polls/2004/data_mining_methodology.htm>

[36] Acosta Aguilera, M. E. *Minería de Datos y Descubrimiento del Conocimiento.* Artículo del Congreso Internacional de Información, 2004.

[37] Witten, Ian H. ; Cunningham, Sally Jo; Holmes, Geoffrey; McQueen, Robert J. ; Smith, Lloyd A. *Practical machine learning and its potential application to problems in agriculture*. Proc New Zealand Computer Conference, volume 1, pages 308-325, Auckland, New Zealand, 1993.

[38] Drummond, S. T.; Joshi, A. *Predictive Ability Of Neural Networks For Site-Specific Yield Estimation*. Second International Geospatial Information in Agriculture and Forestry Conference, Lake Buena Vista, Florida, 10-12 January 2000.

[39] Liu, J.; Goering, C. E.; Tian, L. *A Neural Network for Setting Target Corn Yields*. 1999 ASAE Annual Meeting, Paper No. 99–3040. 1999.

[40] Thomasson, J. A. ; Shearer, S. A. *Water and crop-yield management improvement with data from remote and ground-level sensors*. Remote Sensing Technologies Center, Mississipi State University, 2001.

[41] Canteri, M.G.; Ávila, B.C.; Dos Santos, E.L.; Sanches, M.K. ; Kovalechyn, D. ; .Molin, J.P; Gimenez, L.M. *Application of Data Mining in Automatic Description of Yield Behavior in Agricultural Areas*. World Congress of Computers in Agriculture and Natural Resources , 2002.

[42] Crisler, M.T. ; Strickland, R.M. ; Ess, D.R.; Parsons, S.D. *Data Mining Methods for Use with Geo-Referenced Field Crop Data*. World Congress of Computers in Agriculture and Natural Resources, 2002.

[43] Abdullah, A. ; Brobst, S. ; Pervaiz, I. ; Umer, M.; Nizhar, A. *Agri Data Mining/Warehousing:Innovative Tools for Analysis of Integrated Agricultural & Meteorological Data*. DEA, 2004.

[44] Lobell, D. B. ; Ortiz-Monasterio, J. I. ; Asner, G. P. ; Naylor, R. L. ; Falcon, W. P. *Combining Field Surveys, Remote Sensing, and Regression Trees to Understand Yield Variations in an Irrigated Wheat Landscape*. Agronomy Journal, 2005.

[45] Wikipedia. *Definición del Proceso Unificado de Desarrollo (RUP)* [en línea]. Wikipedia la enciclopedia libre. Fecha de consulta: 12 de abril del 2006. <http://es.wikipedia.org/wiki/RUP>.

[46] Soto Mora, C. *La agricultura comercial de los distritos de riego en México y su impacto en el desarrollo agrícola*. Instituto de Geografía, UNAM, México, Boletín, núm. 11, pp. 45-182, 1981.

[47] Rodríguez Martínez, Andrés F.; Morales Manzanares, Eduardo. *Descubrir conocimiento en bases de datos: minería de datos y aplicaciones*. Boletín IIE, marzo-abril del 2000, pág.75-82. 2000.

[48] Witten, Ian H.; Frank, Eibe; Trigg, Len; Hall, Mark; Holmes, Geoffrey; Cunningham, Sally Jo. *Weka: Practical Machine Learning Tools and Techniques with Java Implementations*. H. Kasabov and K. Ko, eds., ICONIP/ANZIIS/ANNES'99 International Workshop, Dunedin, 1999.

[49] McQueen, Robert J.; Neal, Donna L; DeWar, Rhys; Garner, Stephen R.; Nevill-Manning, Craig G. *The WEKA machine learning workbench: Its application to a real world agricultural database*. Proc. from Canadian Machine Learning Workshop, Banff, Alberta, Canada, 1994.

[50] Wikipedia. *Paradox (base de datos)* [en línea]. Wikipedia la enciclopedia libre. Fecha de consulta: 13 de abril del 2007. < http://es.wikipedia.org/wiki/Paradox_(base_de_datos) >.

[51] Bolaños González, M.; Palacios Vélez, E.; Scott, C.; Exebio García, A. *Estimación del Volumen de Agua Usado en una Zona De Riego Mediante una Imagen de Satélite e Información Complementaria*. Publicado como ENSAYO en Agrociencia 35: 589-597. 2001.

[52] Quintas, Isabel. *ERIC II. Documentación de la base de datos climatológica y del programa extractor*. Documentación adjunta al sistema de cómputo. Manuales IMTA. Coordinación de Desarrollo Profesional e Institucional, Subcoordinación de Desarrollo Institucional, Mayo 2000.

[53] Tukey, John. *Exploratory Data Analysis*. Addison-Wesley. 1977.

[54] Dasu, T.; Johnson, T. *Exploratory Data Mining and Data Cleaning*. WILEY, 2003

[55] Kessler, Mathieu. *Apuntes de Estadística Industrial*. Departamento de Matemática Aplicada y Estadística, Universidad Politécnica de Cartagena. Versión Preliminar, 2003.

[56] Chambers, J.; Cleveland, W.; Kleiner, B.; Tukey, P. *Graphical Methods for Data Analysis*. Wadsworth, 1983.

[57] Kumar T., G.; Ballou, D. P. *Examining Data Quality*. Communications of the ACM, Feb. 1998.

[58] Strong, D. M.; Lee, Y. W.; Wang, R. Y. *Data Quality In Context*. Communications of the ACM, Mayo 1997.

[59] Cappiello, C.; Francalanci, C.; Pernici, B. *Data Quality Assessment from the user's perspective*. Communications of the ACM, 2004.

[60] Pipino, L.L.; Lee, Y. W.; Wang, R. Y. *Data Quality Assessment*. Communications of the ACM, Abril 2002.

[61] Maletic, J. I.; Marcus, A. *Data Cleasing, A prelude to Knowledge Discovery*. The Data Mining and Knowledge Discovery Handbook , Chapter 2. 2005.

[62] Kim, W.; Choi, B.; Hong, E.; Kim, S.; Lee, D. *A Taxonomy of Dirty Data,* Data Mining and Knowledge Discovery, 7, p81-99, 2003.

[63] Reinartz T. *A Unifying View on Instance Selection*. Data Mining and Knowledge Discovery, 6, 191–210. 2002.

[64] John, G. H.; Kohavi, R.; Pfleger, K. *Irrelevant Features and the Subset Selection Problem*. Proceedings of the Eleventh International Conference, 121-129, 1994.

[65] Grzymala-Buse, J. W.; Grzymala-Buse W. J. *Handling Missing Attribute Values*. The Data Mining and Knowledge Discovery Handbook , Chapter 3, 2005.

[66] Xiong, H.; Pandey, G.; Steinbach, M.; Kumar, V. *Enhancing Data Analysis with Noise Removal*. IEE Transactions on Knowledge and Data Engineering, Vol. 18, No. 3. March 2006.

[67] Hall, M. A. *Correlation-based Feature Selection for Machine Learning*. The University of Waikato, Thesis, April 1999.

[68] Quinlan, J. R. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, California. 1993.

[69] Barnett, V.; Lewis, T. *Outliers in Statistical Data*. John Wiley & Sons, 3th edition, 1994.

[70] Aggarwal, Charu C.; Yu, Philip S. *Outlier Detection for High Dimensional Data*. ACM SIGMOD 2001 May 21-24, Santa Barbara, California USA. 2001.

[71] Hodge, V. J.; Austin, J. *A Survey of Outlier Detection Methodologies*. Artificial Intelligence Rev. vol 22, 2004.

[72] Knorr, E. M.; Ng, R. T. *A Unified Notion of Outliers: Properties and Computation*. American Asociation for Artificial Intelligence, 1997.

[73] Knorr, E. M.; Ng, R. T. *Algorithms for Mining Distance-Based Outliers in Large Datasets.* Proceedings of the 24th VLDB Conference New York, USA, 1998.

[74] Breunig, M. M.; Kriegel, H.P.; Ng, R. T.; Sander, J. *LOF: Identifying Density-Based Local Outliers*. ACM, 2000.

[75] Portnoy, L.; Eskin, E.; Stolfo, S. *Intrusion Detection with Unlabeled Data Using Clustering*. ACM Workshop on Data Mining Applied to Security, 2001.

[76] McQueen, J. *Some methods for classification and analysis of multivariate observations*. Computer and Chemistry, 4, 257-272, 1967.

[77] Kaufman, L. Rousseeuw, P. *Finding Groups in Data: An Introduction to Cluster Analysis*. Ed. Wiley, 1990.

[78] Newcomb, S. R. *A Semantic Integration Methodology*. Proceedings of the Extreme Markup Languages Conferences, Montréal, Québec, 2003.

[79] Wang, Y. R.; Madnick, S. E. *The inter-database instance identification problem in integrating autonomous systems*. Proceedings of the Sixth International Conference on Data Engineering, 1989.

[80] Fellegi, I.; Sunter. A. *A Theory for Record Linkage*. American Statistical Association Journal, 1969.

[81] Hernández, M. A.; Stolfo, S. J. *Real-World Data is Dirty: Data Cleansing and The Merge/Purge Problem*. Data Mining and Knowledge Discovery, 2,9-37, 1998.

[82] Lee, M. L.; Hsu, Wynne. *Improving Data Quality Eliminating Dupes & I-D-ing Those Spurius Links*. IEE Potentials, APRIL/MAY 2005.

[83] Han, H.; Giles, L.; Zha, H.; Li, C.; Tsioutsioulikis, K. *Two Supervised Learning Approaches for Name Disambiguation in Autor Citations*. JCDL June 7-11, 2004.

[84] Lee, M. L.; Wang, T.; Low, W. L. *IntelliClean: A Knowledge-Based Intelligent Data Cleaner*. ACM, 2000.

[85] Otero, Fernando; Silva, Monique; Freitas, Alex; Nievola, Julio. *Genetic Programming for Attribute Construction in Data Mining*. Proc. Genetic and Evolutionary Computation Conf (GECCO-2002)*, page 1270,* New York, July 2002.

[86] Markovitch, Shaul.; Rosenstein, Dan. *Feature Generation Using General Constructor Functions*. Machine Learning, Vol. 49, Issue 1, Oct. 2002.

[87] Hirsh, H.; Japkowicz, N. *Bootstrapping training-data representations for inductive learning: A case study in molecular biology*. In Proc. 11th International Conference on Machine Learning, pp. 639-644. Morgan Kaufmann. 1994.

[88] Pagallo, G.; Haussler, D. *Boolean feature discovery in empirical learning*. In Proc. 7th International Conference on Machine Learning, pp. 71-99. Morgan Kaufmann. 1990.

[89] Matheus, C. J.; Rendell, L. A. *Constructive induction on decision trees*. In Proc. 11th International Conference on Artificial Intelligence., pp. 645-650. 1989.

[90] Aha, D. W. *Incremental constructive induction: An instance-based approach*. In Proc. 8th International Conference on Machine Learning, pp. 117-121. Morgan Kaufmann. 1991.

[91] Ragavan, H.; Rendell, L. A.; Shaw, M.; Tessmer, A. *Complex concept acquisition through directed search and feature caching*. In Proc. 13th International Conference on Artificial Intelligence., pp. 946-951. 1993.

[92] Hu, Y. J.; Kibler, D. *Generation of attributes for learning algorithms.* In Proc. 13th InternationalConference on Machine Learning. Morgan Kaufmann. 1996.

[93] Halevy, A.; Rajaraman, A.; J. Ordille. *Data Integration: The Teenage Years*, VLBD'06, Sept. 12-15, 2006.

[94] *Data Integration Manual*. Statistics New Zealand, August 2006.

[95] Weka. *Weka 3: Data Mining Software in Java* [en línea]. Sitio official de Weka en la Universidad de Waikato. Nueva Zelanda. Fecha de consulta: 13 de abril del 2007. < http://www.cs.waikato.ac.nz/~ml/weka/>

[96] Sun. *Java Database Connectivity* [en línea]. Fecha de consulta 13 de abril del 2007. < http://java.sun.com/javase/technologies/database/index.jsp>

[97] Gersten, Wendy; Wirth, Rüdiger; Arndt Dirk. *Predictive Modeling in Automotive Direct Marketing: Tools, Experiences and Open Issues*. KDD 2000, Boston, MA USA. 2000.

[98] Hair, J. F. Jr.; Anderson, R. E.; Tatham, R. L. *Multivariate Data Analysis*. Macmillan Publishing Company, Second Edition. 1987.

[99] Wasserman, Larry. *All of Statistics. A Concise Course in Statistical Inference*. Ed. Springer, 2004.

[100] Zhang, Peter G. *Neural Networks*. The Data Mining and Knowledge Discovery Handbook , Chapter 22 pp. 487. 2005.

[101] García, Jorge L.; Chagolla G., Hernando; Noriega, Salvador M. *Efectos de la Colinealidad en el Modelado de Regresión y su Solución*. CULCyT//Septiembre–Diciembre, 2006. Año 3, No. 16-17. Pág. 26-34. 2006.

[102] Hoerl, A. E.; Kennard, R. W. *Ridge regression: Biased Estimation for Nonorthogonal Problems*. Technometrics, 12: 55-67. 1970.

[103] Liu, H. *A New Class of Biased Estimate in Linear Regression*. Communications in Statistics: Theory and Methods, 22: 393-402. 1993.

[104] Kaciranlar, S. and Sakallioglu, S. *Combining the LIU Estimator and the Principal Component Regression Estimator*. Communications in Statistics: Theory and Methods, 22: 393-402. 2001.

[105] Piña, Manuel R.; Rodríguez Medina, Manuel; Díaz Núñez, Juan J. *Superioridad de la Regresión General Ridge Sobre Mínimos Cuadrados*. CULCyT//Enero-Febrero, 2005. Año 2, No 6. 2005.

[106] Romina, Laura B.. *Data Mining utilizando Redes Neuronales.* Tesis de grado. Facultad de Ingeniería de la Universidad de Buenos Aires. Mayo de 2005.

[107] Chen, T.; Chen, H. *Universal Approximation to Nonlinear Operators by Neural Networks with Arbitrary Activation Functions and its Application to Dynamical Systems*. Neural Networks 1995; 6:911-917.

[108] Cybenko, G. *Approximation by Superpositions of a Sigmoidal Function.* Mathematical Control Signals Systems. 2:303-314. 1989.

[109] Hornik, K.; Stinchcombe, M.; White, H. *Multilayer Feedforward Networks are Universal Approximators. Neural Networks*. 2:359-366. 1989.

[110] López Cruz, Misael. *Codiseño de una Arquitectura para Entrenamiento de Redes Neuronales Usando Retropropagación*. Tesis de Maestría, ITESM. Monterrey, México. 2005.

[111] Wong, B. K.; Bodnavich, T. A.; Selvi, Y. *Neural Networks Applications in Bussines; A Review and Analysis of the Literature (1988-1995).* Decision Support Systems 1997. 19:301-320. 1997.

[112] Santín González, Daniel. *Eficiencia Técnica y Redes Neuronales: Un Modelo para el Cálculo del Valor Añadido en Educación*. Memoria para Optar al Grado de Doctor. Universidad Complutense de Madrid. Facultad de Ciencias Económicas y Empresariales. Madrid. 2005.

[113] Rivas P., Pablo. *Reconocimiento de Rostros Mediante Perceptrones Multicapa*. Instituto Tecnológico de Nogales – Grupo de Inteligencia Artificial.

[114] Electrónica México. *Redes Neuronales Artificiales* [en línea]. Fecha de consulta: 19 de abril del 2007. <http://electronica.com.mx/neural/>

[115] Hopfield, J. J. *Neural Networks and Physical Systems with Emergent Collective Computacional Abilities*. Proc. Natl. Acad. Sci. U.S.A. Vol. 79, 1982. pp. 2554-2558. 1982.

[116] Hong-Liang, Lu. Xi-Jun, Qiu. *Some Exact Results of Hopfield Neural Networks and Applications*. arXiv:physics/9907042v1 [physics.bio-ph] 24 Jul 1999.

[117] Johannet, A.; Personnaz, L. G.; Dreyfus G.; Gascuel, J. D.; Weinfeld M. *Specification and Implementation of a Digital Hopfield-Type Associative Memory With On-Chip Training*. IEEE Transactions on Neural Networks 3, 529. 1992.

[118] Hebb, D. O. *The Organization of Behavior*. New York: Wiley, 1949.

[119] Kohonen, Teuvo. *The Self-Organizing Map*. Proceedings of the IEE, Vol. 78, No. 9, September 1990.

[120] Heiss-Czedik, Dorotea; Bajla, Ivan. *Using Self-Organizing Maps for object classification in Epo image analysis*. Measurement Science Review, Volume 5, Section 2, 2005.

[121] Deboeck G.; Kohonen, T. *Visual Explorations in Finance with Self-organizing Maps*. London: Springer-Verlag, 1998.

[122] Van Laerhoven, K.; Cakmakci, O. *What Shall we Teach Our Pants?*. Proceedings of the fourth International Symposium on Wearable Computers (ISWC 2000), Atlanta (GA), IEEE Press, pp.77-83, 2000.

[123] French, R. M. *Catastrophic Forgetting in Connectionist Networks. Trends in Cognitive Sciences. 3*(4), 1999, pp. 128-135. 1999.

[124] Shearer, S.A.; Burks; T.F. ; Thomasson, J.A.; Mueller, T.G.; Fulton, J.P.; Higgins, S.F; Samson, S. *Yield Prediction Using a Neural Network Classifier Trained Using Soil Landscape Features and Soil Fertility Data*. ASAE Paper No. 993041. Annual International Meeting, Sherton Centre, Toronto, Canada, July 18-21. 1999.

[125] Back, B.; Oosterom, G.; Sere, K.; van Wezel, M. *A Comparative Study of Neural Networks in Bankruptcy Prediction*. Proceedings of the 10th Conference on Artificial Intelligence Research in Finland, Turku, Finland, p. 140-148. 1994.

[126] Lee, Kidong; Booth, David; Alam, Pervaiz. *A Comparative Study of Backpropagation and Kohonen Self-Organizing Feature Map In Bankruptcy Prediction.* Decisión Sciences Institute. 2002.

[127] Pérez, Miguel A. *Predicción Meteorológica basada en Redes Neuronales Artificiales*. Universidad de Las Palmas de Gran Canaria. Ingeniería en Informática. Noviembre de 2003.

[128] Romero M., Cristóbal. *Aplicación de Técnicas de Adquisición de Conocimiento para la Mejora de Cursos Hipermedia Adaptativos Basados en Web*. Tesis de

Doctorado. Universidad de Granada, Departamento de Ciencias de la Computación e Inteligencia Artificial. Granada, España Julio 2003.

[129] Hu, Hong; Jiuyong, Li. *Using Association Rules to Make Rule-based Classifiers Robust*. 16th Australasian Database Conference, University of Newcastle, Australia. 2005.

[130] Michalski, R.; Mozetic, I.; Hong, J.; Lavrac, N. *The AQ15 inductive learning system: an overview and experiments*. Proceedings of IMAL 1986, Université de Paris-Sud, Orsay. 1986.

[131] Clark, P.; Boswell, R. *Rule Induction With CN2: Some Recent Improvements. In Machine Learning* - EWSL-91, pp. 151-163. 1991.

[132] Yin, X.; Han, J. *CPAR: Classification Based on Predictive Association Rules*. In Proceedings of 2003 SIAM International Conference on Data Mining. 2003.

[133] Quinlan, J. R. *Discovering rules from large collections of examples: a case study.* Expert Systems in the Micro-electronic Age, pp. 168--201. Edinburgh University Press, Edinburgh, 1979.

[134] Agrawal, R.; Srikant, R. *Fast Algorithms for Mining Association Rules in Large Databases*. Proceedings of the Twentieth International Conference on Very Large Databases. Santiago, Chile, pp. 487-499. 1994.

[135] Han, J.; Pei, J.; Yin, Y. *Mining Frequent Patterns Without Candidate Generation*. Proc. 2000 ACM-SIGMOD Int. Conf. on Management of Data. pp. 1-12. May, 2000.

[136] Bayardo, R.; Agrawal, R. *Mining the Most Interesting Rules*. Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM Press, N.Y., pp. 145-154. 1999.

[137] Li, W.; Han, J.; Pei, J. *CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules*. Proceedings 2001 IEEE International Conference on Data Mining. IEEE Computer Society Press, pp. 369-376.

[138] Wikipedia. *Algoritmo Id3 de árboles de decisión* [en línea]. Fecha de consulta: 10 de octubre del 2006. <http://es.wikipedia.org/wiki/Algoritmo_id3>.

[139] Kirkby, R. Frank, E. *Weka, Documentación Javadoc*. Versión 3.4.

[140] Polumetla, Aditya. *Machine Learning Methods for the Detection of RWIS Sensor Malfunctions*. Tesis. Faculty of the Graduate School of University of Minnesota. July 2006.

[141] Cohen, William W. *Fast effective rule induction*. Machine Learning: Proceedings of the Twelfth International Conference (ML95), 1995.

[142] Martín, B. *INSTANCE-BASED LEARNING: Nearest Neighbor with Generalization*. Tesis de Grado, Universidad de Waikato, 1995.

[143] Buddhinath, Gaya; Derry, Damien. *A Simple Enhancement to One Rule Classification* [En línea]. Reporte. Fecha de consulta: 22 de Abril del 2007. < http://goanna.cs.rmit.edu.au/~gjayatil/OtherLinks/Extra.php>

[144] Liu, Ying; Kandula, Sasikiran. *A Comparative Study Of Classification Algorithms* [en línea]. Reporte. Fecha de consulta: 22 de Abril del 2006. < http://www.utdallas.edu/~sxk049000/Papers_docs.htm>

[145] Molina, L. C.; Béjar, J. *Integración de Reglas de Asociación y de Clasificación*. Reporte técnico, Universidad de Cataluña, 1999.

[146] Frank, E.; Witten, Ian H. *Gennerate Accurate Rule Sets Without Global Optimization*. Proc. 15th International Conf. on Machine Learning. 1998.

[147] Servente, Magdalenta. *Algoritmos TDIDT Aplicados a la Minería de Datos Inteligente*. Tesis de Grado en Ingeniería Informática. Facultad de Ingeniería, Buenos Aires, Argentina. Febrero 2002.

[148] Quinlan, J. R. *Improved Use of Continuous Attributes in C4.5*. Journal of Artificial Intelligence Research 4 (1996), 77-90. 1996.

[149] Shannon, C. E. *A Mathematical Theory of Communication*. The Bell System Technical Journal, Vol. 27, pp. 379–423, 623–656, July, October, 1948.

[150] Loh, Wei-Yin. *Regression Trees With Unbiased Variable Selection and Interaction Detection*. Statistica Sinica. Vol. 12, pp. 361-386. 2002.

[151] Urbanek, Simon. *Many Faces of a Tree*. Proceedings in Interface 2003: Security and Infrastructure Protection Salt Lake City, Utah. March 12-15, 2003.

[152] Morgan, J. N.; Sonquist, J. A. *Problems in the analysis of survey data, and a proposal*. Journal of the American Statistical Association, 58:415-434, 1963.

[153] Fielding, A. Binary segmentation: *The Automatic Detector and Related Techniques for Exploring Data Structure*. C. A. O'Muircheartaigh and C. Payne, Editors, The Analysis of Survey Data, Volume I, Exploring Data Structures. Wiley, New York, 1977.

[154] Doyle, P. *The use of Automatic Interaction Detector and Similar Search Procedures*. Operational Research Quarterly, 24:465-467, 1973.

[155] Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone. C. J. *Classification and Regression Trees*. Wadsworth, Belmont, 1984.

[156] Chipman, H.; George, E.; McCulloch, R. *Bayesian CART Model Search (With Discussion)*. Journal of the American Statistical Association, 93(443):935–960, Sept. 1998.

[157] Denison, D. G.; Mallick, B. K.; Smith, A. F. M. *A Bayesian CART Algorithm*. Biometrika, 85:363-377, 1998.

[158] Ripley, B. D. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.

[159] Kass, G. V. Significance Testing in Automatic Interaction Detection (A.I.D.). Applied Statistics, 24:178-189, 1975.

[160] Hawkins, D. M. FIRM: Formal Inference-Based Recursive Modeling, PC version, Release 2.1. Technical Report 546, School of Statistics, University of Minnesota, 1997.

[161] Bonferroni, C. E. *Il Calcolo delle Assicurazioni su Gruppi di Teste*. In Studi in Onore del Professore Salvatore Ortu Carboni. Rome: Italy, pp. 13-60, 1935.

[162] Landwehr, Niels; Hall, Mark; Frank, Eibe. *Logistic Model Trees*. Proceedings of the 14th European Conference on Machine Learning. 2003.

[163] Quinlan, J. R. *Learning with Continuous Classes*. In: Proc 5th Australian Joint Conference on Artificial Intelligence. pp. 343–348, World Scientific Publishing Company Incorporated. 1992

[164] Solomatine , Dimitri P. ; Baskara L. A. Siek, Michael. *Flexible and Optimal M5 Model Trees with Applications to Flow Predictions*. 6th International Conference on Hydroinformatics - Liong, Phoon & Babovic (eds). 2004.

[165] Wang, Y.; Witten, I. H. *Induction of model trees for predicting continuous classes*. Proceedings of the poster papers of the European Conference on Machine Learning. University of Economics, Faculty of Informatics and Statistics, Prague.1997.

[166] Gama, J. *Functional Trees*. Machine Learning, 55, 219–250, 2004.

[167] Karalic, A. *Employing linear regression in regression tree leaves*. B. Neumann (Ed.), European conference on artificial intelligence (pp. 440–441). John Wiley & Sons. 1992.

[168] Ciampi, A.; Hogg, S. A.; McKinney, S.; J. Thiffault. *RECPAM: A Computer Program for Recursive Partition and Amalgamation for Censored Survival Data and Other Situations Frequently Occurring in Biostatistics, I: Methods and Program Features*. Computer Methods and Programs in Biomedicine, 26:239-256, 1988.

[169] A. Ciampi, Z. Lou, Q. Lin, and A. Negassa. *Recursive Partition and Amalgamation with the Exponential Family: Theory and Applications*. Applied Stochastic Models and Data Analysis, 7:121-137, 1991.

[170] Chaudhuri, P. Huang, M.-C.; Loh, W.-Y.; Yao. R. *Piecewise-polynomial Regression Trees*. Statistica Sinica, 4:143{167, 1994.

[171] R. J. Marshall. A Program to Implement a Search Method or Identification of Clinical Subgroups. Statistics in Medicine, 14:2645-2659, 1995.

[172] Li, K.-C. Lue, H.-H.; Chen. C.-H. *Interactive Tree-Structured Regression Via Principal Hessian Directions*. Journal of the American Statistical Association, 95:547-560, 2000.

[173] Roel, Alvaro; Plant, Richard E. *Factors Underlying Yield Variability in Two California Rice Fields* . Agron. J. 96:1481–1494 (2004).

[174] Paul, P. A.; Munkvold, G. P. *A Model-Based Approach to Preplanting Risk Assessment for Gray Leaf Spot of Maize*. Phytopathology 94:1350-1357. Publication no. P-2004-1011-4R. 2004.

[175] Arumugam, M; Scott D., Stephen. *EMPRR: A High-Dimensional EM-Based Piecewise Regression Algorithm*. Proceedings of the 2004 International Conference on Machine Learning and Applications (ICMLA '04). pp. 264-271. 2004.

[176] Vens, C.; Blockeel, H. *A simple regression based heuristic for learning model trees*. Intelligent Data Analysis 10 (3), pp. 215-236, 2006.

[177] Rosset , Saharon; Perlich, Claudia; Zadrozny, Bianca. *Ranking-Based Evaluation of Regression Models*. Proceedings of the 5th IEEE International Conference on Data Mining. Houston, Texas, USA. 27-30 November 2005.

[178] Liu, B.; Ma, Y.; Lee, R. *Analyzing the Interestingness of Association Rules from the Temporal Dimension*. IEEE International Conference on Data Mining (ICDM-2001), Silicon Valley, CA. Nov 29 - Dec 2, 2001.

[179] Geng, L.; Hamilton, H. J. *Interestingness Measures for Data Mining: A Survey*. ACM Computing Surveys, Vol. 38 No. 3, Article 9. Septembre 2006.

[180] Bayardo, Roberto J.; Agrawal, Rakesh. Mining the Most Interesting Rules. Proc. If the Fifth ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, 145-154. 1999.

[181] Liu, B.; Hsu, W; Chen, S. *Using General Impressions to Analyze Discovered Classification Rules*. Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining (KDD-97). Newport Beach, CA. 31–36. 1997.

[182] Liu, B.; Hsu, W.; Mun, L.; Lee, H. *Finding Interesting Patterns Using User Expectations*. IEEE Trans. Knowl. Data Eng. 11, 6, 817–832. 1999.

[183] Silberschatz, A; Tuzhilin, A. *On Subjective Measures Of Interestingness In Knowledge Discovery*. Proceedings Of The Ist International Conference On Knowledge Discovery And Data Mining (KDD-95). Montreal, Canada. 275–281. 1995.

[184] Silberschatz, A; Tuzhilin, *A. What Makes Patterns Interesting In Knowledge Discovery Systems*. IEEE Trans. Knowl. Data Eng. 8, 6, 970–974. 1996.

[185] Ling, C.; Chen, T.; Yang, Q.; Chen, J. 2002. *Mining Optimal Actions for Profitable CRM*. Proceedings Of the 2002 IEEE International Conference on Data Mining (ICDM '02). Maebashi City, Japan. 767–770.

[186] Wang, K.; Zhou, S.; Han, J. *Profit Mining: From Patterns to Actions*. Proceedings of the 8th Conference on Extending Database Technology (EDBT 2002). Prague, Czech Republic. 70–87. 2002.

[187] Stone, M. *Cross-validatory choice and assesment of statistical predictions*. J. Roy. Statist. Soc. 36. 111-147. 1974.

[188] Stone, M. *Cross-validation: A review*. Mathematische Operations forschung Statischen, Serie Statistics, 9. 127-139. 1978.

[189] Herrera, Francisco; Hervás, Cesar; Otero, José; Sánchez, Luciano. *Un estudio empírico preliminar sobre los tests estadísticos más habituales en el aprendizaje automático.* R. Giráldez, J.C Riquelme, J.S.Aguilar (Eds.) Tendencias de la Minería de Datos en España, Red Española de Minería de Datos y Aprendizaje (TIC2002-11124-E), 2004, 403-412. 2004.

[190] Y. Wang and I. Witten. *Inducing model trees for continuous classes*. Proceedins of Poster Papers, European Conf. on Machine Learning, 1997.

[191] Sundberg, Rolf. *Collinearity*. Enciclopedia of Environmetrics, Vol. I, pp 365-366. Wiley & Sons, Ltd, Chichester 2002.

[192] Fiszelew, A.; Britos, P.;Perichisky, G.; García-Martínez, R. *Automatic Generation of Neural Networks based on Genetic Algorithms*. Information Systems Electronic Review. Volumen 2 N° 1. 2003.

[193] Yao, Xing; Liu, Yong. Towards *Designing Artificial Neural Networks by Evolution*. Applied Mathematics and Computation Vol. 91-1, pp. 83-90. 1998.

[194] Dow, R. J.; Sietsma, J. *Creating Artificial Neural Networks that Generalize*. Neural Networks, vol. 4, no. 1, pp. 198-209. 1991.

[195] Sexton, Randall S.; Sikander, N. A. *Data Mining Using a Genetic Algorithm Trained Neural Network*. International Journal of Intelligent Systems in Accounting, Finance, & Management, 10, 201-210. 2001.

[196] Dougherty, James; Kohavi, Ron; Sahami, Mehran. *Supervised and Usupervised Discretization of Continuous Features*. Machine Learning: Proceedings of the Twelfth International Conference, 1995, Morgan Kaufmann Publishers, San Francisco CA. 1995.

[197] Catlett, J. *Megainduction: machine learning on very large databases*. PhD thesis, University of Sydney. 1991.

[198] Liu, Huan; Hussain, Farhad; Lim Tan, Chew; Dash, Manoranjan. *Discretization: An Enabling Technique*. Data Mining and Knowledge Discovery, 6, 393–423, 2002.

[199] Ismail K., Michael. *An Empirical Investigation of the Impact of Discretization on Common Data Distributions*. Thesis. A dissertation submitted in partial fulfillment of the requirements for the degree of Master of Technology. Department of Computer Science, Melbourne, AUSTRALIA. 2003.

[200] Schichl, Hermann. *Mathematical Modeling and Global Optimization*. Draft of a Book, submitted to Cambridge University Press, November 2003.

[201] Bhatti, M. Asghar. *Practical Optimization Methods, With Mathematica Applications*. Springer-Verlag New York, Inc. 2000.

[202] Linares, Pedro; Ramos, Andrés; Sánchez, Pedro; Sarabia, Ángel; Vitoriano, Begoña. *Modelos Matemáticos de Optimización*. Universidad Pontificia de Madrid. Escuela Técnica Superior de Ingeniería. Departamento de Organización Industrial. Octubre 2001.

[203] G. Dantzig. *Linear Programming and Extensions*. Princeton University Press, Princeton, NJ, 1967.

[204] Holland, H. John. *Adaptation in Natural and Artificial Systems*. The University of Michigan, 1975.

[205] Marczyk, Adam. Genetic Algorithms and Evolutionary Computation [en línea]. Posted: April 23, 2004. Revisado: 25 de Mayo de 2007. < http://www.talkorigins.org/faqs/genalg/genalg.html>

[206] Goldberg, E. David. *Genetic Algorithms in Search, Optimization, and Machine Learning*.Addisson-Wesley Publishing Company, Inc. 1989.

[207] Gómez Ramos, José Luis. *Algortimos Genéticos con Diversidad Forzada para la Resolución del Problema del Timetabling Educativo*. Tesis de Maestría en Ciencias de la Computación.Tecnológico de Monterrey, Cuernavaca Morelos. Agosto 2005.

[208] Larrañaga, Pedro; Inza, Iñaki; Moujahid, Abdelmalik. *Métodos Matemáticos en Ciencias de la Computación.* Apuntes del Curso. Tema 2. Algoritmos Genéticos. Universidad del País Vasco, Departamento de Ciencias de la Computación. 2005.

[209] Kalashnikov, Dmitri V.; Mehrotra, Sharad. *Domain-Independent Data Cleaning via Analysis of Entity-Relationship Graph*. ACM Transactions on Database Systems, Vol. 31, No. 2, Pages 716–767, June 2006.

[210] Rumelhart, D.E.; Hinton G.E.; Williams, R.J. *Learning Internal Representations by Error Propagation*. In David E. Rumelhart and James A. McClelland, volume 1. The MIT Press, 1986.

Anexo 1. Comprensión del negocio

Tabla A1.1. Tipos de problemas y técnicas de solución en minería de datos según la guía CRISP-DM **[33]**.

Tipo de problema	Descripción	Técnicas de solución recomendadas
Descripción y agrupación[17] de datos	Describe de manera concisa las características de los datos, típicamente en forma elemental y agregada. Esto proporciona al usuario una visión global de la estructura de los datos. En ocasiones, el resolver este problema puede ser el único propósito del proyecto de minería.	▪ Medidas estadísticas (media, desviación estándar, mediana, etc.). ▪ Agregados de datos. ▪ Análisis de correlación ▪ Visualización.
Segmentación	La segmentación aborda el problema de separar los datos en subgrupos o clases interesantes y significantes. Todos los miembros de un subgrupo comparten características comunes.	▪ Técnicas de agrupamiento (clustering). ▪ Redes neuronales ▪ Visualización
Descripción conceptual	Es una descripción entendible de conceptos o clases. El propósito no es desarrollar modelos completos con una alta eficiencia de predicción, si no ganar información.	▪ Métodos de inducción de reglas. ▪ Agrupamiento conceptual.
Clasificación	La clasificación asume que existe un conjunto de objetos – caracterizados por algunos atributos o propiedades – los cuales pertenecen a clases diferentes. La etiqueta de clase es un valor discreto (simbólico) y es conocido para cada objeto. El objetivo es construir modelos de clasificación (llamados algunas veces clasificadores), los cuales asignan la etiqueta correcta de clase a objetos sin etiquetar y no vistos previamente.	▪ Análisis discriminante. ▪ Métodos de inducción de reglas. ▪ Aprendizaje por árboles de decisión. ▪ Redes neuronales. ▪ Vecino más cercano (K Nearest Neighbor). ▪ Razonamiento basado en casos. ▪ Algoritmos genéticos.
Predicción	La predicción es muy similar a la clasificación. La única diferencia es que el atributo objetivo no es un valor cualitativo discreto, sino un valor continuo. El propósito de la predicción es encontrar el valor numérico del atributo objetivo para objetos no antes observados. Si la predicción trata con datos de series de tiempo entonces usualmente se le llama pronóstico.	▪ Análisis de regresión. ▪ Árboles de regresión. ▪ Redes Neuronales. ▪ Vecino más cercano. ▪ Métodos de Box-Jenkins. ▪ Algoritmos genéticos.
Análisis de dependencias	Consiste en encontrar un modelo que describa dependencias significativas (o asociaciones) entre elementos de datos o eventos. Las dependencias pueden ser utilizadas para predecir el valor de un dato dada la información de otros datos. A pesar de que las dependencias pueden ser utilizada para modelado predictivo, son más comúnmente utilizadas para entendimiento. Las dependencias pueden ser estrictas o probabilísticas.	▪ Análisis de correlación. ▪ Análisis de regresión. ▪ Reglas de asociación. ▪ Redes Bayesianas. ▪ Programación lógica inductiva. ▪ Técnicas de visualización.

[17] La palabra original en inglés es "summarization", que no tiene una traducción literal en español, pero que hace referencia a la obtención de totales representativos de los datos.

Anexo 2. Descripción de los datos

Tabla A2.1. Descripción estructural de la serie de datos #1: Estadísticas agrícolas SINHDR

Atributos	Identificador	Formato	Uni-dades	Rango de valores	Comentarios
Distrito	Dist	Alfanumérico (3)		001..092	Clave que identifica a un distrito de riego, 84 claves distintas.
Módulo	Mod	Alfanumérico (3)			Clave que identifica a un módulo de riego[18].
Ciclo	Ciclo	Alfanumérico (12)			Identificadores compuestos, 16 tipos distintos.
Cultivo	Cultivo	Alfanumérico (20)			Nombres de cultivos, 228 valores distintos.
Superficie sembrada	SuperficieSembrada	Numérico	ha	0.00..123,630.00	
Superficie cosechada	SuperficieCosechada	Numérico	ha	0.00..714,096.00	
Lámina de riego	LaminaRiego	Numérico	cm	0.00..78.016	
Costo de producción	CostoProduccion	Numérico	$/ha	0.00..877,750.00	
Rendimiento	Rendimiento	Numérico	ton/ha	0.00..43,914.00	
Precio	Precio	Numérico	$/ton	0.00..25,664.00	
Industria	Industria	Alfanumérico (20)			Identificador de la industria, 385 cadenas distintas en total. No se considera de interés para el proyecto.
Destino	Destino	Alfanumérico (12)			Identificador del destino, 37 cadenas distintas en total. No se considera de interés para el proyecto.

Número de registros: 20,976

Tabla A2.2. Descripción estructural de la serie de datos #2: Estadísticas de producción CONAGUA.

Atributos	Identificador	Formato	Unidades	Rango de valores	Comentarios
Distrito de riego	NA	Alfanumérico			Clave+Nombre del distrito.
Ciclo	NA	Alfanumérico			Cadena del ciclo ("otoño-invierno", "primavera-verano", etc.)
Cultivo	NA	Numérico			
Superficie sembrada	NA	Numérico	ha	NC*	
Superficie cosechada	NA	Numérico	ha	NC*	
Rendimiento	NA	Numérico	ton/ha	NC*	
Producción	NA	Numérico	ton	NC*	
Precio medio rural	NA	Numérico	$/ha	NC*	
Valor de la cosecha	NA	Numérico	$	NC*	

* No calculado.

[18] Un módulo de riego es una asociación de varios usuarios de agua (productores agrícolas), que se constituye así como requisito para acceder al volumen concesionado por la Comisión Nacional del Agua.

Tabla A2.3. Descripción estructural de la serie de datos #3: Información Climatológica ERIC III.
Número de registros: 1,500,000.

Atributos	Identificador	Formato**	Unidades	Rango de valores	Comentarios
Año	D*	Entero		1960-2003	Año de la observación. 43 años en total.
Estación	D*	Alfanumérica			Identificador de la estación meteorológica***. 6063 estaciones en total.
Día	D*	Entero			Los datos pueden observarse por día o por mes.
Temperatura observada	D*	Numérico	ºC		
Temperatura mínima	D*	Numérico	ºC		
Temperatura máxima	D*	Numérico	ºC		
Precipitación	D*	Numérico	mm		
Evaporación	D*	Numérico	mm		
Tormenta	D*	Entero			
Granizo	D*	Entero			
Niebla	D*	Entero			
Cobertura del cielo	D*	Entero			

* Desconocido, es un identificador interno.
** Inferido a través de una muestra obtenida con la herramienta del sistema para exportar en formato de texto.
*** El proceso de consulta proporciona otros datos asociados al identificador de la estación, como el nombre, y el estado, la longitud, latitud y altitud en donde se encuentra localizada.

Tabla A2.4. Descripción estructural de la serie de datos #4: Hojas de datos del distrito de riego 038.
Período temporal: 2006.

Atributos	Identificador	Formato	Unidades	Rango de valores	Comentarios
Ciclo	NA	Alfanumérico (30)			Tres cadenas posibles: primavera-verano, otoño-invierno, perennes.
Cultivo	NA	Alfanumérico (30)			Nombres de cultivos, 57 valores distintos.
Superficie sembrada	NA	Numérico	ha	0.50..11,148.04	
Superficie cosechada	NA	Numérico	ha	0.50..11,148.04	
Lámina de riego	NA	Numérico	cm	8.2..119.66	
Costo de producción	NA	Numérico	$/ha	0.00..80.00	
Rendimiento	NA	Numérico	ton/ha	1.05..78	
Precio	NA	Numérico	$/ton	0.00..13.00	

Anexo 3. Sobre la calidad de datos.

Según Kumar (1998) [57], el concepto de "calidad de los datos" es mejor definido como "dependiendo del uso", lo cual implica que el concepto de calidad de los datos es relativo. La calidad presente en los datos puede ser suficiente para una tarea[19], pero no poseer suficiente calidad para otra. Esto significa que la utilidad y la usabilidad son aspectos importantes de la calidad de los datos (Strong, 1997) [58]. Esta utilidad y usabilidad se da en la medida que los datos satisfagan los requerimientos del usuario o bien, sean adecuados para un proceso específico (Cappiello, 2004) [59].

Una amplia variedad de autores [57][58] [59][60] señalan que la calidad de los datos es un concepto multidimencional, esto en referencia a que una gran variedad de medidas (o dimensiones) pueden ser aplicadas dependiendo del aspecto que se esté evaluando, como por ejemplo, la precisión, la relevancia, el valor agregado, la accesibilidad, etc.. Strong (1997) [58] agrupa dichas dimensiones en categorías (ver tabla A3.1), y define un problema de calidad en los datos como "*una dificultad encontrada a lo largo de una o más dimensiones de calidad que provoca que una gran parte o la totalidad los datos sean incapaces de utilizarse*". La tabla A3.1 muestra las dimensiones agrupadas en categorías según Strong et al. Otra referencia para las dimensiones puede encontrarse en Pipino (2002) [60].

Tabla A3.1. Categoría y dimensiones de calidad de datos por Strong (1997) **[58]**.

Categoría	Dimensiones
Calidad intrínseca	Precisión, Objetividad, Credibilidad, Reputación.
Calidad en la accesibilidad	Accesibilidad, Seguridad de acceso.
Calidad contextual	Relevancia, Valor agregado, Oportunidad, Completes, Cantidad de datos.
Calidad en la representación	Facilidad de interpretación[20], Facilidad de entendimiento, Representación concisa, Representación consistente.

La efectividad del ejercicio de minería de datos depende críticamente de la calidad de los datos [27]. El resolver problemas de calidad de datos requiere de información sumamente específica del dominio y dependiente del contexto (Dasu, 2003)[54].

Para resolver los problemas de calidad en los datos es necesario recurrir a otro proceso definido como "limpieza de datos". Existe una enorme confusión entre "limpieza de datos" y "calidad de datos", pero se sabe que, al menos en minería, la calidad de los datos es un proceso que se encarga de señalar el grado de integridad de los datos en términos de sus dimensiones (ver tabla 9), mientras que "limpieza de datos" es otro proceso que se encarga de maximizar esa integridad para así mejorar su calidad[21]. Este mismo es el enfoque de CRISP-DM, donde ambos procesos están contemplados en etapas previas al modelado.

[19] La palabra original era uso ("use"). Al autor de esta tesis le pareció más adecuado la palabra "tarea", ya que la palabra "uso" encierra una generalidad mayor en el idioma español que en el idioma inglés (además de que se había utilizado en el párrafo anterior). Otras alternativas podrían ser "proceso" u "actividad".
[20] La palabra en el idioma original es interpretability, que no tiene traducción literal en el idioma español.
[21] Una buena referencia para la comparación de ambos conceptos es Maletic (2005) [61].

Muchos son los errores que se pueden presentar entre el origen de un dato hasta su almacenamiento en un archivo digital. Ninguna medida está exenta de error, y los tipos de errores son infinitos, pasando desde la falta de cuidado humano y fallas en la instrumentación hasta una especificación inadecuada de lo que se está midiendo [27]. Un concepto altamente ligado con la calidad de los datos y limpieza de datos es el de "dato sucio". Para Kim et al (2001) [62], un "dato sucio" cae en alguno de los tipos de la taxonomía de la tabla A3.2.

Tabla A3.2. Taxonomía de los "datos sucios" por Kim et al (2001) [62]

1 Dato faltante
1.1 Dato faltante donde no hay una condición de NULO-NO-PERMITIDO
1.2 Dato faltante donde una condición NULO-NO-PERMITIDO debe ser forzosa.

2 Dato existente, pero
2.1 Dato erróneo debido a
2.1.1 No cumplimiento de condiciones forzosas automáticas de integridad
2.1.1.1 Condiciones de integridad soportadas en los sistemas de bases de datos relacionales actuales
2.1.1.1.1 Condiciones específicables por el usuario
2.1.1.1.1.1 Uso de un tipo de dato erróneo (violación del tipo de datos, incluyendo rango de valores)
2.1.1.1.1.2 Datos colgantes (violación de integridad referencial)
2.1.1.1.1.3 Datos duplicados (violación de condición de valores únicos no nulos)
2.1.1.1.1.4 Datos mutuamente inconsistentes (una acción no ejecutada en espera de que una condición se cumpla)
2.1.1.1.2 Integridad garantizada a través del manejo de transacciones
2.1.1.1.2.1 Pérdida de actualización (debido a falta de control de concurrencia)
2.1.1.1.2.2 Lectura sucia (debido a falta de control de concurrencia)
2.1.1.1.2.3 Lectura irrepetible (debido a falta de control de concurrencia)
2.1.1.1.2.4 Transacción perdida (debido a una falta de recuperación de choque adecuada)
2.1.1.2 Condiciones de integridad no soportadas en los sistemas de bases de datos relacionales actuales
2.1.1.2.1 Error de dato categórico (por ej. error del nivel de abstracción, dato fuera del rango de la categoría)
2.1.1.2.2 Dato temporal no actualizado (violando la restricción de rango temporal válido, por ej. la edad de una persona o su salario que no ha sido actualizado)
2.1.1.2.3 Datos espaciales inconsistentes (violan la restricción espacial, por ej. una forma incompleta)
2.1.2 Condiciones no forzosas de integridad
2.1.2.1 Error de entrada de datos involucrando un solo archivo o tabla
2.1.2.1.1 Error de entrada de datos que involucra a un solo campo
2.1.2.1.1.1 Entrada errónea (por ej. una edad mal tecleada, 26 en lugar de 25)
2.1.2.1.1.2 Falta de ortografía (por ej. efecto en lugar de afecto).
2.1.2.1.1.3 Datos extraños (por ej. título y nombre, en lugar de sólo el nombre)
2.1.2.1.2 Error de entrada de datos que involucra a más de un campo
2.1.2.1.2.1 Entrada en el campo equivocado (por ej. la dirección en el campo del nombre)
2.1.2.1.2.2 Dato erróneo de un campo derivado (debido a un error en la función que computa el dato en el campo derivado)
2.1.2.1 Inconsistencia a través de múltiples tablas/archivos (por ej., cuando el número de empleados en la tabla empleados y el número de empleados en la tabla departamento no son iguales)
2.2 No error en el dato, pero es un dato inservible
2.2.1 Un dato diferente para la misma entidad a lo largo de múltiples bases de datos (por ej. un salario diferente para la misma persona en dos tablas diferentes de diferentes bases de datos)
2.2.2 Ambigüedad de datos, debido a
2.2.2.1 Uso de abreviaciones (Dr. para doctor o dirección)
2.2.2.2 Contexto incompleto (por ej. Miami, de Ohio o Florida)
2.2.3 Conformación no estándar de los datos, debido a
2.2.3.1 Diferentes representaciones de un dato no compuesto
2.2.3.1.1 La transformación algorítmica no es posible
2.2.3.1.1.1 Abreviaciones (dep por departamento, carr por carretera)
2.2.3.1.1.2 Alias/apodo (nick name) (por ej. Mopac, Loop 1, y Highway 1; Bill Clinton, President Clinton, William Jefferson Clinton)
2.2.3.1.2 La transformación algorítmica es posible
2.2.3.1.2.1 Formatos de codificación (ASCII, EBCDIC, etc.)
2.2.3.1.2.2 Representaciones (incluyendo números negativos, fecha, hora, precisión, fracción)
2.2.3.1.2.3 Unidades de medición (incluyendo la fecha, la hora, distancia, peso, area, volumen, etc.)
2.2.3.2 Diferentes representaciones de un dato compuesto

2.2.3.2.1 Datos concatenados
- **2.2.3.2.1.1** Versión abreviada (por ej. John Kennedy por John Fitzgerald Kennedy)
- **2.2.3.2.1.2** Uso de caracteres especiales (espacio, no espacio, punto, paréntesis, en el número del seguro social o número de teléfono¨)
- **2.2.3.2.1.3** Orden diferente (John Kennedy vs Kennedy, John)

2.2.3.2.2 Datos jerárquicos (por ej. el concepto jerárquico de la dirección: estado-país-ciudad en lugar estado-ciudad).
- **2.2.3.2.2.1** Versión abreviada
- **2.2.3.2.2.2** Uso de caracteres especiales
- **2.2.3.2.2.3** Orden diferente (ciudad-estado en lugar de estado-ciudad)

Existen muchas técnicas para detectar errores en los datos. En Maletic et al (2005)[61] se proporciona una descripción de los métodos más comunes que pueden ser utilizados para la detección de errores:

a) Estadísticos: Identifica los valores anormales en campos y registros utilizando métricas como la media, la desviación estándar, el rango, basándose en el teorema de Chebyshev y considerando los intervalos de confianza para cada campo. Mientras este enfoque puede generar "falsos positivos", es simple y rápido, y puede ser utilizado en conjunción con otros métodos.

b) Agrupamiento (clustering): Identifica los registros anormales utilizando técnicas de conglomerados basados en la distancia Euclidiana (u otra). Algunos algoritmos de agrupamiento proveen soporte para identificar los valores anormales

c) Basados en patrones: Identifica los campos y registros anormales que no se adaptan a los patrones existentes en los datos. Combinan técnicas (particionamiento, clasificación y agrupamiento) que son utilizadas para identificar patrone sque aplican a la mayoría de los registros. Un patrón es identificado por un grupo de registros que tienen características o comportamientos similares para un p% de los campos en la serie de datos, donde p es un valor definido por el usuario (normalmente, arriba de 90).

d) Reglas de asociación: Las reglas de asociación con un alto valor de confianza y soporte definen un tipo de patrón diferente. Como antes, los registros que no siguen dichas reglas son considerados anormales. El poder de las reglas de asociación es que pueden trabajar con diferentes tipos de datos.

Como se ha visto en etapas previas, el enfoque utilizado en el proyecto para la detección de errores fue el estadístico (consultar sección 4.2.3 dedicada a la exploración de los datos).

Anexo 4. Sobre la limpieza de datos.

Muchas son las técnicas existentes para abordar los errores presentes en los datos. En esta sección de la tesis, se exponen las técnicas más comunes, agrupadas en las siguientes categorías:

a) Técnicas para datos faltantes.
b) Técnicas para datos atípicos (outliers).
c) Técnicas para eliminación de registros duplicados.

1. Técnicas para datos faltantes

Existen una variedad de razones por las que las series de datos son afectadas por atributos con valores faltante [65]. Algunos valores no son almacenados porque resultan irrelevantes, como por ejemplo, preguntar por los datos del cónyuge cuando el sujeto es soltero, o, evaluar la calidad de un aire acondicionado cuando un hogar no cuenta con este equipo. Otras razones pueden ser un olvido de captura, un borrado erróneo u omisiones voluntarias. Son los llamados datos "perdidos".

En Grzymala et al (2005) [65] también se clasifican los métodos para solucionar el problema de los datos faltantes. De manera general, éstos se clasifican en métodos secuenciales o métodos paralelos.

Métodos secuenciales para el error de datos faltantes

- **Eliminación de registros.** Este método se basa en ignorar los registros con atributos con valores faltantes. Todos los registros con valores faltantes son eliminados del conjunto de datos.
- **Uso del valor más común del atributo.** Es uno de los métodos más simples. Se basa en asignar al valor faltante de un atributo el valor más probable, donde la probabilidad está representada por frecuencias relativas de valores del atributo.
- **Asignación de todos los valores posibles a valores faltantes.** Cada registro con valores faltantes es reemplazado por un conjunto de registros, en el cual cada valor faltante es reemplazado por cada uno de todos los valores conocidos.
- **Asignación de la media aritmética.** Aplica sólo a atributos numéricos. Los valores faltantes se reemplazan con la media de los valores existentes. Otra alternativa es el uso de la mediana. La definición de media y mediana se había dado con anterioridad en la sección 4.2.3.1.
- **Asignación global del caso más cercano.** Se basa en la asignación al valor faltante de un valor conocido localizado en otro registro, el cual se asemeja lo más posible al registro con el valor faltante. En la búsqueda del caso más cercano se comparan dos vectores de valores de atributos, uno correspondiente al que posee el valor faltante y otro un posible candidato para el ajuste. La búsqueda es conducida a través de todos los registros (de ahí el nombre de global), el registro para l cual la distancia es la más pequeña es considerado el caso más cercano. La distancia entre los casos $\boldsymbol{x}$ y $\boldsymbol{y}$ es calculada como se muestra a continuación.

$$dist(x,y) = \sum_{i=1}^{n} dist(x_i, y_i) \qquad \textbf{(A3.1)}$$

donde

$$dist(x,y) = \begin{cases} 0 \text{ si } x_i = y_i, \\ 1 \text{ si } x \text{ y } y \text{ son símbolos y } x_i = y_i, \text{ o } x_i = ?, \text{ o } y_i = ? \\ \dfrac{|x_i - y_i|}{r} \text{ si } x_i \text{ y } y_i \text{ son numéricos y } x_i \neq y_i, \end{cases}$$

donde r es la diferencia entre el máximo y el mínimo de los valores conocidos del atributo numérico con el valor faltante. Si existen dos valores con la misma distancia se requiere un tipo de heurística para decidir, por ejemplo, seleccionar el primer registro (Grzymala, 2005) [65].

- **Regresión simple.** El objetivo del análisis de regresión es ayudar a predecir el valor de una variable dependiente por medio del conocimiento de una o más variables independientes. Cuando el problema involucra una sola variable dependiente que es predicha por una sola variable independiente, la técnica estadística es referida como regresión simple (Hair, 1987) [98]. En esta versión de regresión, el modelo toma la forma de la ecuación de la recta:

$$r(x) = B_o + B_1 x \qquad \textbf{(A3.2)}$$

donde:

r(x) es la función de regresión,
B₀ es el coeficiente que indica en donde la función r(x) se intercepta con el origen,
B₁ es la pendiente de la recta.

Los parámetros ***B₀*** y ***B₁*** son aquellos que minimizan la suma residual de cuadrados, definida como $RSS = \sum_{i=1}^{n} \varepsilon_i^2$, donde $\varepsilon_i = Y_i - r(x_i)$, y Y_i es el valor real. Para calcularlos, se emplean las siguientes ecuaciones (Wasserman, 2004) [99] :

$$B_1 = \frac{\sum_{i=1}^{n} (X_i - \overline{X}_n)(Y_i - \overline{Y_n})}{\sum_{i=1}^{n} (X_i - \overline{X}_n)^2} \qquad B_0 = \overline{Y}_n - B_1 \overline{X}_n \qquad \textbf{(A3.3) y (A3.4)}$$

- **Regresión múltiple.** A diferencia de las técnicas anteriores, la regresión lineal múltiple explota las interrelaciones existentes entre los atributos para inferir múltiples valores faltantes, en lugar de un solo valor [54].

La lista anterior no es absoluta ni exclusiva. En Grzymala et al (2005) [65] y en Dasu et al (2003) [54] se puede encontrar información referente a otros métodos para corregir el error de datos faltantes sin interactuar aún con la etapa de representación del conocimiento.

Métodos paralelos para el error de datos faltantes

Para el segundo grupo de técnicas, los valores faltantes se toman en cuenta en el proceso principal donde el conocimiento adquirido es representado [65]. Un ejemplo es la modificación del algoritmo para inducción de reglas LEM2 (Learning from Examples Module, version 2), en el cual las reglas inducidas de un conjunto de datos con atributos que presentan valores perdidos se consideran condiciones "sin cuidado" o valores perdidos. Otro ejemplo es el algoritmo C4.5 de Quinlan (1993) **[68]**, en el cual la división de los registros con valores faltantes se realiza en fracciones, agregando estas fracciones a nuevos subconjuntos de registros.

2. Técnicas para datos atípicos (outliers)

La definición de Barnett y Lewis (1994) [69] de un dato atípico (outlier) es que es una observación (o subconjunto de observaciones) que parece ser inconsistente con el conjunto restante de datos. Otra definición, indica que, además, la diferencia presentada debe estar basada en alguna medida [70]. Pese a que ambas definiciones pueden ser aplicadas a los datos nominales, las técnicas para detectar y corregir outliers están orientadas principalmente a datos numéricos.

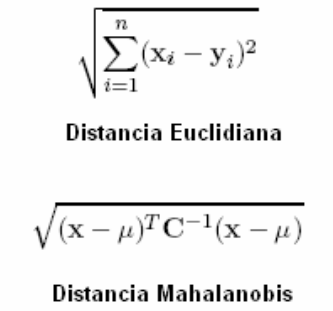
Muchas son las técnicas para detectar *outliers* [71], la mayoría de ellas provienen de la estadística [69]. En Xiong et al (2006) [66] se dividen los métodos para la detección de datos atípicos en tres tipos distintos:

2.1 Métodos de detección de *outliers* basados en distancia.

Son unos de los métodos más simples. Un objeto en el conjunto de datos $\boldsymbol{D}$ es un dato atípico basado en la distancia si al menos una fracción $\boldsymbol{\alpha}$ de los objetos en $\boldsymbol{D}$ está a una distancia más grande que $\boldsymbol{r}$. Esta definición es simple, pero puede caer en problemas cuando un mismo conjunto de datos se divide en regiones con distintas densidades de datos, ya que el criterio de los parámetros globales $\boldsymbol{r}$ y $\boldsymbol{\alpha}$ podría resultar insuficiente para todas las regiones [66]. Algunos ejemplos de algoritmos para la detección de outliers basados en distancia son:

- **La desviación estándar.** Para una distribución normal, los outliers pueden ser considerados como las observaciones que están 3 o más desviaciones estándar arriba de la media (i.e., $> 3\sigma$). En Knorr (1997) [72] se aborda con más detalle este tipo de método para datos atípicos.
- **Algoritmo del vecino más cercano (K-nearest Neighbour, K-NN).** Utiliza una métrica de distancia, como la distancia Euclidiana o la distancia de Mahalanobis (ver figura A4.1) para detectar los vecinos más cercanos de un registro.

Figura A4.1. Métricas de distancia (tomada de Hodge et al (2006) [71]).



$$\sqrt{\sum_{i=1}^{n}(\mathrm{x}_i - \mathrm{y}_i)^2}$$

Distancia Euclidiana

$$\sqrt{(\mathrm{x}-\mu)^T \mathrm{C}^{-1}(\mathrm{x}-\mu)}$$

Distancia Mahalanobis

- **Algoritmo de ciclo anidado (Nested-Loop, NL).** Utiliza un diseño orientado en bloques y ciclos anidados. El algoritmo NL asume un tamaño de buffer de B% del conjunto total de datos, divide el buffer en dos partes (llamados primer y segundo arreglo). Vacía los datos en los arreglos y directamente calcula la distancia entre cada par de registros. Para cada registro en el primer arreglo, se mantiene una cuenta individual de sus D-vecinos. Se detiene el conteo para un registro en particular siempre que el número de D-vecinos exceda una constante M. Esta definición y una descripción más amplia del algoritmo se encuentran en Knorr et al (1998) [73].

2.2. Métodos de detección de *outliers* basados en densidad.

Estos métodos detectan *outliers* en conjuntos de datos con densidades variantes. Uno de los enfoques más influyentes es el uso del factor atípico local (Local Outlier Factor, LOF) de cada objeto [74]. El LOF de un objeto está basado en la densidad local de la vecindad de un objeto, donde la vecindad de un objeto está definida por el parámetro del número mínimo de puntos (MinPts) cercanos al objeto. Los objetos con un alto índice LOF son tratados como outliers, en lugar de una distancia específica o similaridad, lo que da la ventaja de manejar conjuntos de datos con una densidad variable [66].

2.2.1. Métodos de detección de *outliers* basados en agrupamiento (clustering).

Los algoritmos de clustering pueden detectar *outliers* como producto derivado del proceso de agrupamiento [66]. Algunos de los enfoques utilizados para esto es detectar los clusters pequeños (anomalías) que se alejan de los clusters mayores [75], o bien, una vez que se tienen hechos los clusters, detectar los objetos dentro de cada cluster que se alejan de sus correspondientes centroides. Un ejemplo de este último enfoque puede verse en el algoritmo CCleaner [66]. Normalmente, las técnicas para realizar los agrupamientos se basan en los métodos tradicionales como k-medias (k-means) de McQueen (1967) [76] y k-medoides (k-mediods) [77].

En Dasu et al (2003) [54] se describe un enfoque gráfico para la detección de *outliers* llamado gráfica de control (control-chart). Aquí, varios resúmenes estadísticos (media, error estándar, coeficiente de correlación) son aplicados a los registros recopilados. El nombre de la gráfica de control es derivado de la estadística, que denota el valor esperado

de la muestra como un gráfico-$\overline{X}$ ($\overline{X}$-*Chart*) si es la media o un gráfico-*R* **(***R*-*Chart***)** en el caso del coeficiente de correlación. Los valores son entonces graficados en un gráfico de control, el cual típicamente consiste de:

- Una línea horizontal central que denota el "valor esperado", como la media o la suma.
- Dos líneas paralelas a la línea central que representan los límites aceptables para la estadística correspondiente a la línea central.
- La gráfica de la estadística correspondiente a cada punto de datos, con el ID de cada punto en el eje X y la estadística en el eje Y.

3. Técnicas para eliminación de registros duplicados.

Un tipo de error que ocurre con frecuencia se produce en el momento de integrar una base de datos con series de datos que provienen de distintas fuentes. A menudo, registros diferentes de forma pero que representan a la misma instancia son almacenados por separado, ocasionando una duplicidad de información. Este tipo de problema ha aparecido con diversos nombres a lo largo del tiempo, como por ejemplo, integración semántica (Newcomb, 2003) [78], identificación de instancias (Wang, 1989) [79], encadenamiento de registros (Fellegi, 1969) [80], combinar/depurar (Hernández, 1998) [81] y detección y eliminación de registros duplicados (Lee, 2005) [82]. Una instancia especial del problema de registros duplicados es la desambiguación de nombres (Han, 2004) [83], el cual también es abordado con las técnicas aquí expuestas.

La eliminación de registros trata de identificar registros duplicados inexactos que se refieren a la misma entidad del mundo real pero que no son equivalentes de manera sintáctica [82].

La manera más confiable de detectar duplicados inexactos es comparar cada registro con cada uno de los registros restantes [84]. Algunos ejemplos para hacer la eliminación de duplicados de manera optimizada son los siguientes:

- **El método de ordenamiento vecindario (Sorted Neighbourhood Method, SNM) [81].** Éste se divide en tres fases:
  2) **Creación de llaves:** Se crean las llaves para representar de manera única a cada registro. La llave se forma de información extraída de los campos relevantes o de porciones de los campos.
  3) **Ordenamiento de datos:** Se ordenan los registros en base a las llaves creadas en la fase 1.
  4) **Combinación:** Se mueve una ventana de tamaño variable a través de la lista secuencial de registros limitando las comparaciones para encontrar registros con los registros dentro de la ventana. Si el tamaño de la venta es ***w*** registros, entonces cada nuevo registro entrante en la ventana se compara con los previos ***w-1*** registros para encontrar registros que coincidan.

- En Han et al (2004) **[83]** se explorar técnicas derivadas de modelos Naive-Bayes y máquinas de soporte vectorial (Support Vector Machines, SVM) para detectar registros duplicados.
- En Lee et al (2005) **[82]** se utiliza una técnica basada en reglas de asociación para detectar duplicidad de entradas de datos.

Anexo 5. Integración de datos.

Tabla A5.1 Estructura de la entidad resultante de la integración de datos.

Clasificación		#	Atributo	Unidades
Atributos agrícolas		1	Año agrícola	Entero
		2	Cultivo	Cadena (30)
		3	Superficie sembrada	ha
		4	Superficie cosechada	ha
		5	Lámina de riego	cm
		6	Costo de producción	M$/ha
		7	Rendimiento	Ton/ha
		8	Precio	M$/Ton
		9	Ingreso bruto	M$
		10	Costo total de producción	M$
		11	Ingreso neto	M$
Atributos climáticos	Temperatura	12	Temperatura de octubre	°C
		13	Temperatura de noviembre	°C
		14	Temperatura de diciembre	°C
		15	Temperatura enero	°C
		16	Temperatura febrero	°C
		17	Temperatura marzo	°C
		18	Temperatura abril	°C
		19	Temperatura mayo	°C
		20	Temperatura junio	°C
		21	Temperatura julio	°C
		22	Temperatura agosto	°C
		23	Temperatura septiembre	°C
	Temperatura máxima	24	Temperatura máxima octubre	°C
		25	Temperatura máxima noviembre	°C
		26	Temperatura máxima diciembre	°C
		27	Temperatura máxima enero	°C
		28	Temperatura máxima febrero	°C
		29	Temperatura máxima marzo	°C
		30	Temperatura máxima abril	°C
		31	Temperatura máxima mayo	°C
		32	Temperatura máxima junio	°C
		33	Temperatura máxima julio	°C
		34	Temperatura máxima agosto	°C
		35	Temperatura máxima septiembre	°C
	Temperatura mínima	36	Temperatura mínima octubre	°C
		37	Temperatura mínima noviembre	°C
		38	Temperatura mínima diciembre	°C
		39	Temperatura mínima enero	°C
		40	Temperatura mínima febrero	°C
		41	Temperatura mínima marzo	°C
		42	Temperatura mínima abril	°C
		43	Temperatura mínima mayo	°C
		44	Temperatura mínima junio	°C

		45	Temperatura mínima julio	°C
		46	Temperatura mínima agosto	°C
		47	Temperatura mínima septiembre	°C
	Precipitación	48	Precipitación octubre	mm
		49	Precipitación noviembre	mm
		50	Precipitación diciembre	mm
		51	Precipitación enero	mm
		52	Precipitación febrero	mm
		53	Precipitación marzo	mm
		54	Precipitación abril	mm
		55	Precipitación mayo	mm
		56	Precipitación junio	mm
		57	Precipitación julio	mm
		58	Precipitación agosto	mm
		59	Precipitación septiembre	mm
	Evaporación	60	Evaporación octubre	mm
		61	Evaporación noviembre	mm
		62	Evaporación diciembre	mm
		63	Evaporación enero	mm
		64	Evaporación febrero	mm
		65	Evaporación marzo	mm
		66	Evaporación abril	mm
		67	Evaporación mayo	mm
		68	Evaporación junio	mm
		69	Evaporación julio	mm
		70	Evaporación agosto	mm
		71	Evaporación septiembre	mm

Anexo 6. Descripción de las técnicas de modelado seleccionadas

1. Regresión multivariable

El análisis de regresión múltiple es una técnica estadística que puede ser utilizada para analizar las relaciones entre una variable dependiente única (criterio) y varias variables independientes (covariables). El objetivo análisis de regresión múltiple es utilizar varias variables independientes cuyos valores son conocidos para predecir el valor dependiente que el investigador desea conocer [98].

La regresión multivariable tiene varios usos:

1. Determinar lo apropiado que es el uso de la técnica de regresión con un problema dado.
2. Examinar la significancia estadística de la futura predicción.
3. Examinar la fortaleza de la asociación entre la variable dependiente y una o más variables independientes.
4. Predecir los valores de una variable en términos de los valores de las otras.

Descripción del modelo

Nota: La siguiente descripción es tomada de Larry (2004) [99].

Suponiendo que los datos tienen la forma de:

$$(Y_1, X_1),...,(Y_i, X_i),...,(Y_n, X_n)$$

donde
$X_i = (X_{i1},...,X_{ik})$ es el vector de covariables de longitud k para la observación i^{ésima}.
Y_i es el valor de la variable dependiente de la observación i^{ésima}.

El modelo de regresión lineal es entonces

$$Y_i = \sum_{j=1}^{k} B_j X_{ij} + \in_i \qquad \textbf{(A6.1)}$$

donde
i=1,...,n. donde *n* es el número de observaciones.
Bj es el coeficiente de regresión $_j$.
Xij es el valor j de la observación $_i$.
$\in_i$ es el error residual para la observación $_i$.

Usualmente, se desea incluir un punto de intercepción en el modelo, que se puede hacer colocando los valores de $X_{i1} = 1$ para i=1,…,n. Representando el modelo en notación matricial se tiene que:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ \cdot \\ Y_n \end{pmatrix} \qquad X = \begin{pmatrix} X_{11} & X_{12} & \cdot & \cdot & \cdot & X_{1k} \\ X_{21} & X_{22} & \cdot & \cdot & \cdot & X_{2k} \\ & & \cdot & \cdot & \cdot & \\ & & \cdot & \cdot & \cdot & \\ & & \cdot & \cdot & \cdot & \\ X_{n1} & X_{n2} & \cdot & \cdot & \cdot & X_{nk} \end{pmatrix}$$

Cada renglón en X es una observación; las columnas corresponden a las k covariables. Entonces, X es una matriz de (n x k). Utilizando la misma notación para B y para $\in$:

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \cdot \\ \cdot \\ \cdot \\ \beta_n \end{pmatrix} \qquad \in = \begin{pmatrix} \in_1 \\ \in_2 \\ \cdot \\ \cdot \\ \cdot \\ \in_n \end{pmatrix}$$

Entonces, (14) se puede escribir como:

$$Y = X\beta + \in \qquad \textbf{(A6.2)}$$

Asumiendo que la matriz $X^T X$ de (k x k) es invertible:

$$\overline{\beta} = (X^T X)^{-1} X^T Y \qquad \textbf{(A6.3)}$$

$$V(\overline{\beta} | X^n) = \sigma^2 (X^T X)^{-1} \qquad \textbf{(A6.4)}$$

La matriz $\overline{\beta}$ almacena los coeficientes de regresión calculados. V representa el cálculo de la varianza.
La función estimada de regresión es

$$\bar{r}(x) = \sum_{i=1}^{n} \overline{\beta}_j x_j \qquad \textbf{(A6.5)}$$

Un cálculo sin desviación de

$$\overline{\sigma}^2 = (\frac{1}{n-k}) \sum_{i=1}^{n} \overline{\in_i}^{\,2} \qquad \textbf{(A6.6)}$$

donde $\bar{\in} = X\overline{\beta} - Y$ es el vector de errores residuales. Una intervalo de confianza aproximado (1-$\propto$) para β_j esta dado por

$$\overline{\beta_j} \pm z_{\alpha/2}\,\overline{se}(\overline{\beta}_j) \qquad \textbf{(A6.7)}$$

Donde $\overline{se}^2(\overline{\beta}_j)$ es el j^ésimo elemento en la diagonal de la matriz $\sigma^2 (X^T X)^{-1}$.

Existe un "riesgo" inherente al utilizar los mínimos cuadrados ordinarios para la estimación de los parámetros de regresión cuando existe la presencia de colinealidad[22] en las variables independientes. Esto puede generar problemas de inestabilidad en los parámetros, signos incorrectos y frecuentes errores estándar elevados, lo que conduce a generar modelos con muy poco poder explicativo o de difícil interpretación [101]. Diversos autores han realizado diversas propuestas para reducir error en la estimación de los parámetros [102][103][104] . En esta tesis se utiliza la técnica de Hoerl y Kennard (1970) [102] para la reducción del error estándar en los parámetros, la cual se denomina regresión Ridge (Ridge Regression).

Una demostración de la superioridad de Ridge sobre el método de mínimos cuadrados se puede ver en Piña et al (2005) [105]. En la técnica Ridge se sacrifica el sesgo de los parámetros por una reducción de error estándar de los parámetros estimados.

2. Redes Neuronales

Las redes neuronales son modelos computacionales para el procesamiento de información y son particularmente útiles para identificar la relación fundamental entre un conjunto de variables o patrones en los datos. Una red neuronal es una abstracción de las redes neuronales biológicas con las cuales comparten dos características importantes: 1) son capaces de realizar el procesamiento en paralelo de información y 2) aprenden y generalizan de la experiencia [100].

Una definición de red neuronal es la siguiente [106]: las redes neuronales son colecciones de nodos conectados, con entradas, salidas y procesamiento en cada nodo. Entre las entradas y las salidas de la red existe un número de capas ocultas de procesamiento. La red neuronal debe ser entrenada con un conjunto de patrones de entrenamiento (aprendizaje supervisado).

El gran poder de las redes neuronales reside en su capacidad de aplicación en reconocimiento de patrones. Algunas características destacables tomadas de Zhang (2004) [100] son las siguientes:

- No se requieren asumir condiciones a priori de la estructura del modelo o del proceso que generó los datos.
- El proceso de modelado es altamente adaptable y el modelo es enteramente determinado por las características o patrones que la red neuronal adquirió de los datos durante el proceso de aprendizaje.

[22] La colinealidad se define en la sección 4.4.3.3.1. Otra opción es consultar Sundberg (2002) [191].

- La propiedad matemática de la aproximación con exactitud y de representación de relaciones complejas ha sido bien establecida y soportada por mucho trabajo teórico [107][108][109].
- Las redes neuronales son modelos no lineales y no parámetricos. Esto permite modelar con mayor exactitud problemas del mundo real que pocas veces presentan un comportamiento lineal.
- Las redes neuronales son resolver problemas que presentan patrones imprecisos o datos incompletos o con ruido, que también presentan un número grande de variables.

La principal desventaja de las redes neuronales se menciona en López (2005) [110]. *El aprendizaje de las redes neuronales, basado en la presentación iterativa de patrones de entrenamiento, realiza una exploración en un gran espacio de entrada e intenta correlacionar todas las dependencias del conjunto total de patrones. A pesar de que las arquitecturas y velocidades de los procesadores actuales son altas, el proceso de entrenamiento de las redes neuronales artificiales consume un tiempo demasiado largo*.

Descripción del modelo

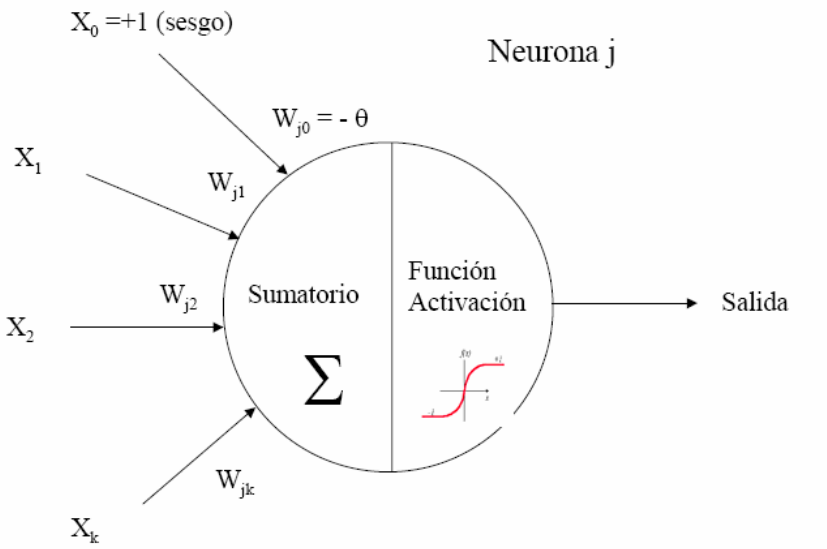
Los orígenes de las redes neuronales artificiales suelen situarse a partir del modelo de neurona básico debido al neurofisiólogo Warren McCulloch y al matemático Warren Pitts. Estos autores escribieron un estudio en el que modelaban el funcionamiento de una neurona simple con circuitos eléctricos. En este trabajo proponían una teoría general del procesamiento de la información basada en redes de elementos binarios, capaces de realizar cómputos similares a los ordenadores digitales pero con una ejecución paralela y no serial, en donde los pesos hacían el papel que el programa juega en un ordenador convencional [112].

McCulloch y Pitts (1943) modelaron la sinapsis en la transmisión de señales al cuerpo de la neurona asociando un factor multiplicativo a cada línea de entrada a la neurona. Las sinapsis en una neurona biológica convierten la actividad del axón en efectos eléctricos que inhiben o excitan la actividad del axón, inhibiendo o excitando la actividad en las neuronas conectadas [106]. De manera análoga, el modelo de McCulloch y Pitts utiliza la sumatoria de los factores multiplicativos de las entradas (que pueden provenir de otras neuronas) y un umbral (θ) para generar una señal de salida que activa o desactiva a otras neuronas. Así, si la neurona tiene *n* entradas, la suma de las entradas ***X*** por los pesos ***W*** queda como:

$$SumXW = X_1 x W_1 + \ldots + X_n x W_n = \sum_{i=1}^{n} W_i X_i \qquad \textbf{(A6.8)}$$

Si $SumXW \geq \theta$ entonces la salida es 1, en caso contrario, la salida es 0. La figura A6.1 muestra la estructura de una neurona artificial básica. Generalmente, el umbral se introduce en el modelo como la suma adicional de otra entrada de valor 1, y cuyo peso asociado tiene el valor del umbral.

Figura A6.1. Estructura de una neurona artificial básica. Ilustración tomada de Santíni (2005) [112]



El propósito de la función de activación es la de limitar la amplitud de la salida de la neurona [113]. Normalmente las funciones no se modifican, de tal forma que el estado de la red neuronal depende del valor de los factores de peso que se aplica a los estímulos de la neurona [114]. Incluso, la función de activación podría no existir, siendo la salida la misma entrada a la neurona (normalmente, la sumatoria de los pesos por las entradas anteriores). Se dice que la función de activación logística o sigmoidal es probablemente la función de activación más empleada en la actualidad [112] (ecuación A6.9).

$$f(x) = \frac{1}{(1 + e^{-\alpha x})} \qquad \textbf{(A6.9)}$$

El proceso de funcionamiento de una neurona simple es el siguiente:

1. Dado un vector de entrada X, se calcula la sumatoria de los pesos por los valores del vector de entrada (R1).
2. A R1 se le resta el valor del umbral (R2).
3. Si R2 es positivo, entonces evalúa la función de activación (R3).
4. Transfiere R3 a la siguiente neurona.

La forma de interconectar las redes neuronales, el número de capas, las funciones de activación y los algoritmos de aprendizaje utilizados son los factores que determinan la existencia de diferentes topologías o modelos en las redes neuronales. Actualmente, existen docenas de diferentes modelos de redes neuronales que son utilizados regularmente para una gran variedad de problemas. Los tres modelos de redes neuronales más conocidos y comúnmente utilizados para propósitos de minería de datos son: la red perceptrón multicapa, la red de Hopfield y el mapa de Kohonen [100].

2.1. Perceptrón multicapa

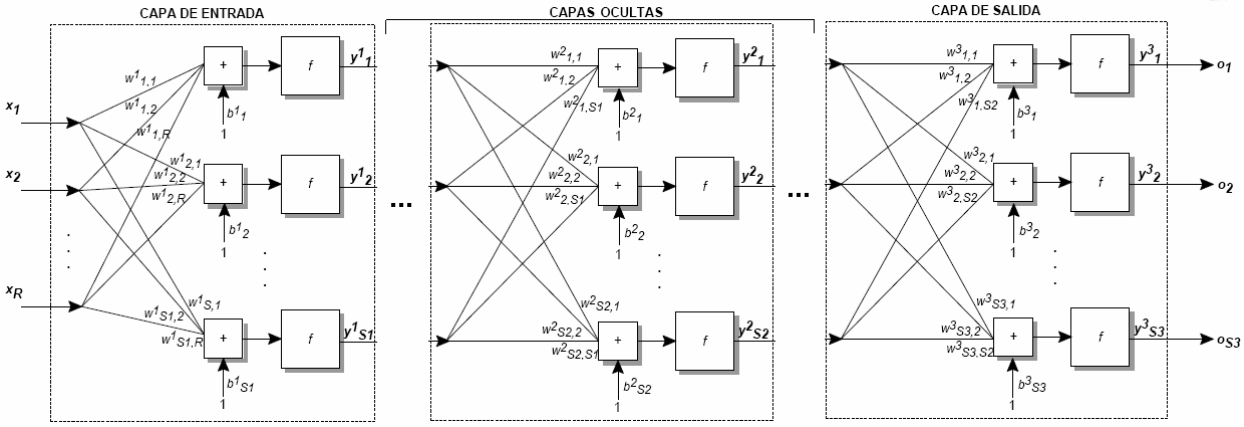
El perceptrón simple fue el primer modelo de red neuronal capaz de asociar patrones. Rosenblatt (1958) propuso el perceptrón simple con tan solo dos capas y un mecanismo de aprendizaje como un sistema de clasificación capaz de interpretar patrones tanto abstractos como geométricos[112]. El mecanismo de aprendizaje consiste en la modificación de los pesos W, de modo de hacer las salidas producidas por la red lo más parecido posible a las salidas de la muestra.

El principal problema que tiene el perceptrón simple es en las clasificaciones de conjuntos que no son linealmente separables, ya que la célula de decisión que utiliza es el hiperplano [112]. Una red utilizada para clasificación que sólo tiene neuronas de entrada y salida es capaz de representar únicamente límites de decisión lineales [106]. Como la mayoría de los problemas complejos no son lineales, las posibilidades de utilizar el perceptrón simple como mecanismo de solución se reducen considerablemente.

Como solución, Rumelhart et al (1986) [210] popularizaron el uso de las redes perceptrón multicapa (Multi-Layer Perceptrón, MLP) y un algoritmo de aprendizaje supervisado denominado de retropropagación (backpropagation). Juntos, la arquitectura de la red neuronal y el algoritmo de aprendizaje, son el modelo de red neuronal más ampliamente utilizada y estudiada en la práctica [100]. Según Wong et al (1997) [111], cerca del 95% de las aplicaciones comerciales de redes neuronales reportadas en la literatura utilizan dicho modelo.

El perceptrón multicapa tiene la estructura que se describe en la figura A6.2.

Figura A6.2. Red perceptrón multicapa.



El perceptrón multicapa puede ser totalmente o localmente conectado. En el primero, todas las salidas de la capa ***i*** se conectan a cada una de las entradas de las neuronas de la capa ***i+1***, mientras que en el segundo, las salidas de las neuronas de la capa ***i*** se conectan sólo con algunas (una región) de la capa ***i+1***.
En la figura A6.2 se pueden distinguir tres niveles de capas. La primera, conformada por la capa de entrada, está constituida por aquellas neuronas que introducen los patrones de entrada en la red. En estas neuronas no se produce ningún cálculo, su función es iniciar la

secuencia de procesamiento a través de la activación/desactivación de las siguientes neuronas. El segundo nivel está conformado por las capas ocultas, dichas capas contienen a las neuronas cuyas entradas provienen de capas anteriores y cuyas salidas van a las capas posteriores. Por esta razón se dice que en el perceptrón multicapa las conexiones entre las neuronas son de alimentación hacia adelante (feedforward). Las neuronas de las capas ocultas. Las neuronas de las capas ocultas usan como regla de propagación la suma ponderada de las entradas con los pesos sinápticos w_{ij} y sobre esa suma ponderada se aplica una función de transferencia de tipo sigmoide, que es acotada en respuesta. Finalmente, la capa de salida incluye a las neuronas cuyos valores de salida se corresponden con las salidas de toda la red.

Pero la red perceptrón multicapa por sí sola no tendría mayor utilidad de no ser por el algoritmo de entrenamiento que se encarga de asignar los pesos más adecuados a los nodos de la red.

El algoritmo de retropropagación (backpropagation)

El algoritmo backpropagation[23] es una generalización del algoritmo LMS[24], que basa su aprendizaje en el error cuadrático medio. Por lo tanto, el ajuste de los pesos que interconectan a las neuronas de toda la red se realiza con el objetivo de minimizar el error cuadrático medio. Backpropagation es un algoritmo de aprendizaje supervisado, dado que necesita un conjunto de patrones de entrenamiento como valores de entrada [110].

En general, el algoritmo de entrenamiento basado en el error backpropagation es el siguiente:

1. Inicializar los pesos y los umbrales iniciales de cada neurona. Hay varias posibilidades de inicialización siendo las más comunes las que introducen valores aleatorios pequeños.
2. Para cada patrón del conjunto de los datos de entrenamiento:
    a. Obtener la respuesta de la red ante ese patrón. Esta parte se consigue propagando la entrada hacia adelante. Las salidas de una capa sirven como entrada a las neuronas de la capa siguiente, procesándolas de acuerdo a la regla de propagación y la función de activación correspondientes.
    b. Calcular la sumatoria de errores asociados según la ecuación:

$$E(\boldsymbol{W}) = \sum_{q=1}^{p} E(\boldsymbol{x}^{(q)}; \boldsymbol{W}) \qquad \textbf{(A6.10)}$$

donde:
$\boldsymbol{W}$.- Representa al conjunto de parámetros (pesos) de la red.
$\boldsymbol{x}^{(q)}$.- Corresponde a los valores de entrada para la observación q.
p es el número total de observaciones.

[23] En español, propagación hacía atrás o retropropagación.
[24] LMS.

E(x(q); W) es el error obtenido para la entrada ***x(q)*** y parámetros ***W***. Este error se calcula de la siguiente manera:

$$E(\boldsymbol{x}^{(q)}; \boldsymbol{W}) = \frac{1}{2}\|N(\boldsymbol{x}^{(q)}; \boldsymbol{W}) - \boldsymbol{t}^{(q)}\|^2 \quad \textbf{(A6.11)}$$

donde:
N(x(q),W) es la función que representa la salida de la red neuronal para la entrada ***x(q)*** y los parámetros ***W***.
t(q) es la salida para la entrada ***x(q)***.
|| ***X - Y*** || es la norma euclidiana [25].

c. Calcular los incrementos parciales (sumandos de las sumatorias). Estos incrementos dependen de los errores calculados en 2.b

$$\Delta \boldsymbol{W} = -\gamma \frac{\partial E(\boldsymbol{W})}{\partial \boldsymbol{W}}, \quad \textbf{(A6.12)}$$

donde:
0 < γ < 1 es un parámetro conocido como factor de aprendizaje, el cual es el encargado de regular la cuantía en la que los pesos serán modificados.
$\frac{\partial E(W)}{\partial W}$ es la derivada parcial de ***E(W).***

3. Calcular el incremento total, para todos los patrones, de los pesos y los umbrales.
4. Actualizar pesos y umbrales. Los cambios en los pesos se producen en la dirección en la que el error caiga lo más rápidamente posible (gradiente negativo descendente). Estas señales de error se propagan a los nodos de las capas, empezando por la de salida y siguiendo con las capas sucesivas. De esta forma, los nodos de cada capa sólo reciben una facción del error global en función de su aproximada contribución relativa a la obtención de la salida. Los pesos se modificarán en base a la señal de error recibida, de manera que se reduzca el error actual de la red y la salida obtenida se vaya aproximando a la deseada **[112]**.
5. Calcular el error actual y volver al paso 2 si no es satisfactorio. Esto hasta que no se cumpla un límite de iteraciones establecido previamente (epochs) ó se alcance un mínimo de error también preestablecido.

Esta es una versión resumida del algoritmo. En Santín (2005) [112] se encuentra una explicación más detallada del algoritmo de aprendizaje.

2.2. Hopfield

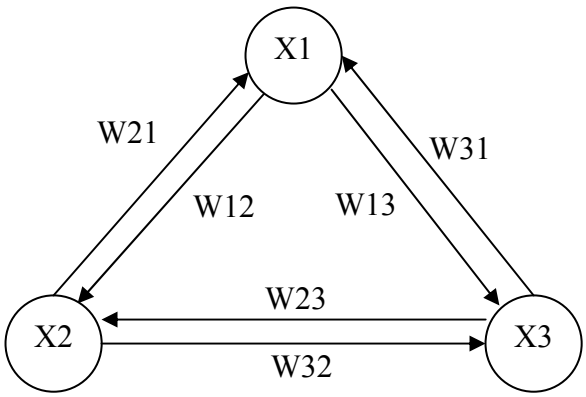
[25] La norma euclidiana || X –Y || está dada por $\sqrt{(x_1 - y_1)^2 + \ldots + (x_n - y_n)^2}$.

J. J. Hopfield [115] propuso a principios de los años 80's estas redes que consisten de una capa de neuronas que están completamente conectadas entre ellas. Hopfield analizó el comportamiento de estas redes, y probó matemáticamente que un comportamiento estable puede lograrse bajo ciertas condiciones [116].

Las redes de Hopfield son redes de adaptación probabilística, recurrentes, funcionalmente entrarían en la categoría de las memorias autoasociativas, es decir, que aprenden a reconstruir los patrones de entrada que memorizaron durante el entrenamiento. Son arquitecturas de una capa con interconexión total, funciones de activación booleana de umbral (cada unidad puede tomar dos estados, 0 o 1, dependiendo de si la estimulación total recibida supera determinado umbral), adaptación probabilística de la activación de las unidades, conexiones recurrentes y simétricas, y regla de aprendizaje no supervisado [114].

Las redes de Hopfield no presentan una estructura organizada en varias capas, en su lugar, presentan una sola capa de neuronas completamente interconectadas. A diferencia de las redes tipo feedforward (como el perceptrón multicapa), donde la información fluye en varias direcciones, en las redes de Hopfield la información fluye en ciclos entre las neuronas. La figura A6.3 proporciona una ilustración de una red tipo Hopfield de tres neuronas.

Figura A6.3. Red Hopfield de tres neuronas [100].



La red es completamente descrita por un vector de estados el cual está en función del tiempo *t*. Cada nodo en la red contribuye a un elemento en el vector de estados y cualquiera o todas las salidas de los nodos pueden ser tratadas como salidas de la red. La dinámica de las neuronas puede ser describa como sigue [100]:

$$u_i(t) = \sum_{j=1}^{n} w_{ij} x_j(t) + v_i$$
$$x_i(t+1) = sign(u_i(t)) \qquad \textbf{(A6.13)}$$

donde:
$u_i(t)$ es el estado interno de la i-ésima neurona.

$x_i(t)$ es la activación o estado de salida de la neurona i-ésima.
v_i es el umbral para la i-ésima neurona.
n es el número de neuronas.
sign(x) = 1, si x=0 y -1 en caso contrario.

Dado un conjunto de condiciones iniciales x(0), restricciones apropiadas para los pesos (como por ejemplo, simetría), la red de la figura A6.3 y con la dinámica modelada en (A6.13) convergerá a un punto de equilibrio adecuado.

Para cada estado en la red, existe una energía asociada con ese estado. Una función común de energía se muestra en (27).

$$E(t) = -\frac{1}{2} x(t)^T W x(t) - x(t)^T v \qquad \textbf{(A6.14)}$$

Donde:
x(t) es el vector de estados.
W es la matriz de pesos.
v es el vector de umbrales.

La idea básica de la función de energía es que siempre disminuye o al menos permanece constante mientras el sistema evoluciona en el tiempo [100].

El uso principal de la red de Hopfield es la memoria asociativa. La memoria asociativa es un dispositivo el cual está entrenado para efectuar asociaciones entre elementos origen y destino; una vez que el entrenamiento está completo, la memoria es capaz de recuperar la información del destino cuando se presente una versión incompleta o distorsionada de la información del origen. Cuando los elementos origen y destino son idénticos, se dice que la memoria es auto-asociativa. En general, el entrenamiento es efectuado de manera no-adaptable, haciendo la fase de aprendizaje independiente de la fase de recuperación [117].

Aprendizaje

Existen muchas maneras para determinar los pesos por medio de un conjunto de entrenamiento el cual es un conjunto de patrones conocidos. Una manera es utilizar el enfoque de prescripción expuesto en Hopfield (1982)[115] donde los pesos están dados por:

$$w = \frac{1}{n} \sum_{i=1}^{n} z_i z_i^T \qquad \textbf{(A6.15)}$$

Donde:
Z_i, para i=1,2,…,p son patrones que serán almacenados en la red.

En Zhang (2005)[100] se describe también otra manera de usar un proceso iterativo e incremental de aprendizaje llamado la regla de aprendizaje de Hebbian (1949) [118]. Ésta define el siguiente proceso:

1. Seleccionar un patrón de manera aleatoria del conjunto de entrenamiento.
2. Presentar un par de componentes del patrón a la salida de los nodos correspondientes de la red.
3. Si dos nodos tienen el mismo valor, entonces se hace un pequeño incremento positivo en el peso de la interconexión. Si tienen valores opuestos, entonces se hace un pequeño decremento negativo. El tamaño del incremento ó decremento puede ser expresado como $w_{ij} = \alpha z_i^p z_j^p$, donde α es una razón constante entre 0 y 1 y z_i^pes el i-ésimo componente del patrón p.

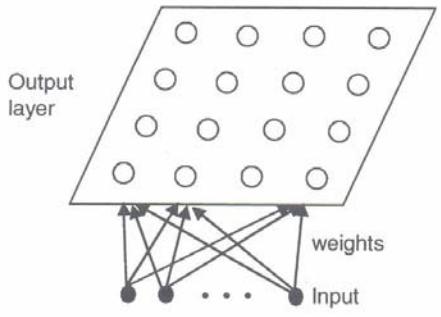
Las redes de Hopfield tienen limitaciones en cuanto al tamaño de patrones que pueden almacenar y recordar. Si demasiados patrones son almacenados, la red puede converger a un falso patrón diferente a los programados. También se corre el riesgo de que la red se vuelva inestable si los patrones almacenados son muy similares.

2.3. Kohonen

Las redes de Kohonen (SOM[26]) [119] son redes neuronales artificiales que se adaptan así mismas en respuesta a señales de entrada y basándose en el algoritmo de Kohonen. Las redes consisten de nodos distribuidos uniformemente en una rejilla y que están funcionalmente conectados cada uno con sus nodos vecinos [120].

Una red típica de Kohonen tiene dos capas de nodos, una capa de entrada y una capa de salida (algunas veces también llamada capa de Kohonen). Cada nodo en la capa de entrada está totalmente conectado a los nodos en dos en la capa bidimensional de salida. La figura A6.4 muestra un ejemplo de una red SOM con varios nodos de entrada en la capa de entrada y una capa de salida de dos dimensiones con un arreglo rectangular de 16 neuronas. El número de nodos en la capa de entrada corresponde al número de variables de entrada mientras el número de nodos de salida depende del problema específico y es determinado por el usuario. Usualmente, este número de neuronas en el arreglo rectangular debe ser lo suficientemente largo como para permitir que se formen un número suficiente de agrupamientos (clusters) [100]. Es recomendable que este número sea 10 veces más la dimensión del patrón de entrada [121].

Figura A6.4. Una red de Kohonen de 4x4 [100].



[26] Regularmente, las redes de Kohonen son llamadas Mapas Autoorganizativos de Kohonen (Self-Organizing Maps, SOM).

Las desventajas más significativas de las redes de Kohonen son enunciadas en Laerhoven et al (2000) [122]. Al parecer, el algoritmo tradicional inicia con un alto nivel de adaptabilidad - con una gran taza de aprendizaje y un amplio rango de vecindad - y se ajusta gradualmente. Después de esta etapa, es muy difícil que la red pueda aprender cualquier cosa más, la cual plantea un obstáculo en el caso de que se necesite que el sistema se adapte a nuevas condiciones de manera constante. Si el algoritmo permanece flexible, puede ocurrir el caso de que se sobrescriban los prototipos previamente almacenados. Este intercambio es conocido en el campo del aprendizaje automático como el "dilema de la estabilidad plástica"[27] o el "olvido catastrófico"[28] [123].

Otro problema es el hecho de que el aprendizaje se hace más lento a medida de que el número de entradas se incrementa. A este problema se le llama la "maldición de la dimensionalidad"[29].

3. Algoritmos de inducción de reglas

El uso de reglas es una de las formas más popular de representación del conocimiento debido, entre otras razones, a su sencillez, capacidad de expresión y escalabilidad [128].

Una tarea de aprendizaje basado en reglas puede ser definida como sigue: dado un conjunto de ejemplos de entrenamiento (instancias para las que ya se conoce una clasificación), encontrar un conjunto de reglas de clasificación que puedan ser utilizadas para la predicción o la clasificación de nuevas instancias (casos que no se hayan presentado antes).

En el ámbito computacional, los ejemplos o instancias son almacenados en una base de datos en forma de atributos y registros. El propósito de las reglas es buscar relaciones entre los atributos de la base de datos, colocando unos en el antecedente de la regla unos y en el consecuente otros.

Las reglas pueden tener formas diversas. Una forma es:

$$P_1, \ldots, P_m \rightarrow Q_1, \ldots, Q_n \qquad \textbf{(A6.16)}$$

Donde:

P_n Es una condición o prueba sobre uno o más atributos de la base de datos.
Q_n Son acciones también llamadas conclusiones que afectan a uno o más atributos.

(29) también puede ser leído como:

$$\textbf{\textit{SI } } P_1 \textbf{\textit{ y } } P_2 \textbf{\textit{ y … y } } P_m \textbf{\textit{ ENTONCES } } Q_1 \textbf{\textit{ y } } Q_2 \textbf{\textit{ y … y } } Q_n \qquad \textbf{(A6.17)}$$

Siendo (30) una forma más común y comprensible para las personas que (29).

[27] Traducción literal del término *Stability-Plasticity Dilemma*.
[28] Traducción literal del término *Catastrophic Forgetting*.
[29] Traducción literal del término *curse of dimensionality*

Existen muchos métodos para la inducción de reglas. En Hu y Li (2005) [129] se presenta la siguiente clasificación:

1. *Métodos basados en algoritmos de cobertura.* Este tipo de algoritmos emplean el enfoque divide y vencerás y trabajan de la siguiente manera: se localiza primero la "mejor regla" y entonces todos los registros cubiertos por la regla se remueven del conjunto de datos. Este procedimento se repite hasta que no quedan registros en el conjunto original. La manera de encontrar la "mejor regla" es usualmente basada en alguna heurística (por ej. la entropía[30]). Algunos ejemplos métodos típicos en esta categoría son AQ15 [130], CN2 [131], y CPA [132].
2. *Métodos basados en árboles de decisión.* Un árbol de decisión es también un enfoque típico del tipo "divide y vencerás". Difiere de los algoritmos de cobertura en que divide el conjunto de entrenamiento en subconjuntos disjuntos de maneara simultánea por medio de los valores de un atributo. Estos subconjuntos son simultáneamente divididos por algún otro atributo de manera recursiva hasta que cada subconjunto contiene únicamente registros de una sola clase o aproximados. La forma de hacer las particiones está guiada por alguna heurística (por ejemplo, la ganancia de información de Quinlan [68]). Cada camino desde la raíz hasta la hoja del árbol puede ser interpretado como una regla. Ejemplos de algoritmos para árboles de decisión son ID3 [133] y C4.5 [68].
3. *Métodos basados en la asociación.* Las reglas de asociación son propuestas para resolver problemas del tipo "cesta de compras"[31] en datos transaccionales. Sin embargo, cuando todos los posibles blancos de las reglas de asociación están contenidas por las etiquetas de clase, las reglas de asociación se convierten en reglas de asociación de clase y pueden ser utilizadas para propósitos de clasificación. Todos los algoritmos de reglas de asociación (por ejem, A priori [134] ó FP-growth [135]) pueden ser fácilmente adaptados para minar reglas de asociación de clase.
4. *Métodos de optimización basados en la asociación.* La principal característica de la minería de reglas de asociación es utilizar la propiedad de soporte para confinar el espacio de búsqueda. Cuando el propósito de la minería de reglas es encontrar una alta confianza (o precisión) como la que se necesita en las aplicaciones de clasificación, el problema se centra en el descubrimiento del conjunto de reglas óptimo. La propiedad de confianza limita más allá el espacio de búsqueda que la propiedad de soporte, y la minería de reglas óptimas de asociación de clase es más eficiente que la minería de reglas de asociación. Algunos algoritmos típicos de minería de reglas óptimas de asociación de clase son la minería de reglas óptimas PC [135] y la minería de conjuntos de reglas óptimas de asociación de clase de Li et al [135].

Algunos ejemplos de algoritmos para la inducción de regla son los siguientes:

[30] La entropía se define en la sección de "elaboración de modelos".
[31] "Market basket problem". Un problema común en minería de datos que consiste en encontrar relaciones entre los productos de una operación de compra en un supermercado.

Id3. Este algoritmo se utiliza para la búsqueda de hipótesis o reglas dado un conjunto de ejemplos. El conjunto de ejemplos deberá estar conformado por una serie de tuplas de valores, cada uno de ellos denominados atributos, en el que uno de ellos (el atributo a clasificar) es el objetivo el cual es de tipo binario (positivo o negativo , si o no , valido o invalido , etc). De esta forma el algoritmo trata de obtener las hipótesis que clasifiquen ante nuevas instancias si dicho ejemplo va a ser positivo o negativo. ID3 realiza esta labor mediante la construcción de un árbol de decisión [138]. Id3 requiere que todos los atributos sean valores discretos.

J48. J48 es una variante del algoritmo de inducción de árboles de decisión C4.5 [68]. Los nodos de prueba en el árbol son seleccionados basándose en su habilidad para producir una separación clara entre las clases. J48 puede trabajar con atributos continuos, pero el atributo a clasificar debe ser discreto.

Random Tree (árbol aleatorio). Un árbol aleatorio es una estructura de datos con un nodo en la cumbre que se ramifica en un número arbitrario de nodos, los cuales a su vez se ramifican en un número arbitrario, y así de manera sucesiva. Un árbol aleatorio se define en un evento elemental ω de un espacio de probabilidad (Ω, F, P). La probabilidad P depende del modelo de árbol utilizado. La forma más simple de un árbol aleatorio es un árbol aleatorio de búsqueda binaria. El término "aleatorio" en este contexto significa que cada árbol en el conjunto tiene igual oportunidad de ser muestreado, por lo que la distribución de árboles es uniforme. Los árboles aleatorios pueden trabajar con atributos continuos, pero el atributo a clasificar debe ser discreto. Una desventaja de estos algoritmos es que no contemplan un mecanismo de poda.

REPTree. Árbol de decisión de aprendizaje rápido que utiliza la poda con error reducido (Reduced Error Prunning, REP). Construye un árbol de decisión/regresión utilizando información de la ganancia/varianza y lo poda utilizando el mecanismo de error reducido (con ajuste hacia atrás). Los valores faltantes son tratados con la división de las estancias correspondientes en pedazos (como en C4.5) [139].

Conjunctive Rule. El algoritmo de regla conjuntiva aprende una sola regla que puede predecir un valor de salida [140]. Una regla conjuntiva consiste de antecedentes y operadores "AND" juntos, más un consecuente (el atributo clasificador). Este algoritmo selecciona un antecedente computando la ganancia de información de cada antecedente y poda la regla general utilizando la poda de error reducido o pre-poda simple basándose en el número de antecedentes. Para clasificación, la información de uno de los antecedentes es el promedio ponderado de las entropías de ambos datos, los cubiertos y los no cubiertos por la regla [139].

JRip. Este clasificador implementa reglas proposicionales utilizando el algoritmo RIPPER (Repeated Incremental Prunning to Produce Error Reduction) propuesto por William W. Cohen como una versión optimizada del IREP [141]. Consiste de dos etapas, una etapa de construcción, que contempla las fases de crecimiento y poda, y la etapa de optimización, donde se seleccionan las reglas y su representación.

NNge. Es un algoritmo que utiliza el concepto de "vecino más cercano" con ejemplares generalizados (donde los hiper-rectangulos pueden ser vistos como reglas si-entonces [142].

OneR. OneR (abreviación de "One Rule"), es un algoritmo simple de clasificación que genera un árbol de decisión de un nivel. OneR es capaz de inferir de manera simple y típica, incluso precisa, reglas de clasificación para un conjunto de instancias [143]. Implementa un clasificador de una regla, utilizando el atributo de error mínimo para predicción, discretizando atributos numéricos.

PART. Genera una lista de decisión PART. Utiliza divide y vencerás, construye un árbol parcial de decisión en cada iteración y hace de la "mejor" rama una regla [139]. El algoritmo PART primero infiere una regla fuerte en el conjunto de datos y entonces elimina todos los ejemplos que satisfacen la regla. Posteriormente, el algoritmo desarrolla reglas que son un poco más débiles en los ejemplos restantes [144].

Ridor. Es la implementación de un clasificador de regla RIpple-DOwn. Genera primero una regla y crea entonces excepciones para la regla con la menor tasa de error. Genera las mejores excepciones para cada excepción e itera hasta obtener un conjunto puro. Así, ejecuta una expansión de excepciones tipo árbol. Las excepciones son un conjunto de reglas que predicen clases. El algoritmo IREP es utilizado para generar las excepciones [139].

Como ejemplo de un algoritmo para inducción de reglas, se presenta aquí la descripción del algoritmo J48.

Descripción del algoritmo J48

J48 es la implementación en Weka del algoritmo C4.5 de Quinlan [68]. Es mucho más común encontrar el algoritmo como C4.5 en la literatura. El C4.5 se basa en el ID3, por lo tanto, la estructura principal de ambos métodos es la misma. El C4.5 construye un árbol de decisión mediante el algoritmo "divide y vencerás" y evalúa la información en cada caso utilizando los criterios de entropía y ganancia o proporción de ganancia, según sea el caso [147]. Estos criterios (cruciales para la comprensión del algoritmo) los proporciona Quinlan (1996) [148] con relación al algoritmo C4.5.

El criterio de facto para particionar utilizado por C4.5 es la proporción de ganancia, una métrica basada en la información que toma en cuenta números diferentes (y probabilidades diferentes) de pruebas de resultados. En el caso de la siguiente ecuación, ***C*** denota el número de clases y ***p(D, j)*** la proporción de casos en ***D*** que pertenecen a la clase ***j***. La incertidumbre residual con relación a la clase para la cual un caso en ***D*** pertenece puede ser expresada como:

$$Info(D) = -\sum_{j=1}^{C} p(D,j) X \log_2(p(D,j)) \qquad \textbf{(A6.18)}$$

La ecuación (A6.18) no es otra que la definición de Shannon (1948) [149] para entropía de la información. Por lo tanto, la ganancia de información correspondiente por la prueba ***T*** con ***k*** resultados posibles es:

$$Gain(D,T) = Info(D) - \sum_{i=1}^{C} \frac{|D_i|}{|D|} X\ Info(D_i) \qquad \textbf{(A6.19)}$$

La información ganada por una prueba es fuertemente afectada por el número de resultados y es máxima cuando existe un solo caso en cada subconjunto D_i. Por otro lado, la información potencial obtenida por el particionamiento de un conjunto de casos está basada en conocer el subconjunto D_i para el cual un caso falla; esta *información de corte* se expresa en la ecuación (A6.20).

$$Split(D,T) = -\sum_{i=1}^{k} \frac{|D_i|}{|D|} x\ log_2(\frac{|D_i|}{|D|}) \qquad \textbf{(A6.20)}$$

Por medio de la ganancia de información y la información de corte es posible definir la proporción de ganancia de información (A6.21).

$$Gain\_Ratio(D) = \frac{Gain(D,T)}{Split(D,T)} \qquad \textbf{(A6.21)}$$

La proporción de ganancia expresa la proporción de información útil generada en la partición. Si la partición es casi trivial, la información de la división será pequeña y esta proporción se volverá inestable. Para evitar este fenómeno, el criterio de proporción de ganancia selecciona una prueba que maximice la expresión anterior, sujeta a la restricción de que la información de la división sea grande, al menos tan grande como la ganancia promedio sobre todas las pruebas realizadas [147].

La figura A6.5 muestra el algoritmo C4.5.

Figura A6.5. Algoritmo C4.5 (tomado de Servente, 2002 [147])

```
Función C4.5
(R: conjunto de atributos no clasificadores,
C: atributo clasificador,
S: conjunto de entrenamiento) devuelve un árbol de decisión;

Comienzo
Si S está vacío,
       devolver un único nodo con Valor Falla;
Si todos los registros de S tienen el mismo valor para el atributo clasificador,
       Devolver un único nodo con dicho valor;
Si R está vacío, entonces
       devolver un único nodo con el valor más frecuente del atributo clasificador en
       los registros de S [Nota: habrá errores, es decir, registros que no estarán bien
       clasificados en este caso];
Si R no está vacío, entonces
       D ← atributo con mayor Proporción de Ganancia(D,S) entre los atributos de R;
       Sean {dj| j=1,2, .., m} los valores del atributo D;
       Sean {Sj| j=1,2, .., m} los subconjuntos de S correspondientes a los valores de
       dj respectivamente;
       Devolver un árbol con la raíz nombrada como D y con los arcos nombrados  d1, d2,
       .., dm que van respectivamente a los árboles
       C4.5(R-{D}, C, S1), C4.5(R-{D}, C, S2), .., C4.5(R-{D}, C, Sm);
Fin
```

La proporción de ganancia mencionada en la última condición del algoritmo corresponde a la ecuación (A6.21).

Ventajas y limitaciones

- Los árboles de decisión generados por C4.5 son relativamente fáciles de entender. La dificultad de entendimiento aumenta de manera proporcional al árbol generado.
- Los árboles de decisión se pueden convertir fácilmente a reglas de producción.
- Los árboles C4.5 pueden clasificar ambos, datos numéricos y categóricos, pero el atributo de salida debe ser categórico.
- No se hacen asunciones a priori acerca de la naturaleza de los datos.
- No se permiten múltiples atributos de salida.
- Los algoritmos basados en árboles de decisión (como C4.5) son inestables. Variaciones omitidas en los datos de entrenamiento pueden resultar en selecciones diferentes de atributos para cada punto de elección dentro del árbol. El efecto puede ser significativo dado que las elecciones de atributos afectan a los subárboles subsecuentes.
- Los árboles creados por conjuntos de datos numéricos pueden ser bastante complejos dado que las divisiones en los atributos numéricos son binarias.

4. Árboles de regresión

Un árbol de regresión es una pieza constante o una pieza de estimación lineal de una función de regresión, construido por el particionamiento recursivo de los datos y del espacio muestral. Su nombre se deriva de la práctica de visualizar las particiones como un árbol de decisión, desde el cual los papeles de las variables predictoras pueden ser inferidas [150].

Los modelos de árbol estadísticos son utilizados en muchos y diferentes dominios científicos. Comparados con otros modelos, éstos son versátiles, porque generan resultados satisfactorios para problemas no lineales, tienen l habilidad de manejar datos con valores faltantes y permite ser menos restrictivo con respecto a la distribución. Una de las propiedades más importantes para el análisis de datos es el hecho de que resultan estructuras jerárquicas fáciles de interpretar, dado de que frecuentemente se deben exponer los resultados a terceras personas. A pesar de que los modelos de árbol son un poco complejos en su estructura, son fáciles de interpretar [151].

El algoritmo AID [152][153] es la primera implementación de un árbol de regresión. El algoritmo busca sobre todas las particiones ortogonales al eje y produce una constante estimada. En cada fase, se selecciona la partición binaria que minimiza la suma total del error cuadrático. El particionamiento se detiene si la suma total del error cuadrático es menor que un valor pre-especificado γ o si el tamaño de la muestra es demasiado pequeño.

Una debilidad de de AID es precisamente la dificultad para la estimación de γ. Un valor demasiado pequeño o demasiado grande puede provocar sobreentrenamiento o falta de (respectivamente). Otra debilidad es que el mecanismo de búsqueda voraz utilizado induce una desviación en la selección de variables [154].

El algoritmo CART de Breiman et al (1984) [155] anula la dificultad de seleccionar γ empleando una estrategia de eliminación en retroceso para determinar el árbol. Este desarrolla un árbol demasiado grande y entonces lo poda algunas ramas, utilizando un muestreo de prueba o validación cruzada a fin de estimar la suma de error cuadrático total. El problema de la desviación en la selección de variables permanece, dado que CART utiliza el mismo mecanismo de búsqueda voraz que AID [150].

La mayoría de los trabajos relacionados con árboles de regresión utilizan el algoritmo CART como base. Desde su creación, muchos enfoques diferentes se han propuesto y utilizado para la poda, construcción y evaluación de los árboles. Hoy en día es muy fácil encontrar modelos desarrollados que utilizan varios algoritmos, como el enfoque Bayesiano de Chipman et al (1998) [156] ó Denison et al (1998) [157] para realizar la búsqueda entre los árboles; las modificaciones propuestas por Ripley (1996) [158]. El método FIRM de Kass (1975) [159], Hawkings (1997) [160] maneja el problema de la desviación utilizando pruebas de significancia con ajuste Bonferroni [161] para seleccionar los predictores a particionar.

Una vez probada la efectividad de los árboles de regresión, lo más natural fue la combinación de estas técnicas con los modelos clásicos de regresión lineal para generar algoritmos de aprendizaje automático que producen una salida de regresión lineal simple si existen pocos datos disponibles, y agregar estructuras de árbol más complejas si existen suficientes datos que soporten dichas estructuras [162]. El más claro ejemplo de dicha combinación es el algoritmo M5 de Quinlan [163].

En el contexto de regresión, si una hoja del árbol está asociada con un valor de salida promedio de las instancias ordenadas bajo la hoja (un modelo de orden cero), entonces el

modelo completo es llamado árbol de regresión (como CART). Si el árbol tiene en sus hojas funciones más complejas de regresión que están en función de las variables de entrada, entonces el modelo completo es llamado árbol de modelos (como M5) [164].

Recientemente, Witten y Wang (1997) [165] presentaron el algoritmo M5', una reconstrucción racional del algoritmo original M5 de Quinlan. M5' primero construye un árbol de regresión de forma recursiva, dividiendo el espacio entre las instancias utilizando pruebas en atributos individuales que maximizan la reducción de varianza en la variable destino (atributo a predecir). Después de que el árbol ha crecido, Un modelo de regresión lineal múltiple es construido para cada nodo interno, utilizando los datos asociados con el nodo y todos los atributos que participan en las pruebas en el subárbol que nace a partir de cada nodo. Entonces, los modelos de regresión son simplificados por el descenso de atributos si eso resulta en una error esperado más bajo en los datos futuros (más específicamente, si la disminución en el número de parámetros de salida supera el incremento en el error observado de entrenamiento). Después de que esto se ha hecho, cada subárbol es considerado para poda. La poda ocurre si el error estimado para el modelo lineal en la raíz de un subárbol es más pequeño o igual al error esperado de todo el subárbol. Después de que la poda termina, M5' aplica un proceso de eliminación de ruido que combina el modelo de cada hoja con los modelos de la ruta a la raíz para formar el modelo final que es colocado en cada hoja [166]. Karalic (1992) [167] estudió la influencia de utilizar modelos de regresión lineal en las hojas de un árbol de regresión. Como en el trabajo de Quinlan, Karatic mostró que el mantener modelos pequeños puede incrementar el desempeño del árbol.

Descripción del algoritmo M5'

La figura A6.6 muestra el algoritmo M5' desde el punto de vista de un algoritmo genérico para inducción de árboles de decisión (lado izquierdo) y las funciones que convierten el algoritmo genérico en una instancia del algoritmo M5' (lado derecho). Desde este punto de vista, lo que diferencia a M5' de otros algoritmos que producen árboles de regresión es la generación del valor de la hoja (representado por la función local_model en la figura A6.6), la función de evaluación de una partición dada (función quality en la figura A6.6) y el criterio de parada (función stop_criterion).

Figura A6.6. Algoritmo M5' (tomado de Vens y Blockeel, 2006 [176])

```
function GROW_TREE(T: set of examples)
                        returns decision tree:
    if stop_criterion(T)
    then return leaf(local_model(T))
    else
        S := set of all possible splits
        s* := arg max_{s∈S} quality(s, T)
        P := partition induced on T by s*
        for all P_j in P: t_j := GROW_TREE(P_j)
        return node(s*, ⋃_j {(j, t_j)})

function PRUNE_TREE(t: tree; T: set of examples)
                        returns decision tree:
    m := local_model(T)
    for all t_j that is a child of t: t_j := PRUNE_TREE(t_j)
    if error(m) > error(t) then return t
    else return leaf(m)
```

```
function local_model(T: set of examples):
    return linear regression model for T
    (based on variables occurring in the subtree)

function quality(s: split; T: set of examples):
    P := partition induced on T by s
    return SD(T) − Σ_{T_j∈P} |T_j|/|T| SD(T_j)
    (SD(T) = target variable's standard deviation in T)

function stop_criterion(T: set of examples):
    return |T| < 4 or SD(T) < 0.05 * SD(all examples)
```

El funcionamiento del algoritmo es el siguiente:

1. Como entrada, se recibe un conjunto de instancias. Se devuelve un árbol de modelos de regresión (un árbol de decisión con modelos en sus hojas).
2. Se evalúa el criterio de parada para el conjunto de instancias recibidas. El criterio de parada se basa en el tamaño del conjunto de instancias (en la figura, el límite es 4 instancias) o bien, en la magnitud de la desviación estándar para el atributo a predecir (cuando ésta es menor al 5 % de la desviación presente en todo el conjunto).
3. Si el criterio de parada indica que se debe detener el crecimiento, entonces se realiza un modelo lineal con las variables que intervienen en el subárbol. Si el criterio de parada no indica que se deba detener el crecimiento, entonces todas las particiones posibles son evaluadas en busca de aquella que maximice la calidad del conjunto de instancias. La calidad se mide por la desviación presente en la partición, a menor desviación, el conjunto tiene una mayor calidad.
4. Una vez determinada la mejor partición, se llama de manera recursiva al algoritmo M5' para cada una de las particiones generadas. Estas particiones se "cuelgan" bajo un nodo único el cual es regresado como resultado del algoritmo. Para la prueba del nodo se utiliza el atributo más discriminante (el que reduce la varianza de la variable a predecir).
5. Finalmente, el árbol generado se somete al proceso de poda. Esto es importante, ya que ayuda a prevenir el sobre-entrenamiento del árbol. La poda se basa en la comparación del error generad por los modelos en las hojas del árbol.

Anexo 7. Atributos de la serie de datos objetivo seleccionados para el modelado.

A7.1. Selección final de atributos para la etapa de construcción de modelos.

#	Atributo	Unidades
1	Superficie sembrada	ha
2	Lámina de riego	cm
3	Rendimiento	Ton/ha
4	Temperatura de octubre	°C
5	Temperatura de noviembre	°C
6	Temperatura de diciembre	°C
7	Temperatura enero	°C
8	Temperatura febrero	°C
9	Temperatura marzo	°C
10	Temperatura abril	°C
11	Temperatura mayo	°C
12	Temperatura junio	°C
13	Temperatura julio	°C
14	Temperatura agosto	°C
15	Temperatura septiembre	°C
16	Temperatura máxima octubre	°C
17	Temperatura máxima noviembre	°C
18	Temperatura máxima diciembre	°C
19	Temperatura máxima enero	°C
20	Temperatura máxima febrero	°C
21	Temperatura máxima marzo	°C
22	Temperatura máxima abril	°C
23	Temperatura máxima mayo	°C
24	Temperatura máxima junio	°C
25	Temperatura máxima julio	°C
26	Temperatura máxima agosto	°C
27	Temperatura máxima septiembre	°C
28	Temperatura mínima octubre	°C
29	Temperatura mínima noviembre	°C
30	Temperatura mínima diciembre	°C
31	Temperatura mínima enero	°C
32	Temperatura mínima febrero	°C
33	Temperatura mínima marzo	°C
34	Temperatura mínima abril	°C
35	Temperatura mínima mayo	°C
36	Temperatura mínima junio	°C
37	Temperatura mínima julio	°C
38	Temperatura mínima agosto	°C
39	Temperatura mínima septiembre	°C

40	Precipitación octubre	mm
41	Precipitación noviembre	mm
42	Precipitación diciembre	mm
43	Precipitación enero	mm
44	Precipitación febrero	mm
45	Precipitación marzo	mm
46	Precipitación abril	mm
47	Precipitación mayo	mm
48	Precipitación junio	mm
49	Precipitación julio	mm
50	Precipitación agosto	mm
51	Precipitación septiembre	mm
52	Evaporación octubre	mm
53	Evaporación noviembre	mm
54	Evaporación diciembre	mm
55	Evaporación enero	mm
56	Evaporación febrero	mm
57	Evaporación marzo	mm
58	Evaporación abril	mm
59	Evaporación mayo	mm
60	Evaporación junio	mm
61	Evaporación julio	mm
62	Evaporación agosto	mm
63	Evaporación septiembre	mm

Anexo 8. Resultados de la regresión multivariable

Tabla A8.1. Coeficientes de la regresión lineal multivariable para cada cultivo.

		Cultivos														
		Alfalfa	Algodón	Cártamo	Chícharo	Forraje	Frijol	Frutales	Garbanzo	Hortalizas	Maíz	Papa	Sorgo	Tomate	Trigo	Zacate
Coeficientes	Superficie-Sembrada	-0.0017	0.0000	0.0000	0.0160	0.0138	-0.0005	-0.0865	-0.0004	0.0046	-0.0002	0.0026	-0.0065	-0.0013	0.0000	0.1016
	LaminaRiego	-0.0035	-0.0127	0.0034	0.0130	-0.0009	0.0000	0.0910	-0.0007	-0.0315	-0.0064	0.0196	0.0042	-0.0227	-0.0047	0.0494
	TempOct	-0.2429	0.0184	0.0441	-0.0037	0.0577	0.0089	0.0248	0.0374	0.5815	0.0899	0.1099	0.0120	-0.4920	0.0050	-0.1124
	TempNov	-0.1065	0.0258	0.0211	0.0433	0.0080	-0.0016	-0.1270	0.0045	0.2268	-0.0152	0.2415	0.0289	0.0295	0.0173	-0.0478
	TempDic	-0.0220	-0.0047	0.0034	0.5504	0.0432	-0.0028	0.0713	-0.0190	0.0837	0.0122	0.2856	0.0046	0.4682	0.0451	-0.0789
	TempEne	-0.0486	0.0085	0.0019	0.1495	0.0507	0.0019	-0.0428	-0.0050	0.0474	0.0389	0.2590	0.0313	0.3265	-0.0016	-0.0757
	TempFeb	0.0363	-0.0033	0.0065	0.1130	-0.0080	-0.0089	0.0799	-0.0296	0.0024	0.0570	0.2660	-0.0288	0.0593	-0.0188	0.0194
	TempMar	-0.0387	0.0074	-0.0077	0.1842	-0.0014	-0.0037	0.1053	-0.0035	0.2034	0.0029	0.0559	0.0294	0.1024	0.0260	-0.0555
	TempAbr	-0.0283	0.0063	0.0031	0.1683	-0.0233	-0.0017	0.0390	-0.0104	0.1184	0.0581	0.2436	0.0235	0.1063	0.0092	0.0121
	TempMay	-0.0056	0.0034	-0.0028	0.0938	-0.0140	0.0008	0.0343	-0.0083	0.0462	0.0448	0.1139	0.0056	0.2143	-0.0039	0.0517
	TempJun	0.0197	0.0070	-0.0200	0.0041	-0.0576	-0.0025	0.0248	-0.0040	0.0342	0.0168	-0.1096	-0.0037	0.0911	-0.0129	0.1109
	TempJul	-0.0024	0.0392	-0.0167	0.1506	0.1368	-0.0049	-0.0338	-0.0025	0.0929	0.0323	-0.0106	0.0296	0.1023	-0.0070	0.1323
	TempAgo	-0.0352	0.0038	0.0077	0.1566	0.2492	-0.0010	0.0430	0.0002	0.1114	0.0129	0.1641	0.0215	0.0368	0.0155	-0.1710
	TempSep	-0.0341	0.0120	0.0119	0.1867	0.2383	-0.0015	0.0184	-0.0026	0.1300	0.0430	0.2569	0.0483	0.0403	0.0175	-0.1326
	TempMaxOct	-0.2001	0.0106	0.0252	-0.3160	-0.0246	0.0298	0.2162	0.0571	0.3868	-0.0014	-0.4199	-0.0448	-0.4773	-0.0556	-0.0691
	TempMaxNov	-0.0536	0.0123	0.0244	-0.2874	-0.1097	0.0081	0.0673	0.0088	-0.0051	0.0648	0.1132	0.0033	-0.1537	-0.0629	-0.0101
	TempMaxDic	0.0612	-0.0039	-0.0188	-0.0991	-0.0647	-0.0017	0.0635	0.0027	0.0009	0.0205	-0.2800	-0.0077	-0.1237	-0.0169	0.0613
	TempMaxEne	0.0225	0.0003	0.0415	-0.0332	-0.0499	-0.0060	-0.0167	-0.0140	-0.1053	0.0495	0.4691	-0.0151	-0.3025	-0.0123	0.0943
	TempMaxFeb	0.0310	-0.0029	0.0145	0.0302	-0.0550	-0.0074	0.0526	-0.0054	-0.0187	-0.0103	0.0220	-0.0179	-0.2459	0.0064	0.0504
	TempMaxMar	-0.0302	-0.0012	0.0220	-0.0289	-0.0270	0.0055	0.0605	0.0118	-0.0519	-0.0176	0.0185	0.0039	0.2277	0.0203	-0.0632
	TempMaxAbr	0.1314	-0.0118	-0.0108	0.1550	-0.1584	-0.0106	0.2393	-0.0091	-0.0562	0.0469	-0.1446	-0.0266	-0.1535	-0.0052	0.4214
	TempMaxMay	0.0015	0.0013	0.0029	-0.1381	-0.0576	0.0038	-0.0141	0.0128	-0.0807	-0.0016	-0.1061	0.0077	-0.0360	-0.0036	0.0560
	TempMaxJun	0.0522	-0.0178	0.0393	-0.2347	0.0482	0.0043	-0.0913	0.0045	-0.3733	-0.0255	0.0894	-0.0420	-0.2028	-0.0195	-0.0567
	TempMaxJul	0.0199	-0.0038	-0.0128	-0.3795	0.1456	0.0173	-0.1204	0.0147	-0.7481	-0.0514	-0.5132	-0.1007	-0.0591	-0.0875	0.0716
	TempMaxAgo	0.0723	-0.0100	-0.0315	-0.4156	0.1266	0.0132	-0.1486	0.0132	-0.4630	-0.0594	-0.6121	-0.0552	-0.1191	-0.0734	0.0491
	TempMaxSep	0.0301	0.0005	-0.0327	-0.2141	0.0488	0.0021	-0.1092	0.0029	-0.0506	-0.0133	-0.3192	-0.0119	-0.1563	-0.0327	0.0884
	TempMinOct	-0.1662	0.0140	0.0279	-0.1847	0.0149	0.0240	-0.0885	0.0955	0.3834	-0.0477	-0.3918	0.0194	-1.0105	0.0508	-0.0400
	TempMinNov	-0.0695	0.0104	0.0432	0.0122	-0.0325	0.0010	-0.0867	0.0073	0.0350	-0.0022	0.4112	0.0274	0.0126	0.0193	-0.0322
	TempMinDic	0.0397	-0.0134	-0.0157	-0.0769	0.0151	0.0017	0.1244	-0.0156	-0.0142	0.0117	-0.2115	-0.0702	-0.3286	-0.0207	0.0194
	TempMinEne	-0.0001	-0.0019	0.0372	0.3602	0.2145	-0.0052	-0.0587	-0.0221	-0.0203	0.1249	0.6272	0.0067	-0.8332	-0.0014	0.0080

	TempMinFeb	0.0311	-0.0069	0.0189	-0.0320	-0.0173	-0.0026	0.0813	-0.0218	-0.0849	0.0406	0.2115	-0.0490	-0.2744	-0.0219	0.0364
	TempMinMar	-0.0042	-0.0087	-0.0136	-0.0938	0.0061	0.0099	0.1214	0.0081	-0.0234	-0.0365	-0.3630	-0.0440	0.4030	0.0050	-0.0597
	TempMinAbr	-0.0005	0.0000	0.0230	0.1425	-0.0738	0.0017	0.1015	-0.0110	0.0222	0.0840	0.3165	-0.0061	0.2256	0.0043	0.1683
	TempMinMay	0.0040	-0.0059	-0.0063	-0.0869	-0.0331	0.0084	0.0652	0.0093	-0.0188	0.0035	-0.1976	-0.0062	0.0135	0.0107	0.0305
	TempMinJun	0.0242	0.0008	-0.0208	-0.0964	-0.0677	0.0017	0.0220	-0.0008	-0.0269	0.0065	-0.2262	-0.0203	-0.0122	-0.0200	0.1243
	TempMinJul	0.4461	0.0076	-0.2917	-0.7972	0.2565	-0.0056	-0.1969	0.0238	-1.2058	-0.1817	-3.8832	-0.1776	-0.0423	-0.2858	0.4813
	TempMinAgo	0.0844	-0.0148	-0.0421	-0.3569	0.2370	0.0023	-0.0144	0.0121	-0.3664	-0.1068	-0.8635	-0.0713	-0.6750	-0.0673	-0.1105
	TempMinSep	0.0299	-0.0117	-0.0313	-0.0754	0.2157	0.0083	-0.0041	0.0093	-0.0924	-0.0207	-0.4396	-0.0199	-0.3200	-0.0181	-0.0873
	PrecipOct	0.0027	-0.0012	-0.0006	0.0019	0.0007	0.0000	-0.0022	-0.0002	0.0133	0.0009	0.0003	0.0008	-0.0015	0.0009	-0.0002
	PrecipNov	-0.0014	0.0001	-0.0008	-0.0121	0.0040	0.0002	-0.0031	-0.0001	-0.0014	-0.0032	-0.0027	0.0018	0.0239	-0.0007	-0.0114
	PrecipDic	0.0010	-0.0002	-0.0012	-0.0061	0.0030	-0.0001	0.0003	-0.0005	-0.0020	-0.0018	-0.0106	-0.0006	0.0120	-0.0011	-0.0039
	PrecipEne	0.0089	0.0023	-0.0045	0.0184	0.0456	-0.0019	0.0347	0.0029	0.0284	-0.0095	-0.0543	0.0041	0.0027	0.0092	-0.0003
	PrecipFeb	0.0130	-0.0015	-0.0054	-0.0023	0.0159	-0.0012	-0.0059	-0.0047	0.0016	0.0041	0.0083	0.0046	0.0124	-0.0053	-0.0156
	PrecipMar	0.1633	0.0148	-0.0455	0.1131	0.0405	-0.0141	-0.1727	0.0121	0.1244	-0.0619	-0.4174	0.0352	0.0428	0.0196	0.0270
	PrecipAbr	-0.0333	0.0104	0.0251	0.0454	0.0142	0.0016	-0.0137	-0.0033	-0.0309	0.0399	0.3124	0.0311	0.0925	0.0050	-0.1061
	PrecipMay	-0.0488	0.0034	-0.0149	-0.1222	0.0266	0.0129	-0.0128	-0.0088	0.0555	0.0465	0.2552	0.0678	0.4832	-0.0016	-0.1591
	PrecipJun	-0.0186	0.0017	0.0054	0.1134	0.0078	-0.0011	-0.0338	0.0067	0.0522	-0.0144	0.0516	0.0185	-0.0753	0.0176	-0.0369
	PrecipJul	0.0021	-0.0003	-0.0001	-0.0004	-0.0037	-0.0001	0.0007	-0.0002	-0.0012	0.0006	0.0001	0.0004	-0.0005	0.0001	0.0003
	PrecipAgo	0.0012	-0.0001	-0.0008	-0.0159	-0.0006	0.0000	0.0008	-0.0005	-0.0044	-0.0003	-0.0037	0.0013	0.0263	-0.0012	-0.0042
	PrecipSep	0.0014	-0.0005	-0.0004	-0.0024	0.0007	0.0001	0.0016	-0.0003	-0.0031	0.0005	-0.0024	-0.0001	0.0007	-0.0006	-0.0010
	EvaporacionOct	-0.0017	0.0004	-0.0017	-0.0140	0.0000	0.0005	0.0055	0.0015	0.0066	-0.0008	-0.0331	0.0001	-0.0091	-0.0050	-0.0057
	EvaporacionNov	0.0015	0.0003	-0.0029	-0.0153	0.0001	0.0005	0.0071	0.0016	0.0082	0.0005	-0.0423	0.0003	-0.0113	-0.0072	-0.0064
	EvaporacionDic	-0.0008	0.0006	-0.0028	-0.0207	0.0002	0.0005	0.0092	0.0017	0.0132	-0.0006	-0.0437	0.0008	-0.0104	-0.0064	-0.0100
	EvaporacionEne	-0.0005	-0.0014	0.0017	0.0165	0.0004	-0.0003	0.0143	-0.0013	0.0030	0.0017	0.0246	0.0020	0.0251	0.0013	-0.0230
	EvaporacionFeb	-0.0030	0.0046	-0.0026	0.0082	0.0049	-0.0004	0.0560	-0.0031	0.0391	0.0033	0.0230	0.0013	0.0264	-0.0002	-0.0241
	EvaporacionMar	0.0135	-0.0006	-0.0001	0.0317	0.0260	-0.0014	0.1744	-0.0038	0.0095	0.0134	0.0380	0.0064	0.0263	-0.0014	0.0071
	EvaporacionAbr	0.0200	0.0007	-0.0016	0.0455	0.0262	-0.0022	0.3239	-0.0024	0.0240	0.0116	0.0199	0.0086	0.0119	0.0013	0.0067
	EvaporacionMay	0.0031	0.0000	-0.0015	0.0016	0.0000	-0.0003	-0.0050	-0.0006	0.0012	0.0022	-0.0062	0.0003	-0.0045	-0.0011	0.0048
	EvaporacionJun	-0.0024	-0.0004	0.0016	0.0018	-0.0001	0.0002	0.0046	0.0007	-0.0014	-0.0024	0.0048	0.0004	0.0043	0.0017	-0.0081
	EvaporacionJul	0.0059	0.0005	-0.0005	-0.0045	-0.0009	-0.0003	0.0154	0.0004	-0.0307	-0.0016	-0.0221	-0.0023	0.0008	-0.0024	0.0061
	EvaporacionAgo	0.0006	-0.0004	0.0009	-0.0013	-0.0001	0.0000	0.0044	0.0004	-0.0110	-0.0012	-0.0032	-0.0003	-0.0012	0.0000	-0.0041
	EvaporacionSep	0.0082	-0.0065	-0.0002	-0.0009	0.0001	-0.0004	0.0223	0.0001	-0.0191	0.0008	-0.0124	-0.0001	-0.0044	-0.0013	0.0255
	(intersección)	0.7419	3.6524	8.6242	70.5330	-30.8209	0.1115	-92.0783	-3.1449	82.5821	4.2337	202.2917	16.8641	107.3789	27.5319	-22.0795

Tabla A8.2. Métricas calculadas para la regresión multivariable (validación simple).

Cultivo	Eficiencia promedio (%)	Error absoluto medio (ton/ha)	Raíz del error cuadrático medio (ton/ha)	Error relativo absoluto (%)	Raíz del error cuadrático relativo (%)	Número de instancias
Alfalfa	85.128	2.43	2.82	126.93	115.73	13
Algodón	82.640	0.55	0.65	83.48	94.01	5
Cártamo	72.805	0.44	0.56	80.43	85.62	30
Chícharo	40.708	5.83	7.29	146.27	133.15	6
Forraje	39.407	4.19	4.70	131.32	133.91	3
Fríjol	75.088	0.43	0.52	100.81	99.55	26
Frutales	54.847	5.63	6.44	103.00	92.41	5
Garbanzo	62.471	0.65	0.80	111.70	118.91	20
Hortalizas	45.343	9.17	11.18	169.34	174.12	21
Maíz	88.874	26.81	33.07	102.75	88.93	33
Papa	85.328	4.67	6.20	112.56	109.59	21
Sorgo	81.087	1.06	1.37	127.26	105.07	11
Tomate	46.699	10.57	13.95	88.41	90.78	16
Trigo	89.268	0.56	0.71	96.07	94.56	33
Zacate	0.000	21.83	25.37	1247.29	922.42	3
Promedios	**63.313**	**6.32**	**7.71**	**188.51**	**163.92**	**246**

Anexo 9. Construcción de modelos de redes neuronales perceptrón multicapa

Notas para la lectura de las tablas A9.1,A9.2,A9.3 y A9.4 del anexo 9:

- Las tablas muestran los coeficientes de correlación entre las estimaciones del modelo de red neuronal y los valores reales del conjunto de prueba (generado por medio de validación cruzada).
- Los resultados se muestran por configuraciones, una configuración de red se establece por el número de nodos en la capa oculta. Los resultados se muestran también por ciclos (epochs) invertidos en el tiempo de aprendizaje.
- Se resaltan en negrillas las mejores aproximaciones en las estimaciones del conjunto de prueba de la configuración de red correspondiente.
- Se resalta en negrillas y cursivas si existe un resultado que haya superado a los resultados de otras configuraciones (el mejor resultado global).

Tabla A9.1 Coeficientes de correlación entre los resultados para las redes perceptrón multicapa con un nodo en la capa oculta.

Nodos ocultos: 1													
Cultivos	Número de instancias	Ciclos de aprendizaje											
		Otra (ciclos:cr)		50	100	125	150	175	200	500	Otra (ciclos:cr)		
Alfalfa	63			**0.176**	0.136	0.123	0.116	0.114	0.113	0.113			
Algodón	23			-0.405	-0.340	-0.232	-0.077	-0.042	**-0.004**	-0.004			
Cartamo	150	20:	***0.244***	0.237	0.228	0.225	0.223	0.221	0.220	0.208			
Chicharo	27	20:	***0.183***	0.138	0.108	0.082	0.064	0.061	0.063	0.088	4500:	0.139	
Forraje	11			**-0.377**	-0.517	-0.544	-0.539	-0.546	-0.540	-0.478			
Frijol	129			-0.220	-0.228	-0.229	-0.231	-0.233	-0.237	-0.198	1200:	**0.078**	
Frutales	23			-0.184	-0.015	0.098	**0.103**	0.094	0.092	0.097			
Garbanzo	96			0.018	0.042	0.051	0.059	0.067	0.074	0.118	'7000:	**0.235**	
Hortalizas	102			**0.323**	0.322	0.318	0.314	0.309	0.305	0.318			
Maiz	161			**0.317**	0.312	0.312	0.309	0.307	0.302	0.289			
Papa	105			0.006	0.013	0.015	0.020	0.026	0.031	0.057	10000:	**0.091**	
Sorgo	51	20:	**0.203**	0.184	0.144	0.112	0.130	0.106	0.104	0.099			
Tomate	77	40:	**0.231**	0.228	0.217	0.206	0.194	0.190	0.187	0.168			
Trigo	156			0.336	0.345	0.350	0.356	0.360	0.362	0.346	225:	***0.364***	
Zacate	11			0.656	0.674	0.557	0.654	0.465	***0.687***	0.647			

Tabla A9.2 Coeficientes de correlación entre los resultados para las redes perceptrón multicapa con tres nodos en la capa oculta.

Nodos ocultos: 3												
Cultivos	Número de instancias	Ciclos de aprendizaje										
		Otra (<)		50	100	125	150	175	200	500	Otra (>)	
Alfalfa	63			0.172	**0.178**	0.176	0.156	0.141	0.141	0.048		
Algodón	23			-0.159	-0.064	-0.010	0.076	0.119	0.119	0.264	1000:	**0.304**
Cartamo	150			**0.233**	0.222	0.217	0.211	0.207	0.207	0.169		
Chicharo	27	30:	**0.164**	0.152	0.135	0.123	0.113	0.103	0.106	0.121	600:	0.131
Forraje	11			-0.510	**-0.475**	-0.543	-0.639	-0.646	-0.616	-0.608		
Frijol	129			-0.130	-0.118	-0.109	-0.103	-0.095	-0.087	-0.011	2100:	**0.103**
Frutales	23			-0.139	0.017	-0.053	-0.044	-0.096	-0.013	-0.020	3500:	**0.272**

Garbanzo	96			0.040	0.058	0.061	0.064	0.063	0.064	0.088	'8000:	**0.135**
Hortalizas	102			0.300	0.364	**0.373**	0.371	0.361	0.356	0.268		
Maiz	161			**0.354**	0.344	0.343	0.338	0.331	..3327	0.287		
Papa	105			-0.032	-0.016	-0.014	-0.016	-0.017	-0.014	0.055	10000:	**0.077**
Sorgo	51	10:	**0.209**	0.143	0.064	0.056	0.051	0.026	-0.027	-0.154		
Tomate	77	40:	**0.274**	0.270	0.252	0.246	0.236	0.234	0.233	0.226	2000:	0.269
Trigo	156			**0.342**	0.334	0.328	0.324	0.321	0.320	0.314		
Zacate	11			**0.612**	0.589	0.586	0.585	0.590	0.587	0.575	2000:	0.573

Tabla A9.3 Coeficientes de correlación entre los resultados para las redes perceptrón multicapa con diez nodos en la capa oculta.

Nodos ocultos: 10											
Cultivos	**Número de instancias**	**Ciclos de aprendizaje**									
		Otra (<)	**50**	**100**	**125**	**150**	**175**	**200**	**500**	**Otra (>)**	
Alfalfa	63		0.226	***0.241***	0.230	0.216	0.190	0.164	0.027		
Algodón	23		-0.129	0.052	0.027	0.069	0.105	0.131	0.323	600:	***0.365***
Cartamo	150		0.204	0.214	0.217	0.219	0.222	0.219	0.231	600:	**0.235**
Chicharo	27		0.161	0.153	0.149	0.160	0.154	**0.163**	0.114		
Forraje	11		**-0.445**	-0.617	-0.582	-0.637	-0.617	-0.602	-0.568		
Frijol	129		-0.071	-0.052	-0.042	-0.032	-0.022	-0.012	0.052	3300:	***0.109***
Frutales	23		-0.047	0.137	0.096	0.086	0.153	0.123	0.182	1500:	**0.393**
Garbanzo	96		0.023	0.023	0.029	0.020	0.032	***0.243***	0.060		
Hortalizas	102		0.341	0.374	***0.380***	0.369	0.357	0.341	0.207		
Maiz	161		0.379	**0.381**	0.379	0.377	0.376	0.371	0.352		
Papa	105		0.028	-0.011	0.000	0.028	-0.009	-0.010	0.076	12000:	**0.131**
Sorgo	51		0.145	0.145	0.143	0.142	0.156	0.171	-0.009	250:	**0.178**
Tomate	77		0.261	0.254	0.237	0.221	0.215	***0.285***	0.206		
Trigo	156		0.247	**0.281**	0.274	0.267	0.250	0.242	0.196		
Zacate	11		0.618	0.588	**0.631**	0.616	0.597	0.629	0.620		

Tabla A9.4 Coeficientes de correlación entre los resultados para las redes perceptrón multicapa con quince nodos en la capa oculta.

Nodos ocultos: 15												
Cultivos	**Número de instancias**	**Ciclos de aprendizaje**										
		Otra (<)		**50**	**100**	**125**	**150**	**175**	**200**	**500**	**Otra (>)**	
Alfalfa	63			0.221	**0.231**	0.216	0.199	0.183	0.176	0.058		
Algodón	23			0.012	0.003	0.031	0.057	0.078	0.095	**0.286**		
Cartamo	150			0.189	0.180	0.180	0.192	0.190	0.186	**0.193**		
Chicharo	27	40:	**0.127**	0.120	0.107	0.101	0.107	0.102	0.106	0.082		
Forraje	11			***0.109***	***0.109***	-0.153	-0.554	-0.554	-0.541	-0.397		
Frijol	129			-0.058	-0.030	-0.011	-0.008	0.012	0.015	0.059	3000:	**0.103**
Frutales	23			0.010	0.061	0.078	0.057	0.042	0.213	0.223	1000:	***0.418***
Garbanzo	96	4000:	0.106	0.020	0.017	0.016	0.020	0.014	0.015	0.024	6000:	**0.115**
Hortalizas	102			0.300	**0.323**	0.319	0.307	0.292	0.277	0.192		
Maiz	161			0.378	0.368	0.381	***0.398***	0.369	0.389	0.345		
Papa	105			0.010	-0.009	-0.017	-0.030	-0.018	-0.041	0.045	1000:	***0.157***
Sorgo	51			0.091	0.169	0.152	0.155	0.170	***0.217***	-0.056	210:	0.214
Tomate	77			**0.189**	0.188	-0.188	0.188	-0.188	0.188	0.035		
Trigo	156			**0.296**	0.257	0.198	0.214	0.193	0.181	0.146		
Zacate	11			**0.675**	0.592	0.603	0.593	0.594	0.592	0.595		

Nota: En el caso del forraje, fue necesaria la configuración de red de 40 nodos y un ciclo de entrenamiento para obtener un coeficiente de correlación de 0.1533.

Tabla A.9.5. Métricas de error calculadas para las redes perceptrón multicapa por el método de validación simple.

Cultivo	Coeficiente de correlación	Eficiencia promedio (%)	Error absoluto medio (ton/ha)	Raíz del error cuadrático medio (ton/ha)	Error relativo absoluto (%)	Raíz del error cuadrático relativo (%)	Número de instancias
Alfalfa	-0.289	66.931	5.51	6.83	287.63	280.59	13
Algodón	0.251	33.423	2.03	2.44	308.42	353.26	5
Cártamo	0.579	77.190	0.48	0.66	87.08	100.82	30
Chícharo	-0.407	56.530	4.56	5.94	114.45	108.38	6
Forraje	0.998	0.000	2.76	2.76	8.65	7.88	3
Fríjol	0.232	0.000	6.47	7.76	1512.41	1488.33	26
Frutales	0.235	44.477	5.51	8.20	100.84	117.66	5
Garbanzo	0.156	66.554	0.55	0.67	94.72	99.87	20
Hortalizas	0.515	20.405	11.54	13.44	213.19	209.19	21
Maíz	0.502	96.844	6.61	8.17	441.94	464.65	33
Papa	-0.252	52.966	13.42	17.50	323.54	309.41	21
Sorgo	0.456	52.895	2.49	3.04	298.72	233.70	11
Tomate	0.371	45.978	11.73	15.14	98.12	98.51	16
Trigo	0.158	88.074	0.58	0.73	100.26	97.71	33
Zacate	-0.995	68.715	3.65	3.66	208.51	132.97	3
Promedios		**51.40**	**5.19**	**6.46**	**279.90**	**273.53**	**246**

Anexo 10. Construcción de modelos de árbol J48

Tabla A10.1. Prueba de la discretización del atributo rendimiento.

Número de intervalos	Tamaño del intervalo (ton/ha)	Hojas (número de reglas)	Tamaño del árbol	Instancias clasificadas (%)		Error de predicción promedio	Raíz de la media del error cuadrático	Error absoluto relativo	Raíz del error relativo cuadrático	Número total de instancias
				Bien	Mal					
10	4.15	65	118	88.80	11.20	0.0331	0.1286	22.5571	47.5482	991
20	2.08	114	216	76.49	23.51	0.0334	0.1293	39.0278	62.5394	991
30	1.38	93	174	74.77	25.23	0.0246	0.1109	43.4755	66.0719	991
40	1.04	138	264	70.64	29.36	0.0206	0.1016	45.9314	67.8853	991
50	0.83	128	244	69.12	30.88	0.0174	0.0932	48.2551	69.6132	991
60	0.69	159	306	67.61	32.39	0.0147	0.0859	48.1692	69.5445	991
70	0.59	171	330	65.89	34.11	0.0132	0.0812	49.7533	70.6802	991
80	0.52	191	370	64.28	35.72	0.0118	0.0768	50.1784	70.9687	991
90	0.46	205	398	64.88	35.12	0.0103	0.0717	49.1976	70.2883	991
100	0.42	189	366	64.38	35.62	0.0095	0.0689	50.0673	70.9066	991
110	0.38	220	428	64.08	35.92	0.0086	0.0655	49.6223	70.5871	991
120	0.35	208	404	62.66	37.34	0.0082	0.0641	51.5563	71.9522	991
120	0.35	203	394	62.26	37.74	0.0077	0.0621	52.5083	72.6213	991
140	0.30	221	430	61.15	38.85	0.0072	0.0601	52.6085	72.6806	991
150	0.28	228	444	61.65	38.35	0.0066	0.0576	51.8126	72.1503	991
160	0.26	227	442	60.54	39.46	0.0064	0.0564	52.8738	72.8714	991
170	0.24	230	448	62.36	37.64	0.0058	0.0537	50.9064	71.5187	991
180	0.23	235	458	60.54	39.46	0.0056	0.0531	52.4453	72.5828	991
190	0.22	228	444	60.54	39.46	0.0054	0.0520	53.1249	73.0640	991
200	0.21	242	472	61.15	38.85	0.0050	0.0498	51.2675	71.7677	991

Figura A10.1. Árbol J48 generado para la información del cultivo Alfalfa

```
SuperficieSembrada <= 142
|   PrecipMay <= 0
|   |   TempJul <= 25.65
|   |   |   SuperficieSembrada <= 26.65
|   |   |   |   SuperficieSembrada <= 21: '(14.25-15.208333]' (2.0/1.0)
|   |   |   |   SuperficieSembrada > 21: '(12.333333-13.291667]' (2.0/1.0)
|   |   |   SuperficieSembrada > 26.65: '(11.375-12.333333]' (5.0/2.0)
|   |   TempJul > 25.65
|   |   |   LaminaRiego <= 84.8
|   |   |   |   TempDic <= 10.3913: '(14.25-15.208333]' (4.0/1.0)
|   |   |   |   TempDic > 10.3913
|   |   |   |   |   LaminaRiego <= 72.12: '(11.375-12.333333]' (3.0/1.0)
|   |   |   |   |   LaminaRiego > 72.12: '(14.25-15.208333]' (3.0/2.0)
|   |   |   LaminaRiego > 84.8
|   |   |   |   SuperficieSembrada <= 19.36: '(15.208333-16.166667]' (3.0)
```

```
|  |  |  |  SuperficieSembrada > 19.36
|  |  |  |  |  TempNov <= 14.5312
|  |  |  |  |  |  SuperficieSembrada <= 45.27: '(15.208333-16.166667]' (3.0/1.0)
|  |  |  |  |  |  SuperficieSembrada > 45.27: '(11.375-12.333333]' (3.0/1.0)
|  |  |  |  |  TempNov > 14.5312
|  |  |  |  |  |  SuperficieSembrada <= 112.04: '(13.291667-14.25]' (6.0/3.0)
|  |  |  |  |  |  SuperficieSembrada > 112.04: '(14.25-15.208333]' (2.0/1.0)
|  PrecipMay > 0
|  |  LaminaRiego <= 131.5
|  |  |  SuperficieSembrada <= 45.27: '(13.291667-14.25]' (3.0/1.0)
|  |  |  SuperficieSembrada > 45.27: '(14.25-15.208333]' (3.0/1.0)
|  |  LaminaRiego > 131.5: '(-inf-9.458333]' (2.0/1.0)
SuperficieSembrada > 142
|  TempDic <= 10.3913: '(12.333333-13.291667]' (4.0)
|  TempDic > 10.3913: '(11.375-12.333333]' (2.0/1.0)
```

Tabla A10.2. Resultados de las métricas aplicadas para la evaluación de la precisión numérica de las reglas (validación simple).

Cultivo	Eficiencia promedio (%)	Error absoluto medio (ton/ha)	Raíz del error cuadrático medio (ton/ha)	Error relativo absoluto (%)	Raíz del error cuadrático relativo (%)	Número de instancias
Alfalfa	82.90	2.988	4.457	259.238	297.053	13
Algodón	76.93	0.735	0.778	145.910	121.386	5
Cartamo	64.33	0.567	0.682	104.537	104.847	30
Chicharo	57.62	4.547	6.357	131.580	169.266	6
Forraje	51.01	2.712	4.144	82.791	118.473	3
Frijol	55.74	0.756	0.855	188.976	173.960	26
Frutales	31.76	6.593	9.315	117.730	136.156	5
Garbanzo	54.09	0.674	0.848	5.417	6.159	20
Hortalizas	47.67	7.627	9.047	53.315	61.962	21
Maiz	78.16	1.684	2.063	186.064	165.317	33
Papa	81.46	6.000	6.879	149.922	136.131	21
Sorgo	67.40	1.750	2.240	224.356	202.051	11
Tomate	14.91	14.687	17.316	142.848	131.551	16
Trigo	84.94	0.729	0.922	122.433	130.121	33
Zacate	81.56	2.346	2.562	105.577	108.678	3
Total	**62.03**	**3.626**	**4.564**	**134.713**	**137.541**	**246**

Anexo 11. Construcción de árboles de regresión M5

Figura A11.1. Árbol de regresión M5 generado para la información del cultivo alfalfa.

```
SuperficieSembrada <= 94 :
|   TempAbr <= 13.796 :
|   |   LaminaRiego <= 87.95 :
|   |   |   LaminaRiego <= 73.4 :
|   |   |   |   LaminaRiego <= 71.06 :
|   |   |   |   |   SuperficieSembrada <= 11.5 : 14.6033 (2/116.139%)
|   |   |   |   |   SuperficieSembrada >  11.5 : 14.6046 (2/0%)
|   |   |   |   LaminaRiego >  71.06 : 14.6262 (2/29.035%)
|   |   |   LaminaRiego >  73.4 :
|   |   |   |   SuperficieSembrada <= 28.635 : 14.604 (2/14.517%)
|   |   |   |   SuperficieSembrada >  28.635 : 14.604 (2/14.517%)
|   |   LaminaRiego >  87.95 :
|   |   |   SuperficieSembrada <= 48.5 :
|   |   |   |   SuperficieSembrada <= 19.43 : 14.3877 (2/11.614%)
|   |   |   |   SuperficieSembrada >  19.43 :
|   |   |   |   |   SuperficieSembrada <= 27.22 : 14.3615 (3/53.432%)
|   |   |   |   |   SuperficieSembrada >  27.22 :
|   |   |   |   |   |   SuperficieSembrada <= 41 : 14.3734 (2/0%)
|   |   |   |   |   |   SuperficieSembrada >  41 : 14.3725 (2/29.035%)
|   |   |   SuperficieSembrada >  48.5 :
|   |   |   |   SuperficieSembrada <= 61 : 14.2665 (2/0%)
|   |   |   |   SuperficieSembrada >  61 : 14.2716 (2/58.069%)
|   TempAbr >  13.796 :
|   |   LaminaRiego <= 89.115 :
|   |   |   LaminaRiego <= 81.12 : 14.2858 (2/0%)
|   |   |   LaminaRiego >  81.12 : 14.3071 (3/72.425%)
|   |   LaminaRiego >  89.115 :
|   |   |   LaminaRiego <= 121.485 :
|   |   |   |   SuperficieSembrada <= 14 : 14.596 (2/23.228%)
|   |   |   |   SuperficieSembrada >  14 :
|   |   |   |   |   SuperficieSembrada <= 26.9 : 14.6152 (2/29.035%)
|   |   |   |   |   SuperficieSembrada >  26.9 : 14.6172 (3/119.321%)
|   |   |   LaminaRiego >  121.485 :
|   |   |   |   TempOct <= 20.525 : 14.4655 (2/31.938%)
|   |   |   |   TempOct >  20.525 : 14.4683 (2/275.829%)
SuperficieSembrada >  94 :
|   TempAbr <= 12.725 :
|   |   LaminaRiego <= 113.825 : 13.9052 (2/58.069%)
|   |   LaminaRiego >  113.825 :
|   |   |   LaminaRiego <= 141.35 : 13.9326 (2/14.517%)
|   |   |   LaminaRiego >  141.35 : 13.9273 (2/29.035%)
|   TempAbr >  12.725 :
|   |   TempDic <= 11.396 : 13.7671 (3/10.95%)
|   |   TempDic >  11.396 : 13.7545 (2/0%)
```

Tabla A11.1. Resultados de las métricas para la evaluación de los árboles de regresión (validación simple).

Cultivo	Número de reglas	Eficiencia promedio (%)	Error absoluto medio (ton/ha)	Raíz del error cuadrático medio	Error absoluto relativo (%)	Raíz del error cuadrático relativo (%)	Número de instancias
Alfalfa	23	87.78	2.08	2.56	108.58	105.05	13.00
Algodón	7	77.54	0.62	0.67	93.74	97.14	5.00
Cártamo	51	68.08	0.54	0.64	97.18	98.12	30.00
Chícharo	9	64.16	3.97	5.49	99.45	100.17	6.00
Forraje	4	52.62	3.34	3.67	104.54	104.69	3.00
Fríjol	44	74.07	0.42	0.52	98.43	99.74	26.00
Frutales	8	51.73	5.18	6.74	94.70	96.78	5.00
Garbanzo	32	64.85	0.56	0.65	96.81	96.97	20.00

Hortalizas	32	65.78	5.18	6.58	95.73	102.50	21.00
Maíz	56	81.28	1.47	1.75	98.57	99.58	33.00
Papa	35	86.82	4.28	5.43	103.19	95.97	21.00
Sorgo	7	87.44	0.78	1.20	93.64	92.23	11.00
Tomate	27	46.25	11.63	15.12	97.32	98.40	16.00
Trigo	53	89.89	0.48	0.64	83.23	85.69	33.00
Zacate	3	80.65	1.94	2.49	111.04	90.45	3.00
Total	**391**	71.93	**2.83**	**3.61**	**98.41**	**97.57**	**246.00**

Anexo 12. Evaluación de resultados

Tabla A12.1. Resultados de la métrica de efectividad promedio para cada una de las técnicas de modelado (validación simple).

Cultivo	Regresión multivariable	Perceptrón multicapa	J48	M5
Alfalfa	85.13	66.93	82.90	87.78
Algodón	82.64	33.42	76.93	77.54
Cártamo	72.80	77.19	64.33	68.08
Chícharo	40.71	56.53	57.62	64.16
Forraje	39.41	0.00	51.01	52.62
Fríjol	75.09	0.00	55.74	74.07
Frutales	54.85	44.48	31.76	51.73
Garbanzo	62.47	66.55	54.09	64.85
Hortalizas	45.34	20.41	47.67	65.78
Maíz	88.87	96.84	78.16	81.28
Papa	85.33	52.97	81.46	86.82
Sorgo	81.09	52.90	67.40	87.44
Tomate	46.70	45.98	14.91	46.25
Trigo	89.27	88.07	84.94	89.89
Zacate	0.00	68.71	81.56	80.65
Promedios	63.31	51.40	62.03	71.93

Tabla A12.3. Resultados de la métrica de error absoluto medio para cada una de las técnicas de modelado (validación simple).

Cultivo	Regresión multivariable	Perceptrón multicapa	J48	M5
Alfalfa	2.43	5.51	2.99	2.08
Algodón	0.55	2.03	0.74	0.62
Cártamo	0.44	0.48	0.57	0.54
Chícharo	5.83	4.56	4.55	3.97
Forraje	4.19	2.76	2.71	3.34
Fríjol	0.43	6.47	0.76	0.42
Frutales	5.63	5.51	6.59	5.18
Garbanzo	0.65	0.55	0.67	0.56
Hortalizas	9.17	11.54	7.63	5.18
Maíz	26.81	6.61	1.68	1.47
Papa	4.67	13.42	6.00	4.28
Sorgo	1.06	2.49	1.75	0.78
Tomate	10.57	11.73	14.69	11.63
Trigo	0.56	0.58	0.73	0.48
Zacate	21.83	3.65	2.35	1.94
Promedios	**6.32**	**5.19**	**3.63**	**2.83**

Tabla A12.3. Resultados de la métrica de error absoluto relativo para cada una de las técnicas de modelado (validación simple).

Cultivo	Regresión multivariable	Perceptrón multicapa	J48	M5
Alfalfa	126.93	287.63	259.24	108.58
Algodón	83.48	308.42	145.91	93.74
Cártamo	80.43	87.08	104.54	97.18
Chícharo	146.27	114.45	131.58	99.45
Forraje	131.32	8.65	82.79	104.54
Fríjol	100.81	1,512.41	188.98	98.43
Frutales	103.00	100.84	117.73	94.70
Garbanzo	111.70	94.72	5.42	96.81
Hortalizas	169.34	213.19	53.31	95.73
Maíz	102.75	441.94	186.06	98.57
Papa	112.56	323.54	149.92	103.19
Sorgo	127.26	298.72	224.36	93.64
Tomate	88.41	98.12	142.85	97.32
Trigo	96.07	100.26	122.43	83.23
Zacate	1,247.29	208.51	105.58	111.04
Promedios	**188.51**	**279.90**	**134.71**	**98.41**

Anexo 13. Resultados de la aplicación del algoritmo GenSM5

Tabla A13.1 Rendimientos accesibles para cada cultivo calculados por el algoritmo dadas las restricciones iniciales y la información de los árboles de regresión M5. También se presentan las posiciones oculpadas por el cultivo en la representación del cromosoma utilizada por el algoritmo genético.

Posición en el cromosoma (Gen)	#	ID del rendimiento	Cultivo	Superficie (ha)		Lámina de riego (mm)		Rendimiento asociado (ton/ha)
				Límite inferior	Límite superior	Límite inferior	Límite superior	
0	1	0	ALFALFA	84.09	94	89.65	123.96	14.33
	2	1	ALFALFA	84.09	94	82.64	89.65	14.3
	3	2	ALFALFA	94	126.14	82.64	123.96	13.96
1	4	0	ALGODON	227.06	262.5	68.11	91.2	3.05
	5	1	ALGODON	262.5	340.6	68.11	91.2	2.96
	6	2	ALGODON	227.06	340.6	91.2	102.17	2.66
2	7	0	CARTAMO	564.43	654	29.19	36.84	2.08
	8	1	CARTAMO	537	564.43	29.19	36.84	2.08
	9	2	CARTAMO	513.58	537	29.19	36.84	2.07
	10	3	CARTAMO	654	770.36	29.19	36.84	2.07
	11	4	CARTAMO	513.58	770.36	43	43.79	1.98
	12	5	CARTAMO	513.58	770.36	36.84	38	1.96
	13	6	CARTAMO	513.58	770.36	38	43	1.96
3	14	0	CHICHARO	67.78	101.66	61.99	92.98	5.4
4	15	0	FORRAJE	109.7	164.55	76.83	87.67	10.55
	16	1	FORRAJE	109.7	164.55	87.67	115.25	10.51
5	17	0	FRIJOL	156.1	219.18	91.45	92.53	2.1
	18	1	FRIJOL	146.12	156.1	91.45	92.53	2.1
	19	2	FRIJOL	146.12	161.5	79.35	91.45	2.01
	20	3	FRIJOL	161.5	219.18	79.35	91.45	1.99
	21	4	FRIJOL	146.12	219.18	61.69	79.35	1.98
6	22	0	FRUTALES	17.26	21	66.46	99.69	18.78
	23	1	FRUTALES	21	25.89	66.46	99.69	18.69
7	24	0	GARBANZO	218.87	327.3	32.84	49.26	1.93
	25	1	GARBANZO	327.3	328.31	32.84	49.26	1.91
8	26	0	HORTALIZAS	240.94	361.41	66.55	93.75	15.99
	27	1	HORTALIZAS	240.94	361.41	93.75	99.83	15.97
9	28	0	MAIZ	638.04	957.06	71.91	98.66	6.32
	29	1	MAIZ	638.04	957.06	65.77	71.91	6.31
10	30	0	PAPA	348.37	522.56	66.45	76.99	32.91
	31	1	PAPA	348.37	522.56	79.78	89.75	31.84
	32	2	PAPA	348.37	522.56	76.99	79.78	31.82
	33	3	PAPA	348.37	522.56	63.32	66.45	30.76
	34	4	PAPA	348.37	522.56	89.75	94.98	30.39
11	35	0	SORGO	44.32	66.47	73.4	74.81	4.74
	36	1	SORGO	44.32	66.47	70.67	73.4	4.73
	37	2	SORGO	44.32	66.09	74.81	95.22	4.64
	38	3	SORGO	66.09	66.47	74.81	95.22	4.64
	39	4	SORGO	44.32	66.47	63.48	70.67	4.46
12	40	0	TRIGO	2,778.50	3,606.91	61.88	73.6	5.52
	41	1	TRIGO	3,307.00	3,606.91	73.6	92.83	5.51
	42	2	TRIGO	2,778.50	2,925.50	73.6	92.83	5.51
	43	3	TRIGO	2,925.50	3,307.00	73.6	92.83	5.51

	44	4	TRIGO	3,894.62	4,110.10	61.88	92.83	5.5
	45	5	TRIGO	3,606.91	3,894.62	61.88	92.83	5.5
	46	6	TRIGO	2,740.06	2,778.50	61.88	82.82	5.49
	47	7	TRIGO	2,740.06	2,778.50	82.82	92.83	5.49
13	48	0	ZACATE	32.19	48.28	120.95	129.81	12.93
	49	1	ZACATE	32.19	48.28	86.54	120.95	12.76
14	50	0	TOMATE	81.43	122.15	82.48	92.6	18.2
	51	1	TOMATE	112.84	122.15	116.95	123.72	17.77
	52	2	TOMATE	112.84	122.15	92.6	116.95	17.42
	53	3	TOMATE	81.43	112.84	119.45	123.72	16.33
	54	4	TOMATE	81.43	112.84	92.6	119.45	16.31