Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Monterrey

School of Engineering and Sciences

Evaluating Teaching Performance of Teaching-only and Teaching-and-Research Professors in Higher Education through Data Analysis

A thesis presented by

Mario Daniel Chávez López

Submitted to the
School of Engineering and Sciences
in partial fulfillment of the requirements for the degree of

Master of Science

in

Computer Science

Monterrey, Nuevo León, June, 2020

Dedication

To my family, friends and colleagues.

To those who have carried a little bit of me in them.

To myself from two years ago and to myself 10 years from now.

Acknowledgements

First of all I would like to thank Dr. Francisco Cantú, my thesis advisor, for always providing an excellent example inside and outside the institution, for his willingness, for pushing, for demanding, for teaching, for understanding, for an excellent company, for the great talks, for his great heart and for letting me see his real essence which reminded me of people who have impacted my life and hence motivate me everyday.

Also, I would like to thank Dr. Héctor Ceballos, my thesis co-advisor, for his orientation, patience and guidance during this process.

To my teachers who in some way helped me getting to know and develop topics of interest that could contribute to this research.

I would like to thank CONACyT for the support for living expenses. Also to Tecnologico de Monterrey for the great opportunity it gave me to be able to study here. For demanding, promoting, fulfilling and approaching much bigger dreams. For opening my eyes in many ways and for the great people who put me on the road.

On the other hand, I would like to thank to the greatest gift I have receieved in my life, my family, especially my parents and siblings, for the unconditional love and support they have given me throughout my life. For their orientation and for being the best team anyone can have.

Also to the family that I chose, my friends, those who are currently in my hometown and those who are in other parts of Mexico and the world. I always carry you with me.

To my life partner, she who has been supporting me for more than half of what I have lived until now. I will always thank you for the good you have done to my life. Day by day you make me forget of my fears and make me see my dreams... our dreams, more closely.

I would also like to thank God for all the blessings he has given me during these last two years and throughout all my life. He has given me a wonderful life without deserving it.

And finally I would like to thank myself, Mario, for these 2 years, because our constant search for meaning has led us here. Thank you for redefining your direction, for your hunger, dedication, honesty and love for yourself and others. This is just the start, keep on going.

Evaluating Teaching Performance of Teaching-only and Teaching-and-Research Professors in Higher Education through Data Analysis

by

Mario Daniel Chávez López

Abstract

We present a study that compares the teaching performance of Teaching-only versus Teaching-and-research professors at higher education institutions. It is a common belief that, generally, Teaching-only professors outperform Teaching-and-Research professors in teaching and research universities according to student perception reflected in student surveys. We present a case study which demonstrates that, in the vast majority of the cases, it is not necessarily true. Our work analyzes these two type of professors at their ability to function as an intellectual challenger, learning guide and their tendency to be recommended to other students. The case study takes place at Tecnológico de Monterrey (Tec), a teaching and research private university in Mexico that has developed a research profile during the last two decades with a mix of teaching-only and teaching and research faculty members and shows a growing accomplishment on world university rankings. We use five datasets from a student survey called ECOA which accounts observations from 2016 to 2019. We present the results of statistical and machine learning methods applied when the taught courses of more than nine thousand professors are taken into account. Methods include Analysis of Variance, Logistic Regression, Recursive Feature Elimination, Coarsened Exact Matching and Panel Data. Contrary to common belief we show that, for the case presented, teaching and research professors perform better or at least the same as teaching-only professors. We also document the differences found on teaching with respect to attributes related to courses and professors.

Keywords: Student Evaluation of Teaching (SET), Teaching Professor, Research Professor, Teaching Performance, Educational Innovation, Higher Education, Logistic Regression, Recursive Feature Elimination, Panel Data, ANOVA, Coarsened Exact Matching, Data Science.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Teaching evaluation is an active field of research and innovation spanning from elementary grades to higher education levels [29]. The digital transformation phenomenon has brought into place new technologies that are revolutionizing modern organizations, including higher education institutions in which teaching and learning are being transformed with novel pedagogical methods that require new ways to assess the quality of education from the administration, the teacher and the student standpoints [47]. Digital technologies and Artificial Intelligence methods has made it possible to combine traditional physical presence with online and virtual approaches in teaching delivery as a worldwide phenomenon. Technologies beyond the internet such as the Internet of Things (IoT), Cloud Computing, Big Data, Machine Learning and Data Analytics, Speech Recognition, Natural Language Understanding, Intelligent Tutoring Systems, and Virtual Reality, to name just a few, are transforming educational systems in general [33].

Aside from cultural and technological concerns, there exist several methods to conduct teacher evaluation that include the development of pedagogical materials, multimedia resources, scoring techniques, book publication, and learning platforms, among others. Nonetheless, one of the schemes more frequently used to evaluate teachers is to take into account the opinion of students about the quality of the service they receive, and this is commonly done by asking students to fill in surveys of student satisfaction [37].

1.1 Problem statement and context

Evaluation processes at any organization are very important. When it comes to universities, it is very important to evaluate students and professors. A lot of research has been made about student's evaluations, but not necessarily the other way around. Our research aims to evaluate teaching at Tecnologico de Monterrey's undergraduate and graduate levels. Student satisfaction has been an important feature in academic's management approach, which is based on key concepts such as accountability, visibility, and transparency. How teachers are being evaluated nowadays can be translated to the quality of their teaching and the extent to which student expectations are being met. Since academic institutions started applying these type of questionnaires many years ago, faculties have been concerned from the differences between ways in which students and teachers perceive effective teaching [37] and the high variations among

their scores. For this reason, many faculty members have been questioning about the specific variables and characteristics that can make a teacher's evaluation resulting good or bad. This identification process has been a major drawback of the academic institution. There is lack of information when analyzing comments made towards teacher's performance. Tecnologico de Monterrey applies a student satisfaction survey at the end of every semester in order to evaluate teacher's performance. This form is answered by approximately 90% of enrolled students of the 26 different campuses located around Mexico. This survey has been applied for the past five years and it has served as a way of measuring actual performance, student's satisfiability and effective decision-making related to new trainings and hiring strategic teaching staff. The objective of this survey system is to obtain fast, reliable and unique information of the opinions that Tecnológico de Monterrey's students make about their courses, professors, directors and services offered by their Campus. ECOA is organized as follows: Professors' Teaching Evaluation which can cover theoretical groups, Laboratory Courses, Project courses, Thesis Groups and co-curricular subjects. Also, it covers evaluation and comments on a specific career director. In addition, general items are covered at the end of the survey related to learning processes, student affairs, services and facilities, managerial performance and organizational pulse, and comments about the main library. With this research, we intend to take a complete advantage of what Tec's main clients are saying about their institution at those surveys only focusing on the information related to professor's teaching evaluation. The amount of generated data through years can be manipulated in certain ways in order to extract important knowledge from them, identify patterns of behavior and understand how its context affect teaching evaluation at Tec.

Tecnologico de Monterrey struggles on identifying the specific characteristics that teaching staff need to have in order to satisfy University's quality measures and students' needs. We believe that, by paying more attention to ECOA's information specifically in a Data Analytic way, several academic activities could be done differently and improvements on academic services can be made. Tec de Monterrey needs to be constantly changing their teaching strategies in order to meet their necessities at different levels. Since university rankings have become a very relevant factor in the past decade, there is a need of measuring quality and academic performance effectively. One way of doing this is by making an in-depth analysis of what are their main assets (students) reviewing about their teaching staff, available academic resources and courses. At Tec, we want to know what are the main reasons why are there still discrepancies between what the institution thinks is best among their offered services and what their alumni are perceiving and evaluating after they finish specific courses. By approaching this problem in this specific way, we believe we can find those main reasons and apply effective solutions in future periods. The results of our study can aid in identifying teaching evaluation approaches that truly respond to the needs of those who evaluate teaching performance [37].

1.2 Motivation

The idea of carrying out this research project came mainly from the members of the Tec de Monterrey Scienciometric Department. They had heard and seen for a while comments from different people which stated or commented that, normally the research professors do not usually perform better in the classes than the non-research teachers. This seemed quite strange

to them since, according to their criteria, normally a research professor knows more about the subjects he is teaching, is more up-to-date, has a better command, contributes new knowledge in those areas, etc. Therefore, this knowledge should be translated in a simple way in the classes. But apparently, taking into account comments from people and the state of the art, it was thought not to be true. Shortly after, they came up with the idea of approaching this topic focusing on the analysis of the opinions of the students exposed in the satisfaction surveys of each semester. So, they decided to invite me to carry out this analysis. It seemed quite interesting to me since I already had more than 4 years of experience evaluating my teachers and I found that debate very relevant because I had already heard certain rumors. I also liked this topic very much because unlike me the members of this department and my advisers had a long time experiencing on the other side of the problem, that is, being a teacher and a researcher. Hence, when it comes to teaching evaluation at academic institutions, we know that it is of common belief that generally, teaching-only professors outperform teaching-and-research professors in teaching and research universities according to student perception reflected in student surveys. We believe this rumor has been developed because of inferences related to the idea that teaching-only professors are better at teaching because they are 100% dedicated to students' learning by preparing their classes, designing new courses, being available for tutoring and they usually have a certificate that supports their way of teaching. In the case of teaching-and-research professors, since they dedicate closely half or more than half of their time to research, they neglect the development of other skills such as teaching. And, an additional inference is that, most people who are research-oriented and highly knowledgeable are not good at teaching or sharing information in an effective way to students. So, our motivation lies in the fact of presenting a case study, applied in Tecnologico de Monterrey, that will show experimental evidence that demonstrates if this idea is true or not. In other words, present a study that compare the teaching performance of teaching-only versus teaching-and-research professors at higher education institutions.

1.3 Hypothesis and Research Questions

As working *hypothesis* we assume that teaching-only professors are better evaluated by students compared to teaching-and-research professors (Ho). The alternate hypothesis is that teaching-only professors are evaluated either worse or the same by students than teaching and research professors (Ha).

Hence, our main research question is whether or not teaching-only professors perform better than teaching and research professors according to student opinion in teaching and research institutions. We contribute to the discussion of this issue by making a detailed analysis of data collected from the teaching evaluation survey "ECOA", (tudent Opinion Survey, in English), which is answered by students at Tecnológico de Monterrey, a teaching and research private university in Mexico. Teacher evaluation through survey that gathers student degree of satisfaction has a long tradition at Tecnológico de Monterrey and has been applied since the middle 70s up to now and has evolved over the years. We believe that the results of this analysis could be relevant to academic institutions with a teaching and research profile and could support the development of efficient research and teaching strategies in those institutions.

We consider the following as our research questions:

- Are Teaching-only Professors better evaluated by students than Teaching-and-Research Professors?
- Regarding the ECOA's Intellectual Challenge Factor towards the student, are Teaching-only Professors better evaluated by their students than those teachers that are not considered researchers?
- With respect to the ECOA's Learning Guide Factor, are Teaching-only Professors better evaluated by students compared to Teaching-and-Research Professors?
- Taking into account ECOA's Recommended Professor Factor, are Teaching-Only Professors more recommended by students than Teaching-and-Research Professors?
- Can we state in a statistical way that the teaching performance of Male Professors is higher than Female Professors?
- Does Professors' teaching experience determines their good or bad evaluation?
- How different is teaching evaluation given by Senior students than Junior ones?
- Does a higher researcher proficiency level ensures higher performance in teaching?
- Does aging affects negatively teaching quality in teaching-only and teaching-and-research professors?
- Is there a significant difference in teaching evaluation's scores between distinct academic levels?

1.4 Objectives

Our general objective is to identify patterns in the behavior of the variables that affect the teacher's evaluation when they are identified as researcher and also when they are not, understand how to segment them and create analytical models that help improve the quality of teaching in the institution through the discovery of knowledge and efficient decision making. The particular goals to achieve as we conduct this research are:

- Determine if Teaching-Only Professors have a better teaching performance than Teaching-and-Research professors.
- State what are the main characteristics which makes a professor being good evaluated.
- Analyze the correlations that certain variables have to determine the evaluation of these two types of professors.
- Analyze how Teaching-only and Teaching-and-Research Professors are evaluated at the different schools of Tec de Monterrey.
- Determine if student's maturity level, measured by their student class standing, is key for a teacher to be rated highly or lower.

- Establish corrective actions in order to improve educational services.
- Facilitate teaching staff hiring by developing predictive models that will ensure student satisfaction.

1.5 Solution Overview

First of all, we decided to use the Cross Industry Standard Process for Data Mining Methodology which helped us to face this problem by understanding the business, understanding the data, preparing the data, decide which models to implement, evaluate our results and deploy them. Our solution consists of preprocessing our data with the Coarsened Exact Matching Algorithm, clean and prepare data, compute the means of the ECOA results of the two types of professors. Filter the results in several ways, apply ANOVA in each of the cases in order to identify statistical differences. Moreover, apply Recursive Feature Elimination to identify those features who have a stronger correlation to the ECOA score. Make use of Logistic Regression in order to create the best predictive model. Finally, apply Panel Data Modelling with the intention of observing the influence of certain group and professor features among time.

1.6 Main Contributions

Our research contributes to the discussion of the comparison of teaching performance between Teaching-only and Teaching-and-Research Professors. We present forceful answers to this question by taking Tec de Monterrey as our study case. We present a study in which we report what makes a professor being good or bad evaluated. And, we present an example of a data analytic framework that other academic institutions could follow and apply in order to answer these same type of research questions and get benefited by their results. If this were the case, we could finally get a more general answer to this question that could apply to the behavior of vast majority of universities.

1.7 Summary

In this first chapter, we have discussed the main factor by which this thesis will be developed around, teaching evaluation. Specifically, we want to analyze the teaching performance of professors from Tecnologico de Monterrey who are solely dedicated to teaching and compare them to those professors who also perform research. We identify the first type of professors as Teaching-only Professors and second ones as Teaching-and-Research Professors. The foregoing, by making use of data collected from ECOA, Tec's Student Survey Opinion. We want to see if Teaching-only Professors are better evaluated by students than Teaching-and-Research Professors through three specific academic factors that are covered in this survey, Intellectual Challenge, Learning Guide and Recommended Professor. In addition, by applying methods of data analysis we hope to identify professor's features that are positively correlated to the score of the evaluation made by students.

To address these issues, we organized this thesis as follows: in Chapter 1 we introduce you to our problem and the main goal of our research. Then, in Chapter 2 we present ways in which similar studies have been implemented related to this area and the theory behind several data analytic techniques. Moreover, the thesis follows CRISP-DM's structure at the beginning of Chapter 3 in which we understand our business and the data to use, then we prepare the data and we finish it by stating the specific models we will apply. Later on, in Chapter 4 we present the results of those experiments and their respective evaluations. After that, through Chapter 5 we discuss the findings we obtained through the interpretation of the results we got and we give way to the last phase of the CRISP methodology when we mention several ways how those results could be used to improve academic strategies. Finally, Chapter 6 comes to be the last part of the thesis where we state our conclusions when taking into account the implemented work.

Chapter 2

Background

2.1 Teacher Evaluation in Teaching and Research Universities

Teaching evaluations performed by students have been used as a measure of teaching performance in several higher education institutions all over the world [37]. These evaluations have developed relatively complex procedures and instruments for gathering, analyzing and interpreting data as the main indicator of teaching quality [48]. Teaching evaluations have been utilized for faculty personnel decisions. This type of evaluations follow three purposes, improving teaching quality, provide input for appraisal exercises and demonstrating the presence of adequate procedures for ensuring quality. The specific results help to improve the quality of their teaching.

Teaching effectiveness has received much attention in research literature. In the decision-making process of higher education, teaching effectiveness plays an important role. Usually, teaching evaluation is measured through some student questionnaire designed to measure observed teaching behaviors and styles. The vast majority of universities use student assessments to evaluate teaching performance at their institution, since there is no other validated alternative that is practical to implement regularly [4]. These evaluations have served as formative and as concise measurements of teaching. It goes from improving professor's teaching, improving course's content and format, influencing professor's tenure, promotion and salary rises and making these results available for students to use in the selection of teachers and concerts [37]. Moreover, these type evaluations are a way of acquiring information about the experience of the student through a specific course. When an academic institution is aware of the satisfaction of their students during their learning process, more efficient academic strategies can be implemented which can maintain student's motivation that will lead to achieve precise and meaningful learning [5].

One of the objectives of the *Student Evaluation of Teaching* work is to examine whether an inappropriately designed teaching evaluation that, in the perception of students, hinders students from providing valid or meaningful feedback will affect their motivation to participate in the evaluation. In order to prove this, a study applied expected theory which has successfully predicted behaviour in various context [26]. The choice of the amount of effort that the person exerts is based on a systematic analysis of the values of the rewards from these

outcomes, the likelihood that rewards will result from these outcomes and the likelihood of reaching these outcomes through his or her actions and efforts. Expectancy theory is composed of two related models, the valence and the force model. In this case application of the theory, the valence model shows that the overall attractiveness of a teaching evaluation system to a student, which is expressed as Vj, is the summation of the products of the attractiveness of those outcomes associated with the system (Vk) and the probability that the system will produce those outcomes (Ijk). Vj is the the attractiveness of a teaching evaluation , Vk is the valence, of outcome k (second level outcome) and Ijk is the perceived probability that the teaching evaluation will lead to outcome k. This can bee seen in equation 2.1.

$$V_j = \sum_{k=1}^{n} (V_k \times I_{jk}) \tag{2.1}$$

For this study, the four potential outcomes ($k = 4$), are the four uses of teaching evaluations that are described in the literature. On the other hand, the force model shows that a student's motivation to exert effort into a teaching evaluation system (F_i) is the summation of the products of the attractiveness of the system (V_j) and the probability that a certain level of effort will result in a successful contribution to the system (E_ij):

$$F_i = \sum_{j=1}^{n} (E_{ij} \times V_j) \tag{2.2}$$

F_i is the motivational force to participate in a teaching evaluation at some level i, E_ij is the expectancy that a particular level of participation (or effort) will result in a successful contribution to the evaluation and V_j is the valence, or attractiveness, of the teaching evaluation, derived in the previous equation of the valence model. In the valence method, each participant in a teaching evaluation system evaluates the system's outcome and subjectively assesses the likelihood that these outcomes will occur. Student uses the force model to determine the amount of effort he or she is willing to exert in the evaluation process.

This study was conducted for a population between 15 000 to 20 000 total enroll students. The instrument was administered at the beginning of a regularly scheduled class around the middle of the quarter to all the students who were present on that particular day. Students other than freshmen and seniors were eliminated from the sample as were the instruments with incomplete data [2]. This resulted in 208 usable instruments completed by 105 freshman and 103 senior students. The main question asked was, *in general, how do you describe the professor you have had at this institution?*. They used a scale with a range of 0 to 10 where 0 represents very bad and 10 represents very good. The question asked was, *What is your general impression about the course evaluation system?*. Again they used a scale with a range of 0 to 10. 0 represented 'useless' and 10 represented 'very useful'. Basically, they used the freshmen versus seniors design to examine whether freshmen and seniors have different motivations to participate in the teaching evaluation system.

The participants were presented with 16 hypothetical situations. They were supposed to detach themselves from their past experiences and evaluate the hypothetical situations from a third party perspective. If the respondents were successful in doing this, they would expect to find no correlation between their actual experiences with student-generated evaluations or background and their responses. In order to test this, they calculated Pearson's correlations

between the R^2 value of the valence model and four selected demographic factors. These factors are gender, grade point average (GPA), impression of professors and perception about the evaluation system. The coding of gender was 1 for male and 0 for female. The impression of professors and perception about the evaluation system were measured by two 11 point scale demographic questions. They also calculated the correlation between the R^2 value of the force model and the four demographic factors. They used those correlations to assess whether the subjects were able to evaluate the 16 hypothetical situations objectively without bias and thus were appropriate for the study. They applied multiple regression analysis to determine each student's perception of the attractiveness of participating in the evaluation. They have also used t-tests to investigate whether there is a difference between the freshman and senior groups in their perception of the attractiveness of the four second level outcomes [26]. This was one example of other researchers approached this type of problem with data analysis techniques. We intend to have a similar approach in order to contribute to this issue.

Student evaluation of teaching (SET) at higher education institutions spans undergraduate and graduate academic levels and has been an active field of research for the last few decades. The evaluation of teachers by students and the consequences that those evaluations impinges on teacher wages, faculty career and promotion has been widely criticized and keeps being a vigorous area of study. However, appealing to student evaluation of teachers through surveys and questionnaires arguably provides an objective and measurable means of lifting the voice of the learners, who are recipients of teaching delivery. Tsinidu et al present a study in which they identify the quality determinants for education services provided by higher education institutions in Greece and to measure their relative importance from the students' points of view.

They show a multi-criteria decision-making methodology that was used for assessing the relative importance of quality determinants that affect student satisfaction. They use the analytical hierarchical process (AHP) in order to measure the relative weight of each quality factor that contributes to the quality of educational services as it is perceived by students was measured. Their study can be used in order to quantify internal quality assessment of higher education institutions [68]. In order to acquire a broader panorama, it is important to analyze what evidence does exists on the relationship between research activity and teaching performance. This will allow us to have a better knowledge on the relationship between the effective production of research results and student satisfaction [55]. On the one hand, Steve Stack analyses the relationship between research productivity and student evaluations of teaching and reports that it has been marked by several shortcomings. He argues that research typically fails to check and adjust for nonlinear distributions in research productivity and that approximately 15% of researchers account for most articles and citations. Then, he highlights that the unit of analysis is typically the teacher and not the class and that top researchers might disproportionately teach small classes at the graduate level, and student evaluations are usually higher in such classes.

His study intends to correct those issues using data from 167 classes in the social sciences and on 65 faculty. He finds that the quality of research productivity measured in citations per year is not related to student evaluation of teaching. And he finds that when the distribution of citations is corrected for skewness, a significant positive relationship between research productivity and student evaluation of teachers emerges. And he concludes that this is the first systematic investigation to demonstrate a significant relationship between the quality of

research (measured by citations) and student evaluation of teachers [48]. In addition, we have also found that those teaching-and-research professors, who are more active in their research activities, have less favorable aptitudes to teaching, and teachers who are less productive in research are most the most committed to teaching [38]. In other words, we have found some evidence in the state of the art which conclude that there is no relationship between teaching performance and the development of research.

On the other hand, Spooren et al presented an overview of the state of the art on student evaluation of teaching in higher education. Their study is based upon research reports published in peer-reviewed journals since 2000. They consider the traditional topics such as the dimensionality debate, the bias question, and questionnaire design, and some recent research trends in student teaching evaluation, such as online and some other bias including professors' character and personality, thus allowing researchers to formulate suggestions for future research. Spooren at al continue arguing that teacher evaluation through student survey remains a current yet delicate topic in higher education, as well as in education research. They add that many stakeholders are not convinced of the usefulness and validity of student evaluation of teachers. They conclude that research on student evaluation of teaching has thus far failed to provide clear answers to several critical questions concerning the validity of teacher assessment [65].

The studies above present the state of the art in teacher evaluation at higher education. This overview allows us the focus on the sort of universities we are interested. These institutions are the teaching and research universities on which teaching is typically done by professors that follow either a teaching-only track or a teaching and research path. Those universities are frequently included in world university rankings that display tables of top-1000 universities in the world like Shanghai, QS, Times Higher Education, U-Multirank, or US News and World Report. World university rankings have been in the landscape of higher education for the last two decades and have become a relevant factor to measure quality and performance of universities as well as public perception and reputation worldwide [17]. Some rankings focus on research intensive universities which is the case of the Shanghai ranking (Academic Ranking of World Universities) which takes into account Nobel prices awardees as well as publications in the journals Nature and Science. The Times Higher Education World University Ranking (THE WUR) also favors universities with a high research profile, and US News and World Report's Best Global University Ranking also is biased towards research intensive universities. On the other hand, the QS world University Ranking (QS WUR) calculates university scores based on a more balanced combination of teaching and research indicators the include academic reputation, employer reputation, citation per faculty, students per faculty, and international students and professors which is more suitable for teaching and research universities without excluding research intensive institutions. Thus, this thesis is focused on higher education institutions ranked by QS WUR.

Student education in research-intensive universities which are typically ranked on the Top 100 of the QS WUR and other world university rankings is mostly done by professors that educate their students following a research-based teaching approach. However, universities that are teaching and research and not only research intensive found on the 101-1000 QS rankings band, combine both teaching-only and teaching and research professors in teaching delivery [18]. Since the impact of the academic research output is one of the most important

metrics for this type of ranking, universities who have a place at a certain rank spend a significant amount of resources on funding research. A higher impact and an increase in research output can guarantee a better position in world university rankings. We have notice that there is research that focuses on how that strategy could work [16]. Research can create positive collateral effects in teaching by having updated courses [6], when the committee in charge of a specific course counts with a frequent research performance, it is easier to deliver updated courses since they maintain familiarize with the state of the art. Thus, universities aim at balancing the proportion of teaching-only and research professors to be listed in the ranking tables and at the same time, fulfil the mission of educating and preparing students for the professional life. We have found that the relationship between research activity and teaching performance has been part of a strong questioning. Depending on the benefits provided by the educational institution, there may be a selection of professors for research and teaching-only [28][8]. However, as found in the literature where the relationship quality of teaching and research has been seen as a synergy [54], we think that research and teaching match perfectly. Most research professors, on top of their research activities, frequently teach a proportion of at least half course lectures of their teaching load at graduate and undergraduate academic levels.

We also have seen that we can highlight three types of relationships between quality of research and teaching performance. The first one relates to a positive relationship when the skills developed through a professor's research experience complement their teaching skills. Teaching-and-Research Professor can boost student's critical thinking and research skills when high impact problems are attacked. The second one discusses that there could be a negative connection as we know that teaching-and-research professors require to make a harder effort and time to preform efficiently in both areas. The negative side of this is that time and effort dedicated to one area reduces the same for the other one, unless they have similar activities that makes one not necessarily quarrel with the other. Finally, the third side of this story is a more neutral one when we base on assuming that these activities do not relate at all [69].

Nevertheless, the academic education and activities of a research professor may be different from the ones a professor who is fully dedicated to teaching students do. For many years, there has been a debate about the role and importance of research activity with respect to teaching performance. In many cases, there is a general belief that at least in teaching-oriented universities, teaching-only professors outperform teaching and research professors based on student opinion, in this study, we want to provide evidence that this could not be necessarily the case. Additionally, research has not been a characteristic that students evaluate when they assess professor performance through opinion surveys at the end of an academic period. However, this theme is a subject of enduring debate.

Another similar work to the one that is intended for this research was found in the literature. This article was developed with the intention of validating Student Valuation of Teaching as many stakeholders were not convinced of its usefulness for formative purposes. The paper provided a complete overview from 2000 to 2013, on Student Evaluation of Teaching (SET) in higher education [48]. These type of student assessments have been the most commonly used measure to estimate the quality of teaching at higher education institutions [55]. In Table 2.1 we illustrate an overview of studies performed in the last 19 years related to student, teacher and course characteristics that are closely related to affect SET. It was found that these

characteristics are significant and logically related to effective teaching performance.

One of the major aspects why certain findings were contrary to each other is due to generalization. There are a great variety of methods, measures, variables, instruments and populations on all the studies. That can be translated to high degree of variation which makes it almost impossible to make statements concerning. This research literature has shown us the existence of correlations between student achievement and SET scores. The results provided evidence to validate SET. On the other hand, this work states there is still variety in stakeholders' views due to variety in the measurement of student achievement so, Student Evaluation Teaching can not be the only indicator of teaching effectiveness. In summary, the research literature revealed the existence of (small to strong) positive correlations between SET scores and student achievement, expert ratings of teaching behavior, self-ratings, and alumni ratings. These results provide evidence of the convergent validity of SET. However, due to the variety in stakeholders' views concerning good teaching and due to the variety in the measurement of student achievement, SET should not be the only indicator of teaching effectiveness in personnel files [64].

Characteristics	Measure	Interpretation
Student's Cognitive Background	Student's Major and Year of Enrollment	Mature students majoring in the same subject as the course, give a higher score [66].
Student's Effort	Student Effort	Teachers who encourage students to make more effort in the course, get a higher score in SET [39].
Student's Gender	Student's gender	Gender preferences: Female students give higher ratings to female teachers. Female students give higher SET than male students [11, 22, 45].
Student's Age	Age	The greater the age, the higher the SET [47].
Grade Discrepancy	Difference between expected grade and believed deserved grade	Students tend to punish teachers when expected grades are lower than they believed to deserve [36].
Teacher Instructor's Gender	Gender	Two studies showed that Female Teachers receive higher SET. Another study showed that Male Teachers receive higher SET [10, 61, 51, 50].
Instructor's Age	Teacher's Age	Younger teachers receive higher SET [51].
Instructor's Language Background	English as a second language (ELS) vs. Native Speakers	ELS speakers receive lower SET than native speakers especially in the science faculties [53].
Instructor's Rank	Full-time Professors vs. Professors, Associate Professors, Lecturers, and Junior Lecturers	Full-time professors receive higher SET than associate professors and professors [51, 63, 66].
Course Level	Course's year level	SET in higher year level are more positive [59].

Table 2.1: SET Results based on students' and professors' features.

2.2 Theoretical framework

2.2.1 Data Analytic Approach

The past twenty years have been composed of extensive investments in business infrastructure, which have propitiated an extensive generation, collection and sharing of data. In fact, nowadays about 2.5 exabytes of data are being created and every three years that number is being doubled [49]. At the same time, this information is being widely available not only for companies but for society, in order to show transparency and congruence. As we know, a good interpretation of data can be translated to generation of knowledge, which can be use to benefit companies and organizations in several ways. With this knowledge being created and being available at all times, companies in almost every industry are focusing more on exploiting this data with the intention of gaining important competitive advantage [56]. Since data have surpassed our capacity for manual analysis, several data mining techniques and data analytic tools have been developed which has let us learn from the behaviour of our business performance. A data analytic approach to business problems help us understand the principles of extracting useful knowledge from data. This data thinking and perspective provides structure that will be transformed to framework that will allow us to systematically analyze problems [56] Facing up problems data-drivenly we left as side solving situations intuitively. It has been proved that the more data-driven a company is, the more productive it is as they tend to identify risk situations before they actually happen [49].

As the Big Data era have arise, new processing technologies are being used for implementing data mining techniques. At this work, we will acquire this precise data-analytic approach in order to understand business problem, understand data, prepare data, develop predictive models, evaluate results and establish corrective actions that will improve teacher's performance.

There are several data analysis and machine learning techniques which are applied to manipulate data and answer research questions. Here are some utilized and helpful ones found in the literature review.

2.2.2 Logistic Regression

Logistic regression is a classification learning algorithm. It is called regression because the mathematical formulation of it, is similar to the linear regression one. We can explain logistic regression on the case of binary classification but it could also extended to a multiclass classification. In a logistic regression, we want to model y_i as a linear function of x_i, however, with a binary y_i this is not directly possible. The linear combination of features such as $wx_i + b$ is a function that spans from minus infinity to plus infinity, while y_i has only two possible values. When we define negative labels as 0 and positive labels as 1, we would just need simple continuous function that goes from 0 to 1. In that case, when the returned value of the model for input x is closer to 1, then we assign a negative label to x, if not, it is labeled as positive. The function that has such a property is the sigmoid function shown in equation 2.3. ϵ is the base of the natural logarithm [15].

$$f(x) = \frac{1}{1 - \epsilon^{-x}} \tag{2.3}$$

The logistic regression model is illustrated in equation 2.4

$$f_{w,b}(x) = \frac{1}{1 + \epsilon^{-(wx+b)}} \tag{2.4}$$

The values of w and b can be optimized appropriately, we interpret the output of $f(x)$ as the probability of y_i being positive. Typically our threshold is 0.5, if it's higher than that, the model tells us that the class of x is positive, otherwise it would be negative. In this algorithm we maximize the likelihood of our training set according to the model. This likelihood function defines how likely the observation is according to our model. The optimization criterion in logistic regression is called maximum likelihood. Instead of minimizing the average loss, like in linear regression, we now maximize the likelihood of the training data according to our model. In practice, it's more convenient to maximize the log-likelihood instead of likelihood. The log-likelihood is defined like follows:

$$LogL_{w,b} = ln(L(_{w,b}(x)) = \sum_{i=1}^{N} y_i ln f_{w,b}(x) + (1-y) ln(1 - f_{w,b}(x)) \tag{2.5}$$

Because ln is an increasing function, when we maximize the function we miximize its argument, so the solution to the new optimization problem is the same as the solution of the original problem [15].

Once we have a model with a learning algorithm built with a training set, we need to verify how good is our logistic regression model. Hence, we use the test set to asses it. In this case, for classification, the most widely used metrics and tools to assess the classification model are: confusion matrix, area under the ROC curve, accuracy, among others. We will explain the first two.

Confusion Matrix

The confusion matrix is a table which summarizes how successful the classification model is at predicting examples belonging to various classes. One axis of the confusion matrix is the label that the model predicted, and the other axis is the actual label. Table 2.2 helps us understand the concept behind it by showing an example when a model wants to predict if an email is spam or not spam [15].

	Spam (Predicted	Not Spam (Predicted)
Spam (Actual)	23 (TP)	1 (FN)
Not Spam (Actual)	12 (FP)	556 (TN)

Table 2.2: Confusion Matrix Example

In other words, a confusion matrix can tell us that a model trained to recognize species of animals tends to mistakenly predict "dog" instead of "cat," or "whale" instead of "shark." In this case, you can decide to add more labeled examples of these species to help the learning algorithm to "see" the difference between them. Additionally, with confusion matrix we can also calculate the precision and recall, two other performance metrics. Precision is define as the ratio of correct positive predictions to the overall number of positive predictions. We identify recall as the ratio of correct positive predictions to the overall number of positive examples in the dataset [30] [15].

Receiver Operating Characteristics

The ROC curve is a very common method which assesses the performance of classification models. In order to build up a summary picture of the classification performance, it is defined by the combination of the true positive rate and false positive rate. They both are defined as the following.

$$TPR = \frac{TP}{TP + FN} \tag{2.6}$$

$$FPR = \frac{FP}{FP + TN} \tag{2.7}$$

The higher the area under the ROC curve (AUC), the better the classifier. A classifier with an AUC higher than 0.5 is better than a random classifier. If AUC is lower than 0.5, then something is wrong with your model. A perfect classifier would have an AUC of 1 [15]. Figure 2.1 illustrates an example of a ROC Curve.

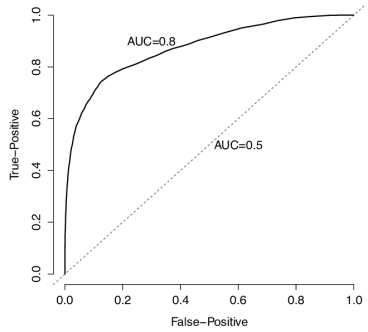

Figure 2.1: ROC Curve Example

2.2.3 Recursive Feature Elimination

In a classification problem with relatively low training samples and high dimensionality, feature selection is a critic process which has an important role when avoiding overfitting. One of the most used methods for feature selection of dataset with lots of features is called Recursive Feature Elimination. It utilizes the generalization capability presented in Support Vector Machines (SVM). It removes the least important features whose removing will have the least effect on training errors [25].

2.2.4 Panel Data

Panel data is also referred as longitudinal data, in other words, data containing time series observations of a number of individuals. Hence, the observations presented in this type of dataset involve at least two dimensions, a cross-sectional dimension, and a time series dimension. It presents a variety of advantages, it is more accurate inference of model parameters, it has a greater capacity for capturing the complexity of human behavior than a single cross-section, it

controls the impact of omitted variables and uncovers dynamic relationships[41]. Panel data also gives more informative data, more variability, a lower collinearity between features, and more degrees of freedom that can lead to more efficiency [60].

Ordinary Least Squares in Panel Data

There are several methods that can be applied when our datasets have a panel data format. The first one is the Pooled regression model, this type of model has constant coefficients, which refers to both intercepts and slopes. The data can be pooled and an ordinary least squares regression model can be applied to it. Our second model is called Fixed effects, it is the differences across cross-sectional observations that can be captured in differences in the constant term and the intercept term of the regression model varies across the cross sectional units [40]. Finally, the third most used method we identify it as Random Effects model. In this case, the individual effects are randomly distributed across the cross-sectional units and in order to capture the individual effects, the regression model is specified with an intercept term representing an overall constant term [60].

2.2.5 Matching

Matching is a statistically very powerful technique developed by Gary King, Stefano M. Iacus, and Giuseppe Porro. It was developed in R Statistical Programming Language and it is called Coarsened Exact Matching (CEM). It is relatively new and simple to use. Gary King refers to it as a method that should be introduced in casual inference courses, before even teaching regression. For other matching methods, usually students learn causal inference, then regression and its type of problems, then they analyze the results and try to correct the problems and they they might do matching. These three Harvard researchers state that it is much simpler to start with matching because it conveys very clearly what the control and treated groups are. CEM optimizes balance between these two groups, one of its main goals is to get a treated and a control group that are the same prior the application of a statistical or machine learning method. Most of matching methods such as Propensity Score Matching and Mahalanobis Distance matching either optimize imbalance or they try to get the largest number of possible observations and we end up seeing whether if we achieved any balance at the end of the procedure. We should be doing both, that is why CEM was created [44].

Gary King mentions that it should rather been called pruning instead of matching, because basically that is what it does. It prunes away certain observations under specific conditions so we do not create selection bias. One of the most important aspects of this method is that the final dataset results with less model dependence, meaning less changes to the results due to small decisions. CEM does not create selection bias because it is a function of the explanatory variable and not the dependent one. Coarsened Exact Matching is a preprocessing method. Once data has been preprocessed we can apply to it any method we want. Without applying this method imbalance between the treated and control groups leads to model dependence which leads to researcher discretion because when there is model dependence, we get to decide which model to write up. It can be completely automated and it is fast computationally. This process involves pruning observations that have no close matches on pre-treatment covariates in both the treated and control groups. It basically involves three steps. First it

temporarily coarsen each control variable in X as much as we are willing to. It sorts all units into strata, each of which has the same values of the coarsened X. Finally, it prunes from the original dataset the units in any stratum that does not include at least one treated and one control unit [42].

2.2.6 Additional Approaches

Decision Trees

Is a decision support system expressed as a tree-like graph illustrating their possible after-effect. It includes a root node, leaf nodes that represent any classes, internal nodes that represent test conditions which are applied to attributes. This technique is usually easy to understand by the end user. This technique makes possible to handle a variety of input data, either nominal, numeric or textual [46].

Random Forest

Is one of the most efficient classification methods. It is the collection of tree-structured classifiers. In this case, random forest splits each node using the best among a subset of predictors randomly chosen at that node. A new training dataset is created from the original dataset with replacement. Then, a tree is grown using random feature selection [2]. This technique has shown excellent performance in datasets where the number of variables is much larger than the number of observations, it also can cope with complex interaction structures as well as highly correlated variables and return measures of variable importance [13].

Neural Networks

They provide a more suitable inductive bias than competing techniques. They tend to have a more appropriate restricted hypothesis space bias than other learning algorithms. In some cases they are the preferred learning method because they induce hypotheses that generalize better than other algorithms. There are certain type of problem domains in which this technique provide superior predictive accuracy to commonly used learning algorithms [27].

2.3 Actionable Knowledge Discovery

Data mining has been facing challenging problems when it comes to real-life business needs[20]. Among those challenges, there are often several patterns mined but it gets complex when we try to make them informative and transparent to business management people.

Those patterns often can be either commonsense or of no interest to business needs. In addition, it is common for business people to get confused while analyzing data mining results, and they struggle on why and how they should care and act regarding those findings [21].

Business departments are often not well trained in order to interpret important results. It has been found that there is a gap between business expectations and research and development results, as well as between data miners and business analysts. Both intermediaries

have realized the importance of domain knowledge among both sides in order to close this gap and start developing actionable knowledge for decision makers [20]. Most techniques are represented as algorithms which summarize training-data distributions in one way or another. Their output models are typically mathematical formulas or classification results describing test data. This means that they are data centric, those models do not correspond to actions that will bring desirable states. Data mining models should generate actions that can be performed either automatically or semiautomatically. In that way data mining system will be truly considered actionable [19]. Regarding this, there is a need of developing a general, effective and practical methodology in order to achieve actionable knowledge discovery (AKD). This system follows Domain-Driven Data Mining's Methodology. AKD is critical in promoting and releasing data mining's and knowledge discovery's productivity for decision making and business operations. Actionability refers to measuring the ability of a pattern to suggest the user on making specific actions to gain advantage in the real problem. In other words, it measures the ability of suggesting business decision-making actions[21]. Since traditional Knowledge Discovery in Databases (KDD)is only a data-driven consisted of a trial and error process which targets automated hidden knowledge discovery, it can not satisfy business problems. Literature shows that the goals of traditional data mining is to let data to create and verify research innovation, demonstrate and motivate the use of novel algorithms discovering knowledge of interest to researchers [19]. That is why, some type of KDD needs to support commercial actions, support business requirements for trustworthy, cost-effective and reliable performance.

When it comes to Domain-driven data mining we say it is consisted of the following key components. Understanding, defining and involving domain intelligence. Data mining where there is a constraint-based context. In addition, a pattern discovery targets mining in-depth patterns. It presented as a loop-closed iterative refinement process. Mined results must be actionable in business [19].

According to research, Actionable Knowledge Discovery involves the activities illustrated in the following figure 2.2. These highly correlated ideas are critical for the success of a data mining process in the real world problem.

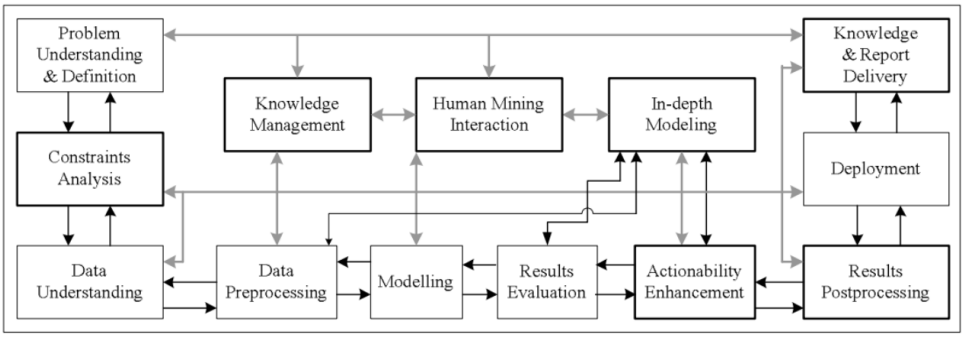

Figure 2.2: Actionable Knowledge Discovery Process Model

Nowadays, data mining must be closely related to business knowledge and technical knowledge at all times. It has been found that Data miners often lack business domain knowledge when they perform mining and modeling tasks. Business experts are the ones who can help throughout the data mining life cycle. In order to create an efficient model, they must

guide the exploration process, acting as navigators while data miners do the driving. In that way, Knowledge-driven data mining and business experts will identify important results and interpret them in the form of metaknowledge [35].

Moreover, we found several Actionable Knowledge Discovery's Frameworks that can be applied to this project such as [21]:

- *Postanalysis-Based AKD*: Is a two-step pattern extraction and refinement exercise.
- *Unified-Interestingness-Based AKD*:It develops unified interestingness metrics capturing and describing both business and technical concerns.
- *Combined-Mining-Based AKD*: Extract actionable knowledge in a progressive way, comprise multisteps of pattern extraction and refinement on the dole dataset.
- *Multisource + Combined-Mining-Based AKD*: It discovers actionable knowledge either in multiple datasets or data subsets, through partition.

2.4 Summary

This section represented our second chapter of this thesis. Our main objective was to discuss the main findings based on the state of the art related to Teacher Evaluation in Teaching and Research Universities. We presented several case studies which had a Data Analytics Approach, same as the one we will have in this research work. We discussed the relationship between research activity and teaching performance among professors at higher education institutes. Some studies stated that there is no relationship between one another and some other mentioned that these two features are negatively correlated, or that it is still an open debate. Other studies focused on the fact that teaching evaluation through student surveys remains a current delicate topic in higher education and research. In addition, we also covered the utilized metrics to evaluate universities at university world rankings. We saw how it is related to research activities and the interests of academic institutions. Moreover, since this thesis has a data analytic approach, we covered a theoretical framework. We briefly discuss some of the methods that could help us answer our research questions.

Chapter 3

Methodology

In order to carry out this study, we utilized a very well known and used method presented in the Data Science field. This data analytic methodology is called Cross-Industry Standard Process for Data Mining, best known as CRISP-DM. It shares a general vision of the usual life cycle of a data mining project [71]. This cycle is composed of six principal phases, which are not necessarily sequenced, but each of them are equally important and have a specific role throughout this process. CRISP-DM was conceived in 1996 and became a European Union project under the ESPRIT funding initiative in 1997 under the leadership of several companies that included Integral Solutions Ltd, Teradata, Daimler AG, NCR Corporation, and OHRA. The first version of the methodology was presented at the 4th CRISP-DM SIG Workshop in Brussels in March 1999 and published as a step-by-step data mining guide later that year [70]. While many non-IBM data mining practitioners use CRISP-DM, IBM is the primary corporation that currently uses the CRISP-DM process model and it has incorporated it into its SPSS (Statistical Package for the Social Science) modeler product [23]. This process model provides a framework for carrying out data mining projects in a more manageable way, less costly and more reliable. According to the state of the art, this method is very useful for planning, documentation and communication. All the stages are properly organized, structured and defined, this allows that any data mining project can be easily analyzed, replicated and understood, in other words, it guides people to know how data analysis can be applied in practice in real systems [7]. Additionally, this data mining method reduces the abilities required for knowledge discovery. Its performance is very stable, meaning that it can be applied to problems with different scopes, it is independent from the tools and techniques needed to be applied in order to satisfy a problem's need and it is insensible to changes in the environment [24]. Thanks to all these peculiarities, we decided to make CRISP-DM a substantial part of the structure of this study.

Figure 3.1 illustrates the life cycle of a data mining project. As we can see, the outer part of the figure represents the cycle part of the process that we were discussing before. Those arrows indicate us that, we could get to the last phase and it would not necessarily mean that we got to the last part of our project. All these stages can be presented several times until we satisfy all of the objectives of a specific project. When we take a look to the inside of the outer circle, we can identify the six stages that we mentioned before. These are represented as Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment.

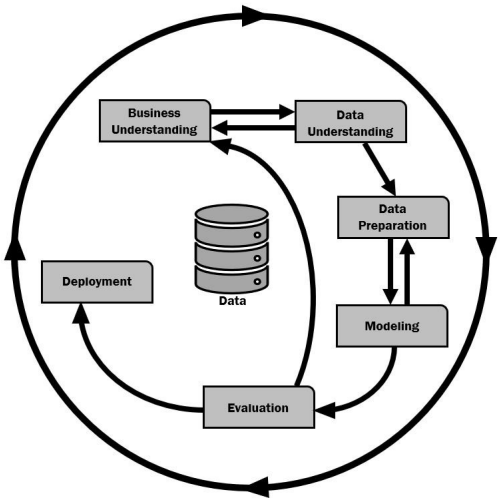

Figure 3.1: Cross-Industry Standard Process for Data Mining (CRISP-DM)

Data is the mainly used element presented in this methodology. Those phases which have more than one arrow, emphasize a stronger relationship with another stage. In the following subsections, we will explain each of the phases, the activities that have to be covered in each of them, and finally, we will show how did we applied this method in order to answer the research questions of our study.

3.1 Business Understanding

CRISP-DM's first phase covers the identification and comprehension of the main objectives of the project alongside with all the requirements of the business or management department. After having a clear understanding of the actual situation and what is causing pain to the organization, that knowledge needs to be converted into a data mining problem definition and a preliminary plan is designed to accomplish objectives [71]. Business understanding involves describing the organization's background, determining criteria, inventory of resources, assumptions, constraints, risks, specific terminology utilized in the problem, costs and benefits, data mining goals and production of project plan.

Comprehensive and profound business understanding is a key factor for yielding a successful project and delivering useful results. Business understanding is about apprehending and deciphering the problem domain on which the study is conducted. In this case, the evaluation of teaching performance in teaching and research universities typically ranked in the 101-1000 rank of world university rankings is the problem and the domain on which we do our analysis. More specifically, we take Tecnologico de Monterrey, a teaching and research private university in Mexico that has developed a research profile during the last two decades, with a mix of teaching-only and teaching and research faculty members and had shown a growing accomplishment on world university rankings. This university ranked in position 158 of QS World University Ranking 2020, and will function as our case study to conduct the teacher performance analysis [1].

Tec has also recently been identified as the best university of Mexico and 5th in the region by Times Higher Education (THE) in its new Latin America ranking. In addition, Tec was ranked number 1 in Latin America in the category of international professors. Nonetheless, the issues raised in previous sections regarding teaching evaluation come up to surface when we get into the context and cultural aspects of particular institutions. The author of this thesis has more than five-year experience evaluating teachers as a student. Thus, we believe that there is a fairly good understanding of the problem domain and the relevance of the problem we are trying to solve [32].

On the other hand, it is worth to mention that, Tecnologico de Monterrey has been recently implementing their newest educational model, Tec 21. Within its main innovations, their strategy integrates training and development of research professors, this being part of a series of initiatives presented in their 2030 vision. In the past years, Tec has been strongly committed to the idea that scientific and applied research should be implemented to add value to this educational institution and the society that surrounds them in an efficient, quicker and measurable way. This can be seen from the continuously raising of questions and strategies for continuous improvement, like Tec 21, and excellence that this institution has been striving for. In addition, Tec de Monterrey has always pursued its own evaluation, in this specific case, through research-professors in relation to the quality of the teaching they provide to their students. In addition, this academic institution holds itself to the highest standards of collaboration with partner universities and contributions to academic progress, as well as commitment to society and resolution of socio-economic problems in Mexico and around the world. This constant effort has been translated to an improvement in the QS World University Rankings of more than 20 positions with respect to the previous year, standing out as the number 1 private university in Mexico. On the other hand, Tec is now identified by the QS ranking employers' opinion as one of the universities that produces the best graduates for the labor market and it has advanced to the 53th place in employers' reputation in the world.

During the August-December 2018 semester, Tecnologico de Monterrey had around 9,900 professors which more than 1,700 where listed as Teaching-only professors and nearly 600 as Teaching-and-research professors. Moreover, Tec had approximately 400 active PhD students, around 6,500 master's students, 76 postdoctoral students, 18 distinguished professors and 42 research groups. Additionally, during 2018 to the January-May 2019 semester, more than 8 thousand students were doing research. Furthermore, in the 2014-2018 period, almost 4,400 scientific projects were published which translated to approximately 18,000 citations, 107 patent applications and 102 granted patents.

As we can see, at Tec the research professor model is known as a teacher who dedicates at least half of their schedule to research. Consequently, these strategies have caused an increase in the number of scientific publications and the number of citations . Additionally, this academic institution has created strategic alliances with highly recognized academies internationally. We see an example of this in the nanoscience and nanotechnology area where Tecnologico de Monterrey and the Massachusetts Institute of Technology (MIT) have a strong alliance. In a broader perspective, Tec is constantly improving and, with the intention of being an international leading university, recently they have exploiting their research activities and collaborations.

Tecnologico de Monterrey wants to compare the teaching performance of teaching-only versus teaching-and-research professors. At Tec and other institutions there is a common

belief that, generally, teaching professors outperform research professors in teaching and research universities, according to student perception reflected in student surveys. In that context, it is necessary to think about the relationship between research and teaching illustrated as a mutually beneficial relationship, reciprocality and alliance [31] [57] [58]. Since we want to identify the relationship that could exist in this institution between its quality of teaching and its research activities, we seek to focus on identifying the satisfaction of its students by taking only into account student's opinions towards their teachers. First, based on the intellectual challenge imposed by their teacher in a given course, then on how such a teacher developed as a learning guide according to student's experience in that course, and finally by taking into account whether they would recommend the course with that teacher again or not. By focusing on these three factors, we could identify how different the evaluations are between the professors who investigate and those who do not. We believe that in this way we can contribute to this debate since, according to what has been investigated, few institutions do so in this way, making their differences difficult to show [48].

When it comes to this specific study, the understanding of the business consists of a series of discussions between the Department of Human Development for Research, the Vice-Rectory for Research and Technology Transfer at Tecnologico de Monterrey and the Data Science and Applied Mathematics Research Group. These first two mentioned departments are the ones that have identified, gathered and analyzed these informations. Since they are in charge of the implementation of the ECOA Survey at the end of each semester, and hence, they store and manage the specific data we need. By gathering all those datapoints, making a profound literature review, understanding their business and having the knowledge on how data mining techniques can help on making business decisions, we are intending to visualize and map the situation in a clearer way.

Tecnologico de Monterrey seeks to identify the specific characteristics that their professors must have to satisfy university's quality standards and students' needs. We believed that, by paying more attention to the ECOA database, specifically in a data analytic way, interesting facts could be identified that can justify future academic decisions and improvements in academic services can be made. Since university rankings have become a very relevant factor in the last decade, there is a need to measure quality and academic performance effectively. One way to do this is by making an in-depth analysis of students' main comments about their teaching staff, their academic resources and the courses offered to them.

Tec wants to know what are the main reasons why there are still discrepancies between what the institution believes it is best and what their students are perceiving and evaluating after completing specific courses. By addressing this problem in a data-driven way, we believe that we can find those main reasons and apply effective solutions in the future. The way in which this research seeks to address this problem is to make a comparative analysis of the academic performance of professors dedicated solely to teaching and professors dedicated to teaching and research. Since most of the indicators taken into account to qualify a university are related to research activity present in the institution and to the academic reputation exhibited by students and industry, Tec is constantly forced to implement innovative academic strategies. The intention of this is to increase the teaching quality of the institution and that these same efforts can be reflected in its worldwide reputation.

Today, this institution seeks to know who has been better evaluated over time, either full-time professors or research professors, and to identify which aspects have made one be better

evaluated than the other. By making this discovery, we intend that many institutions with several characteristics as Tec de Monterrey will know what how can they approach to their data, which techniques they could use and which measures they could take when hiring new academic staff taking into account a balance between the different profiles existing among professors and what important higher education networks are demanding.

3.2 Data Understanding

Data understanding derives from a good comprehension of the business problem domain. According to the CRISP-DM Methodology, this second phase starts with an initial data collection in order to get familiar with the translation of the problem to data and start proceeding with activities. At this section, we identify data quality problems, discover insights, detect interesting behaviours and start constructing hypothesis of what actually is happening behind this information [71]. Since this project will analyze university teaching evaluation from a data analytic approach, there has to be a close relationship between business and available data. The Department of Human Development for Research is intended to provide us more than 4 years of information gathered from ECOA's student survey which is applied at Tec at the end of each semester. This can be translated into more than 120,000 instances which contain the evaluation score of each professor at a given course with certain characteristics. At this stage of the project we will be focusing in understanding all the variables that can affect teaching evaluation in several ways. At this moment there is a recognition of feature's meanings and an interpretation of specific results. Basically the activities to perform involve collecting initial data, describe it, perform a complete exploration and report its quality.

For the case study, data is taken from the ECOA student satisfaction survey administered by the Registry office at Tecnológico de Monterrey. ECOA is answered at the end of each semester by students from 26 university campuses that conform Tecnológico de Monterrey, to get feedback about the professors' teaching performance of various academic periods that include semesters and quarters. Student satisfaction is measured through three questions of the ECOA survey on a 1 - 10 scale (where 10 is the best and 1 is the worst score). The survey questions comprise: (1) Intellectual challenge (RET), (2) Learning guide (APR), and (3) Recommended Teacher (REC). The first question asks how intellectually challenging the teacher was during his course. The next one refers to how good was the teacher as a learning guide during your course. By last, the third question refers asking the student, from 1-10 how likely is it that you would recommend some other student to take a course with that professor. A record of the survey contains the average score and the number of students who answered each of those questions about their satisfaction for a given teacher in a given class. This score is the most accurate measure we have of teaching quality from the students' perspective.

We use three data sets in this study. The first dataset, which we identify as "Dataset A", is constituted by responses for the semester August - December 2017 and contains 15,781 records, each of them being a specific professor's performance in a certain course during a specific semester. This dataset consists of 60 features that describe the professors' teaching and research activities and the attributes of the class. Professor's attributes include Nationality, Gender, Age, Campus, School, Department, Maximum Professional Degree, Teaching Certificate, Semesters of Experience, membership on the National Researcher System (SNI),

Total Teaching Hours, and the number of papers published in Scopus in the last 5 years (Conference Proceedings, Journal Articles, and Total publications). Class features describe the attributes of the class, such as the number of credits, number of students, the level of students (undergraduate or graduate) and most importantly, the answers to ECOA questions where students evaluate the professor performance for each course they take.

We performed an initial exploratory analysis using Dataset A . In this analysis, student's teaching satisfaction was analyzed calculating the average of the score obtained by the professor in the three questions of the survey. Later on Tec de Monterrey provided us a second dataset, which we identify as "Dataset B", with the intention of performing a more complete analysis. It contains responses of five semesters, January 2017 to June 2019, this means that it is 5 times much larger than the former. This can allow us to perform a longitudinal analysis over five periods of time, analyze the three questions separately, have a better understanding of teaching performance at Tec in the past few years and develop a comparison with higher quality between teaching-only professors and teaching-and-research professors. However, this dataset lacks of the variety of features that Dataset A presented. Because of this, we would not be able to perform more or the exact same analysis applied to the Dataset A.

Our third dataset is very similar to the previous one. Actually, we could say that dataset B is a subset of this dataset which we identify as "Dataset C". The reason of this is that our third dataset contains the responses of an extra year, 2016. That year is constituted of two semiannual and quarterly academic periods. The records of 2016 present only the average answers of two questions, Intellectual Challenge (RET) and Recommended Teacher (REC), this means that it does not present the instances coming from the Learning Guide (APR) questions. Dataset C has more than 150,000 records and it will be used for performing statistical models with a different focus. It is worth to mention that all the work related to this dataset was developed in conjunction with Dr. Gabriela Torres Delgado, who is part of the evaluation committee.

3.2.1 Dataset A: Exploratory Data Analysis

We did an exploratory analysis of the first dataset using statistical techniques to compare teaching-only with teaching-and-research professors. As shown in Figure 3.2, the mean ECOA score for the three questions, RET, APR and REC of teaching-and-research professors, called SNI and identified in green, is greater on average by 0.32 than the one for teaching-only professors or NO SNI, illustrated in orange.

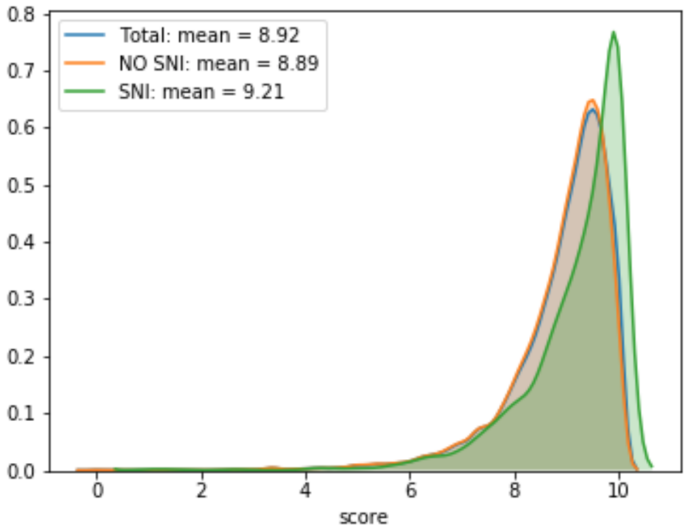

Figure 3.2: Score distribution of teacher-only professors (NO SNI) and teaching-and-research professors (SNI)

On the other hand, if we split Dataset A and analyze the responses of undergraduate and graduate students, we observe that graduate students evaluate professors better compared to undergraduate students 3.3, in other words, according to this dataset, graduate students give higher scores to professors in the ECOA survey.

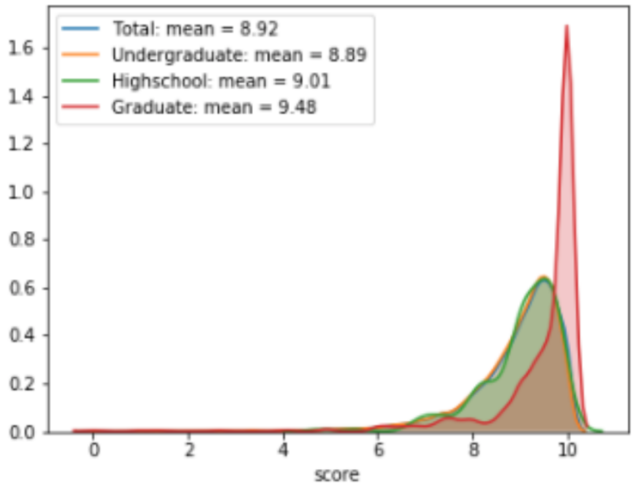

Figure 3.3: Score distribution of professors at the three academic levels (Highschool, Undergraduate and Graduate), and overall (Total).)

Moreover, Figure 3.4 illustrates how our results look when we split our data by SNI/NO SNI and academic levels. Dataset A helps us understand that teaching-and-research professors received a higher score from graduate students, 9.21 over 8.98 for the teaching-only professors, whereas undergraduate students almost do equal evaluation of both type of professors.

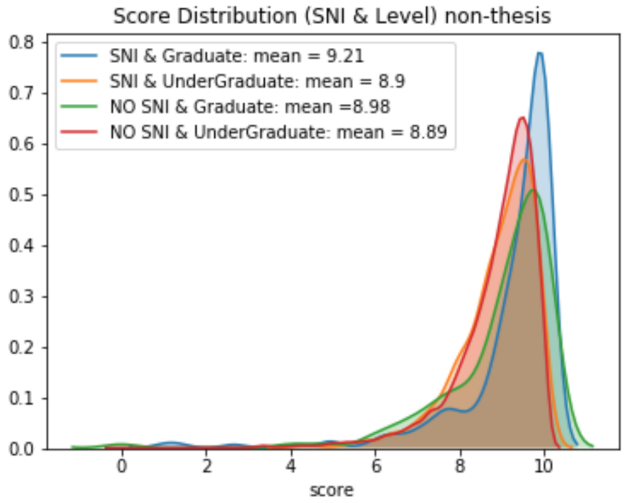

Figure 3.4: Score distribution of Teacher-only Professors (NO SNI) and Teaching-and- Research Professors (SNI) at graduate and undergraduate group.

3.2.2 Dataset B: Exploratory Data Analysis

The second dataset contains responses collected during 5 semesters, from January 2017 to June 2019, with 118,818 records in which 9,469 professors were evaluated several time. From them, 626 are considered as researchers, given that they were classified in the National Research System (SNI) of Mexico during the period of 2016-2019. As we mentioned before, dataset B presents less features than Dataset A, however, one of the most important features that it does include is the maximum proficiency level of researchers, in a scale 1-4. Scale 1 is for those teaching-and-research professors who are candidates to be part of Mexico's National Research System, in other words, they have been doing research for a while and the system are evaluating them. Scale 2 is where Level 1 SNI Researcher belongs. Scale 3 is for Level 2 SNIs and finally Scale 4, which the upper proficiency level, is where teaching-and-research professors with Level 3 of SNI belong. We believe that this feature is of high importance since we could see if this level affects professor's evaluation in a positive or negative way.

In order to make a deeper analysis of student satisfaction in the three dimensions measured by the ECOA survey we used this second dataset. As we mentioned before, this dataset contains responses from 5 semesters. Analyzing the average score for the three questions, in this case separately, on the five semesters, we have observed that both Intellectual Challenge (05. RET) and Learning Guide (06. APR) have similar scores, 9.06 and 9.01, respectively, whereas Professor Recommendation (08. REC) is below with 8.73 in average. Nevertheless, when we compare the responses for full-time professors and researchers, we observe differences. As shown in Figure 3.5, the average score obtained by researchers is higher than those obtained by full-time professors, in the three questions.

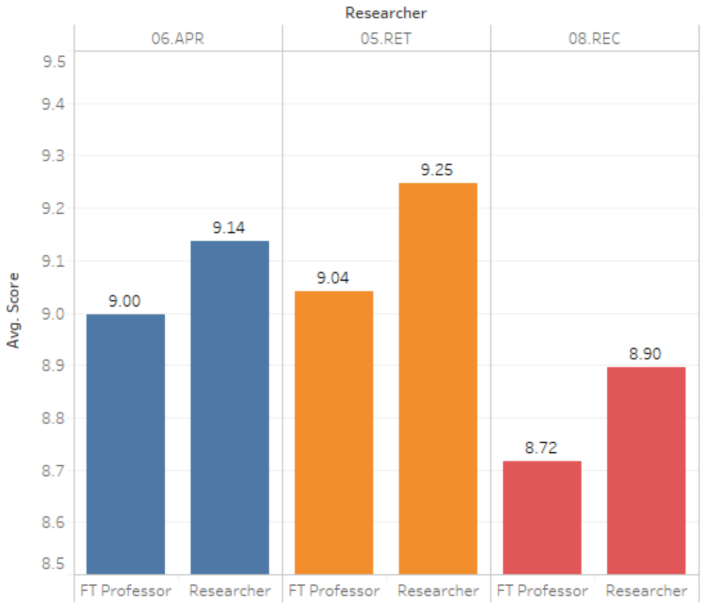

Figure 3.5: Differences between the average score of full-time professors (FT Professor) and researchers (Researcher) at the three dimensions: Learning Guide (06. APR), Intellectual Challenge (05. RET) and Professor Recommendation (08. REC).

Next, we analyzed the evolution of the three scores along the five semesters. Figure 3.6 shows the different average scores obtained. It can be observed that the three of them increase over time. Question 08. REC went from 8.70 in the January-June 2017 semester to 8.81 two years later in 2019 ($\Delta = 0.11$). In the case of the intellectual challenge question (05. RET), it went from 9.02 to 9.12 ($\Delta = 0.10$). Finally, a lower delta was produced for the Learning Guide question (06. APR) since it passed from 8.98 to 9.06 ($\Delta = 0.8$).

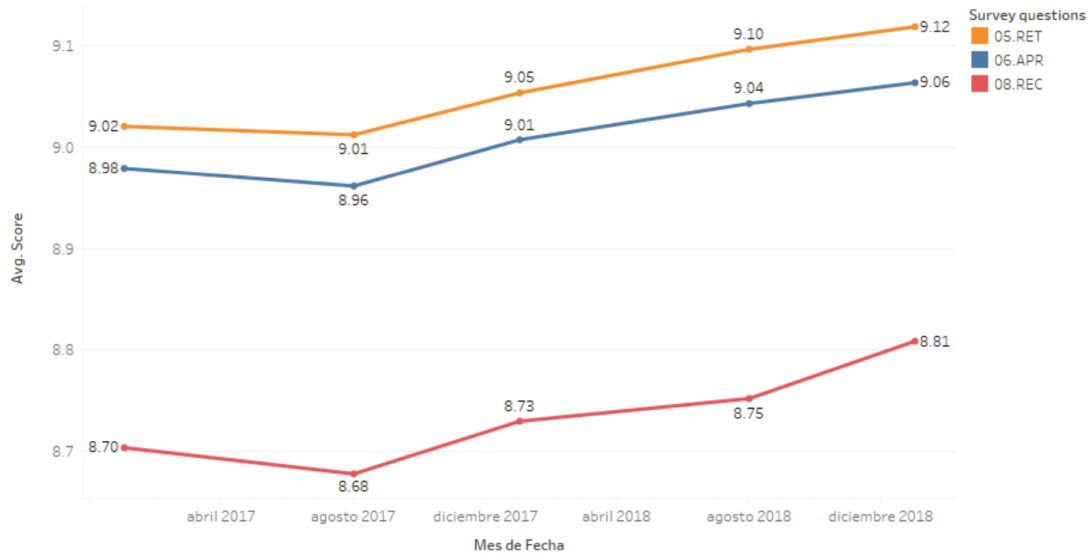

Figure 3.6: Temporal evolution of the three dimensions: Learning Guide (06. APR), Intellectual Challenge (05. RET) and Professor Recommendation (08. REC).

As shown in the previous subsection, graduate students score higher than undergraduate students to their professors. Furthermore, when we analyze Dataset B, as Learning Guide (06. APR), undergraduate students score professors with 9.00 in average, in contrast to graduate students that score professors with 9.33 ($\Delta = 0.33$). In terms of Intellectual Challenge (05. RET), undergraduate students score with 9.05 whereas graduate students score 9.42 in average ($\Delta = 0.37$). Moreover, the difference is still more notorious on the Professor Recommendation score (08. REC), where undergraduate students give professors an 8.71 score, whereas graduate students score them with 9.28 ($\Delta = 0.57$).

Nevertheless, when we break down both dimensions, students' academic level and research activities, the differences between researchers and full-time professors vanish at undergraduate courses. Figure 3.7 shows the differences in satisfaction scores between teaching-and-research professors (identified in the figure as Researcher) and teaching-only professors (identified in the figure as FT Professor), for both undergraduate and graduate students. Evidently, graduate students score higher to researchers at the three scores. Nevertheless, undergraduate students score only slightly higher to researchers on the Intellectual Challenge factor, but in Learning Guide and Professor Recommendation both are scored almost the same. Once again, the biggest difference is in Professor Recommendation question (08. REC), where the teaching-and-research professors are scored 0.68 points higher than teaching-only professors.

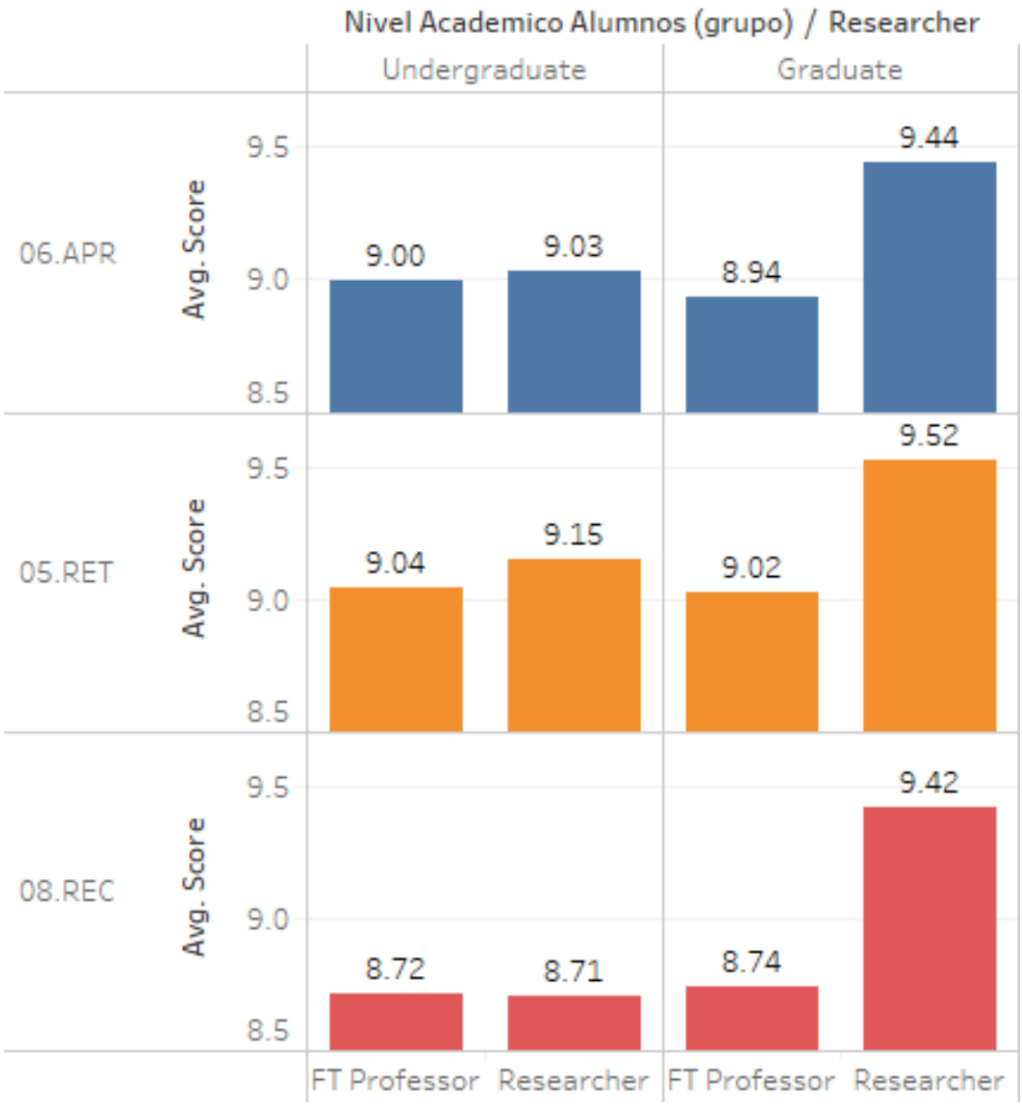

Figure 3.7: Differences between teaching-only professors (FT Professor) and teaching-and-research professors (Researcher) with undergraduate and graduate students, at the three dimensions: Learning Guide (06. APR), Intellectual Challenge (05. RET) and Professor Recommendation (08. REC)

Figure 3.8 shows the average scores obtained by women and men when they are evaluated as professor in the three dimensions. Whereas at Intellectual Challenge there is practically no difference, in Learning Guide and Professor Recommendation women slightly overcome to men by 0.07 and 0.04 points, respectively.

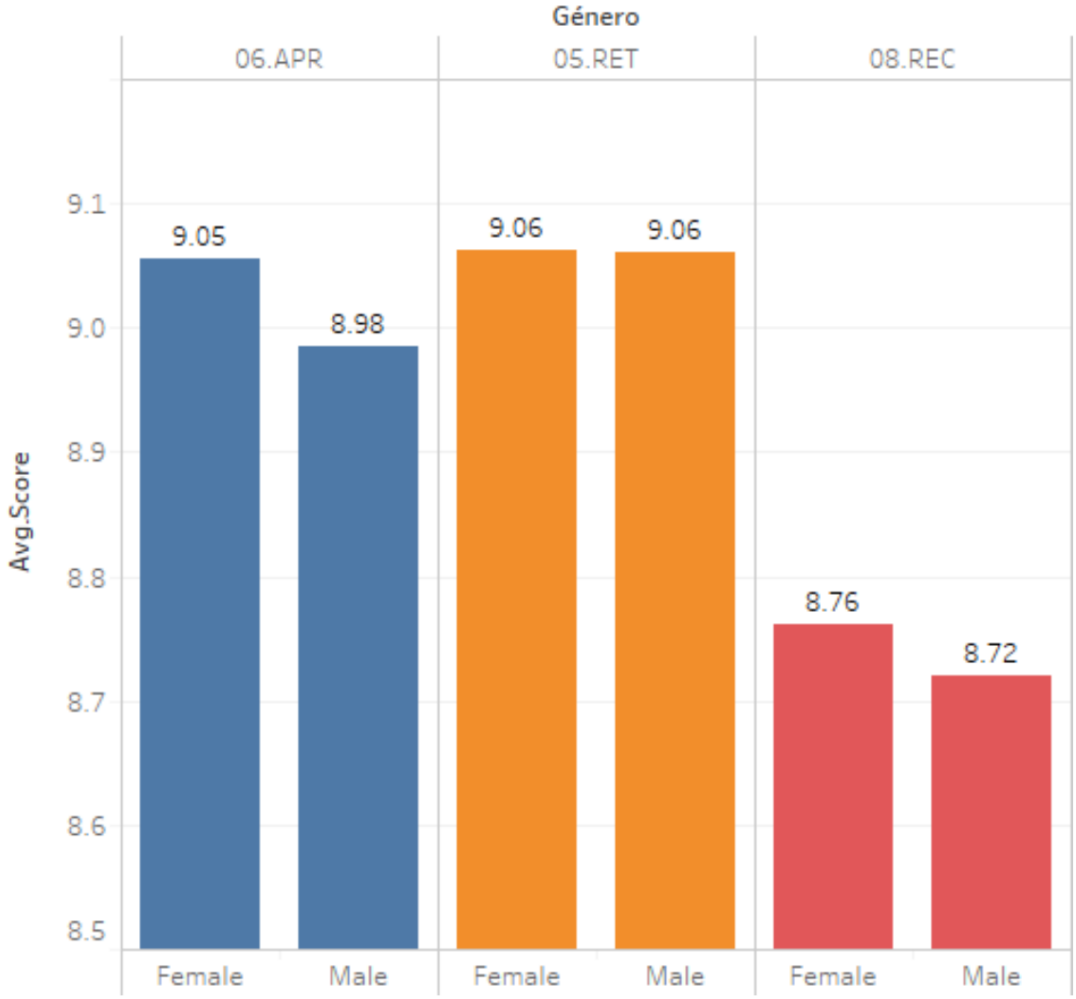

Figure 3.8: Differences on gender at the three dimensions: Learning Guide (06.APR), Intellectual Challenge (05.RET) and Professor Recommendation (08.REC).

Nevertheless, by splitting the sample between researchers and full-time professors the differences become more notorious. Figure 3.9 shows the differences between women and men once they are divided into researchers and full-time professor; bar sizes indicate the number of professors evaluated. On one hand, male researchers overcome to female researchers in the three dimensions, being Intellectual Challenge and Professor Recommendation those with the highest differences, 0.10 and 0.09 points respectively). On the other hand, female full-time professors overcome to male professors only two dimensions: Learning Guide ($\Delta = 0.09$) and Professor Recommendation ($\Delta = 0.07$); at Intellectual Challenge, both men and women are evaluated the same [29].

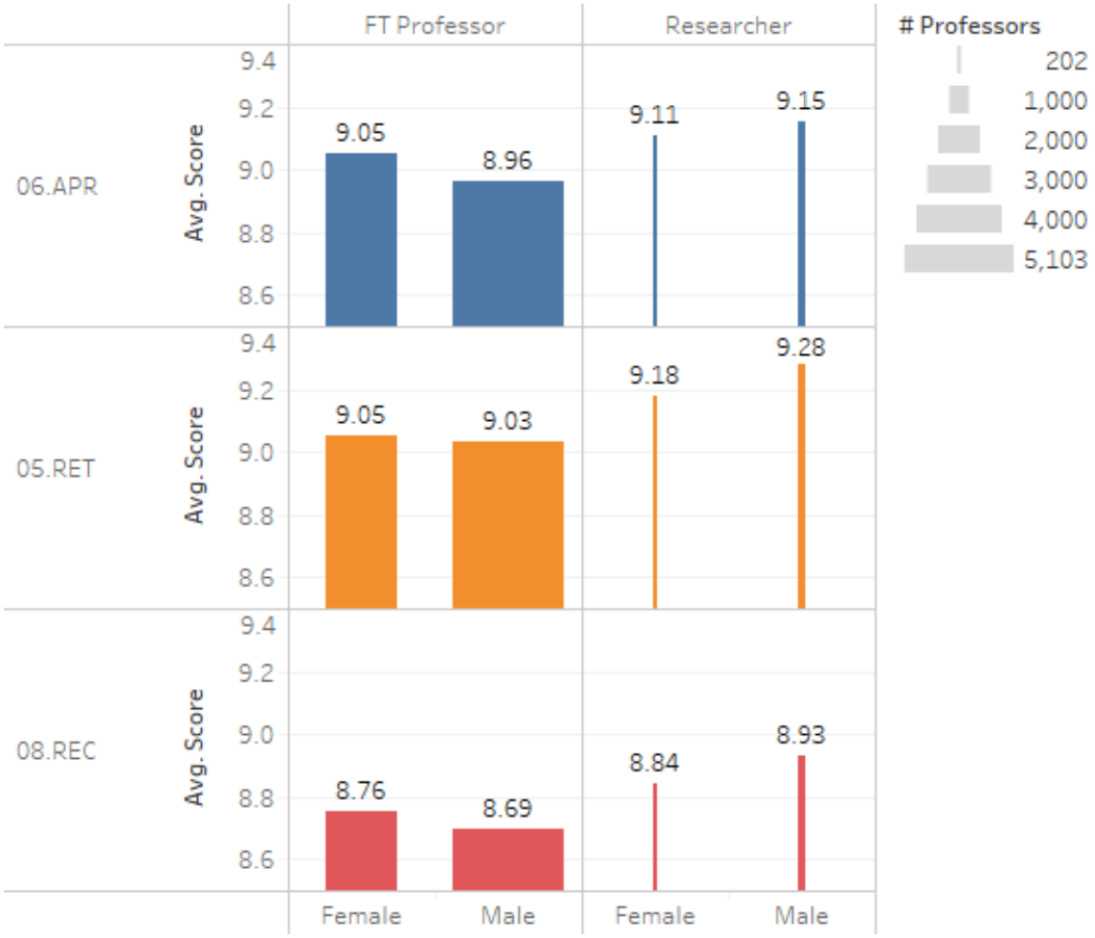

Figure 3.9: Differences by gender between teaching-only professors (FT Professor) and teaching-and-research professors (Researcher), at the three dimensions: Learning Guide (06. APR), Intellectual Challenge (05. RET) and Professor Recommendation (08. REC). Bar size indicates the number of professors at each group.

Next, we analyzed how professor's age influences student satisfaction, either for teaching-only professors (Researcher), and teaching-only professors (FT Professor). Figure 3.10 shows the distribution of professors by age, divided into segments of 10 years. It can be observed that most of the professors are between 30 and 49 years old. Also, we can see that it is not very common for Tecnologico de Monterrey to hire teaching-and-research professors younger than 30 years old, or it is not usual for professors to be considered as a SNI candidate while in their 20s.

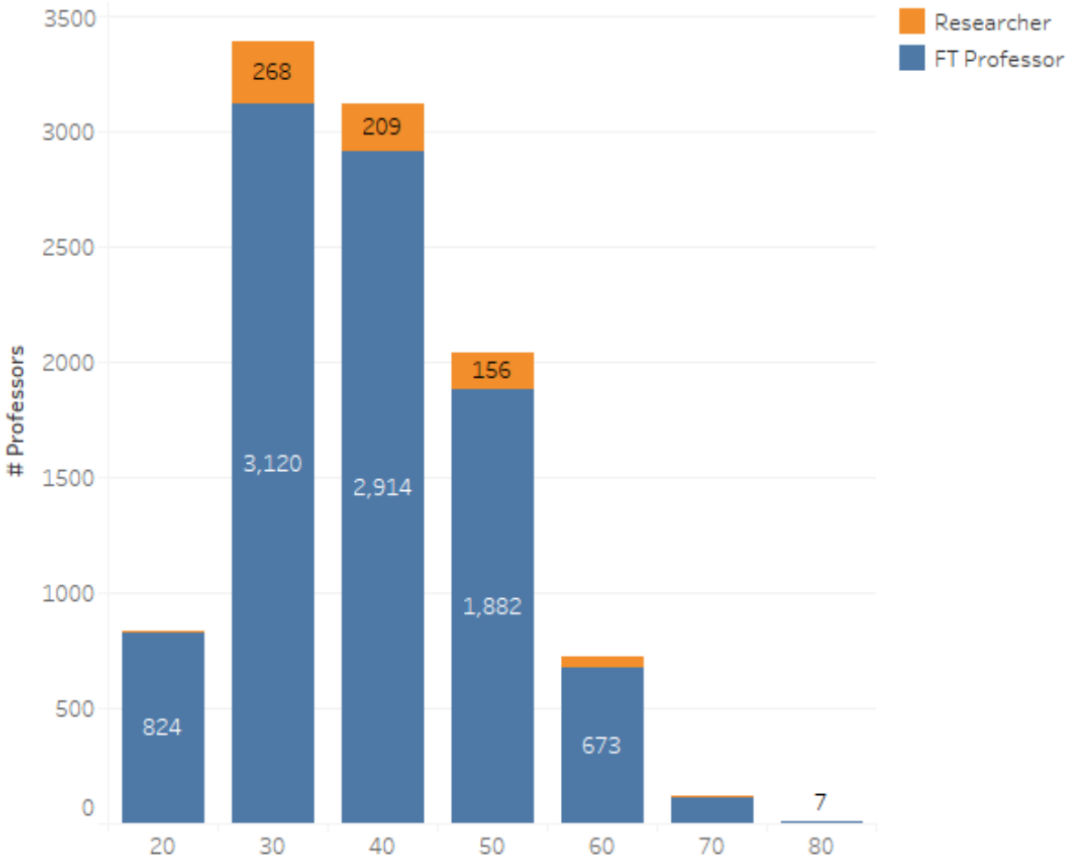

Figure 3.10: Distribution of professors by age (in 10 years bins).

Next, we split our sample between teaching-only professors and researcher in order to observe how aging affects teaching quality for both types of professors. For teaching-only professors, aging affects negatively student's perception of teaching in the three dimensions, as shown in Figure 3.11. For the Learning Guide factor, the difference average score between the youngest and oldest teaching-only professors is only -0.07 points. On the other hand, for the Intellectual Challenge question we recorded a difference of -0.13 points. Finally, for question 08. REC (Professor Recommendation) our figure illustrates us a negative difference of 0.34 points.

Figure 3.11: Average satisfaction scores of teaching-only professors by age, in the three dimensions: Learning Guide (06.APR), Intellectual Challenge (05.RET) and Professor Recommendation (08.REC).

In contrast, for teaching-and-research professors, aging improves teaching quality according to student's ECOA responses. As shown in Figure 3.12, for the Intellectual Challenge factor, the difference between the youngest and the oldest researchers is +0.22 points. For the Learning Guide question, the average increases by 0.21 and finally, for question 08. REC (Professor Recommendation), the average score of the ECOA of oldest professors is 0.27 higher than the younger ones. In addition, we need to have in consideration the amount of professors that conform each bin.

Figure 3.12: Average satisfaction scores of teaching-and-research professors by researcher age, in the three dimensions: Learning Guide (06. APR), Intellectual Challenge (05. RET) and Professor Recommendation (08. REC).

The last exploratory analysis we performed for Dataset B is shown in the following figure. We examined the differences in student satisfaction due to the proficiency level of researchers. For that purpose, we used the four levels of proficiency conferred by the National Research System of Mexico (SNI) to our researchers. Figure 3.13 makes evident that teaching quality grows with the proficiency level of the researcher at the three first dimensions. For the last to dimensions, SNI Level 2 and SNI Level 3, we can say that they present the same teaching quality since students give them almost the same scores. For the Intellectual Challenge factor, the difference between the lowest and the highest proficiency level is +0.35, whereas for Learning Guide is +0.32 and for Professor Recommendation is +0.46.

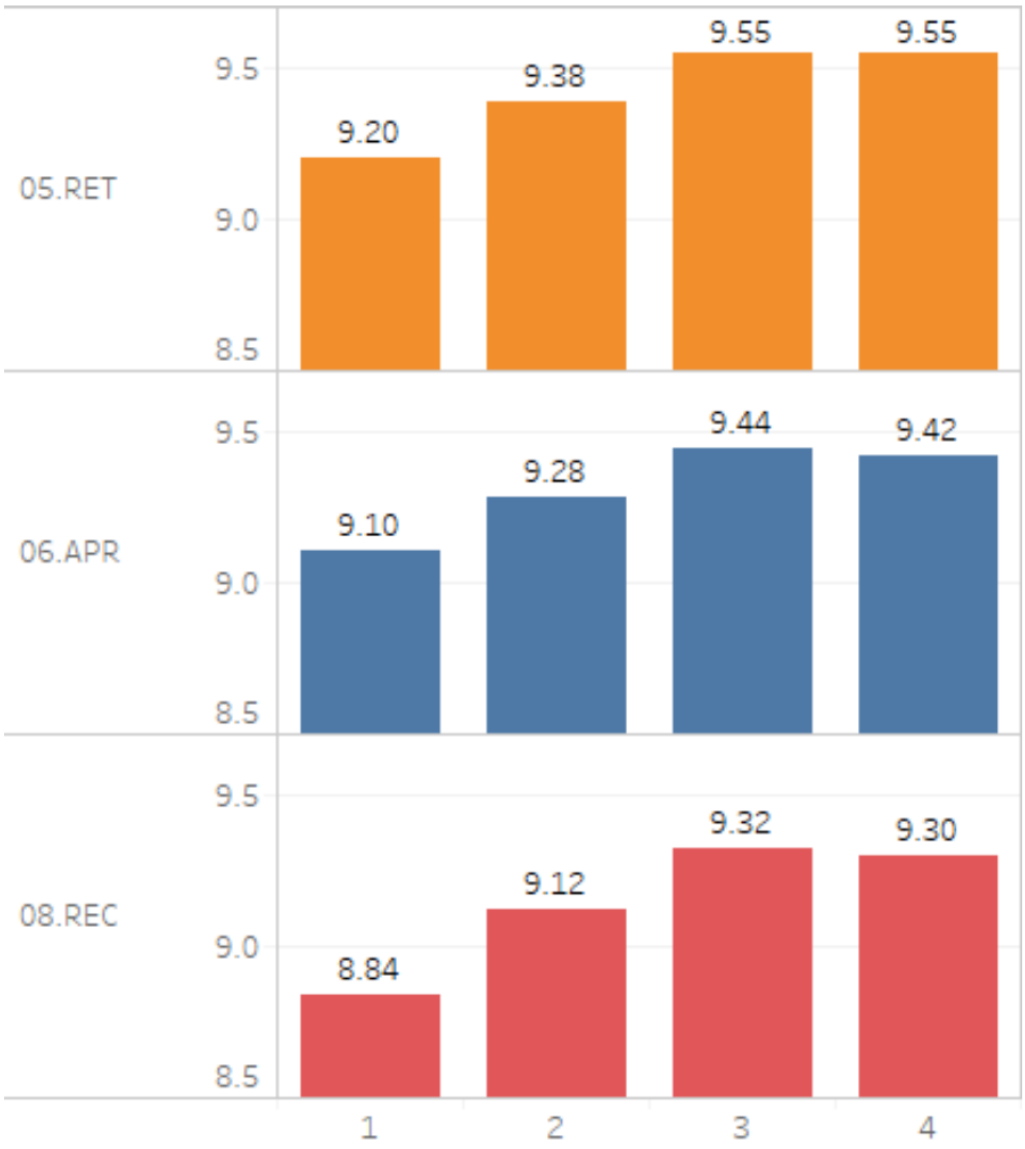

Figure 3.13: Average satisfaction scores of teaching-and-research professors by proficiency level in the three dimensions: Learning Guide (06. APR), Intellectual Challenge (05. RET) and Professor Recommendation (08. REC).

3.2.3 Dataset C: Exploratory Data Analysis

For our third dataset, in which we consider 4 years of ECOA's records and only Intellectual Challenge (05. RET) and Recommended Teacher (08.REC) factors, we got the following analysis. An important note here is that we are only taking into consideration the similar courses given by both type of professors, Teaching-only and Teaching-and-Research Professors, in each period. First, we carried out the comparison of means 2016, 2017, 2018 and Jan-May 2019 in biannual and quarterly periods. We can see that teaching quality is as well in favour of teaching-and-research-professors in the four periods about student satisfaction with positive differences. Figure 3.14 helps us visualizing this.

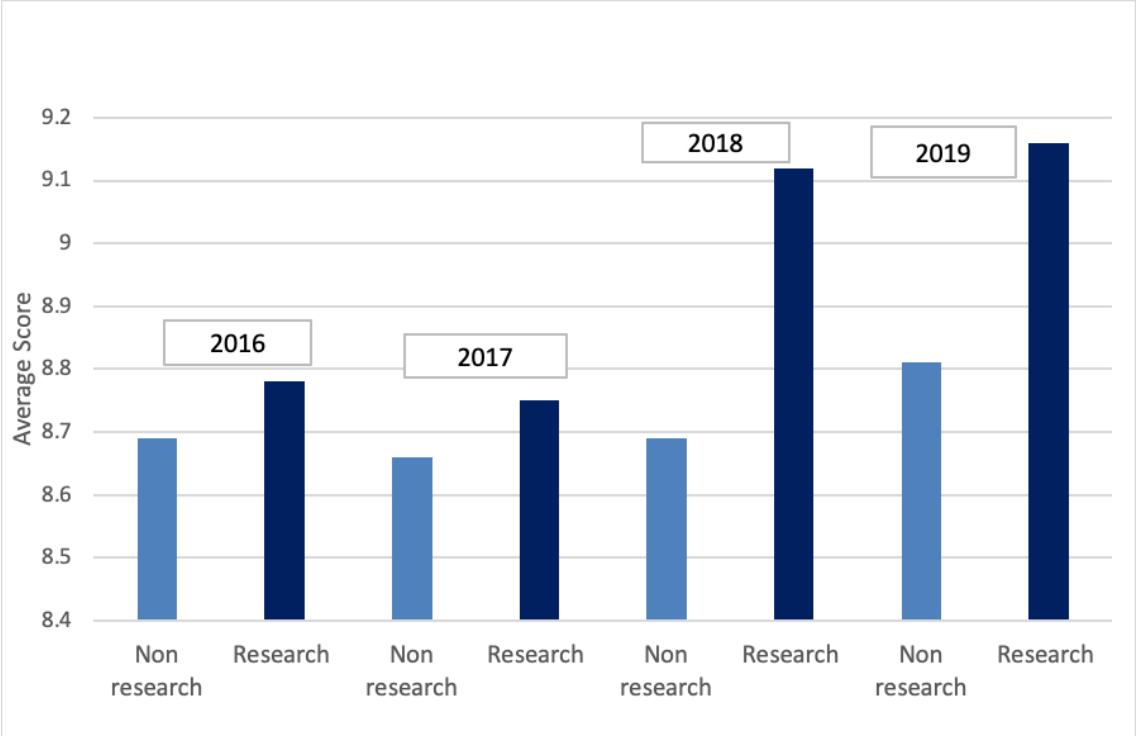

Figure 3.14: Professor Recommendation (08.REC) average score between Non research professors and Research professors across years

On the other hand, when we consider the intellectual challenge factor, specifically in 2018 and 2019 periods, we identify a great difference in student satisfaction scores. Figure 3.15 illustrates that teaching-and-research professors challenge more intellectually to their students rather than teaching-only professors. In other words, the intellectual challenge factor is in favor for the research professors.

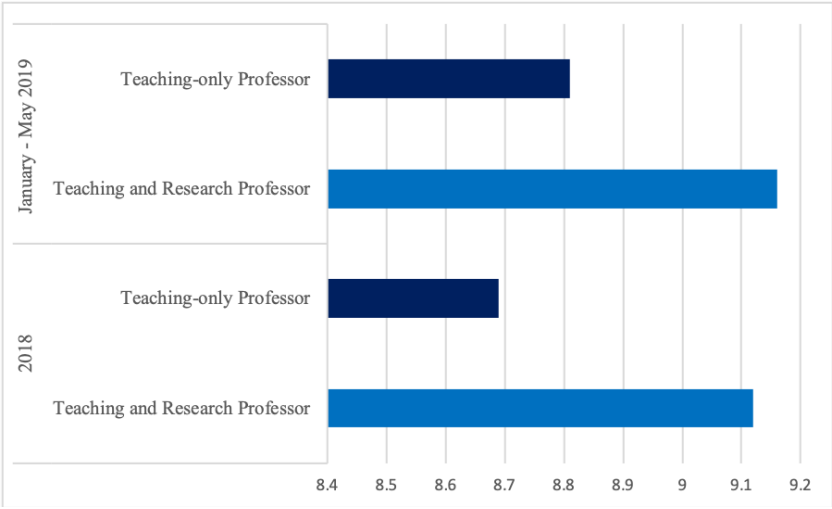

Figure 3.15: Intellectual Challenge (05.RET) scores of Teaching-only and Teaching-and-Research Professor between 2018 and 2019

When we filter our data by the 6 schools that are part of Tecnologico de Monterrey's academic programs, we got the following results. For our first school, Humanities and Education, satisfaction scores look way higher for teaching-and-research professors (identified in the figure as *Research*) than teaching-only professors (identified as *Non researcher*) in both of our factors. We consider that, in this particular school, researchers get very high scores since almost all of them are above 9.5 out of 10. In addition, we can identify that for both type of professors, the average records got lower through time.

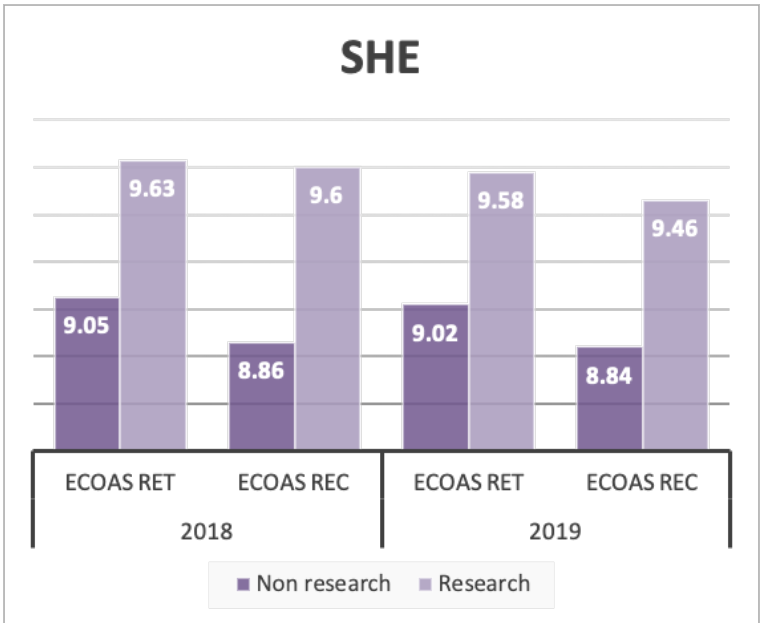

Figure 3.16: School of Humanities and Education scores in Intellectual Challenge (05.RET) and Professor Recommendation (08.REC)

In the case of the School of Engineering and Sciences, the satisfaction scores granted to teaching-and-research professors are higher than teaching-only professors as well, for the two type of questions. We can see this in Figure 3.17. In general, averages are not that high as our last figure but we can see a remarkable difference between each type of professor. Moreover, every score increased with time except for the Professor Recommendation factor in teaching-and-research professors.

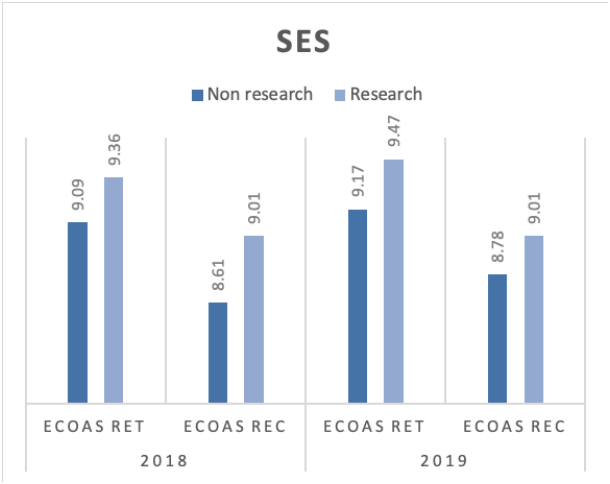

Figure 3.17: School of Engineering and Sciences scores in Intellectual Challenge (05.RET) and Professor Recommendation (08.REC)

Furthermore, we analyzed Tec's School of Medicine and Health Sciences, Figure 3.18 shows our results. Once again, for both of our questions and in both periods, teaching-and-research professors were scored higher than teaching-only ones. The recommended professor factor was the only one who suffered a decrease from one year to another in the section of teaching-and-research professors.

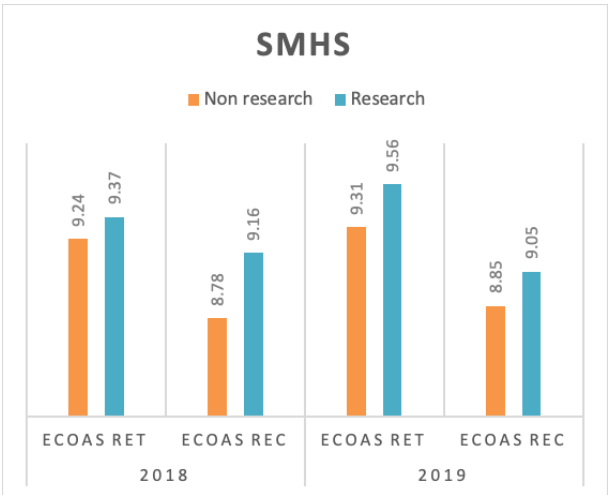

Figure 3.18: School of Medicine and Health Sciences scores in Intellectual Challenge (05.RET) and Professor Recommendation (08.REC)

The School of Business was not an exception in which research-and-teaching professors performed better. In this case, one of the main findings we got through Figure 3.19 was that none of the average scores for teaching-only professors surpassed the 9 over 10 points. Up to this point, the School of Business has been the one with the lowest scores for these type of professors.

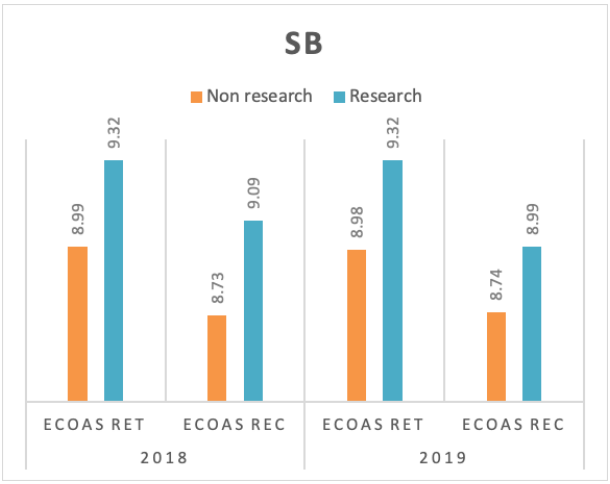

Figure 3.19: School of Business scores in Intellectual Challenge (05.RET) and Professor Recommendation (08.REC)

In contrast, the School of Architecture, Art and Design presented the lowest scores of all. 2018 was not a good year for them in the ECOA's results. Figure 3.20 represents the first exception, this means that teacher-only professors got higher scores in the ECOA than teaching-and-research ones, only for the 2018 year in both questions. When we take a look at the 2019 year, we can see the same type of results as the past figures. It is worth mentioning that only 1 average score was recorded to be higher than 9.

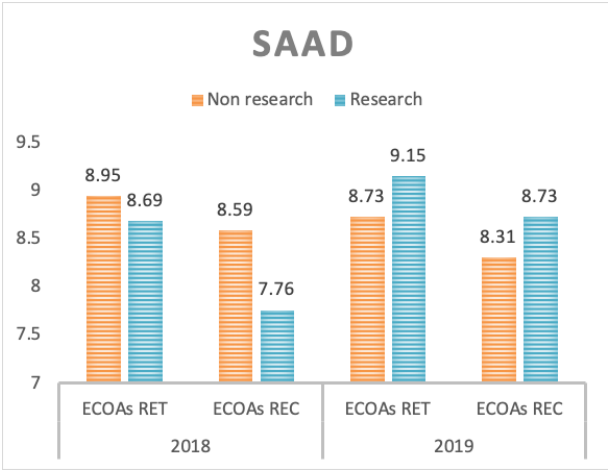

Figure 3.20: School of Architecture, Art and Design scores in Intellectual Challenge (05.RET) and Professor Recommendation (08.REC)

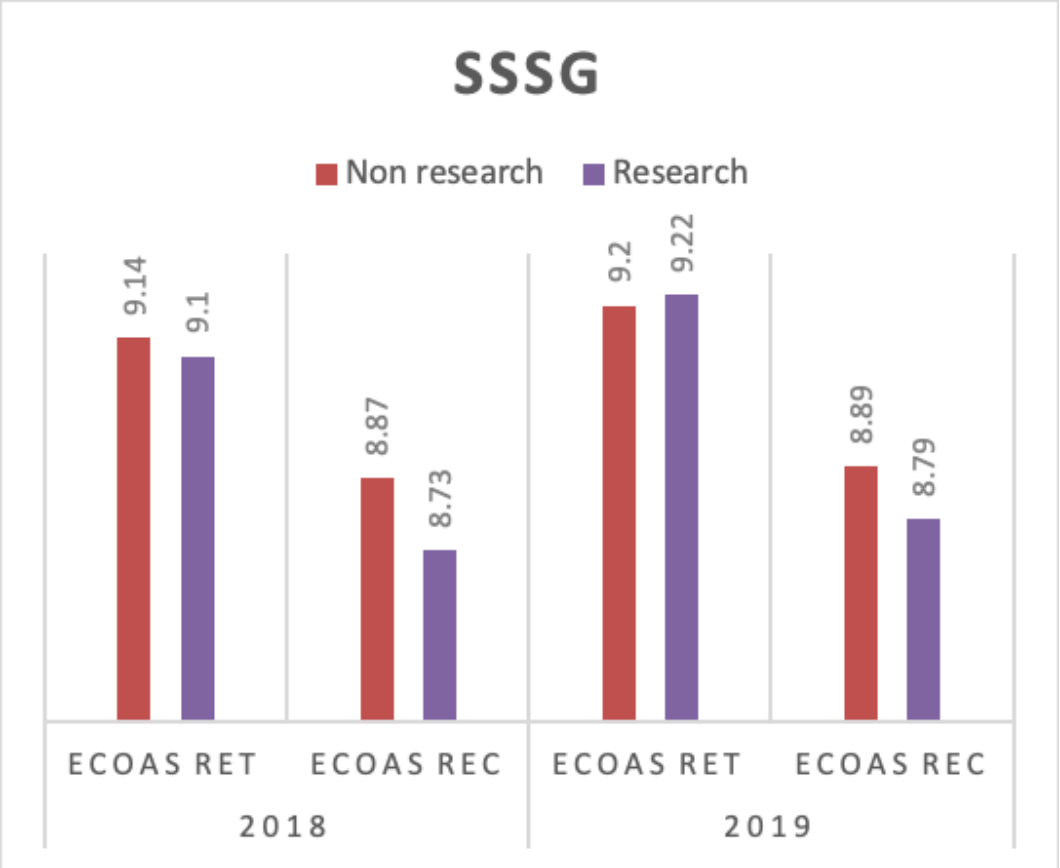

Figure 3.21: School of Social Sciences and Government scores in Intellectual Challenge (05.RET) and Professor Recommendation (08.REC)

The last figure for the exploratory analysis of Dataset C is focused on the Social Sciences and Government School, our second and last exception to cases in favor of teaching-and-research professors. Once again, scores recorded for both questions were lower than other schools, but in this case these scores are more similar between them in both questions for both years. This means that, for this specific school, it seems that students' satisfaction of teaching performance does not depend that much on whether if he or she is a teaching-only professor or a teaching-and research professor. Figure 3.21 illustrates that scenario.

Now that we have understood the different types of data sets that we have, how each of them are composed, what the data tells us about the comparison between teaching-only and teaching and all the features, we will focus in describing how did we prepare them in order to get those results and to further be able to develop several models which will be covered in future sections.

3.2.4 Dataset E: Exploratory Data Analysis

Through this thesis we will see that from Dataset A, B and C, two more were created. We will cover Dataset D in the next subsection, meanwhile, our last dataset is defined as Dataset E. It is characterized for being the preprocesses version of Dataset B. We applied the Coarsened Exact Matching algorithm which essentially pruned almost half of our observations in order to balance our control and treatment variables and reduce model dependence. The following figures cover the Exploratory Data Analysis of this new dataset. It accounts 33,253 records from Teaching-only Professors and 4,572 records from Teaching-and-Research professors.

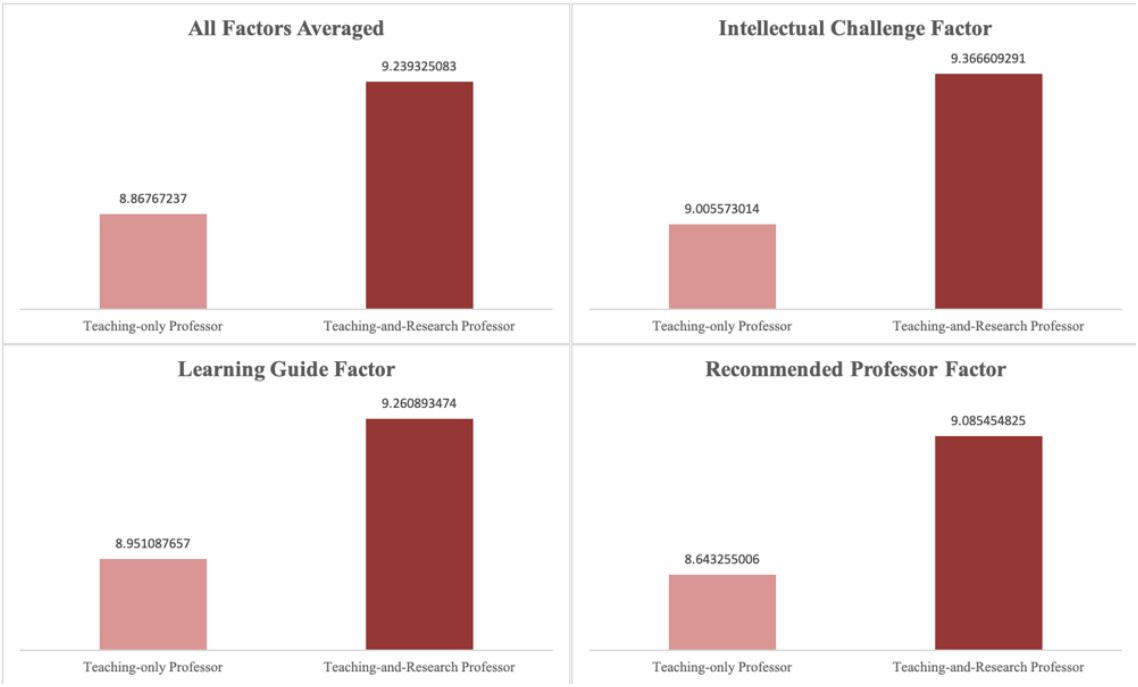

Figure 3.22: Matched Dataset Teaching-only vs Teaching-and-Research Professors

Figure 3.22 illustrates the ECOA scores of the both types of professors in every academic factor, including the averaged one. We can see that at all cases, Teaching-and-Research Professors surpass Teaching-only ones. It is worth mentioning that thesis classes are not included, which tend to create a higher difference between the means of these professors.

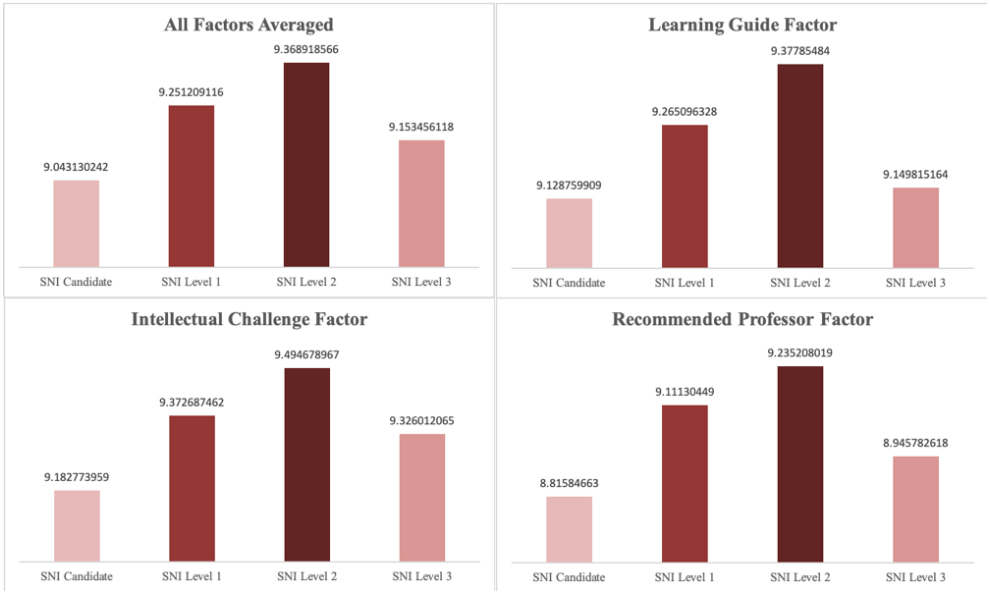

Figure 3.23: Matched Dataset Teaching-and-Research Professors by Proficiency Level

In Figure 3.23 we see the means registered for each of the four different proficiency levels of researchers. At all four cases, Researchers SNI Level 2 were the highest evaluated of all. On the other hand, SNI Candidates were the ones who got the lowest evaluations.

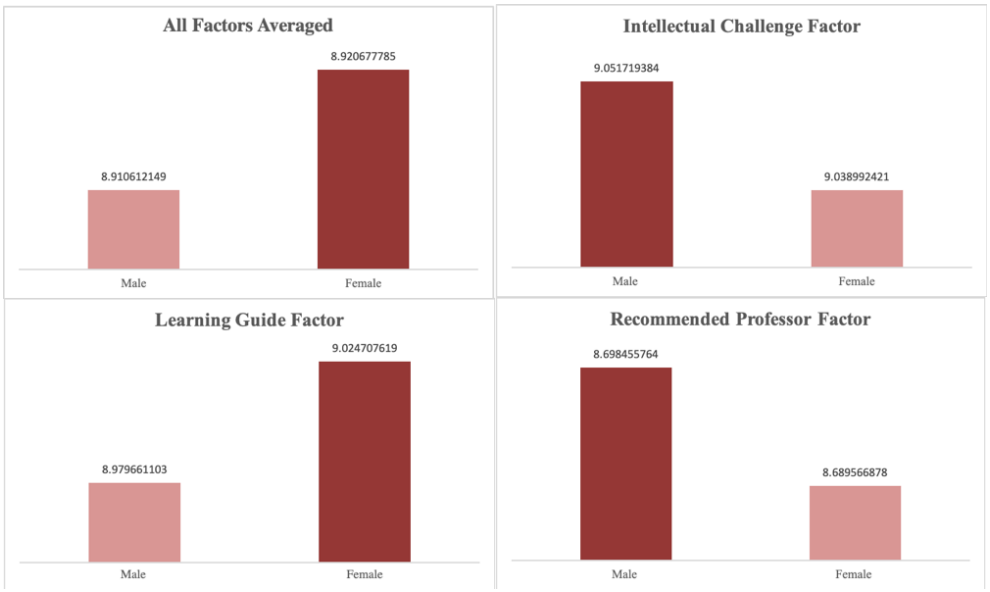

Figure 3.24: Matched Dataset Male vs Female Professors

When looking at Figure 3.24 we can see that when all factors are averaged it seems that there is just a slightly advantage for Female Professors, but the scores are actually very close to each other. Factors such as Intellectual Challenge and Recommended Professor got a similar behavior but in those cases, Male Professsors got a higher mean in the ECOA. We will continue to analyze these figures in the following sections.

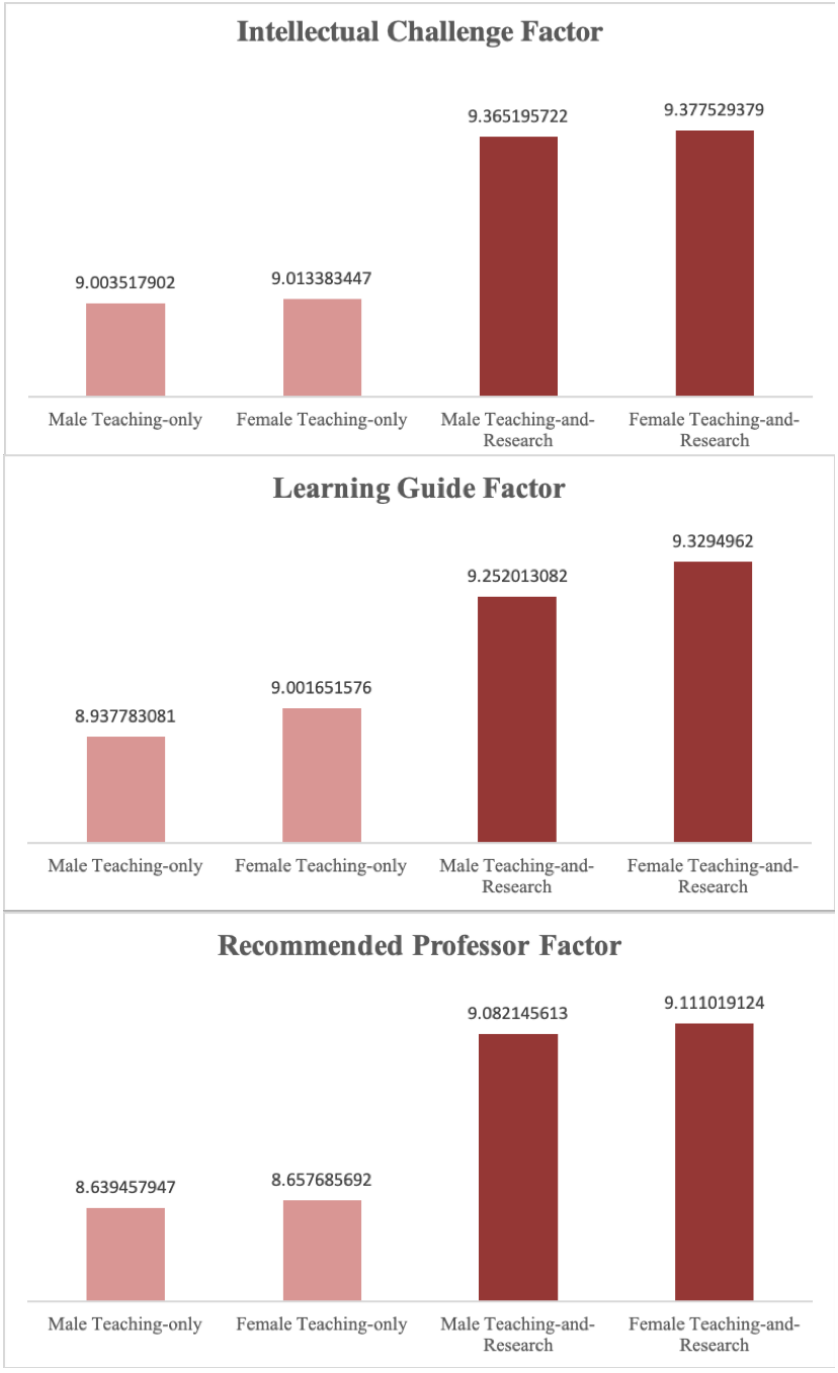

Figure 3.25: Matched Dataset Teaching-only and Teaching-and-Research Professors by Gender

In the case where both characteristics are considered we can see that Teaching-and-Research Female Professors are the ones who are best evaluated of all. In addition, at all three factors, Teaching-and-Research Male professors where better evaluated than Female Teaching-only professors but, Teaching-only Male professors where the worst evaluated by some minimum difference.

3.3 Data Preparation

Usually, when acquiring large data sets, they do not come in a proper format or in a manipulable way. This data preparation phase is part of all the stages of the CRISP-DM methodology, it involves all the necessary activities to construct the final dataset, the one that is ready to be imported into modeling data mining tools. This activity usually is performed several times along the whole process and it goes from data cleaning, eliminating features that do not aggregate value to what we are trying to find, attribute selection, construction of new variables, transformation of data and so on [71].The three data sets described and understood before, will be given a proper format which will make its modelling more feasible. Data requires to be in a specific format, in order to properly read it and perform its manipulation. Basically, to complete this step, data needs to be correctly selected, cleaned, constructed, integrated and reformatted.

There are different data mining tools that are helpful to analyze datasets; these include RStudio, Python, Rapid Miner, Tableau, Weka, Jupyter Notebooks, SPSS, among others. In order to satisfy the constraints of the problem, we decided to manipulate data using Python's programming language on Jupyter Notebooks, RStudio, Tableau and Excel. We used several python and r packages to pre-process data, manage data, visualize it and test classification models.

3.3.1 Dataset A Preparation

To treat this data, first, we omitted variables that had the same value in all the records and features that had a significant amount of null values. For example, the variable age had 5,284 null values, which is one-third of the data. There is no way to fill those values in a good manner, so we decided to drop them. Other features have a high amount of null values, but some of them actually represent a 0, such as the number of published documents by a professor. Next, survey responses were aggregated by professor, and four main variables were created:

- **sni_yn**: We defined a binary variable that is equal to 1 when the professor is a researcher and 0 when he/she is not. A professor was considered researcher if he/she was a member of the National Researcher System (SNI) during the period 2016-2019.
- **score**: We calculated the average of the three student satisfaction scores of a course and decided to use it as the target variable.
- **weighted_score**: To make an analysis of individuals and not of courses, we summarized survey results by professor, so we took the arithmetic mean to summarize the scores that a professor gets in all of his/her courses, weighted by the number of students per group.
- **score category**: Is a binary variable that takes the value 1 if it is above the median and 0 otherwise. This is the classification feature that we use in the logistic regression model. We say that a professor above the median belongs to the alpha type of professors and the ones who are below it belong to beta professors.

3.3.2 Dataset B Preparation

In the second dataset, responses were also aggregated at professor level. Since we received a dataset per each academic period, quarterly and semi annual for each year, we needed to merge all of them. We defined another four variables, one for distinguishing the researcher proficiency and three another for breaking down the weighted score into each one of the questions that comprise student satisfaction.

- **sni**: We classified researchers according to the maximum proficiency level they reached in the period 2016 - 2019. This variable is in the scale 1 - 4, where 1 represents the lower proficiency level and 4 represents the upper one.
- **RET**: It is the professor weighted average score for the Intellectual Challenge question, in scale 1 - 10 where 10 is the maximum score.
- **APR**: It is the professor weighted average score for the Learning Guide question, in scale 1 - 10 where 10 is the maximum score.
- **REC**: It is the professor weighted average score for the Recommended Teacher question, in scale 1 - 10 where 10 is the maximum score.

3.3.3 Dataset C Preparation

We saw in our previous section that Dataset C is indeed very similar to Dataset B, its main difference is that this presents additional records which correspond to the 2016 year. We prepared this dataset with the intention of only considering question 5 and 8, which correspond to how much did the student felt intellectually challenged though the course and how likely is for the student to recommend that class with that specific professor to another student. We filtered this dataset by identifying the weighted averages of both of the questions of the similar courses that were given by teaching-only professors and teaching-and-research professors in a given period. Each year, 2016, 2017, 2018 are constituted by two periods, period A and period B. These periods correspond to the records registered in the semiannual and quarterly academic periods. We prepared the dataset in this way because we wanted to make a deep analysis only on those similar courses imparted between teachers and researchers in the same period. In this way we could say that now our dataset only presents records that are comparable to each other and that this will help us to generate this analysis between research professors and non-research professors. This data segmentation represented a considerable reduction in the total number of records from more than 150,000 to only 15,312 records. In other words, almost 90% of our data was discarded.

At the end this dataset got composed of four variables per semi annual period recorded. Each year was divided in, period A and B, teaching-only and teaching-and-research professor and, intellectual challenge and recommended professor factors. Since we have data of 3 and a half years and we got 4 features in each semi annual period, we got a total of 28 variables.

3.3.4 Dataset D Preparation

Dataset D will be a new dataset for us, the fourth one in this thesis, which is a subset of Dataset B. At first, we analyzed the type of the features in order to see if a variable had wrong values or a different format. Then we identified the amount of null values each feature had. In the case of the average scores, we found that 1% of the values were null, we decided to replace those values with the mean of each feature. For other features, since less than 1% of their values were null, we decided to drop them. Then, we arranged our data in a way that allowed us to see the average scores of the three questions of each professor's evaluations per academic period. This means that if a professor taught more than one lecture in a specific period, we could know in overall how did he or she served in that period and immediately at the following periods as well. In other words, we converted our data into a Time-Series Data which involves measurements over time. When we make the transformation of our normal Dataset B to a Time-Series Dataset D, we got the following format. The first column represented the ID of the professor, then we have the academic period. This was a very important new variable that we had to create since we needed a feature that determine the period to which the teacher's recorded evaluation corresponds. This feature can go from 1 to 5, 1 being data corresponding to the January-May 2017 Semester and 5 the January-May 2019, so each number represented a specific semester in order. So, for example if a professor taught in any of those 5 semesters, the dataset will show its ID the same number of academic periods in which he or she taught a course. Later on the next columns represent our independent variables, which we will mention later and at the end we have our target feature, the averaged score. This new feature represents the averaged score of a certain professor at a specific academic period when all the three factors Learning Guide, Intellectual Challenge and Recommended Professor are taken as one. It is worth mentioning that if a professor taught more than one course in the same academic period the evaluations are averaged regardless of whether they have been similar or different courses. Therefore, the average score can be the average of the evaluation of one taught course or more than one in the same academic period. This is how we converted our dataset into a panel data.

At the end, we hided several features and we left only the ones we wanted to take into account to perform our analysis, these are: Professor ID, Academic Period, Gender, class of the professor's age which can go from class 2 to class 8, this represents that if a professor is 43 years old he is represented in class number 4 or if someone's age is 57 their class would be represented as 5. We also left the variable which say if a professor is also a researcher and finally we added the score variable which represented the average score of that specific professor in a certain academic period. At the end, the purpose of this exercise was to convert this database into panel data. This mean that it contain observations obtained over multiple time periods for same individuals with several characteristics that could or could not change over time.

3.3.5 Dataset E Preparation: Matching

Dataset E is the last one we will use throughout this thesis. It is as well a subset of Dataset B becoming the 5th one of this work. We identify it as *E* when we apply to it the pre-processing method that we saw on our Background section called Matching. The goal of this

was to transform Dataset B into a new one which would present less model dependence in our future experiments. When we say less model dependence we mean prevent that small changes in specifications of our models produce big changes in the substantive results. Model dependence is the variation in our causal effects across the results, which could become a big problem for us. Dataset E is the product of Dataset B when being matched, this results in a new pruned dataset. Matching algorithm will give us a lot more confidence in the kinds of results that we will finally get in the next sections. This 5th and last dataset is approximately half size Dataset B, meaning that when we applied this algorithm nearly half of our data got pruned and features stood the same.

3.4 Modelling

The next step of the CRISP-DM methodology was the modelling task with the purpose of finding interesting patterns and hidden knowledge that may shed light in answering the research question of the study. In order to do this task, we will use statistical methods and machine learning algorithms to identify significant differences between researchers and professor evaluations, and the more meaningful variables that predict the assessment of a professor. The main goal of the selection and application of models is to be able to accept or reject the hypothesis we have presented in previous chapters. By taking into account the type of features and values that our different data sets cover, we are making specific model selections which can help us to contribute to the discussion whether if teaching-only professors are better evaluated by students than teaching-and-research professors. Since our data is now well prepared, the next stage is to select which statistical and machine learning experiments will be applied in order to test our hypothesis.

We will conduct a careful design of experiments that were carried out using various modelling methods. We expect to apply certain techniques which will help us to identify the attributes of the student groups on which researchers and teaching-only professors are best evaluated, Regarding the use of tools, there are different data mining packages both open source and proprietary, that have proved helpful in conducting modelling with datasets; these include Rapid Miner, Weka, Jupyter, SPSS, among others. In order to satisfy the constraints of the problem, we decided to mine data using Python's programming language on Jupyter Notebooks. We used Pandas package to manage data as data frames and the Scikit-learn package to test classification models and to pre-process the data. In the following subsections we will mention the specific models that were chosen to help us in this study among their respective adaptations needed to be applied in each model.

3.4.1 Analysis of Variance (ANOVA)

In the past subsections we have seen a broad application of EDA, Exploratory Data Analysis, among our three different data sets. In order to validate the results that were illustrated from Figures 3.2 to Figure 3.21, we are in need of applying one of the most important and used statistical methods. The method we are taking into account is called ANOVA, and as we explained in the Background Chapter, this analysis of variance will help us to efficiently

compare the difference of the means that we have reported before for each type of professors when several features where taken into account in a specific period. If we pay a little bit more of attention to the past figures, there are several of them which report very similar scores from both type of professors. So, the way in which we plan to know whether if they are indeed statistical different, we intend to apply this method in every dataset for every figure illustrated. The type of ANOVA we will be applying is the Single Factor and we will be using an alpha of 0.05, this means that we will expect our models to present results with a 95% level of confidence.

3.4.2 Logistic Regression

We have discussed before in this work that we would also like to know what makes that a specific type of professor with certain characteristics result on being better evaluated than others. Additionally, we have seen the creation of new categorical variables in the Data Preparation subsection. These categorical variables are denoting the classification of the evaluated professor. Depending on the score obtained by the professor on a certain question, he or she can be classified to either the alpha professors, meaning those who excelled more in the ECOA, or they can be classified as the beta professors. As seen in the Background Chapter, Logistic Regression is used when the dependent variable of the model is of type categorical. Its main goal is to model the relationship between the probability that a response Y (target) is equal to 1 given a set of features. So in our case we will apply logistic regression in the following ways.

Dataset A

As we saw earlier, Dataset A presents a feature called *score category* which consider only two values, 0 or 1. We will use this feature as our target variable of the Logistic Regression model. Now, we said before that this dataset was composed of more than 60 features, we expect that our model can identify which of them have a more important role when we classify a professor as alpha. On the other hand, when a dataset presents several variables, there is a specific method which helps on deciding which features should we take into consideration in our model. This method is called Recursive Feature Elimination. We will discuss about it later in this Modelling Section.

Dataset B

On the other hand, we want to use Dataset B with the goal of sustaining Dataset A's results. We know that in this case we do not have many features but still our objective is the same, have a way in which we can predict if a certain professor with specific characteristics would be classified as an alpha professor by taking into account a more complete dataset and starting by identifying if that professor is either a teacher-only or a teaching-and-research professor. In this case, since this dataset does not present too many features, we are not applying the model discussed before.

It is worth mentioning that, when it comes to Dataset B, several Logistic Regressions are going to be performed. First, we will apply several logistic regressions when our dataset is

either balanced or imbalanced. In past sections we mentioned that Tecnologico de Monterrey have way a lot more teaching-only professors than teaching-and-research ones, so when we talk about balancing our dataset, we mean increasing our dataset by adding records which only correspond to teaching-and-research ones. The next variation we will see in our Logistic Regressions is in whether how do we determine if a professor is alpha or beta. Some model are going to be applied when we consider as alpha to professors whose scores correspond to the last quartile of all and we will present other logistic regression models when we consider alpha professors to those who got average scores belonging to the last decile. After applying those regressions in both balanced and imbalanced data sets, we will apply this same model classifying alpha professors when having the dataset balanced, with the last decile level but for each question separately, not the average of the three questions as we saw before in Dataset A.

The logistic regression equation for Dataset B will be determined by:

$$Profab = \beta_0 + \beta_1 ALvl + \beta_2 G + \beta_3 RAge + \beta_4 SNI + \beta_5 R + e \tag{3.1}$$

In formula 3.1, *Profab* is our target variable and it refers to Professor alpha/beta, which can be 0 or 1. The beta are the coefficients of each of the features we are taking into account. *ALvl* corresponds to the Academic Level in which the course was being offered, 0 denotes Undergraduate Level and 1 indicates Graduate Level. The next feature *G* refers to Gender, and we have chosen a value of 0 to account male professors and 1 for female professors. Additionally, *RAge* corresponds to the group of ages to which the professor corresponded in a specific academic period. This feature can present values from 2 to 8, meaning that if a professor is 33 years old he or she would have a value of 3 at this feature. Or if his or her age is of 57, the professor would present a 5 in this category. Moreover, the *SNI* variable denotes the research proficiency level of the professor. And our 5th feature corresponds to a binary variable which accounts if the professor is either a teaching-only (0) or a teaching-and-research professor (1).

3.4.3 Recursive Feature Elimination

Since Dataset A presents more than 60 features, we have decided to apply the Recursive Feature Elimination Algorithm (RFE), which we explained before in the Background Chapter. This means that we want to rank the variables which are susceptible for providing better results to our desirable model. RFE will export the list of the variables with their respective rank to denote which ones seem to have a heavier weight when we consider the score variable as our target. Once we have identified those variables we can apply several Logistic Regressions until we get a desirable result which can contribute directly to this discussion. These models will also determine us which of those few variables have a heavier presence in the analysis, so we will be able to compare it against the results of Dataset B. This method will be the one to who define the composition of the Logistic Regression equation applied for Dataset A.

3.4.4 Panel Data Modelling

We have seen before that Dataset D involves measurements over time for specific entities, in this case, professors. Now that we have our database ready, our main goal is to perform a

multiple regression on our Panel Data. In order to achieve this, the famous statistical computer program, RStudio, will be used alongside with its plm library. Panel Data Regression can be done by applying several methods, such as Random Effects, Fixed Effects, Pooled Regression by OLS, First regression and the Between Algorithm. This stage will comprise all these practices in order to perform a broad comparison and identify which is or best model and based on what results. Each of these methods share equation 3.2 where the intercept of the regression is being represented as β_0 and the regression beta coefficients of our independent variables are represented as $\beta_1, \beta_2, \beta_3$. These variables are presented as G for Gender, SNI for declaring if the professor is also a researcher or not and Age to denote what age range does the teacher belongs to. The last term e corresponds to the error term.

$$Score = \beta_0 + \beta_1 G + \beta_2 SNI + \beta_3 Age + e \tag{3.2}$$

As seen in our Background Chapter, there are several ways to evaluate panel data results. For our case, we will be using the Lagrange Multiplier Test, the pFtest and the pHtest in order to determine which of our panel data models best fits our data.

3.5 Summary

The third chapter of this Master's thesis consisted of making a profound use of the CRISP-DM Methodology. This chapter covered the first four steps which were identified as Business Understanding, Data Understanding, Data Preparation and Modelling. We discussed that Tecnologico de Monterrey wants to compare the performance of teaching-only and teaching-and-research professors and identify the specific characteristics that their professors need to have in order to satisfy university's and students' quality standards. In order to know who performs better, either researcher professors or non researcher professors, we needed to make a comparative analysis of the academic performance. In the Data Understanding subsection, we discussed the several data sets that were going to be utilized to perform our experiments and their specific characteristics and performed their Exploratory Data Analysis. These data sets are based on Tecnologico students' survey ECOA ("Encuestas de Evaluacion de Profesores") which is answered at the end of each semester by students from all the campuses. Then, we presented the third section of the data mining methodology, in it we mentioned how the data sets were manipulated in our convenience in order for them to be prepared for their respective experiments. Finally, we saw the different models that are going to be applied with the intention of answering all of our research questions. It is important to mention that the last two sections of the CRISP-DM will be covered in Chapter 4 and Chapter 5, the first one will present the Evaluation part of the results of our experiments and the last part will discuss the deployment.

Chapter 4

Results

As we mentioned before, the fifth stage of the CRISP-DM Methodology corresponds to Evaluation. When the method accounts this step, it refers to the evaluation of the results obtained by the application of statistical and machine learning models. This chapter will be presenting the results obtained in each of our data sets when a certain technique was applied. Then the evaluation part of the CRISP Methodology will be presented followed by each result. A common approach employed in data mining studies is the splitting of the dataset into a training dataset and a testing dataset. The model is trained using the training set and one of various modelling techniques. Once the model has been trained, we test the accuracy of the resulting model making predictions over the testing dataset, and since we know the real answers, we are able to compare model predictions with the real answers in order to evaluate the accuracy of the trained model.

We followed this approach to evaluate the models obtained by applying logistic regression, analysis of variance to validate the statistical significance in comparing various attributes of teaching-only versus teaching and research professors, Recursive Feature Elimination and Panel Data. In order to avoid biases and making fair comparisons, we deleted the theses courses from the dataset. These courses are taught by teaching and research professors and typically, they receive high scores. We did the same with graduate courses, so that the comparison is made using undergraduate courses taught by teaching-only and teaching and research professors. The results obtained are further explained.

4.1 Analysis of Variance ANOVA

4.1.1 ANOVA Dataset A

We applied analysis of variance ANOVA in order to determine if our results were forceful, i.e. statistically significant. The mean scores shown in the distributions of, teacher-only professors and teaching-and-research professors (Figure 3.2), professors at the three academic levels (High school, Under-graduate and Graduate), and overall (Total) (Figure 3.3), and Teacher-only Professors and Teaching-and-Research Professors at graduate and undergraduate groups (Figure 3.4), are statistically different between each other. We present these results in Table 4.1, which illustrate that the level of significance of these tests show high values.

Parameter	SNI - NO SNI	Academic Level	SNI/ NO SNI Academic Levels
F-statistic	6.678	20.233	8.498
P-value	0.0097676	0.0000069	0.0000123
R-squared	0.0004397	0.0013470	0.0016968

Table 4.1: ANOVA results for the comparison of average scores of Teaching-and-Research professors (SNI) versus Teacher-only professors (NO SNI), in groups of different Academic Levels, and comparing both dimensions.

Evaluation Table 4.1: When we look closer to the table we identify that our p-values are all lower than 0.05. We have mentioned before that our experiments were going to use a confidence level of 95%, and since our p-values are lower than our alpha, our results are totally justifiable for the three cases, SNI - NO SNI, Academic Level, and SNI/NO SNI when Academic Levels are taken into account.

4.1.2 ANOVA Dataset B

When it comes to this second dataset, we performed several analysis of variance after applying our exploratory data analysis. We applied ANOVA to the results shown in Figure 3.5 in order to discover if they were statistically different between each other. According to a 1-way ANOVA test, the differences are significant. Table 4.2 illustrates the corresponding results.

ECOA Question	F-statistic	P-value	F-crit
05. Intellectual Challenge	403.54	1.59E-89	3.841
06. Learning Guide	136.36	1.76E-31	3.841
08. Recommendation	135	3.50E-31	3.841

Table 4.2: ANOVA: Teaching-Only vs Teaching-and-Research Professors at the three dimensions: Learning Guide (06. APR), Intellectual Challenge (05. RET) and Professor Recommendation (08. REC)

Evaluation Table 4.2: As we can see in Table 4.2, the difference between the means of teaching-only and teaching-and-research professors for the three different dimensions are statistically significant. The three cases show a F-statistic higher than F-crit and a p-value lower than alpha.

For Figure 3.6, the 1-way ANOVA test was applied in several ways. First, we wanted to know if they scores registered in each period were different between each other in the three dimensions. Table 4.3 illustrate the results of this experiment, we applied ANOVA in each dimension.

ECOA Question	F-statistic	P-value	F-crit
05. Intellectual Challenge	45.71	2.03E-38	2.37
06. Learning Guide	29.46	1.60E-24	2.37
08. Recommendation	23.36	2.48E-19	2.37

Table 4.3: ANOVA: Temporal Evolution of the three dimensions.

Evaluation Table 4.3: According to a 1-way ANOVA test, these differences are significant. The means of the five semesters are significantly different between each other. Intellectual Challenge, Learning Guide and Recommendation factors present an F-statistic higher than F-crit, and p-values lower than 0.05.

If we take a closer look to 3.6 we can see that the scores for the first and last two periods are very similar. That is why we decided to apply ANOVA test in the three dimensions for those two cases.

ECOA Question	F-statistic	P-value	F-crit
05. Intellectual Challenge	0.70	0.4	3.84
06. Learning Guide	2.43	0.11	3.84
08. Recommendation	3.11	0.077	3.84

Table 4.4: ANOVA: Temporal Evolution of the three factors for the first two academic periods.

Evaluation Table 4.4: According to a 1-way ANOVA test, the differences between the scores for the first to periods are not statistically significant. In each factor we can see that F-crit is higher than the F-statistic and their p-values are higher than 0.05.

Now, lets take a look at the ANOVA results for each dimension at their last two academic periods.

ECOA Question	F-statistic	P-value	F-crit
05. Intellectual Challenge	5.44	0.019	3.84
06. Learning Guide	3.51	0.06	3.84
08. Recommendation	15.4	8.7E-05	3.84

Table 4.5: ANOVA: Temporal Evolution of the three factors for the last two academic periods.

Evaluation Table 4.5: In this case, p-values are not too far away from alpha. The results of the Intellectual Challenge factor tell us there is a statistical difference between the means of the last two academic periods (9.10 and 9.12), since p-value is lower than alpha and F-statistic is a little bit higher than F-crit. On the other hand, the results for the Learning Guide factor were different. By using an alpha of 0.05 as reference we can not conclude that the means of their last two academic periods are different (9.04 and 9.06). At the same time, F-statistic is not higher than F-crit. Additionally, the Recommendation factor did presented a stronger statistical difference by registering high values in F-statistic and a result very close to 0 for the p-value.

On the other hand, in our Exploratory Data Analysis, Figure 3.7 presented some close average scores between teaching-only professors and teaching-and-research professors in undergraduate and graduate academic levels, we performed analysis of variance in order to verify their statistical differences. Table 4.6 and Table 4.7 illustrate the ANOVAs' results.

ECOA Question	F-statistic	P-value	F-crit
05. Intellectual Challenge	81.499	1.79E-19	3.841
06. Learning Guide	4.241	0.039	3.841
08. Recommendation	0.346	0.556	3.842

Table 4.6: ANOVA: Undergraduate Level Teaching-only vs Teaching-and-Research Professors.

Evaluation Table 4.6: For the undergraduate level, P-values are indeed lower than our alpha (0.05) and the F-statistics are higher than F-crit, only in the first two dimensions, Intellectual Challenge and Learning Guide. This means that the differences in means are statistically different. On the other hand, our last dimension Recommendation presents a very high p-value (0.556) and the F-crit is higher than the F-statistic, so there is no statistical difference in that factor.

ECOA Question	F-statistic	P-value	F-crit
05. Intellectual Challenge	96.304	2.475E-22	3.841
06. Learning Guide	63.448	2.471E-15	3.845
08. Recommendation	97.801	1.196E-22	3.846

Table 4.7: ANOVA: Graduate Level Teaching-only vs Teaching-and-Research Professors.

Evaluation Table 4.7: P-value is so close to 0 in the three dimensions. Additionally, the F-statistics are way higher than the F-crit. That is why we say that the means of the figure of Teaching-only and Teaching-and-Research are statistically different.

Furthermore, Figure 3.8 illustrated us the average score differences on gender at the three dimensions, Intellectual Challenge, Learning Guide and Professor Recommendation. Table 4.8 helps us understand in a statistical way who performs better as a learning guide, either males or females. It also shows the ANOVA results to know if there is a significant difference at all between female or male as intellectual challenge and to conclude who do actually get a higher score as a recommended professor.

ECOA Question	F-statistic	P-value	F-crit
05. Intellectual Challenge	0.0403	0.8407	3.841
06. Learning Guide	92.782	6.011E-22	3.841
08. Recommendation	20.3817	6.352E-06	3.841

Table 4.8: ANOVA: Intellectual Challenge, Learning Guide and Professor Recommendation

Evaluation Table 4.8: Different results are being illustrated in the table. When we compare male professors with female professors we get significant differences in two of our three dimensions, Learning Guide and Recommendation. Both of their p-values are very close to zero and their F-statistics present higher values than the F-crit. The ANOVA results for the Intellectual Challenge could have been somehow been deducted since Figure 3.8 illustrated the same mean for male and female professor. Table 4.8 confirms that there is indeed no statistical significance between their recorded means. Its F-statistic is very low (close to zero) and its p-value somehow close to 1, which is the least thing we need to prove a statistical difference.

Additionally, in Table 4.9 and Table 4.10 we show the ANOVA results related to figure 3.9. This helps us knowing who gets better evaluated at the three dimensions either male or female when they are teaching-only or teaching-and-research professor.

ECOA Question	F-statistic	P-value	F-crit
05. Intellectual Challenge	7.868	0.005	3.841
06. Learning Guide	130.664	3.124E-30	3.842
08. Recommendation	39.329	3.6E-10	3.841

Table 4.9: ANOVA: Teaching-only Male vs Female Professors.

Evaluation Table 4.9: The three factors considered in this study present a p-value lower than alpha (0.05) and F-statistics higher than the F-crit. These parameters help us evaluate the results illustrated in figure 3.9, there is statistical difference between the registered means of the ECOA results between Male and Female Teaching-Only Professors.

ECOA Question	F-statistic	P-value	F-crit
05. Intellectual Challenge	20.946	4.8E-06	3.842
06. Learning Guide	2.2929	0.13	3.842
08. Recommendation	6.724	0.009	3.842

Table 4.10: ANOVA: Researchers Male vs Female.

Evaluation Table 4.10: When it comes to Research-and-teaching Professors, we saw a slightly difference in their results. The Learning Guide factor did not presented a statistical difference, we got an F-statistic lower than F-crit and a p-value higher than alpha. The means for the other two factors are statistically different according to their F-statistics and p-values.

In the Exploratory Data Analysis performed to Dataset B, we presented the ECOA means for each of the dimensions when taking into consideration the age range of the professor. Figure 3.11 illustrated the case of Teaching-only professors and Figure 3.12 corresponded to case of Teaching-and-Research professors. The following tables show the ANOVA results of these two cases.

ECOA Question	F-statistic	P-value	F-crit
05. Intellectual Challenge	30.72	1.349E-25	2.37
06. Learning Guide	35.57	9.83E-30	2.37
08. Recommendation	91.16	1.96E-77	2.37

Table 4.11: ANOVA: Aging Teaching-Only Professors.

Evaluation Table 4.11: The ANOVA results corresponding to the Teaching-only aging case, helps us conclude that there is indeed statistical difference between the ECOA means for each of the category ages of this type of professors. All of the p-values are very close to zero and F-statistics higher than F-crit.

Now that we have presented the ANOVA results for our first type of professors, we will see if there is any difference between the results of the teachers who also perform research at Tecnologico de Monterrey. Figure 3.12 presented less variance in the means between each of the categories of Teaching-and-Research Professors. Table 4.12 will let us know if their differences are statistically significant.

ECOA Question	F-statistic	P-value	F-crit
05. Intellectual Challenge	2.78	0.04	2.37
06. Learning Guide	2.28	0.07	2.37
08. Recommendation	1.624	0.18	2.37

Table 4.12: ANOVA: Aging Teaching-and-Research Professors

Evaluation Table 4.12: The ANOVA results for this case only identified one factor which presented statistical difference in their means, Intellectual Challenge. F-crit is 0.41 lower than the F-statistic and the p-value is 0.01 lower than our alpha. It is worth mentioning that these tests are not considering the first age range (2) since its population is very small compared to the other ranges. For the case of the Learning Guide factor, its results are very debatable since p-value is very close to alpha and F-statistic to F-crit. The Recommendation factor did not presented any statistical difference in its means since its p-value is higher than 0.05 and F-statistic lower than F-crit.

Finally, in the past section we saw the average scores in the three dimensions for the different proficiency levels of teaching-and-research professors. Table 4.13 contain the ANOVAs' results of the comparison between all of the levels at the three questions which correspond to the results illustrated in Figure 3.13.

ECOA Question	F-statistic	P-value	F-crit
05. Intellectual Challenge	56.822	2.16E-36	2.605
06. Learning Guide	35.001	1.741E-22	2.605
08. Recommendation	49.287	1.356E-31	2.605

Table 4.13: ANOVA:Average satisfaction scores of teaching-and-research professors by all proficiency levels in the three dimensions: Learning Guide (06. APR), Intellectual Challenge (05. RET) and Professor Recommendation (08. REC).

Evaluation Table 4.13: According to this table, there is significant difference between the ECOA's means recorded for each proficiency level in which researchers can be identified. P-values are lower than 0.05, in fact, they are almost zero and F-statistics are way higher than F-crits.

However, since the scores of the last two proficiency levels among the three questions are very similar (Figure3.13) , we decided to apply an additional analysis of variance in order to identify whether if they where statistical different or not from each other. Table 4.14 represents the results of the ANOVA.

ECOA Question	F-statistic	P-value	F-crit
05. Intellectual Challenge	0.002	0.960	3.848
06. Learning Guide	0.081	0.774	3.848
08. Recommendation	0.077	0.780	3.848

Table 4.14: ANOVA: Average satisfaction scores of teaching-and-research professors in proficiency levels 3 and 4.

Evaluation Table 4.14: It is very clear that the ECOA means of these proficiency research levels are not statistically different between each other in any of the three factors, Intellectual Challenge, Learning Guide and Recommendation. F-statistics are very low, small compared to F-crits and P-values are very close which is the opposite of what we wanted to get.

4.1.3 ANOVA Dataset C

As we mentioned in previous sections, Dataset C additionally contains the ECOA's records for the 2016 Academic year. Figure 3.14 presented the average scores between Non Research Professors and Research Professors across the years. It takes into account only the common courses given by these professors and it is prepared to make an emphasis on only two factors, Professor Recommendation and Intellectual Challenge. Now, we would like to see whether if the means of these two groups in each year are statistically different between each other or not. Table 4.15 and Table 4.16 will allow us to end with this doubt. The tables presented in this subsection indicate the size of the populations (Teaching-only and Teaching-and-Research Professors), the result of the mean the value of the t-statistic and the P-value. First, we will start with the Recommendation factor results.

ECOAs 8. REC	Research/ Non Research	N	Mean	t	P-value
2016	No	8384	8.69	4.028	0.028
	Yes	2139	8.78		
2017	No	1333	8.66	3.057	0.001
	Yes	3171	8.75		
2018	No	7177	8.69	11.15	0.000
	Yes	1327	9.12		

Table 4.15: Recommendation Factor scores between Research and Non-research professors.

Evaluation Table 4.15: This table illustrates difference in means for research and non research professors in each of the years. In order to evaluate these results, we need to take a closer look to the p-values. In order to identify a statistical difference in means, we need a p-value lower than 0.05. Each of the three values shown here are below that quantity, hence, there is significant difference between the ECOA results of Researcher and Non Researcher professors.

ECOAS Intellectual Challenge	Research/ Non Research	N	Mean	t	P-value
2018	No	7177	8.69	11.15	0.000
	Yes	1327	9.12		
Jan - May 2019	No	1333	8.81	11.14	0.000
	Yes	3171	9.16		

Table 4.16: Intellectual Challenge Factor scores of Teaching-and-Research Professors and Teaching-only Professors.

Evaluation Table 4.16: This table helps us to denote that there is as well statistical different between the means of Researchers and Non Researcher Professors when the Intellectual Challenge factor is considered. Both of the p-values are close to zero, or lower than alpha. Records from previous years were not considered since we did not currently had the information of Teaching-and-Research professors corresponding to those academic periods.

On the other hand, we have presented before the same comparison, ECOA's results of Teaching-only and Teaching-and-Research Professors but filtered by Tecnologico de Monterrey's schools. That is why we are forced to present the following table.

School	ECOAs 5. RET	Research/ Non Research	N	Mean	t	P-value
SHE	2018	No	920	8.69	8.02	0.000
		Yes	156	9.12		
	Jan - May 2019	No	938	8.81	10.13	0.000
		Yes	186	9.16		
SES	2018	No	3270	9.09	8.38	0.000
		Yes	729	9.36		
	Jan - May 2019	No	2789	9.17	11.71	0.000
		Yes	981	9.47		
SMHS	2018	No	317	9.24	1.3	0.1
		Yes	110	9.37		
	Jan - May 2019	No	304	9.31	3.51	0.000
		Yes	118	9.56		
SB	2018	No	912	8.99	3.51	0.000
		Yes	88	9.32		
	Jan - May 2019	No	1171	8.98	4.28	0.000
		Yes	136	9.32		
SAAD	2018	No	134	8.95	-0.47	0.32
		Yes	4	8.7		
	Jan - May 2019	No	132	8.73	0.62	0.27
		Yes	4	9.15		
SSSG	2018	No	394	9.14	-0.39	0.35
		Yes	95	9.1		
	Jan - May 2019	No	516	9.2	0.31	0.38
		Yes	116	9.22		

Table 4.17: Intellectual Challenge Factor ANOVA between Teaching-and-Research Professors and Teaching-only Professors by School

Evaluation Table 4.17: We see significant difference in means in the first four schools described in this table, the School of Humanities and Education, Engineering and Sciences, and the School of Business. All of their p-values are close to 0. The School of Medicine and Health Sciences presents two different cases, one in which there is statistical difference (2018) and one where the difference of means is not significant (2019). However, the School of Arts, Architecture and Design and the School of Social Sciences and Government presented p-values higher than 0.05 and their t-statistics are very low, less than zero. This means that there is no statistical difference between the means of teaching-only and teaching-and-research professors at those academic periods, specifically in those schools.

Now, since we have seen the statistical results of the Intellectual Challenge factor, let us take a look to the comparison of means between Research and Non Research Professors when the Recommendation Factor is taken into account. Those results correspond to the ones presented in the Exploratory Data Analysis performed to Dataset C illustrated from Figure 3.16 to Figure 3.21.

School	ECOAS 8. REC	Research/ Non Research	N	Mean	t	P-value
SHE	2018	No	920	8.86	7.78	0.000
		Yes	155	9.6		
	Jan - May 2019	No	937	8.84	7.42	0.000
		Yes	191	9.46		
SES	2018	No	3270	8.61	7.44	0.000
		Yes	728	9.01		
	Jan - May 2019	No	2787	8.78	10.20	0.000
		Yes	975	9.2		
SMHS	2018	No	317	8.78	2.37	0.01
		Yes	106	9.16		
	Jan - May 2019	No	304	8.85	1.51	0.07
		Yes	118	9.05		
SB	2018	No	911	8.73	2.82	0.000
		Yes	89	9.09		
	Jan - May 2019	No	1166	8.74	2.17	0.01
		Yes	136	8.99		
SAAD	2018	No	134	8.59	-0.9	0.16
		Yes	4	7.76		
	Jan - May 2019	No	133	8.31	0.48	0.32
		Yes	5	8.73		
SSSG	2018	No	394	8.87	-1.0	0.15
		Yes	95	8.73		
	Jan - May 2019	No	516	8.89	-0.8	0.21
		Yes	116	8.79		

Table 4.18: Recommendation Factor ANOVA between Teaching-and-Research Professors and Teaching-only Professors by School

Evaluation Table 4.18: For the case of the Recommendation Factor, basically we got the same results. Both the School of Arts, Architecture and Design and the School of Social Sciences and Government presented p-values higher than our alpha, additionally their t-statistics are lower than zero. That is how we can say that, even though their registered means are different when comparing these two type of professors, in reality they are not, statistically for this academic periods we can not say that at these schools one type of professor is better than the other one. On the other hand, the rest of the schools, Humanities and Education, Engineering and Sciences, Medicine and Health Sciences and Business, presented a p-value lower than 0.05, stating that the means of these professors in each of those schools, are statistically different between each other.

4.1.4 ANOVA Dataset E

This subsection corresponds to the ANOVA results when comparing the ECOA means of professors with certain characteristics coming from the dataset which was pre-processed with the Coarsened Exact Matching Technique. Thus, we will see if these results are different from Dataset B, which was not preprocessed at the beginning.

ECOA Question	F-statistic	P-value	F-crit
05. Intellectual Challenge	1.2982	0.2545	3.8417
06. Learning Guide	12.5552	0.0004	3.8417
08. Recommendation	0.2867	0.5923	3.8417
Overall	0.6198	0.4310	3.8417

Table 4.19: ANOVA: Matched Male vs Female Professors

Evaluation Table 4.20: We saw the comparison of Male and Female Professors in Figure 3.24. Now we want to know if those results are statistically significant or not. By taking a look to this table we can see that there is only one statistical difference in our analysis, this corresponds to the Learning Guide Factor. Its p-value is lower than 0.05 and F-statistic is higher than F-crit.

ECOA Question	F-statistic	P-value	F-crit
05. Intellectual Challenge	0.729	0.393	3.841
06. Learning Guide	23.338	1.055E-06	3.841
08. Recommendation	1.117	0.290	3.841

Table 4.20: ANOVA: Matched Teaching-Only Professors Male vs Female

Evaluation Table 4.20: We obtained two different results when we took into consideration the comparison between Male Teaching-Only Professors and Female Teaching-only Professors. The only statistical difference we identified was in the Learning Guide Factor, since we got a p-value lower than alpha and an F-statistic higher than F-crit. The other two factors did not met the criteria to identify any significant difference in the means of these groups. This ANOVA corresponds to Figure 3.25.

Now let us see what kind of results we got for the professors who also were identified as researchers at these academic periods.

ECOA Question	F-statistic	P-value	F-crit
05. Intellectual Challenge	0.095	0.757	3.843
06. Learning Guide	2.562	0.109	3.843
08. Recommendation	0.2398	0.624	3.843

Table 4.21: ANOVA: Matched Teaching-and-Research Professors Male vs Female

Evaluation Table 4.21: In this case, we did not get any satisfactory result. Apparently there is no different in means in any of the three factors we are considering between Male and Female Teaching-and-Research Professors, when a matched dataset is being studied. P-values are too high and F-crit is higher than F-statistic in our three cases. This ANOVA corresponds to Figure 3.25.

We continue with our analysis, Table 4.22 will let us know if the ECOAs' means of researchers with different proficiency levels are different from one another.

ECOA Question	F-statistic	P-value	F-crit
05. Intellectual Challenge	15.455	5.3E-10	2.606
06. Learning Guide	7.540	4.967E-05	2.606
08. Recommendation	14.266	2.986E-09	2.606
Overall	13.169	1.48E-08	2.606

Table 4.22: ANOVA: Matched Researcher Proficiency Levels

Evaluation Table 4.21: As we can see, all of our factors present a significant difference in their means. Intellectual Challenge, Learning Guide and Recommendation factors meet the requirements, P-values are way lower than 0.05 and F-statistics are higher than F-crit. At the end of the table, we present the ANOVA results when the evaluations of the three questions are considered as one. There is statistical difference between the means of professors with different proficiency levels. P-value and F-statistic meet the requirements as well. This ANOVA corresponds to Figure 3.23.

Moreover, we present the results for the general comparison between Teaching-only and Teaching-and-Research Professors when we take into account the CEM dataset.

ECOA Question	F-statistic	P-value	F-crit
05. Intellectual Challenge	715.085	4.455E-156	3.841
06. Learning Guide	403.054	3.479E-89	3.841
08. Recommendation	482.824	2.407E-106	3.841
Overall	576.34	2.07E-126	3.841

Table 4.23: ANOVA: CEM Teaching-only Professors vs Teaching-and-Research Professors.

Evaluation Table 4.23: Our analysis of variance presented forceful results. We have identified that there is a notorious difference in ECOA means of professors who are either Teaching-only or Teaching-and-Research. When we take each of the factors individually, we found p-values way lower than 0.05. Additionally, this table contain the lowest F-statistics registered of all of the analysis that we have mentioned before. The difference is still significant when we consider each factor as one, it meets the requirements since we can say that its p-value is zero and the F-statistic is superior than F-crit. The ANOVA corresponds to the Figure 3.22.

4.2 Logistic Regression

In the following subsections, we will cover how we applied Logistic Regression to some of the different data sets we are studying, with the respective evaluations of their results.

4.2.1 LR Dataset A

In order to apply data mining techniques to find patterns on the datasets, we determined if some specific characteristics prevail when professors are evaluated as good or bad (above/below the mean) by student opinion. To accomplish this, we used Logistic Regression, also called logit regression, a technique used to find the probability that a binary goal variable takes the value yes or no, based on the attributes that define the goal variable. It uses a logistic function whose coefficients are calculated from the data to best fit the classification of goal variable as a yes or no. Next, since this dataset contains more than 60 features, we ranked the variables that provide better accuracy individually running a Recursive Feature Elimination algorithm, that recursively removes features using the remaining attributes to work out the combination of attributes that contributes to the prediction of the goal variable. Making combinations of the attribute variables and using 80% of the data for training and 20% for testing, we found multiple models with accuracy ranging between 50% and 80%. Later on, we will see how this algorithm helped us to get the best model.

Recursive Feature Elimination

First, we will see how does the features behave when we separate courses given to graduate students only. Let us remember that among the more than 60 features presented in this dataset, some of them are related to the professor and some others to the group, the following tables make this indication as well. In first instance, we have found that the ranking of variables on groups of graduate students show that these students score better to professors who: 1) have a higher responsibility percentage in the group, and 2) are researchers. Also, professors receive a higher grade when the course is being transmitted to multiple campuses of Tecnologico de Monterrey and when there are more senior students in that certain course. We illustrate the first 15 ranked features in Table 4.24.

On the other hand, at groups with undergraduate students, we have found that the average score of the ECOA for the three factors combined is positively correlated for professors that: (1) also teach at high school, (2) have a foreign nationality, (3) teach more groups, (4) have a greater percentage of responsibility of the group, and (5) is a teaching-only professor. The full list of ranked professor and group features are shown in Table 4.25.

Table 4.26 shows the ranking for professor features only at undergraduate courses. In both cases, at graduate courses and at undergraduate courses when we focus only on the professor features, the percentage of senior students in the group or number of graduate students attended, is ranked among the top 6 features, which indicate that the maturity of students contributes most to scoring professors higher. Additionally, observing that percentage of responsibility is in the top 5 features in both populations indicates us that team teaching seems to affect the professors' evaluations in a negative way.

Rank	Feature	Professor / Group
1	Percentage of responsibility of the group	Professor
2	Is a teaching-and-research professor	Professor
3	Class transmitted to multiple campus	Group
4	Number of senior students	Group
5	Number of hours at classroom	Professor
6	Number of credits	Group
7	Foreign nationality	Professor
8	Number of undergraduate students attended	Professor
9	Number of scientific publications	Professor
10	Percentage of participation on the survey	Group
11	Number of graduate students attended	Professor
12	Number of laboratory hours	Group
13	Main professor	Group
14	Total number of students attended	Professor
15	Number of teaching hours	Group

Table 4.24: Professor and group features ranked by the Recursive Feature Elimination algorithm on groups of graduate students.

Rank	Feature	Professor / Group
1	Percentage of participation on the survey	Group
2	Number of high school students attended	Professor
3	Foreign nationality	Professor
4	Number of teaching hours	Professor
5	Percentage of responsibility of the group	Professor
6	Main professor	Professor
7	Number of credits	Group
8	Is a terminal group	Group
9	Is a teaching-only professor	Professor
10	Total number of students attended	Professor
11	Number of undergraduate students attended	Professor
12	Certified in the teaching abilities program	Professor
13	Number of hours at classroom	Professor
14	Number of senior students	Group
15	Class transmitted to multiple campus	Group
16	Number of graduate students attended	Professor
17	Number of laboratory hours	Group
18	Number of scientific publications	Professor

Table 4.25: Professor and group features ranked by the Recursive Feature Elimination algorithm on groups of undergraduate students.

Rank	Feature
1	Number of high school students attended
2	Has a PhD
3	Certified in the teaching abilities program
4	Number of undergraduate students attended
5	Number of graduate students attended
6	Total number of students attended
7	Has a masters
8	Number of scientific publications
9	Is a teaching-only professor
10	Foreign nationality

Table 4.26: Professor features ranked by the Recursive Feature Elimination algorithm on groups of undergraduate students.

Logistic Regression

Next, we tried to identify if some specific characteristic prevail when professors are evaluated as alpha or beta (above/below the mean). In order to do this, we used Logistic Regression. The results of the Recursive Feature Elimination where taken into account several times. We performed several logit experiments until, we tried a considerable amount of combinations to our model until we could get an acceptable result. Among all the features and combinations we only found one model which gave us a good result. We achieved this by applying this method to sample of graduate students only, this model presented the highest accuracy, 80%, and it only had a single variable; namely, the percentage of senior students. With this single variable, it was possible to forecast if a professor would be qualified above or below the average in 80% of the cases. In all of our other cases, we got accuracies hovering around 40 and 50 percent.

The confusion matrix of this model is illustrated in Figure 4.2, where it can be observed that it is more likely to predict bad professors as good ones (38) than to predict good professors as bad ones (15). In other words, the total sample population consisted of 267 professors. Out of it 145 are considered as alpha, this model could identify 130 of them leaving 15 incorrect predictions. On the other hand, this sample population consisted of 122 beta professors, this model could identify 84 of them correctly, leaving 38 incorrect predictions. Figure 4.2 shows the ROC curve (Receiver Operating Characteristic) of this model, it covers 80% of the area and shows that the model built is a reliable approximation of the unknown true model.

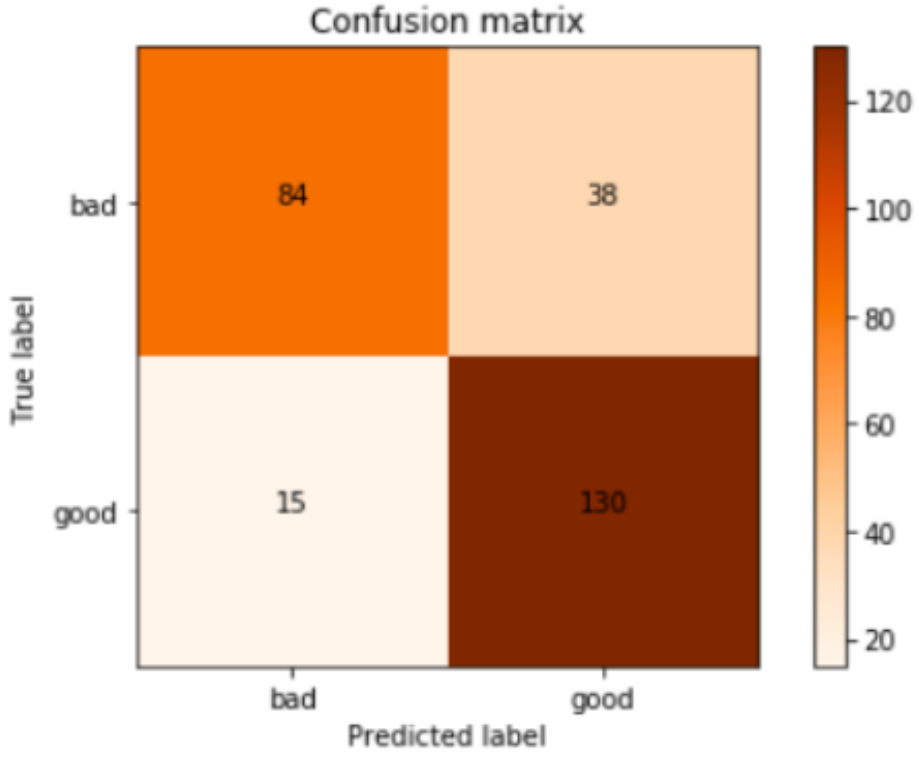

Figure 4.1: Confusion matrix for the Logistic Regression model based on the number of senior students in the graduate group.

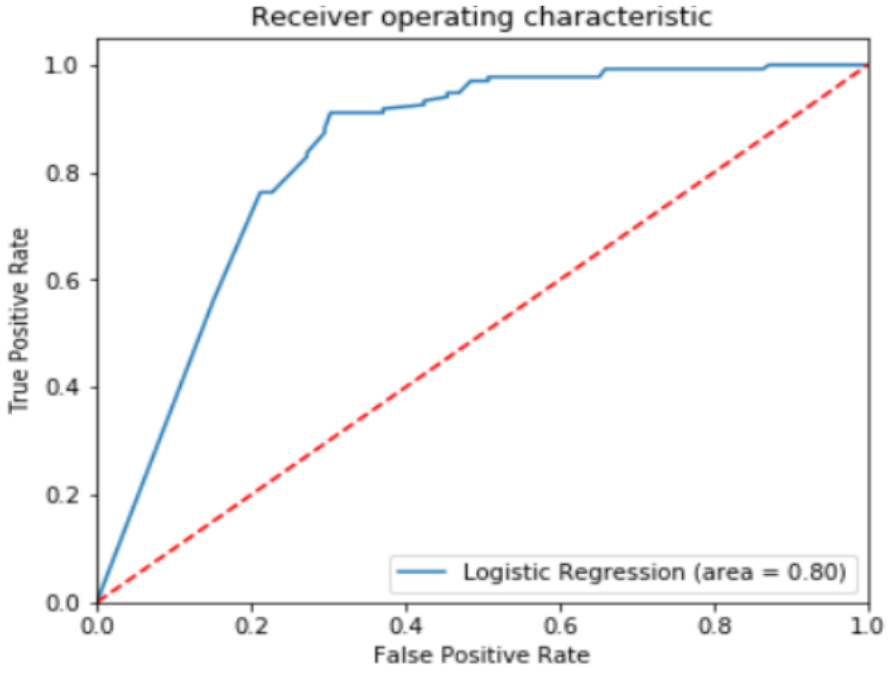

Figure 4.2: Receiver Operating Characteristic (ROC) Curve for the Logistic Regression models built using Feature Elimination on the variables of the first dataset

4.2.2 LR Dataset B

Now that we have seen the results of the Logistic Regression and the Recursive Feature Selection, we will see how much do they differ from the ones applied to Dataset B. Since Dataset B does not present much features it was unnecessary to apply RFE. In order to perform logit regression to this data, we decided to create models when we manipulated our dataset in several ways. First, since our dataset presents a big difference in records, meaning that more than 68,000 records correspond to Teacher-only Professors and only more than 9,000 corresponds to Teaching-and-Research Professor, we wanted to see which model fits the data in a better way, when we use it just the way it is or when we balance it by applying an algorithm in python. Then, a second variation we will try is changing the criteria to determine if a professor is be classified as alpha or beta when taking into account their score in the ECOA Survey. In the following experiments we will consider two criteria. The first one is when we identify alpha professors to those whose score is in the top 10% of all, this means that we split the data in deciles, and beta to those whose score falls in the rest of the percentage. The second case is when we identify alpha professors to those whose score fall in the top 25%, meaning that we now use quartiles, and beta to those whose score fall in the other 75%. Finally, the last variation we will be applying in this dataset is the question or questions of the student survey that will be analyzed. In the last subsection, we saw that Logistic Regression was applied when the three factors were considered as one, meaning that we only deal with the average of the three questions. However, in this dataset, we will be doing the same analysis, the average of the three questions, and a logistic regression for each of the questions separately. Let us continue with the results of each case. The results of the most important metrics for each case will we presented at the end using a comparative table.

A. Logistic Regression: Decile / Balanced dataset

When we finish balancing the dataset, we got the following. The number of Teacher-Only Professors stood in 68,203 and the number of Research-and-Teaching Professors went up approximately 7.5 times higher until equalize the other. Hence, we got 23,569 alpha professors and 112,837 professors considered as beta. Figure 4.5 illustrates the ROC Curve with a result of 0.75 and Figure 4.4 corresponds to the confusion matrix.

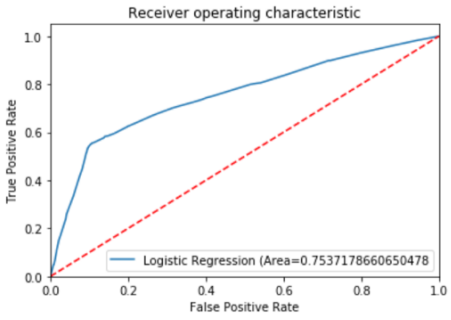

Figure 4.3: Receiver Operating Characteristic (ROC) Logistic Regression: Decile / Balanced dataset

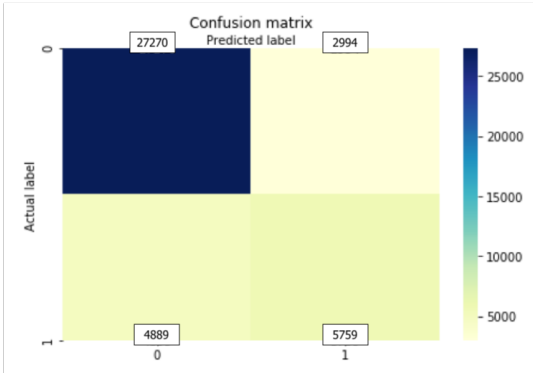

Figure 4.4: Confusion Matrix Logistic Regression: Decile / Balanced dataset

B. Logistic Regression: Decile / Imbalanced dataset

In the case of the imbalanced dataset, the number of teacher-only professors is 68,203 (same quantity as the balanced one) but, 9,327 is the amount of teaching-and-research professors. In addition, once we classify our alpha and beta professors we got 7,651 and 70,659 respectively. Figure 4.5 corresponds to its ROC Curve with a result of 0.70, and the Confusion Matrix is illustrated in Figure 4.6.

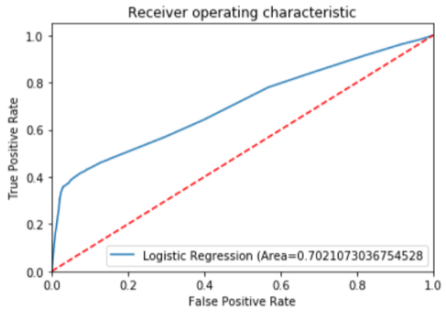

Figure 4.5: Reciever Operating Characteristic (ROC) Logistic Regression: Decile / Imbalanced Dataset

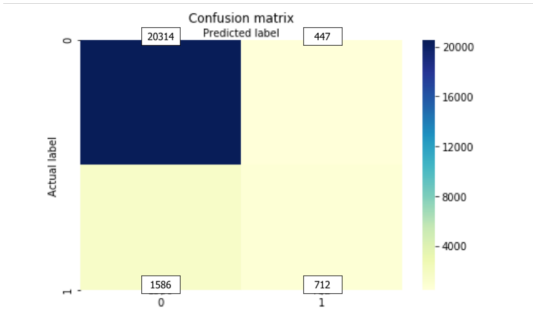

Figure 4.6: Confusion Matrix Logistic Regression: Decile / Imbalanced Dataset

C. Logistic Regression: Quartile / Balanced Dataset

When we are considering quartiles instead of percentiles, the number of alpha professors increase and the amount of beta ones decrease. In this case, for a balanced dataset, we got a total of 35,120 alpha professors and 101,286 beta professors. Later on, we will see what was resulted from this experiment in the summary table.

D. Logistic Regression: Intellectual Challenge Decile / Balanced Dataset

When the only factor we are studying is the Intellectual Challenge in a balanced dataset, we get 28,020 alpha professors and 108,836 beta ones. Let us take a look to Figures 4.7 and 4.8 to analyze its ROC Curve and Confusion Matrix.

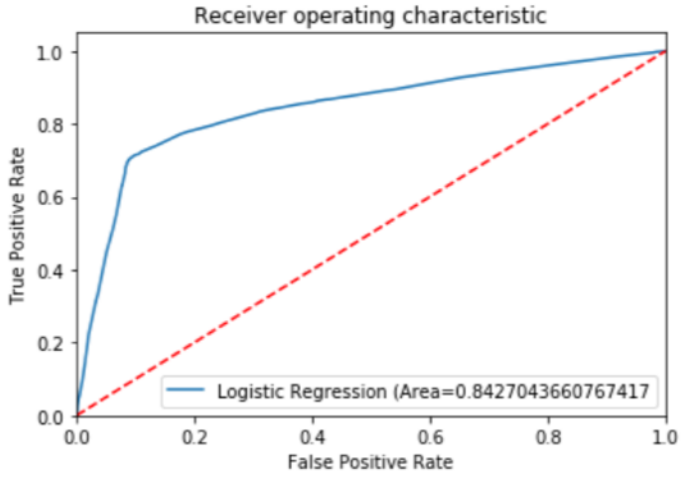

Figure 4.7: Receiver Operating Characteristic (ROC), Intellectual Challenge: Decile / Balanced Dataset

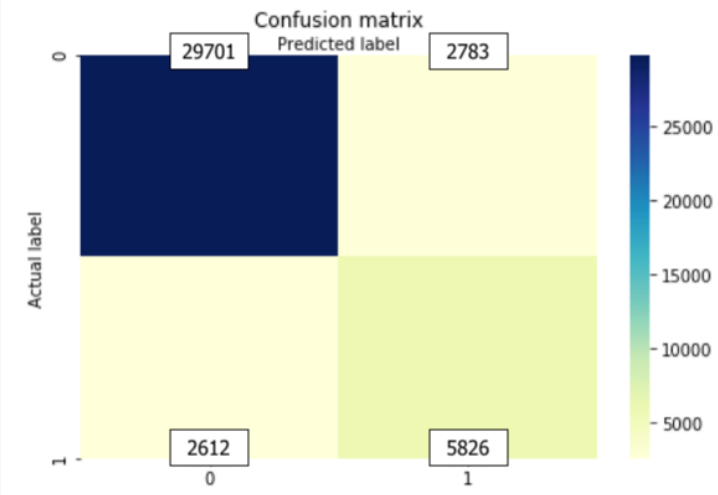

Figure 4.8: Confusion Matrix, Intellectual Challenge: Decile / Balanced Dataset

E. Logistic Regression: Learning Guide Decile / Balanced Dataset

The next factor to analyze is the Learning Guide when having our dataset balanced. This factor registered 28,068 alpha professors and 108,339 beta professors. Its ROC Curve and Confusion Matrix are illustrated in Figure 4.9 and Figure 4.10. The first one covered 83% of the data.

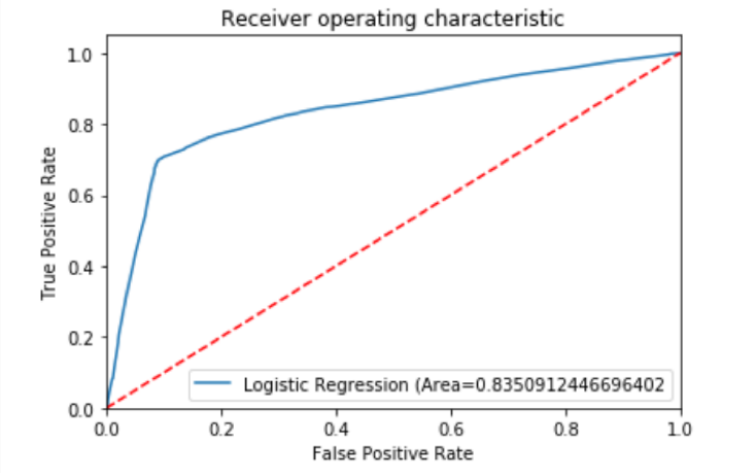

Figure 4.9: Receiver Operating Characteristic (ROC) Learning Guide: Decile / Balanced dataset

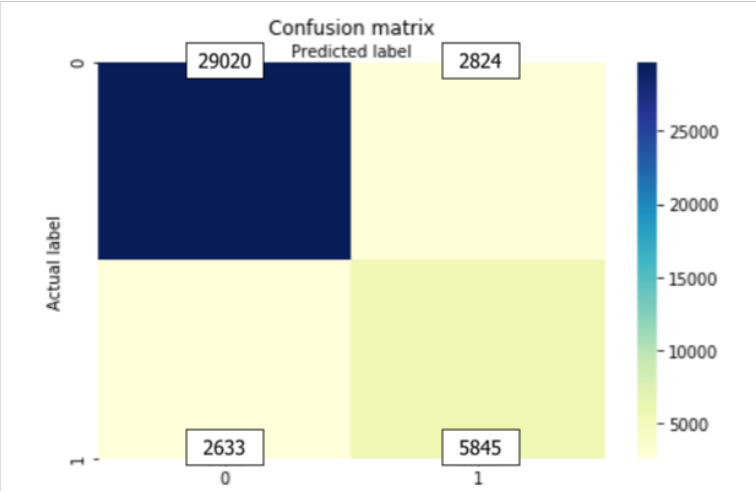

Figure 4.10: Confusion Matrix Learning Guide: Decile / Balanced dataset

F. Logistic Regression: Recommended Professor Factor Decile / Balanced Dataset

Finally, we applied Logistic Regression to the Recommendation Factor while making use of a Balanced dataset. Here, more than 106,000 professors fall in the beta category and almost 30,000 in the alpha one. Figure 4.11 shows how in this case we obtained 0.82 in the ROC Curve. Besides this, we present its Confusion Matrix in Figure 4.12, let us remember that 0 stands for beta professor and 1 stands for alpha.

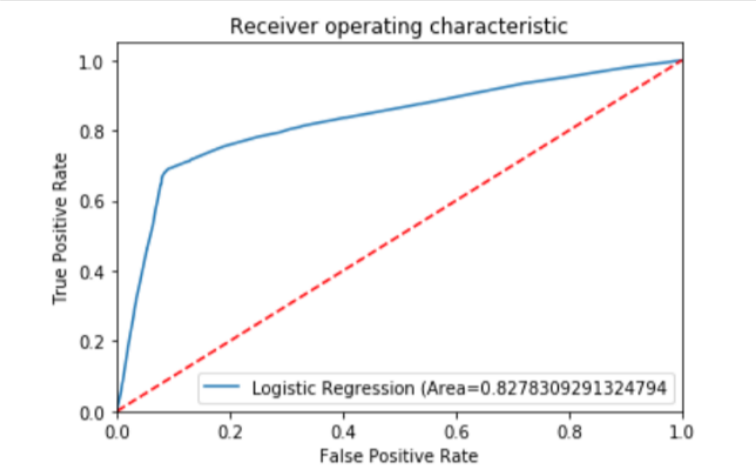

Figure 4.11: Receiver Operating Characteristic (ROC) Question 8: Decile / Balanced Dataset

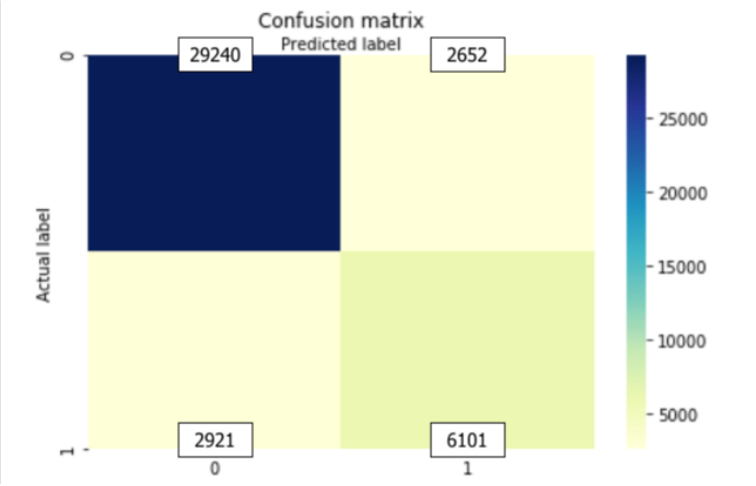

Figure 4.12: Confusion Matrix Question 8: Decile / Balanced dataset

Now that we have made all our possible models we present in the next table the results of the most important metrics we follow in order to identify which model fitted better our data and what type of contributions does it make. Table presents the six cases we discussed before. Let us first remember that accuracy is given by the number of correctly classified examples divided by the total number of classified examples [15]. In terms of the confusion matrix, precision is known as the ratio of true positives to the sum of true and false positives, in other words, this determines the percentage of instances that were correctly classified as positive, when true positives and false positives are taken into account. That is why, while we get a higher value, we say that our model fits data in a better way. On the other hand, recall helps us knowing the percentage of professors correctly classified as alpha when all the true positive and false negative instances are taken into account. In our next Chapter, Discussion, we will broadly discuss what can we conclude from Table 4.27.

Case	Accuracy	Precision	Recall
A	81.71%	65.79%	54%
B	91.25%	61.43%	30.94%
C	80.71%	65.79%	54.03%
D	86.81%	67.67%	69.04%
E	86.66%	67.42%	68.94%
F	86.38%	69.70%	67.62%

Table 4.27: Logistic Regression Results Dataset B

4.3 Panel Data Modelling

4.3.1 Dataset D

In this section, the outcomes of five different Panel Data Regressions are exposed along with their tests. When it comes to a Panel Data Regression, there are several methods that can help us evaluating which Panel Data Regression technique fits better to our data base. These methods are composed by plmtest which evaluates the Pooled Regression, the pFtest which compares the Fixed Effects Model with Pooled and the phtest which compares the Random Effects with the Fixed Effects. First, we will see what kind of results we got from those five regressions, Pooled Regression by OLS, Between Estimator by OLS, First Difference Estimator, Fixed Effects Estimator and Random Effects Estimator. Later on, we will cover the Evaluation part of the Cross Industry Standard Process for Data Mining when we show the results of those tests we mentioned before.

Pooled Regression by OLS

When it comes to Panel Data, Pooled Regression by OLS usually is not a good technique to use because is a bad fit of the data, pooled is actually ignoring that we are managing a panel data. We are confirming that by taking a look to our results in Figure 4.13, R-squared is so low because the error terms are highly correlated and this estimator is not taking care of that, a lot of information is being lost in this particular model.

```
> pooling <- plm(Score ~ Genero+RangoEdad+TypeProfessor, data=pdata, model="pooling")
> summary(pooling)
Pooling Model

Call:
plm(formula = Score ~ Genero + RangoEdad + TypeProfessor, data = pdata,
    model = "pooling")

Unbalanced Panel: n = 9038, T = 1-5, N = 30969

Residuals:
    Min.  1st Qu.   Median  3rd Qu.     Max.
-8.95189 -0.32634  0.21116  0.57811  1.21335

Coefficients:
               Estimate Std. Error  t-value  Pr(>|t|)
(Intercept)   9.0576154  0.0195144 464.1495 < 2.2e-16 ***
GeneroM      -0.0396279  0.0101581  -3.9011 9.596e-05 ***
RangoEdad    -0.0330480  0.0045825  -7.2118 5.646e-13 ***
TypeProfessor  0.2041844  0.0181544  11.2471 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    23530
Residual Sum of Squares: 23385
R-Squared:      0.0061597
Adj. R-Squared: 0.0060635
F-statistic: 63.9728 on 3 and 30965 DF, p-value: < 2.22e-16
```

Figure 4.13: Pooled Regression by OLS

Between Estimator by OLS

In the second and last estimator that applies OLS we found similar results as the Pooled Regression one. As shown in Figure 4.14, our outcomes are very poor, it does indicate that the three features are significant but the R-squared is very low. The Between estimation model by OLS is definitely not a fit to this data.

```
> between <- plm(Score ~ Genero+RangoEdad+TypeProfessor, data=pdata, model="between")
> summary(between)
Oneway (individual) effect Between Model

Call:
plm(formula = Score ~ Genero + RangoEdad + TypeProfessor, data = pdata,
    model = "between")

Unbalanced Panel: n = 9038, T = 1-5, N = 30969
Observations used in estimation: 9038

Residuals:
    Min.  1st Qu.   Median  3rd Qu.     Max.
-8.24858 -0.31358  0.19322  0.56335  1.27756

Coefficients:
               Estimate Std. Error  t-value  Pr(>|t|)
(Intercept)    8.9392809  0.0347923 256.9327 < 2.2e-16 ***
GeneroM       -0.0484097  0.0186227  -2.5995 0.0093511 **
RangoEdad     -0.0280725  0.0083935  -3.3445 0.0008275 ***
TypeProfessor  0.2685383  0.0361394   7.4306 1.179e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    6756.5
Residual Sum of Squares: 6702.9
R-Squared:      0.0079322
Adj. R-Squared: 0.0076028
F-statistic: 24.0776 on 3 and 9034 DF, p-value: 1.6321e-15
```

Figure 4.14: Between

First Difference Estimator

Figure 4.15 corresponds to First difference which exploits the features of panel data and finds the association between the individual specific changes in the dependent variable. In this case, we did not get any significant variable, every feature got dropped. The R-squared is extremely low. We found two main differences in this estimator, the first one is that none of the variables from the dataset are being identified as significant. The second one is the poor value of the R-squared. We did not expected this since we had consider that the professor's age, type, and academic level could be affecting directly their scores in student evaluations.

```
> firstdiff <- plm(Score ~ Genero+RangoEdad+TypeProfessor, data=pdata, model="fd")
> summary(firstdiff)
Oneway (individual) effect First-Difference Model

Call:
plm(formula = Score ~ Genero + RangoEdad + TypeProfessor, data = pdata,
    model = "fd")

Unbalanced Panel: n = 9038, T = 1-5, N = 30969
Observations used in estimation: 21931

Residuals:
     Min.   1st Qu.    Median   3rd Qu.      Max.
-8.718164 -0.353164  0.011836  0.368503  6.796836

Coefficients:
             Estimate Std. Error t-value Pr(>|t|)
(Intercept) -0.0118362  0.0059874 -1.9768  0.04807 *
RangoEdad    0.0148017  0.0257145  0.5756  0.56488
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    16306
Residual Sum of Squares: 16306
R-Squared:      1.5109e-05
Adj. R-Squared: -3.0492e-05
F-statistic: 0.331335 on 1 and 21929 DF, p-value: 0.56488
```

Figure 4.15: First Difference

Fixed Effects Estimator

As shown in Figure 4.16, Fixed Effects dropped every variable but the class of the professor's age. Since Academic Level, gender and type of professor are time invariant, they all got cancelled by the algorithm. We also got a very low R squared. Before going to the next subsection, it is important to know that one of the main differences between Random effects and this estimator is that the first one assumes that individual specific effects are independent of the regressor, this means that the correlation between the alpha and Xi is taken as 0. However, in Fixed Effect, none zero correlation is assumed.

```
> fixed <- plm(Score ~ Genero+RangoEdad+TypeProfessor, data=pdata, model="within")
> summary(fixed)
Oneway (individual) effect Within Model

Call:
plm(formula = Score ~ Genero + RangoEdad + TypeProfessor, data = pdata,
    model = "within")

Unbalanced Panel: n = 9038, T = 1-5, N = 30969

Residuals:
     Min.   1st Qu.    Median   3rd Qu.      Max.
-6.710000 -0.200000  0.013333  0.250000  3.392500

Coefficients:
           Estimate Std. Error t-value Pr(>|t|)
RangoEdad 0.0014295  0.0192637  0.0742   0.9408

Total Sum of Squares:    8541.7
Residual Sum of Squares: 8541.7
R-Squared:      2.5112e-07
Adj. R-Squared: -0.41213
F-statistic: 0.00550701 on 1 and 21930 DF, p-value: 0.94084
```

Figure 4.16: Fixed Effects

Random Effects Estimator

As for our last Panel Data Experiment, Figure 4.17 illustrates somehow similar results compared to the ones we got in the Beetween Estimator Model. Same features resulted highly significant to our goal variable. Later on, in the next chapter we will discuss broadly these results and why we consider them important.

```
> random <- plm(Score ~ Genero+RangoEdad+TypeProfessor, data=pdata, model="random")
> summary(random)
Oneway (individual) effect Random Effect Model
   (Swamy-Arora's transformation)

Call:
plm(formula = Score ~ Genero + RangoEdad + TypeProfessor, data = pdata,
    model = "random")

Unbalanced Panel: n = 9038, T = 1-5, N = 30969

Effects:
                 var std.dev share
idiosyncratic 0.3895  0.6241 0.516
individual    0.3658  0.6048 0.484
theta:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.2819  0.4882  0.5810  0.5330  0.5810  0.5810

Residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-7.5682 -0.2022  0.1589  0.0157  0.4095  2.0271

Coefficients:
               Estimate Std. Error  z-value  Pr(>|z|)
(Intercept)    8.9707457  0.0287289 312.2555 < 2.2e-16 ***
GeneroM       -0.0470412  0.0159547  -2.9484 0.0031940 **
RangoEdad     -0.0260449  0.0067972  -3.8317 0.0001272 ***
TypeProfessor  0.2438554  0.0300062   8.1268 4.407e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    24049
Residual Sum of Squares: 12955
R-Squared:      0.47273
Adj. R-Squared: 0.47268
Chisq: 26517.2 on 3 DF, p-value: < 2.22e-16
```

Figure 4.17: Random Effects

Summary Panel Data Table

Now, since these models take into account same type of estimators, we decided to expose them in the following table. First, we have the name of the five models we applied followed by the different and most estimators to take into account in order to decide which was our most effective experiment. Let us remember that for the Adjusted R^2 the higher the value, the better it is, and it can go from 0 to 1. The Residual Sum of Squares (RSS) is a discrepancy measure between the data and our estimation model. In this case, the lower the value, the better our model since it indicates a better fit to the data. Our third estimator, the Relative Standard Error follows the same criteria, the lower the value the better, and it goes from zero to one. Finally, as we have seen in other experiments, the p-value is a very important estimator to follow, in this case we are still expecting for a value lower than 0.05.

Models	Adjusted R2	RSS (error)	RSE	P-value
Random Effects	0.47	12955	0.6468168	2.22E-16
Feature Effects	2.5E-5	8541.7	0.5251903	0.94084
First Difference	3.04E-5	16306	0.8623135	0.56488
Between	0.076	67029	0.8613704	1.63E-15
Pooled	0.006	23885	0.8690306	2.22E-16

Table 4.28: Panel Data Models Results

Panel Data Statistical Tests

By now, we cannot say which one is the best model for our data just by looking at the R-squared or the significant features identified, we need to verify it statistically, several test will tell us which estimate is more suitable for our panel data. As we mentioned before, in order to evaluate the fitness of the Panel Data Models, we need to apply three different stistical methods, figures 4.18, 4.19 and 4.20 correspond to the results of those tests show their results. We will discuss these evaluation methods broadly in the next Chapter.

```
> plmtest(pooling)

        Lagrange Multiplier Test - (Honda) for unbalanced panels

data:  Score ~ Genero + RangoEdad + TypeProfessor
normal = 70.677, p-value < 2.2e-16
alternative hypothesis: significant effects
```

Figure 4.18: Lagrange Multiplier Test, Random Effects vs OLS

```
> phtest(random,fixed)

        Hausman Test

data:  Score ~ Genero + RangoEdad + TypeProfessor
chisq = 2.3234, df = 1, p-value = 0.1274
alternative hypothesis: one model is inconsistent
```

Figure 4.19: Random and Fixed Effects Comparison

```
> pFtest(fixed,pooling)

        F test for individual effects

data:  Score ~ Genero + RangoEdad + TypeProfessor
F = 4.2179, df1 = 9035, df2 = 21930, p-value < 2.2e-16
alternative hypothesis: significant effects
```

Figure 4.20: Fixed Effects and OLS Comparison

4.4 Summary

Chapter Four has been one of the broadest sections of this thesis. First of all, we divided it in three subsections, Analysis of Variance (ANOVA) Logistic Regression and Panel Data. The first part, basically evaluated the results of Exploratory Data Analysis that we did in Chapter 3 for all the different data sets. We did this in order to identify which results were definitely valid statistical findings. We applied it to Dataset A, B, C and E. In the second part of this chapter we explained how did we manage to apply logit regression in Dataset A and B. In A we saw how we implemented Recursive Feature Elimination and in both we evaluated our results with different estimators and figures. Finally, we ended up by illustrating how we took advantage of our time-series dataset (E) by applying different Panel Data Models. In the next chapter, we will discuss how our experiments have contributed to answer our research questions.

Chapter 5

Discussion and Deployment

5.1 Discussion

5.1.1 Dataset A

We show a summary of the mean score for both teaching-only and teaching-and-research professors at different levels and obtained from the first dataset (August-December 2017) in Figure 5.1. The advantage of teaching-and-research professors is observed at all levels, except in undergraduate groups. The ANOVA analysis helps us by stating that these means are indeed significantly different from each other. Additionally, note that graduate students evaluate the teaching-and-research professors even higher when thesis groups are considered.

Course Level	Teaching-Only Professors	Teaching-and-Research Professors
Undergraduate	8.89	8.90
Graduate	9.89	9.21
Undergraduate and Graduate	8.9	9.21
Graduate and Thesis	9.18	9.57

Table 5.1: Professors' evaluation means results classified by different levels. Note that a professor might be teaching simultaneously in undergraduate and graduate groups, but his/her evaluation is accounted for in the corresponding level.

Ranking of professor and group characteristics using the Recursive Feature Elimination (RFE) algorithm permitted to identify the relevance that students would assign to them (see Table 4.24 to Table 4.26). Table 5.2 shows the comparison of the rankings obtained by RFE by analyzing the responses of undergraduate and graduate students. The column Agreement has the absolute difference of both rankings: the lower this number, the most similar relevance both students assigned to a given feature. In the top of table 4 are the features with more agreement, meanwhile features ranked by only one type of students are at the bottom.

Feature Type	Feature	Under-graduate Rank	Graduate Rank	Agreement
Group	Number of credits	7	6	1
Professor	Number of undergraduate students attended	11	8	3
Professor	Foreign nationality	3	7	4
Professor	Percentage of responsibility of the group	5	1	4
Professor	Total number of students attended	10	14	4
Professor	Number of graduate students attended	16	11	5
Group	Number of laboratory hours	17	12	5
Professor	Main professors	6	13	7
Professor	Number of hours at classroom	13	5	8
Group	% of participation on the survey	1	10	9
Professor	Number of scientific publications	18	9	9
Group	Number of senior students	14	4	10
Group	Number of teaching hours	4	15	11
Group	Class transmitted to multiple campus	15	3	12
Professor	Number of high school students attended	2	-	
Group	Is a terminal group	8	-	
Professor	Is a teaching-only professor	9	-	
Professor	Certified in the teaching abilities program	12	-	
Professor	Is a teaching and research professor	-	2	

Table 5.2: Agreement/Disagreement between graduate and undergraduate students ranking of Professor and Group features based on the Recursive Feature Elimination algorithm.

The features that undergraduate students rank high and graduate students rank low are: Foreign nationality (3 vs 7), Being the main professor at team-teaching classes (6 vs 13), Percentage of participation on the survey (1 vs 10), and the number of professor's teaching hours (4 vs 15). Additionally, four features were ranked only in undergraduate groups: the number of high school students attended by the professor (ranked 2nd), terminal groups (ranked 8), teaching- only professors (ranked 9), and being certified in the teaching abilities program (ranked 12).

The features that graduate students rank high and undergraduate students rank low are: the percentage of responsibility of the group in team-teaching groups (1 vs 5), the number of hours spent in the classroom (5 vs 13), the number of scientific publications made by the professor (9 vs 18), the number of senior students in the group (4 vs 14), and whether the class is transmitted to multiple campuses (3 vs 15). Additionally, the unique feature that was ranked only at graduate groups was: being a teaching-and-research professor (ranked 2nd).

As it can be seen and could be expected, features related to the professor's research activities are more appreciated by graduate students: 1) being a teaching-and-research professor (ranked 2), 2) classes with more hours of theory (ranked 5), and 3) the number of scientific publications made by the professor (ranked 9). The prominence of the ranking obtained by classes transmitted to multiple campuses at graduate groups can be explained by the relative higher number of these groups at graduate programs.

On the other hand, undergraduate students appraise professor features such as: 1) being a teaching-only professor (ranked 9), 2) the number of high school students attended by the professor, and 3) being certified in the teaching abilities program (ranked 12). Another feature ranked high by undergraduate students was having a foreign nationality (ranked 3), which seems to be appealing to Mexican students. And finally, features related to maturity of students was also ranked high on undergraduate groups: being a group of the last semesters of the program (ranked 8th).

Our results confirm the correlation found by both Stack [65], Ting [67] and Spooren [63] between student perception of teaching quality and research activities. Stack and Ting correlate teacher's productivity, in terms of papers and citations, with higher scores on student evaluation. As for the results of the Recursive Feature Elimination, we have seen that being a teaching-and-research professor is one of the most highly positively correlated features with the professor's score in a student survey, see Table 4.24. This variable proves the statement which indicates that professor's research productivity is a valuable indicator of a teacher's educational skills and knowledge of the subject matter and is reflected on student evaluation of teaching [63] . In addition, we saw in Tables 4.24, 4.25 and 4.26 that the maturity of students contributes most to scoring professors higher. A similar study also indicated that the older the student evaluating the professor, the higher score will be given [12]. Moreover, observing that percentage of responsibility in both populations indicates that team teaching seems to affect the professors' evaluations in a negative way.

5.1.2 Dataset B

We confirmed the previous result by the analysis of the second dataset even when we split the weighted score into the three satisfaction dimensions of the ECOA survey. In the first place, we observed that Professor Recommendation is scored lower than Learning Guide and Intellectual Challenge overall. And when we compare the scores obtained by teaching-only and teaching-and- research professors, we observed again that the latter obtain a higher score in all the three dimensions.

Our results are consistent with their findings as long as those professors we classified as teaching- only have very few scientific publications in comparison to those classified as teaching-and- research professors. The latter has at least one research product by year and depending on their discipline this number varies. Nevertheless, the differences of both kinds of professors given the average scores of the three teaching dimensions is more evident. We can attribute this pronounced difference to the definition of researcher we used, which is supported by CONACYT (the Mexican Council for Science and Technology), an external agency. Furthermore, our results contribute to this discussion by analyzing the effect of gender on teaching quality. In our analysis, female professors are better evaluated than male professors as Intellectual Challenge (05. RET), Learning Guide (06. APR) and Professor Recommendation (08. REC), confirming the observations of Basow and Montgomery [9] and Smith and colleagues [62]. And whereas this difference is also observed among teaching-only professors, it is reversed among teaching-and- research professors, where male researchers are better evaluated by students in two teaching dimensions, Intellectual Challenge and Professor Recommendation.

On the other hand, we observed that aging seems to improve teaching quality of teaching-and-research professors but when ANOVA was applied we identified that it only happens in the Intellectual Challenge Factor. On the other hand, aging negatively affects the teaching quality of teach-only professors. The pattern found for teach-only professors is consistent with the results of Spooren [63], McPherson and Jewell [65] and McPherson and colleagues [52], on which they notice that younger teachers receive higher evaluations of students. Nevertheless, when we analyze teaching-and-research professors we found the opposite trend. This could be attributed to the experience accumulated through years of research, as pointed out by McPherson [52]. In both cases, the dimension Professor Recommendation showed the higher difference between the youngest and the oldest professors. In terms of proficiency level, we also found differences among teaching-and-research professors. In this case, an increase in the proficiency level is correlated with the quality of teaching in the three dimensions. An important finding is that researchers classified in the two upper levels of proficiency have similar scores, they are not statistically different between them but they are significantly higher than the lower two. Finally, we would like to point out that we found evidence of an improvement in teaching quality of the professors overall (see Figure 3.6). The causes of this improvement must be further studied. By now, we can only hypothesize that this variation is attributed to institutional efforts on educational innovation and training for all the professors.

5.1.3 Dataset C

As for our third dataset, we have identified that when considering periods from 2016 to 2019, Teaching-and-Research Professors are better evaluated in the Recommended Professor Factor, whereas for the 2018 and 2019 periods, Teaching-and-Research Professors are evaluated better in the Intellectual Challenge one. Later on, when we take into consideration the schools of Tec de Monterrey, we got interesting results. At the Humanities and Education School, Teaching-and-Research Professors had higher scores in both academic factors, in fact these type of professors got the highest evaluations at this precise school. Moreover, in the Engineering and Sciences School, Teaching-and-Research Professors were better evaluated in both factors and in all periods.

On the other hand, in the School of Medicine and Health Sciences, Teaching-and-Research and Teaching-only professors' scores were not statistically different in 2018 for the Intellectual Challenger Factor, it was only in the next year where researchers surpass full-time professors. In the other dimension, we saw that Teaching-only professors got higher scores in the first period, the second one did not favored any type of teacher. In general, this school presented very close scores between both type of teachers. In addition, Teacher-only professors received higher scores at this one compared to any other school. In Business School, the advantage was very clear, again for Teaching-and-Research professors at both dimensions in every period. Despite all these results, in the last two schools we did not found the same results. The School of Architecture and Arts did not presented any statistical difference in the means of professors. It was at this school where we identified that Researchers developed the worst. The Social Sciences and Government also did not presented any statistical differences in the mean scores of professors.

5.1.4 Dataset D

When it comes to Figure 4.17 and by taking into account equation 3.2, we can make the following statements. First, the model gives us a result of 8.97 for the intercept. This means that if Tecnologico de Monterrey wants to know the average score of a professor for the next academic period, just taking into account their specific average scores of the last five semesters, and without considering if he or she is a teaching-only or researcher professor, his or her age, and their gender, the teacher will get a score of 8.97. Additionally, if we take a look at the regression beta coefficients or slopes, for each of the independent variables, it can be said that if the professor is a woman, the average score will be affected by 0.047 points negatively. In the case of the type of the professor the 0.24 coefficient means that, when a teacher is also a researcher it is expected that his or her score will be benefited by that amount of points. For the last variable, a coefficient of -0.26 was gotten by the model. This can be interpreted as whenever a professors' age belongs to a higher range, the score will be affected negatively by that number. On the other hand, it is worth mentioning that the model shows us that all of the independent variables mentioned before are statistically significant to the average score, which is the dependent variable. This is concluded by taking a look to their p-values, each of them are lower than 0.05. The next column of the coefficients table represents the standard errors which define the accuracy of each of the beta coefficients and it reflects how the coefficient varies under repeated sampling. Since both of the p-values for the

intercept and the predictor variables are highly significant, the null hypothesis can be rejected which means that, there is a significant association between the predictor and these outcome variables.

Once this has been identified we can continue checking how well the model fits the data through the goodness-of-fit process. According to Table 4.28 this model got a Residual Standard Error of 0.646818, meaning that the observed score values deviate from the true regression by that number. In other words, if we divide the mean value of the scores is 8.91878297 by the RSS we get a value 13.7887%. We identify this as our percentage error and since it is considerably low, it is an acceptable prediction error for this specific problem context. Another point to emphasize is that the R-squared is 0.4723. This is a value that can be considered as low but in this case we consider it as significant given the nature of the problem, an R-squared of that magnitude in the are of Social Sciences represents a strong significant result.

For instance, the Lagrange Multiplier Test shown in Figure 4.18 decides between random effects regression and an OLS Pooled Regression by applying the plmtest in R. The null hypothesis is that there is no significant difference across cross-sectional units, this implies that Random Effects model cannot be rejected as the best model. Figure 4.18 illustrates a p-value lower than 0.05, this means that our test is significant, so the null hypothesis in favor of OLS is rejected. OLS is not a better fit than Random Effects.

On the other hand, the LM test, shown in Figure 4.20 to choose between the Fixed effects and the OLS can be done by utilizing the pFtest in R. This method implies that there are no time invariant effects so OLS should not be used. Since Figure 4.20 illustrates a p-value close to 0, we reject the null hypothesis and we can imply that the Fixed Effects Estimator is a more suitable method than OLS.

Now, since our tests did not favored OLS, we need to verify statistically which of the other estimates, either Random Effects or Fixed Effects, fit better into our model. In order to do that we performed the Hausman Test, phtest in R. Figure 4.19 shows a p-value higher than 0.05 which means it is not significant. We found out that the null hypothesis can not be rejected meaning that we fail to reject that the preferred model is random effects.

5.1.5 Dataset E

As for our last dataset, when we manage preprocessed data we found that Teaching-and-Research Professors are better evaluated when students analyze them as Intellectual Challenger, Learning Guides and Recommended Professor. This result was the same one as the one we got in Data Set B when data was not preprocessed. As for the study of Teaching-and-Research Professors when we take into account their SNI Proficiency Level, we got almost the same results as Data Set B. The higher the proficiency level the higher you are evaluated as a professor. This applies until Level 2, our results indicate that SNI Level 3 does not guarantee a higher evaluation since the mean was lower than the group of SNI Level 1. Dataset B results were different in this, we saw that Level 2 and 3 SNI got the highest evaluations and they were not statistically different among them.

In general, when all three factors were averaged,we did not find any difference between Male and Female Teachers. When we analyze the three questions separately, Female Professors were better evaluated in Learning Guide Factor, at other dimensions, means were not different. In the case of teaching-only professors we saw that Female are as well better evaluated in Learning Guide Factor than Male teachers. We did not find any other statistical difference at the rest of the dimensions. Finally, when we studied Teaching-and-Research Professors we also did not find any difference when the gender was also being considered.

5.2 Deployment

This section comprises the last part of the CRISP-DM Methodology. We related it somehow to the essentials behind the theory of Actionable Knowledge Discovery that we covered at the beginning. By taking into consideration all the results we got in the past few sections, we would like to offer Tecnologico de Monterrey the following recommendations. The way in which we think our findings could be put into action is through taking into consideration the following advises.

1. Ensure offering courses given only by Teaching-and-Research Professors at Graduate Levels.
2. At Graduate Levels, make sure to offer all classes at a national scheme.
3. Certificate of Teaching abilities does not guarantee higher ECOA scores at Undergraduate Levels, actually it has a negative correlation with the professors' evaluations.
4. Team teaching is not the most effective strategy to follow. Keep offering courses where one professor is fully responsible of the class, this applies for both academic levels.
5. Keep hiring professors with PhDs, they usually are better evaluated at both academic levels. Masters is not sufficient.
6. Take into consideration that students tend to evaluate lower to Foreign professors.
7. In general, Teaching-and-Research Professors have a higher teaching performance than Teaching-only Professors. Find the way to increase the number of this type of professors.
8. The presence of younger teaching-only professors or older Teaching-and-Research Professors at courses can ensure student satisfaction.
9. Allocate more Female Teaching-only and Teaching-and-Research Professors at Undergraduate Level courses, according to students they are better as Learning Guides.
10. Keep hiring Teaching-and-Research Professors at the School of Humanities and Education and Teaching-only Professors at the school of Medicine and Health Sciences.

5.3 Summary

We have discussed into detail each of the results we got for every single dataset we were managing in this research work. We followed that order covering the interpretation of the ANOVA's results, Logistic Regression results, Recursive Feature Elimination broad interpretation, the assessment of our Panel Data experiments, and finally, the comparison of results between the preprocessed data set and the original one. Later on, we gave some ideas on how we could take those results into consideration in order to improve Tecnologico de Monterrey's future academic strategies in order to ensure student satisfaction.

Chapter 6

Conclusions

We presented a study in which we try to answer the research question regarding teaching performance of teaching-only and teaching and research professors: Does the former perform better than the latter according to student opinion in teaching and research institutions? The context in which this question is addressed is given by teaching and research universities ranked on the band 101 - 200 of QS world university rankings, although the same question can be asked for universities with a teaching and research orientation, independently of the ranking band.

We approached the research question by applying the methodology Cross-Industry Standard Process for Data Mining, known as CRISP-DM, which is a common method used in data analytics studies. We applied the six steps which define the methodology that go from business understanding to system deployment, including data understanding and preparation, as well as modelling and model evaluation. The modelling phase of the CRISP-DM methodology applied data mining and statistical methods that included Logistic Regression and Analysis of Variance, Recursive Feature Elimination, Panel Data and Coarsened Exact Matching. When we considered a relatively small data set with high dimensionality, we discovered that overall, researchers have higher teacher performance that non researchers, graduate students evaluate higher and that both type of professors have similar performance only at Undergraduate courses. The calculation of coefficients of the logistic function using the training data yielded a model that was applied to the holdout data that was set apart to test the model. The accuracy of the resulting model was evaluated using procedures like ROC curves and confusion matrix which showed a statistically significant prediction capacity. On the other hand, when a way bigger data set is considered, we also found that the scores for researchers are way higher than teaching-only professors, in this case at all three academic factors. Also, for the last 3 periods scores have presented a tendency of getting higher. Graduate students indeed give higher scores than Undergraduate students in all three dimensions. We also were able to create 6 additional Logistic Regression Models making an analysis for the overall and for each of the academic factors achieving better results and comparisons among them. Once the model was found, the experiments were carried out using different variables to perform runs with data drawn from the undergraduate level, graduate level, the age of teachers, and teacher gender. Then, when we consider only similar courses given by these two type of professors

and we add an extra period of data we also found that researchers perform better than non researchers at Intellectual Challenge Factor and Recommended Professor Factor, and we could also see which schools favor the evaluations of both professors. Moreover, by transforming our observations into Panel Data we have learned that, if we use the right estimation technique or model, we will could expect a better outcome on our R-squared value and it will be a better fit to the model. Even though we did not get a very high R-squared, the obtained value for the Random Effects Estimator was sufficient to be considered. We statistically prove that from all of our estimators and experiments, the one that fits this data the best is Random Effects. We also proved this when illustrating and comparing its results with the rest of the models. So, Panel Data Modelling is a method that allowed us to estimate data which is both time series and cross sectional. By studying the same cross-sectional unit over a period of time we could found that features such as Type of Professor, professor's age and gender are statistically significant, these features do affect the average score of a professor in the ECOA.

By taking into consideration these results, we can state that there is statistical significant evidence to reject the Null Hypothesis of our study. We have provided evidence in favor of research professor yielding in teaching activity (The Alternate Hypothesis). In other words, and as for our main research questions, Teaching-and-Research professors have better teaching performances than Teaching-only ones. When we only analyze the Intellectual Challenge Factor we found that the ones who performs better in general are the Teaching-and-Research professors. Moreover, they are also considered better Learning Guides and additionally, this type of teachers consequently tend to be more recommended by their students that Teaching-only professors.

We have also found that Male professors are only more preferred at Graduate Levels and that Female teachers have a slightly higher performance than Male Professors in Undergraduate Courses and they are for sure considered better as Learning Guides at both Academic Levels. As for the comparison of Teaching-and-Research Professors between Male and Female we saw two scenarios, in the first one, where our dataset has not been preprocessed with the Coarsened Exact Matching Algorithm, we only found that Male Research Professors are more recommended than Female ones. However, when we preprocessed our data we did not find any differences. Since it is still uncommon to see Female Research Professors at Tec, we believe that either Male Research Professors will still have a certain advantage but it is worth mentioning that we do not believe this will for much longer. We have seen the tendency that the number of Female Research Professors will approach in a considerable way the total number of Male Research Professors in the near future since. The School of Sciences and Engineering is the one with the largest number of research professors and the participation of woman in STEAM research areas is becoming more common. That is why we believe these results could easily change in the next academic periods. Furthermore, we can state that Professor's teaching experience does not determine their good or bad evaluations. We have seen that, other features such as type of professor, gender, research proficiency level, school, among others, are some of the characteristics that have a stronger correlation with the professors' evaluation.

Our research helped us identify that senior students usually evaluate their teachers better. At the same time, one of the most important findings we had is that a higher researcher proficiency level does not ensure higher performance in teaching. A professor with the last level of

SNI usually have a worse performance than Level 1 professors, but having Level 1 or Level 2 does ensure being better evaluated at the ECOA. When it comes to aging, we found that it only affects negatively to Teaching-Only Professors, in the case of Teaching-and-Research Professors, aging favors when the professor is being evaluated at the Intellectual Challenge Factor. We expected that aging in research professors and research proficiency level could show a similar behavior since it is still uncommon for Tec to have researchers who have gained a SNI Level at an early age. Finally, to answer our last research question we can say that there is indeed statistical differences in teacher evaluation's scores between the academic levels. We identified higher scores for graduate courses.

The results showed that in general, teaching and research professors perform better or at least the same as teaching-only professors, on graduate or undergraduate academic levels using data of student survey results of five semesters at Tecnológico de Monterrey, a teaching and research university ranked in position number 158 of QS World University Rankings 2020. We hope that these results contribute to revising the believe that research professors in teaching and research institutions are not good teachers in general.

For all of these reasons we can firmly say that we achieved all of our specified goals. We were able to determine if Teaching-only Professors have better or worse teaching performance than Teaching-and-Research Professors. We could also state what are the main characteristics which makes a professor being good evaluated. In addition, we were able to analyze the correlations of several features affecting professors' scores. It was also possible to know how these two type of professors were evaluated at the different schools of Tecnologico de Monterrey. Our results also helped us to determine that student's maturity level is key for a teacher to be rated highly. At the final sections of the thesis we proposed some corrective actions that can improve Tec's educational services. We believe that research and results of this type can surely facilitate Tecnologico de Monterrey teaching staff hiring with our predictive models that can ensure student satisfaction.

Independently of teaching-only versus research and teaching professor behavior, we believe that the results obtained may be useful information in scheduling teachers to classes in running academic periods of higher education institutions. By identifying the group attributes where researchers are best evaluated, we could recommend a better group assignment for them, i.e., graduate courses and undergraduate courses of terminal semesters. In this way, we would improve the learning process and satisfaction of students.

6.1 Future work

There is still much to investigate on this complex and broad relationship between research and teaching. As we mentioned before, Tecnologico de Monterrey's educational model is continually evolving [43]. The new education model called Tec21 started in August 2019 introducing a set of innovations on pedagogical methods whose elements have been smoothly introduced since 2012 are now fully implemented and deployed [3]. The role of a teacher either teaching-only or teaching and research has been fully revised and new features of teacher activity need to be evaluated in order to determine the level of teaching quality and satisfaction [34]. Thus, a new model for evaluation of teaching performance needs to be designed and implemented, and once it is running, we will need new data generated from student evaluation of teaching in

order to conduct new experiments of the theme of teacher-only versus teaching-and-research professors [14].

As future work we would like to perform teaching performance analysis over a more extensive dataset. At least we would like to double the number of academic periods considered in Dataset B, which were five. Besides this, it would be really important for us that this more complete dataset at least counts with the approximately 60 features that Dataset A presented. We think that this will give us the opportunity to apply additional data science, machine learning and deep learning techniques. For example, we wanted to apply clustering algorithms to our second dataset but its properties would not allow us to present strong results since it presented few variables. Through that future research we would like to apply algorithms such as K-Means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Expectation–Maximization (EM) Clustering using Gaussian Mixture Models (GMM). Additionally, we have identify other feature selection and extraction algorithms than just Recursive Feature Elimination that we could apply when having a dataset vast in features. These are Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Autoencoders, Embedded and Filter Methods.

In order to be able to present concise results that can apply not only to this institution, we propose to first approach this problem in the same way but in other academic institutions. We believe that, if at some moment we would like to say in a more forceful way who performs better as a teacher, teaching-only professors or teaching-and-research ones, we are required to perform a much broader experiment which involves data from universities with characteristics similar to those of Tec and which utilizes a very similar tool for evaluating teaching performance.

Another analysis that could be done in the future could be related to the comparison of evaluations to teachers depending on the region where they are located. We could do this by considering the location of the Tec de Monterrey campuses and thus define clusters. We believe that by taking into account demographic aspects, we could identify the different ways or similarities in which a professor is evaluated at different regions, and this could be related to the culture that prevails in each of those locations.

Moreover, we have seen that some universities are focusing their teaching performance analysis in a different manner. In addition to contrasting the satisfaction of the students or the evaluations of the students towards their teachers, they are considering how was that student graded at that specific course that he or she has evaluated. We consider that there could be a strong relationship between these two evaluations, students' and professors' ones. It could be really interesting if we add specific students' features like the one we mentioned to our existing models. We believe that this could complement our findings.

Furthermore, in the near future this research can be extended by considering the COVID-19 scenario that we are currently living. Given that academic institutions have been forced to offer their courses in fully online format, we believe that the behavior of professors' evaluations results will be very different than previous academic periods where lectures were imparted mainly on campus. We could start with this analysis by taking the advantage that ECOA is answered by students in the middle of the semester and at the end of the semester. As Tec de Monterrey switched to the online format in the middle of the semester, we will be able to analyze these two ECOA databases since they clearly represent the evaluations of the teachers performed by students in a certain semester before the coronavirus contingency

measures were applied and after they were applied. By making this analysis we could also identify the characteristics that makes a professor being good or bad evaluated when he or she teaches online. Additionally, we could see which type of schools are the ones who struggle more or less to teach in that format effectively.

We would also like to continue with this research by making an in depth analysis to the team teaching case. Let us remember that we define and measure Team Teaching as the case in which two or more professors share a certain percentage of responsability in a course. We saw before that it seems that team teaching does affect professors' evaluations, meaning that the lower the percentage of responsibility of a professor in a course, its teaching performance tends to decreaso. However, we would like to know in which specific cases does it affect negatively or positively. We believe that team teaching can be a benefit in some cases since it can involve factors such as the chemistry that exists between professors, the time they have been working together, whether they belong to the same department or not, their age, among others.

Appendix A

Appendix

A.1 Publications

This Thesis motivated three scientific papers, which were developed during its execution. We attach them in this first Appendix and consist of the following:

- Héctor G. Ceballos, Mario D. Chavez, Francisco J. Cantu-Ortiz. "A Data Analytics Approach to Contrast the Performance of Teaching (only) vs. Research Professors" Original Paper accepted for publication at the International Journal on Interactive Design and Manufacturing (IJIDeM). November 1st 2019. The article is about to be published.

- Luis A. Sedas, Mario D. Chávez, Héctor Ceballos, Francisco J. Cantú-Ortiz. "Comparing the Performance of Teaching (only) and Research Professors". Conference Article Accepted and Presented at the 4th Workshop on Educational Innovation in Engineering and Sciences:Technologies for the Future of Learning. Virtual Concept Workshop 2019. Monterrey, Mexico, 16-18 December. Access Article

- Dr. Héctor Ceballos, Mario D. Chávez, Luis A. Sedas, Francisco J. Cantú-Ortiz. "Comparing researchers' teaching performance with full time professors". Abstract (1st version of Virtual Concept Workshop Article) presented at EduData Summit at the Research Analytics Stream. QS Quacquarelli Symonds. British Museum, London, UK, June 11th 2019. Slides

- Gabriela Torres Delgado, Neil Hernandez Gress, Héctor Céballos, Mario D. Chávez. "In quality of teaching, satisfaction, and intellectual challenge, do students prefer research professors to nonresearch professors?". Original Article waiting to be accepted by a Journal.

Appendix B

Appendix

Nómina	Crn	Ejercicio_Académico	Nivel_Académico_Alumnos	Clave_Materia	Género	Edad	RangoEdad	Thesis	SNI	Researcher
L00000177	8611	201911	Profesional	TE1011	M	29.0	2	0	0	0
L00000177	15090	201813	Profesional	TE1011	M	28.0	2	0	0	0
L00000177	15099	201813	Profesional	TE3060	M	28.0	2	0	0	0
L00000177	15410	201811	Profesional	TE1011	M	28.0	2	0	0	0
L00000177	15415	201811	Profesional	TE3060	M	28.0	2	0	0	0

Figure B.1: Example of Data Set B cleaned and prepared for Logistic Regressions Part 1.

Nómina	Crn	Ejercicio_Académico	Nivel_Académico_Alumnos	Clave_Materia	Género	Edad	RangoEdad	Thesis	SNI	Researcher
L00000177	8611	201911	Profesional	TE1011	M	29.0	2	0	0	0
L00000177	15090	201813	Profesional	TE1011	M	28.0	2	0	0	0
L00000177	15099	201813	Profesional	TE3060	M	28.0	2	0	0	0
L00000177	15410	201811	Profesional	TE1011	M	28.0	2	0	0	0
L00000177	15415	201811	Profesional	TE3060	M	28.0	2	0	0	0

Figure B.2: Example of Data Set B cleaned and prepared for Logistic Regressions Part 2.

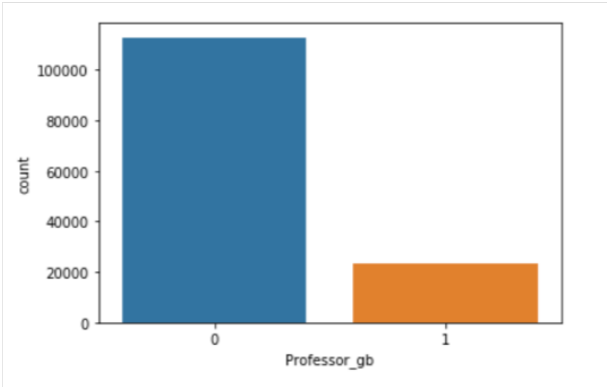

Figure B.3: Amount of Alpha (1) and Beta (0) Professors / Decile/ Balanced Data set B.

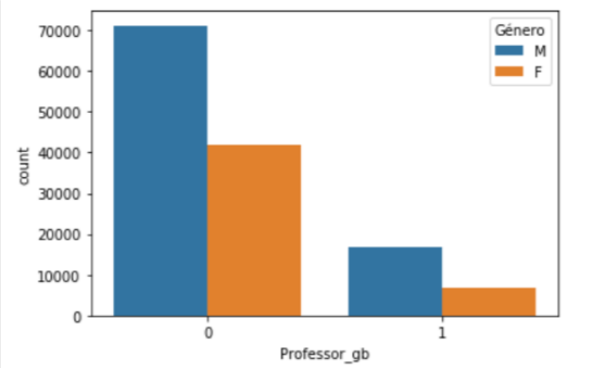

Figure B.4: Amount of Alpha (1) and Beta (0) Professors Classified by Gender/ Decile/ Balanced Data set B.

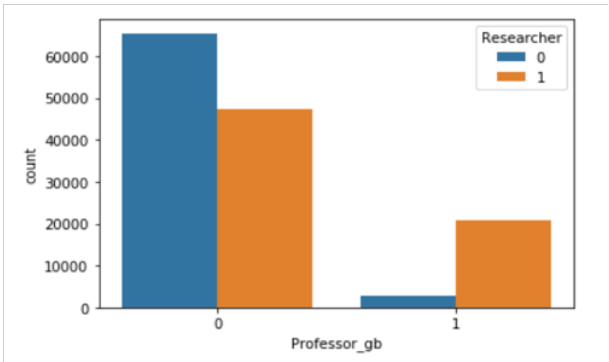

Figure B.5: Researcher (Orange) vs Non Researcher (Blue) Alpha (1) and Beta (0) Professors / Decile/ Balanced Data set B.

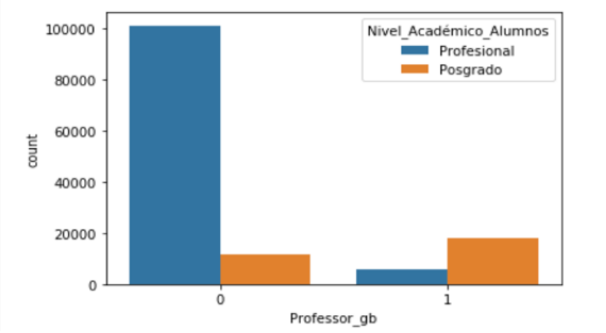

Figure B.6: Alpha (1) and Beta (0) Professors by Academic Level Undergraduate (Blue) Graduate (Orange) / Decile/ Balanced Data set B.

	precision	recall	f1-score	support
0	0.85	0.90	0.87	30264
1	0.66	0.54	0.59	10658
accuracy			0.81	40922
macro avg	0.75	0.72	0.73	40922
weighted avg	0.80	0.81	0.80	40922

Figure B.7: Detailed Logistic Regression results by Alpha (1) and Beta (0) Professors/ Decile/ Balanced Data set B.

```
Optimization terminated successfully.
         Current function value: 0.491202
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:           Professor_gb   No. Observations:               136406
Model:                          Logit   Df Residuals:                   136401
Method:                           MLE   Df Model:                            4
Date:                Sat, 07 Dec 2019   Pseudo R-squ.:                  0.1388
Time:                        20:17:15   Log-Likelihood:                -67003.
converged:                       True   LL-Null:                       -77805.
Covariance Type:            nonrobust   LLR p-value:                     0.000
============================================================================
                              coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------
Nivel_Académico_Alumnos     1.6434      0.017     94.282      0.000       1.609       1.678
Género                     -0.7822      0.013    -60.529      0.000      -0.808      -0.757
RangoEdad                  -0.5792      0.006   -103.119      0.000      -0.590      -0.568
SNI                         0.5895      0.012     47.168      0.000       0.565       0.614
Researcher                 -0.6414      0.027    -23.757      0.000      -0.694      -0.589
============================================================================
```

Figure B.8: R-studio LR Results Decile/ Balanced Data set B.

```
Nivel_Académico_Alumnos    5.172470
Género                     0.457395
RangoEdad                  0.560374
SNI                        1.803084
Researcher                 0.526545
dtype: float64
```

Figure B.9: Estimated Coefficients of Features /Decile/ Balanced Data set B.

	precision	recall	f1-score	support
0	0.93	0.98	0.95	20961
1	0.61	0.31	0.41	2298
accuracy			0.91	23259
macro avg	0.77	0.64	0.68	23259
weighted avg	0.90	0.91	0.90	23259

Figure B.10: Detailed Logistic Regression results by Alpha (1) and Beta (0) Professors/ Decile/Imbalanced Data set B.

```
Optimization terminated successfully.
         Current function value: 0.291262
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:           Professor_gb   No. Observations:                77530
Model:                          Logit   Df Residuals:                    77525
Method:                           MLE   Df Model:                            4
Date:                Fri, 15 Nov 2019   Pseudo R-squ.:                0.09064
Time:                        11:52:21   Log-Likelihood:                -22582.
converged:                       True   LL-Null:                       -24833.
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                             coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Nivel_Académico_Alumnos     1.1159      0.039     28.822      0.000       1.040       1.192
Género                     -0.9608      0.025    -38.062      0.000      -1.010      -0.911
RangoEdad                  -0.6155      0.011    -56.650      0.000      -0.637      -0.594
SNI                         0.7851      0.033     23.942      0.000       0.721       0.849
Researcher                 -0.1505      0.070     -2.151      0.031      -0.288      -0.013
==============================================================================
```

Figure B.11: R-studio LR Results Decile/ Imbalanced Data set B.

```
Nivel_Académico_Alumnos    3.052213
Género                     0.382598
RangoEdad                  0.540385
SNI                        2.192540
Researcher                 0.860250
dtype: float64
```

Figure B.12: Estimated Coefficients of Features /Decile/ Imbalanced Data set B.

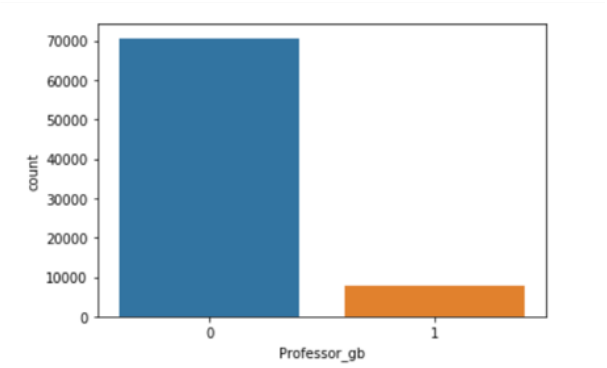

Figure B.13: Amount of Alpha (1) and Beta (0) Professors / Decile/ Imbalanced Data set B.

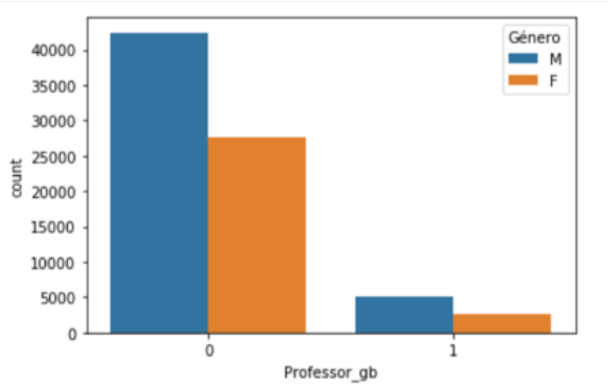

Figure B.14: Amount of Alpha (1) and Beta (0) Professors Classified by Gender/ Decile/ Imbalanced Data set B.

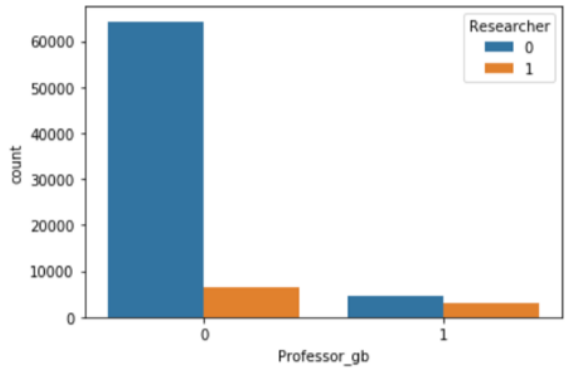

Figure B.15: Researcher (Orange) vs Non Researcher (Blue) Alpha (1) and Beta (0) Professors/ Decile/ Imbalanced Data set B

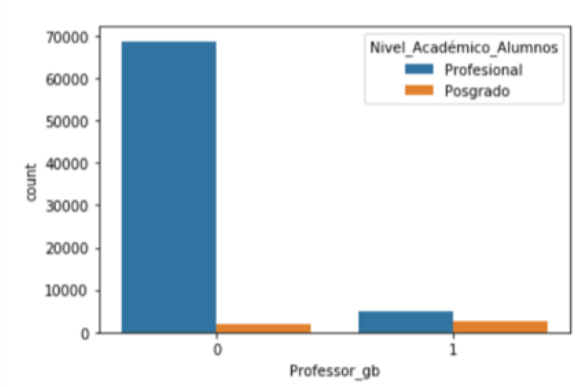

Figure B.16: Alpha (1) and Beta (0) Professors by Academic Level Undergraduate (Blue)Graduate (Orange) / Decile/ Imbalanced Data set B.

	precision	recall	f1-score	support
0	0.85	0.90	0.87	30264
1	0.66	0.54	0.59	10658
accuracy			0.81	40922
macro avg	0.75	0.72	0.73	40922
weighted avg	0.80	0.81	0.80	40922

Figure B.17: Detailed Logistic Regression results by Alpha (1) and Beta (0) Professors/ Quartile/ Balanced Data set B.

```
Optimization terminated successfully.
         Current function value: 0.491202
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:           Professor_gb   No. Observations:               136406
Model:                         Logit   Df Residuals:                   136401
Method:                          MLE   Df Model:                            4
Date:                Sat, 07 Dec 2019   Pseudo R-squ.:                 0.1388
Time:                        20:17:15   Log-Likelihood:                -67003.
converged:                       True   LL-Null:                       -77805.
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Nivel_Académico_Alumnos     1.6434      0.017     94.282      0.000       1.609       1.678
Género                     -0.7822      0.013    -60.529      0.000      -0.808      -0.757
RangoEdad                  -0.5792      0.006   -103.119      0.000      -0.590      -0.568
SNI                         0.5895      0.012     47.168      0.000       0.565       0.614
Researcher                 -0.6414      0.027    -23.757      0.000      -0.694      -0.589
==============================================================================
```

Figure B.18: R-studio LR Results Quartile/ Balanced Data set B.

```
Nivel_Académico_Alumnos    5.172470
Género                     0.457395
RangoEdad                  0.560374
SNI                        1.803084
Researcher                 0.526545
dtype: float64
```

Figure B.19: Estimated Coefficients of Features /Quartile/ Balanced Data set B.

	precision	recall	f1-score	support
0	0.92	0.91	0.92	32484
1	0.68	0.69	0.68	8438
accuracy			0.87	40922
macro avg	0.80	0.80	0.80	40922
weighted avg	0.87	0.87	0.87	40922

Figure B.20: Intellectual Challenge Factor/ Detailed Logistic Regression results by Alpha (1) and Beta (0) Professors/ Percentile/ Balanced Data set B.

```
Optimization terminated successfully.
         Current function value: 0.386627
         Iterations 7
                          Logit Regression Results
==============================================================================
Dep. Variable:           Professor_GB   No. Observations:               136406
Model:                          Logit   Df Residuals:                   136401
Method:                           MLE   Df Model:                            4
Date:                Sun, 08 Dec 2019   Pseudo R-squ.:                  0.2387
Time:                        19:07:27   Log-Likelihood:                -52738.
converged:                       True   LL-Null:                       -69270.
Covariance Type:            nonrobust   LLR p-value:                     0.000
============================================================================================
                               coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------
Nivel_Académico_Alumnos      2.1998      0.020    111.162      0.000       2.161       2.239
Género                      -1.1557      0.016    -73.923      0.000      -1.186      -1.125
RangoEdad                   -0.7969      0.007   -116.121      0.000      -0.810      -0.783
SNI                          0.8291      0.014     60.120      0.000       0.802       0.856
Researcher                  -0.7596      0.031    -24.689      0.000      -0.820      -0.699
============================================================================================
```

Figure B.21: Intellectual Challenge Factor / R-studio LR Results Decile/ Balanced Data set B.

```
Nivel_Académico_Alumnos    9.023227
Género                     0.314849
RangoEdad                  0.450726
SNI                        2.291163
Researcher                 0.467856
dtype: float64
```

Figure B.22: Intellectual Challenge/ Estimated Coefficients of Features /Decile/ Balanced Data set B.

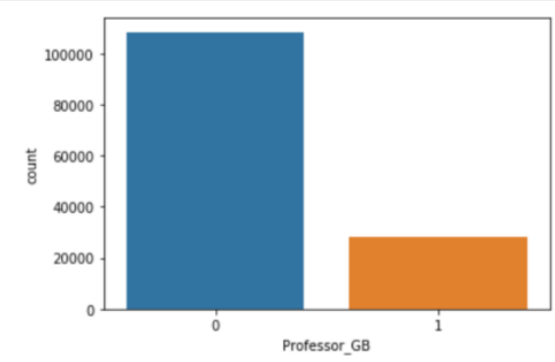

Figure B.23: Intellectual Challenge Factor / Amount of Alpha (1) and Beta (0) Professors / Decile/ Balanced Data set B.

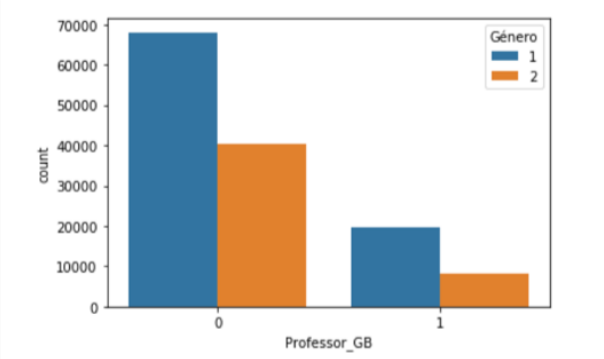

Figure B.24: Intellectual Challenge Factor / Amount of Alpha (1) and Beta (0) Professors Classified by Gender Male (1) Female (2)/ Decile/ Balanced Data set B.

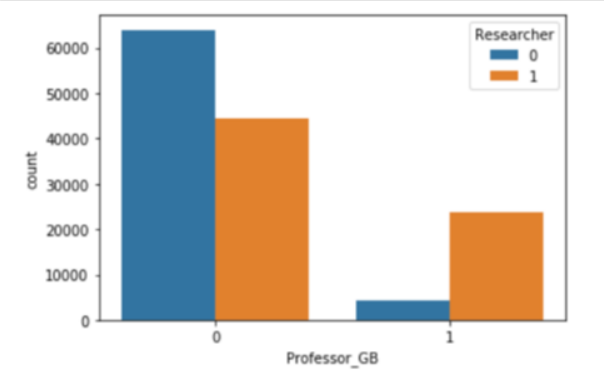

Figure B.25: Intellectual Challenge Factor/ Researcher (Orange) vs Non Researcher (Blue) Alpha (1) and Beta (0) Professors / Decile/ Balanced Data set B.

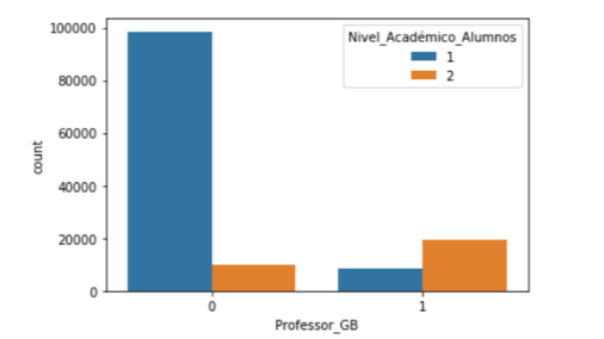

Figure B.26: Intellectual Challenge Factor / Alpha (1) and Beta (0) Professors by Academic Level Undergraduate (Blue)Graduate (Orange) / Decile/ Balanced Data set B.

	precision	recall	f1-score	support
0	0.92	0.91	0.92	32444
1	0.67	0.69	0.68	8478
accuracy			0.87	40922
macro avg	0.80	0.80	0.80	40922
weighted avg	0.87	0.87	0.87	40922

Figure B.27: Learning Guide Factor /Detailed Logistic Regression results by Alpha (1) and Beta (0) Professors/ Decile/Balanced Data set B.

```
Optimization terminated successfully.
         Current function value: 0.391408
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:          Professor_GB6   No. Observations:               136406
Model:                          Logit   Df Residuals:                   136401
Method:                           MLE   Df Model:                            4
Date:                Sun, 08 Dec 2019   Pseudo R-squ.:                  0.2299
Time:                        19:49:11   Log-Likelihood:                -53390.
converged:                       True   LL-Null:                       -69333.
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Nivel_Académico_Alumnos      2.2325      0.020    112.538      0.000       2.194       2.271
Género                      -1.1516      0.016    -74.145      0.000      -1.182      -1.121
RangoEdad                   -0.7910      0.007   -116.240      0.000      -0.804      -0.778
SNI                          0.7521      0.014     54.913      0.000       0.725       0.779
Researcher                  -0.7097      0.030    -23.297      0.000      -0.769      -0.650
==============================================================================
```

Figure B.28: Learning Guide Factor / R-studio LR Results Decile/ Balanced Data set B.

Nivel_Académico_Alumnos	9.323085
Género	0.316119
RangoEdad	0.453409
SNI	2.121499
Researcher	0.491786

Figure B.29: Learning Guide Factor / Estimated Coefficients of Features /Decile/ Balanced Data set B.

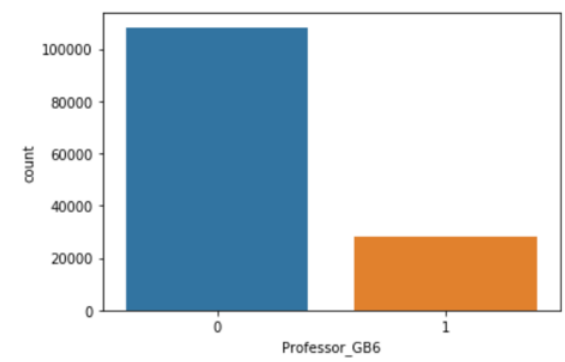

Figure B.30: Learning Guide Factor / Amount of Alpha (1) and Beta (0) Professors / Decile/ Balanced Data set B

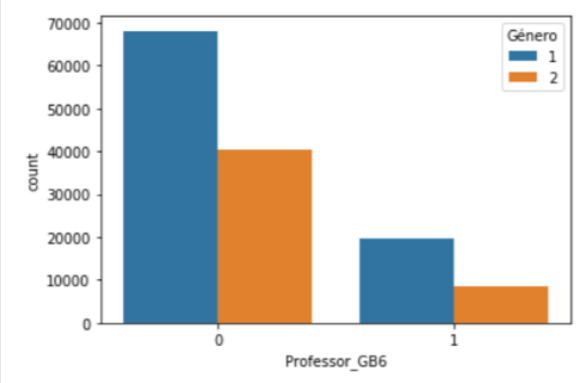

Figure B.31: Learning Guide Factor / Amount of Alpha (1) and Beta (0) Professors Classified by Gender/ Decile/ Balanced Data set B.

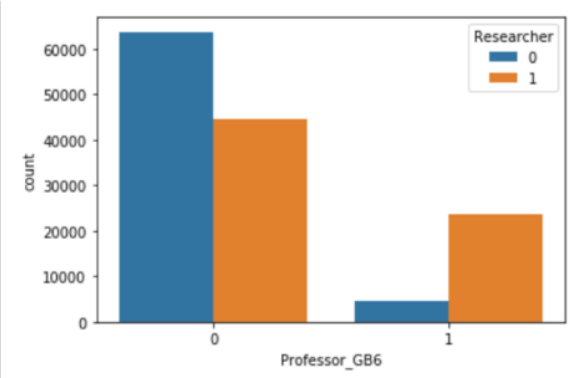

Figure B.32: Learning Guide Factor/ Researcher (Orange) vs Non Researcher (Blue) Alpha (1) and Beta (0) Professors / Decile/ Balanced Data set B.

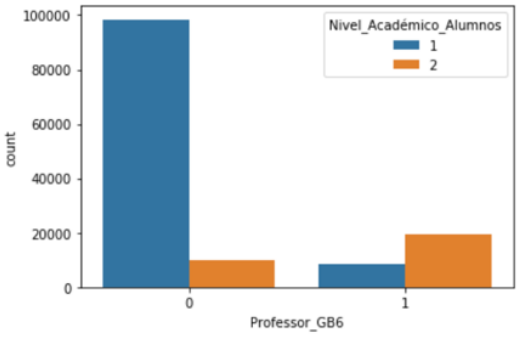

Figure B.33: Learning Guide Factor / Alpha (1) and Beta (0) Professors by Academic Level Undergraduate (Blue)Graduate (Orange) / Decile/ Balanced Data set B.

	precision	recall	f1-score	support
0	0.91	0.92	0.91	31900
1	0.70	0.68	0.69	9022
accuracy			0.86	40922
macro avg	0.80	0.80	0.80	40922
weighted avg	0.86	0.86	0.86	40922

Figure B.34: Recommended Professor Factor /Detailed Logistic Regression results by Alpha (1) and Beta (0) Professors/ Decile/Balanced Data set B.

```
Optimization terminated successfully.
         Current function value: 0.404684
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:          Professor_GB8   No. Observations:               136406
Model:                          Logit   Df Residuals:                   136401
Method:                           MLE   Df Model:                            4
Date:                Sun, 08 Dec 2019   Pseudo R-squ.:                  0.2314
Time:                        20:14:38   Log-Likelihood:                -55201.
converged:                       True   LL-Null:                       -71825.
Covariance Type:            nonrobust   LLR p-value:                     0.000
============================================================================================
                               coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------
Nivel_Académico_Alumnos      2.2756      0.020    115.853      0.000       2.237       2.314
Género                      -1.0818      0.015    -72.051      0.000      -1.111      -1.052
RangoEdad                   -0.7914      0.007   -118.831      0.000      -0.804      -0.778
SNI                          0.7642      0.014     55.948      0.000       0.737       0.791
Researcher                  -0.7813      0.030    -25.978      0.000      -0.840      -0.722
============================================================================================
```

Figure B.35: Recommended Professor Factor / R-studio LR Results Decile/ Balanced Data set B.

```
Nivel_Académico_Alumnos    9.733641
Género                     0.338970
RangoEdad                  0.453195
SNI                        2.147373
Researcher                 0.457813
dtype: float64
```

Figure B.36: Recommended Professor Factor/ Estimated Coefficients of Features /Decile/ Balanced Data set B.

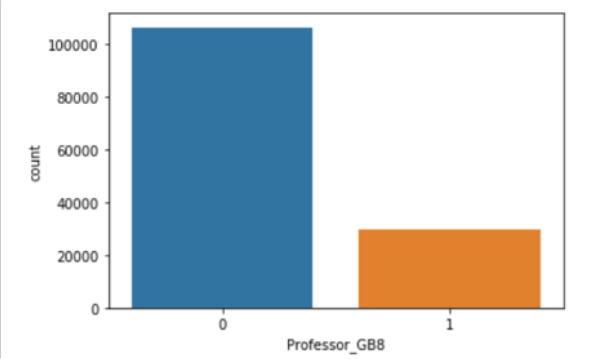

Figure B.37: Recommended Professor Factor / Amount of Alpha (1) and Beta (0) Professors / Decile/ Balanced Data set B

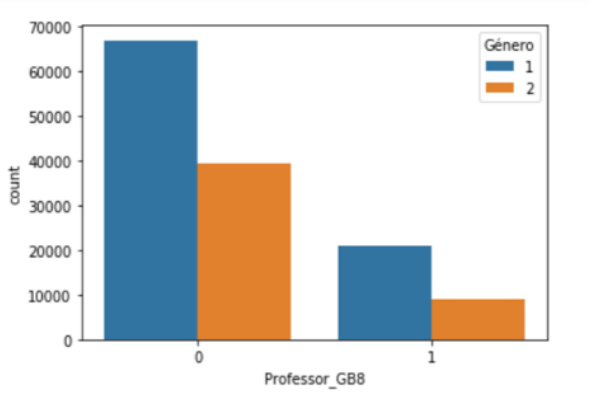

Figure B.38: Recommended Professor Factor / Amount of Alpha (1) and Beta (0) Professors Classified by Gender/ Decile/ Balanced Data set B.

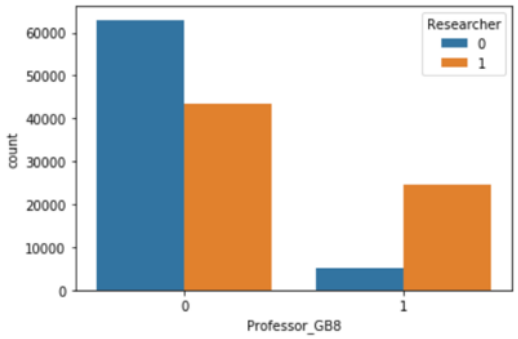

Figure B.39: Recommended Professor Factor/ Researcher (Orange) vs Non Researcher (Blue) Alpha (1) and Beta (0) Professors / Decile/ Balanced Data set B.

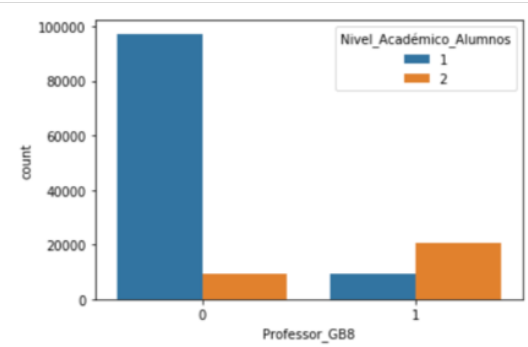

Figure B.40: Recommended Professor Factor / Alpha (1) and Beta (0) Professors by Academic Level Undergraduate (Blue)Graduate (Orange) / Decile/ Balanced Data set B.

	Nomina	Ejercicio_Academico	Genero	RangoEdad	TypeProfessor	Score
1	L00000177	3	M	2	0	9.050000
2	L00000177	4	M	2	0	8.450000
3	L00000177	5	M	2	0	5.570000
4	L00001839	4	F	5	0	9.120000
5	L00001839	5	F	5	0	9.300000
6	L00002399	1	F	4	0	8.320000
7	L00002399	2	F	4	0	9.365000
8	L00002399	4	F	4	0	6.430000
9	L00002399	5	F	4	0	8.880000
10	L00002816	1	M	7	0	8.900000
11	L00002816	2	M	7	0	8.430000
12	L00002816	3	M	8	0	9.230000
13	L00002816	4	M	8	0	9.100000
14	L00002816	5	M	8	0	8.300000
15	L00002980	1	M	7	0	9.070000
16	L00002980	2	M	7	0	8.770000
17	L00002980	3	M	7	0	8.700000
18	L00002980	4	M	7	0	8.430000
19	L00002980	5	M	7	0	7.630000
20	L00003400	1	M	7	1	7.800000
21	L00003400	2	M	7	1	7.833333

Figure B.41: Overlook of Panel Data Set D

Bibliography

[1] Qs world university rankings 2020, Mar 2020.

[2] AKAR, Ö., AND GÜNGÖR, O. Classification of multispectral images using random forest algorithm. *Journal of Geodesy and Geoinformation 1*, 2 (2012), 105–112.

[3] AKELLA, D. Learning together: Kolb's experiential theory and its application. *Journal of Management Organization - J MANAG ORGAN 16* (03 2010), 100–112.

[4] ALLGOOD, S., WALSTAD, W. B., AND SIEGFRIED, J. J. Research on teaching economics to undergraduates. *Journal of Economic Literature 53*, 2 (June 2015), 285–325.

[5] ARGUÍS, R., BOLSAS, A., HERNÁNDEZ, S., AND SALVADOR, M. Programa "aulas felices". *Universidad de Zaragoza. Recuperado de https://www. educacion. navarra. es/documents/27590/203401/Aulas+ felices+ documentación. pdf/3980650d-c22a-48f8-89fc-095acd1faa1b* (2012).

[6] ARTÉS, J., PEDRAJA-CHAPARRO, F., AND DEL MAR SALINAS-JIMÉNEZ, M. Research performance and teaching quality in the spanish higher education system: Evidence from a medium-sized university. *Research policy 46*, 1 (2017), 19–29.

[7] AZEVEDO, A. I. R. L., AND SANTOS, M. F. Kdd, semma and crisp-dm: a parallel overview. *IADS-DM* (2008).

[8] BACON, D. R., PAUL, P., STEWART, K. A., AND MUKHOPADHYAY, K. A new tool for identifying research standards and evaluating research performance. *Journal of Marketing Education 34*, 2 (2012), 194–208.

[9] BASOW, S., AND MONTGOMERY, S. Student ratings and professor self-ratings of college teaching: Effects of gender and divisional affiliation. *Journal of Personnel Evaluation in Education 18* (05 2005), 91–106.

[10] BASOW, S. A., AND MONTGOMERY, S. Student ratings and professor self-ratings of college teaching: Effects of gender and divisional affiliation. *Journal of Personnel Evaluation in Education 18*, 2 (2005), 91–106.

[11] BASOW, S. A., PHELAN, J. E., AND CAPOTOSTO, L. Gender patterns in college students' choices of their best and worst professors. *Psychology of Women Quarterly 30*, 1 (2006), 25–35.

[12] BENTON, S., AND CASHIN, W. Idea paper no. 50: Student ratings of teaching: A summary of research and literature.

[13] BOULESTEIX, A.-L., JANITZA, S., KRUPPA, J., AND KÖNIG, I. R. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2*, 6, 493–507.

[14] BRIDGSTOCK, R., AND TIPPETT, N. *Higher Education and the Future of Graduate Employability*. Edward Elgar Publishing, 2019.

[15] BURKOV, A. *The hundred-page machine learning book*. Andriy Burkov Quebec City, Can., 2019.

[16] CADEZ, S., DIMOVSKI, V., AND ZAMAN GROFF, M. Research, teaching and performance evaluation in academia: the salience of quality. *Studies in Higher Education 42*, 8 (2017), 1455–1473.

[17] CANTU-ORTIZ, F. J. *Research analytics: boosting university productivity and competitiveness through scientometrics*. CRC Press, Taylor Francis Group, 2018.

[18] CANTÚ, F. J., BUSTANI, A., MOLINA, A., AND MOREIRA, H. A knowledge-based development model: the research chair strategy. *Journal of Knowledge Management 13*, 1 (2009), 154–170.

[19] CAO, L., AND ZHANG, C. The evolution of kdd: towards domain-driven data mining. *IJPRAI 21* (06 2007), 677–692.

[20] CAO, L., ZHANG, C., YANG, Q., BELL, D., VLACHOS, M., TANERI, B., KEOGH, E., YU, P. S., ZHONG, N., ASHRAFI, M. Z., TANIAR, D., DUBOSSARSKY, E., AND GRACO, W. Domain-driven, actionable knowledge discovery. *IEEE Intelligent Systems 22*, 4 (July 2007), 78–88, c3.

[21] CAO, L., ZHAO, Y., ZHANG, H., LUO, D., ZHANG, C., AND PARK, E. K. Akd5. *IEEE Transactions on Knowledge and Data Engineering 22*, 9 (Sep. 2010), 1299–1312.

[22] CENTRA, J. A., AND GAUBATZ, N. B. Is there gender bias in student evaluations of teaching? *The journal of higher education 71*, 1 (2000), 17–33.

[23] CHAPMAN, P., CLINTON, J., KERBER, R., KHABAZA, T., REINARTZ, T., SHEARER, C., AND WIRTH, R. Crisp-dm 1.0. step-by-step data mining guide. spss (2000).

[24] CHAPMAN, P., CLINTON, J., KERBER, R., KHABAZA, T., REINARTZ, T., SHEARER, C., AND WIRTH, R. The crisp-dm user guide. In *4th CRISP-DM SIG Workshop in Brussels in March* (1999), vol. 1999.

[25] CHEN, X.-W., AND JEONG, J. C. Enhanced recursive feature elimination. In *Sixth International Conference on Machine Learning and Applications (ICMLA 2007)* (2007), IEEE, pp. 429–435.

[26] CHEN, Y., AND HOSHOWER, L. B. Student evaluation of teaching effectiveness: An assessment of student perception and motivation. *Assessment & evaluation in higher education 28*, 1 (2003), 71–88.

[27] CRAVEN, M. W., AND SHAVLIK, J. W. Using neural networks for data mining. *Future generation computer systems 13*, 2-3 (1997), 211–229.

[28] CRETCHLEY, P. C., EDWARDS, S. L., O'SHEA, P., SHEARD, J., HURST, J., AND BROOKES, W. Research and/or learning and teaching: a study of australian professors' priorities, beliefs and behaviours. *Higher Education Research & Development 33*, 4 (2014), 649–669.

[29] DANIELSON, C. The framework for teaching evaluation instrument: The newest rubric enhancing the links to the common core state standards, with clarity of language for ease of use and scoring. *Copyright material by Charlotte Danielson* (2013).

[30] DEVORE, J. L. *Probability and Statistics for Engineering and the Sciences*. Cengage learning, 2011.

[31] ELTON, L. Research and teaching: Symbiosis or conflict. *Higher Education 15* (05 1986), 299–304.

[32] FAWCETT, T., AND PROVOST, F. *Data Science for Business*. 08 2013.

[33] FOR THE ADVANCEMENT OF ARTIFICIAL INTELLIGENCE (2016)., A. Artificial intelligence and life in 2030. one hundred year study on artificial intelligence.

[34] GARCÍA VÁZQUEZ, N., GUAJARDO-LEAL, B., AND VALENZUELA GONZÁLEZ, J. R. Blueprint de un sistema de innovación educativa en las instituciones de educación superior: el caso del tecnológico de monterrey y su modelo al 2021.

[35] GRACO, W., SEMENOVA, T., AND DUBOSSARSKY, E. Toward knowledge-driven data mining. In *DDDM '07* (2007).

[36] GRIFFIN, B. W. Grading leniency, grade discrepancy, and student ratings of instruction. *Contemporary educational psychology 29*, 4 (2004), 410–425.

[37] GRUBER, T., REPPEL, A., AND VOSS, R. Understanding the characteristics of effective professors: The student's perspective. *Journal of Marketing for Higher Education 20* (07 2010), 175–190.

[38] HATTIE, J., AND MARSH, H. W. The relationship between research and teaching: A meta-analysis. *Review of educational research 66*, 4 (1996), 507–542.

[39] HECKERT, T. M., LATIER, A., RINGWALD-BURTON, A., AND DRAZEN, C. Relations among student effort, perceived class difficulty appropriateness, and student evaluations of teaching: is it possible to" buy" better evaluations through lenient grading? *College Student Journal 40*, 3 (2006), 588–597.

[40] HIESTAND, T. Using pooled model, random model and fixed model multiple regression to measure foreign direct investment in taiwan. *International Business & Economics Research Journal (IBER) 4*, 12 (2005).

[41] HSIAO, C. Panel data analysis—advantages and challenges. *Test 16*, 1 (2007), 1–22.

[42] IACUS, S. M., KING, G., AND PORRO, G. Cem: software for coarsened exact matching.

[43] JUÁREZ, E., CORTÉS, R., AND LABORDE, F. Retos institucionales del modelo tec21 para garantizar el desarrollo de competencias de egreso. In *CIIE Revista del Congreso Internacional de Innovación Educativa* (2015), vol. 1, pp. 47–53.

[44] KING, G. Gary king on simplifying matching methods for causal inference. , 77 (2018), 1–32.

[45] KOHN, J., AND HATFIELD, L. The role of gender in teaching effectiveness ratings of faculty. *Academy of Educational Leadership Journal 10*, 3 (2006), 121.

[46] KÖRTING, T. C4.5 algorithm and multivariate decision trees.

[47] LANGER, A. *Information technology and organizational learning: Managing behavioral change in the digital age, third edition*. 01 2017.

[48] LINDSAY, R., BREEN, R., AND JENKINS, A. Academic research and teaching quality: The views of undergraduate and postgraduate students. *Studies in Higher Education - STUD HIGH EDUC 27* (08 2002), 309–327.

[49] MCAFEE, A., BRYNJOLFSSON, E., DAVENPORT, T. H., PATIL, D., AND BARTON, D. Big data: the management revolution. *Harvard business review 90*, 10 (2012), 60–68.

[50] MCPHERSON, M. A., AND JEWELL, R. T. Leveling the playing field: Should student evaluation scores be adjusted? *Social Science Quarterly 88*, 3 (2007), 868–881.

[51] MCPHERSON, M. A., JEWELL, R. T., AND KIM, M. What determines student evaluation scores? a random effects analysis of undergraduate economics classes. *Eastern economic journal 35*, 1 (2009), 37–51.

[52] MCPHERSON, M. A., JEWELL, R. T., AND KIM, M. What determines student evaluation scores? a random effects analysis of undergraduate economics classes. *Eastern Economic Journal 35*, 1 (2009), 37–51.

[53] OGIER, J. Evaluating the effect of a lecturer's language background on a student rating of teaching form. *Assessment & Evaluation in Higher Education 30*, 5 (2005), 477–488.

[54] ON UNIVERSITY AFFAIRS. TASK FORCE ON RESOURCE ALLOCATION, O. C. *Undergraduate Teaching, Research and Consulting/community Service: What are the Functional Interactions? A Literature Survey: Background Paper*. Ontario Council on University Affairs, Task Force on Resource Allocation, 1994.

[55] PALALI, A., VAN ELK, R., BOLHAAR, J., AND RUD, I. Are good researchers also good teachers? the relationship between research quality and teaching quality. *Economics of Education Review 64* (2018), 40–49.

[56] PROVOST, F., AND FAWCETT, T. *Data Science for Business*. O'Reilly, 2013.

[57] RAMSDEN, P. Describing and explaining research productivity. *Higher Education 28*, 2 (1994), 207–226.

[58] SAHAY, A. The influence of teaching, research and consultancy services on efficiency assessment: Experience from tanzanian universities. *Amity Business Review Vol. 18,* (06 2017).

[59] SANTHANAM, E., AND HICKS, O. Disciplinary, gender and course year influences on student perceptions of teaching: Explorations and implications. *Teaching in Higher Education 7*, 1 (2002), 17–31.

[60] SEDDIGHI, H., LAWLER, K. A., LAWLER, K., AND KATOS, A. V. *Econometrics: A practical approach*. Psychology Press, 2000.

[61] SMITH, S. W., YOO, J. H., FARR, A. C., SALMON, C. T., AND MILLER, V. D. The influence of student sex and instructor sex on student ratings of instructors: Results from a college of communication. *Women's Studies in Communication 30*, 1 (2007), 64–77.

[62] SMITH, S. W., YOO, J. H., FARR, A. C., SALMON, C. T., AND MILLER, V. D. The influence of student sex and instructor sex on student ratings of instructors: Results from a college of communication. *Women's Studies in Communication 30*, 1 (2007), 64–77.

[63] SPOOREN, P. On the credibility of the judge. a cross-classified multilevel analysis on student evaluations of teaching. *Studies In Educational Evaluation 36* (12 2010), 121–131.

[64] SPOOREN, P., BROCKX, B., AND MORTELMANS, D. On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research 83*, 4 (2013), 598–642.

[65] STACK, S. Research productivity and student evaluation of teaching in social science classes: A research note. *Research in Higher Education 44* (10 2003), 539–556.

[66] TING, K.-F. A multilevel perspective on student ratings of instruction: Lessons from the chinese experience. *Research in Higher Education 41*, 5 (2000), 637–661.

[67] TING, K.-F. A multilevel perspective on student ratings of instruction: Lessons from the chinese experience. *Research in Higher Education 41* (01 2000), 637–661.

[68] TSINIDOU, M., GEROGIANNIS, V. C., AND FITSILIS, P. Evaluation of the factors that determine quality in higher education: An empirical study.

[69] UZ ZAMAN, M. Q. *Review of the academic evidence on the relationship between teaching and research in higher education*. Department for Education and Skills London, 2004.

[70] WIKIPEDIA CONTRIBUTORS. Cross-industry standard process for data mining — Wikipedia, the free encyclopedia, 2019. [Online; accessed 2-March-2020].

[71] WIRTH, R., AND HIPP, J. Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (2000), Springer-Verlag London, UK, pp. 29–39.

Curriculum Vitae

Mario Daniel Chávez López was born in Mexicali, Baja California, México on February 8, 1996. He earned an Industrial Engineering with option in Strategic Manufacturing Management Degree from CETYS University in June 2018. He was accepted in the Masters of Computer Science Program by Tecnologico de Monterrey in 2018. He expects to graduate in June 2020.

This document was typed in using LaTeX 2_ε[a] by Mario Daniel Chávez López.

[a]The style file `phdThesisFormat.sty` used to set up this thesis was prepared by the Center of Intelligent Systems of the Instituto Tecnológico y de Estudios Superiores de Monterrey, Monterrey Campus