Instituto Tecnologico y de Estudios Superiores de Monterrey

Monterrey Campus

School of Engineering and Sciences

TECNOLÓGICO DE MONTERREY®

Analyzing Fan Avidity for Soccer Prediction

A thesis presented by

Ana Clarissa Miranda Peña

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, May, 2022

Instituto Tecnologico y de Estudios Superiores de Monterrey

Campus Monterrey

The committee members, hereby, certify that have read the thesis presented by Ana Clarissa Miranda Peña and that it is fully adequate in scope and quality as a partial requirement for the degree of Master of Science in Computer Sciences.

Dr. Miguel González Mendoza
Tecnológico de Monterrey
Principal Advisor

Dr. Laura Hervert-Escobar
Tecnológico de Monterrey
Co-Advisor

Dr. Neil Hernandez Gress
Tecnológico de Monterrey
Committee Member

Dr. Joanna Alvarado Uribe
Tecnológico de Monterrey
Committee Member

Dr. Rubén Morales Menéndez
Associate Dean of Graduate Studies
School of Engineering and Sciences

Instituto Tecnologico y de Estudios Superiores de Monterrey

Campus Monterrey

The committee members, hereby, certify that have read the thesis presented by Ana Clarissa Miranda Peña and that it is fully adequate in scope and quality as a partial requirement for the degree of Master of Science in Computer Sciences.

Dr. Miguel González Mendoza
Tecnológico de Monterrey
Principal Advisor

Dr. Laura Hervert-Escobar
Tecnológico de Monterrey
Co-Advisor

Dr. Francisco Javier Cantú Ortiz
Tecnológico de Monterrey
Committee Member

Dr. Héctor Gibrán Ceballos Cancino
Tecnológico de Monterrey
Committee Member

Dr. Rubén Morales Menéndez
Associate Dean of Graduate Studies
School of Engineering and Sciences

Declaration of Authorship

I, Ana Clarissa Miranda Peña, declare that this thesis titled, Analyzing Fan Avidity for Soccer Prediction and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Ana Clarissa Miranda Peña
Monterrey, Nuevo León, May, 2022

Dedication

Thank you to my family for their unconditional love, for always encouraging me, supporting my decisions, and motivating me to aim for more.

Acknowledgements

Thank you mom for your effort in giving me the best quality education, and my brother who aided me in difficult times in every aspect of life, and to my father from whom I have always felt his blessing.

I want to thank my advisors Dr. Miguel and Dr. Laura, also Dr. Neil for your trust in letting me build this project from scratch and combine my passion for computer science with soccer. Also to the members of this committee Dr. Joanna, Dr. Cantú and Dr. Ceballos who every semester made thoughtful recommendations, guided me, and provided me with concepts that I later applied on this thesis, you all helped me to accomplish my masters' goals. We are proof of the relevance of co-working between the Data Science and the Machine Learning paths.

I will thank my friends, of course, and hope that I am not missing anyone. One of the reasons I stayed in Monterrey to continue my studies was because during my 4.5 years of undergrad, Monterrey became my home, and I got to establish friendships with different groups. First, my local friends Sofia Landeros and Cynthia Castillo, who accompanied me in numerous adventures. In the intersection between local and bachelors' friends Valeria Rocha, Laura Santacruz, Flor Barbosa, Fabiana Serangelli, German Villacorta, Juan Manuel and Diego Contreras who opened their houses and hearts during this time. In the intersection between non-local and bachelors' friends Ali Ghahraei, Andrea Chacon, Manuel Torres, Enrique Garcia, Juan José López and Sofia Cedillo. In the non-local subset I will thank Diana Jacobo, Monserrat Contreras and family, Carolina Cid, Kitzia Rodriguez thank you all for being my sisters. To the friends I made in the exchange program and made me feel like I was in Mexico even though we were not Luis Erick Zul, Sergio Fuentes, Erika Llano, Marko Vasic, Elias Aouad and Constanza Vega. To my hometown friends who helped me a lot in this pandemic Laura Dávila, Valeria de la Serna, Melisa Ramírez, Ximena Pérez, Jessica Esparza, Daniel Esparza, and Fernanda Vázquez and family. And finally to the classmates that later became my friends Jorge Ayala, Gregorio Reyes, Jessica Salinas and Carmina Pérez.

I would like to express my deepest gratitude to Tecnologico de Monterrey, which through its Academic Scholarship Program for Graduate Students provided technical and financial support for the development of this research. In addition, I am grateful to CONACyT for the financial support awarded through the National Scholarship for PNPC and SNI programs designed for promoting quality research and closing the existing gap between industry and academia. Finally, I would like to highlight the research development at my hometown Aguascalientes and the support that was given by IDSCEA.

Analyzing Fan Avidity for Soccer Prediction
by
Ana Clarissa Miranda Peña

Abstract

Beyond being a sport, soccer has built up communities. Fans showing interest, involvement, passion and loyalty to a particular team, something known as Fan Avidity, have strengthen the sport business market. Social Networks have made incredibly easy to identify fans'commitment and expertise. Among the corpus of sport analysis, plenty of posts with a well substantiated opinion on team's performance and reliability are wasted. Based on graph theory, social networks can be seen as a set of interconnected users with a weighted influence on its edges. Evaluating the spread influence from fans' posts retrieved from Twitter could serve as a metric for identifying fans' intensity, if adding sentiment classification, then it is possible to score Fan Avidity. Previous work attempts to engineer new key performance indicators or apply machine learning techniques for identifying the best existing indicators, however, there is limited research on sentiment analysis. In order to achieve the Master's Degree in Computer Science, this thesis aims to strengthen a machine learning model that applies polarity and sentiment analysis on tweets, as well as discovering factors thought to be relevant on a soccer match. The final goal is to achieve a flexible mechanism which automatizes the process of gathering data before a match, with the main objective of quantifying credit on fans' sentiment along with historical factors, while evaluating soccer prediction. The left alone sentiments' model could accomplish independence from the type of tournament, league or even sport.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

1.1 Problem

Sports are unpredictable and constantly changing over time. Even though, sports analytics are becoming more productive in generating datasets based on statistics. Obtaining a set of identical events for detecting patterns on the state of the game is a task difficult to achieve since teams are constantly displaced between divisions, as well as players and staff. Also, in football, common statistics such as possession percentage and pass accuracy could lead to nonrepresentative conclusions when compared to the final score. However, many factors outside the field which are not explicitly extracted from statistics may relate to the results of the upcoming match or affect the rest of the season. Thus, these factors could be contextualize by analyzing fans in Social Networks applying Information Retrieval and Natural Language Processing techniques.

Chen [5], who applied a Convolutional Neural Network for predicting match results by using the player ability indexes as a principal feature, showed the gap between the computer science field and soccer experts. Neural Networks can turn into a black box when providing insufficient guidance on how the model is learning and making it harder to interpret the results from a technical point of view, as well as requiring large sets of data.

Machine learning along with Data Science could provide a better insight into the problem. For example, the Prediction learning model for soccer matches [15] analyzes 200,000 matches gathered from 52 leagues. The Bayesian Model first ranks the team in a ratio of 80/20 where 80% is the value given to the current season and 20% on the previous season, and second by the historical probability of the given two teams to win, lose or draw. Some limitations on Data Science are overfitting the behavior of the data and punishing generalization in other disciplines.

Herold [14] reviews several studies that apply Machine Learning for improving the attacking play in football. This summary is oriented either on tracking data or consuming logs from game events, both of them aim to increment tactical knowledge and performance through Machine Learning. Herold criticizes the lack of subjectivity when rating the pass and trajectory quality as good or bad. He also highlights the absence of divergence on those methodologies while excluding psychological and contextual factors, for example, team adaptability and communication.

Chassy [4], from a psychology perspective, proposes the concept of self-organization as the way dynamics at the local level determine cohesion and coordination at the system level. Chassy defines team performance as how well the organization is working as a team. The formula states that making frequent and accurate passes will acquire possession to generate shooting opportunities. The correlation between passing density and precision fit an R^2 of 0.99. The value of this model relates to finding common patterns for addressing different sports.

1.2 Motivation

Sports in the contemporary era are characterized by their relationship with new environments and technologies. Also, they promote commercial and economic development through marketing and new businesses. Sports have become a distraction and a lifestyle to contemporary society since they let society find harmony in daily dynamics as a way of entertainment [13].

Lovemarks development through sports marketing [7] declares football as a sport for the multitude because it is accessible to everybody. Bloomberg supports this argument in World Cup Russia 2018 where 4 out of 10 sports fans considered themselves soccer fans, making this sport the most popular in the world. During that year, FIFA had a 4.64 billion dollars revenue, FIFA doubled its earnings when compared to Brazil 2014, where revenue was about 2.1 billion dollars [11].

A very attractive aspect in contemporary sports is fanaticism, which is seen as a cultural niche construction that has solidified the concept of fan avidity. Fan avidity means to the consumer a subgroup that shares beliefs and values, jargon, rituals, and symbolism to express themselves [8]. Fan avidity is part of today's sociocultural structure by organizing hierarchically the closeness between a fan and the sports institution. Another concept for fan avidity is the given intensity of a fan to follow and encourage, which justifies the reason a fan watches a particular team [50].

Football fans have proven relevance in the business world, they have valued the soccer industry at least as a 22.8 billion dollars industry when considering fan engagement [27]. The Union of European Football Associations (UEFA) Financial Report 2018/19 [51] showed incomes around 3,857,191 (000 euros), where 3,787,405 belong to media and commercial rights, and the rest is due to tickets and hospitality. This amount does not include the day of the event, nor the team's license rights. The same report showed a positive cumulative profit of 1.58 billion dollars for European clubs once they added sponsorship, broadcasting, gate receipts, etc.

On social media, the most popular account for a league is England's Premier League with 32 million followers on Instagram and 41 million likes on Facebook, followed by Spain's LaLiga with 23 million followers and 53 million likes. The amount of followers indicates a football club's popularity, however the number of interactions per follower measures engagement. Even though FC Barcelona and Real Madrid have over 200 million followers, local teams from the south and southeast Europe have shown engagement with more than 10 million followers with 10 interactions each, Liverpool FC and Juventus belong to this group.

Machine learning is an expensive task, where precision relates severely to the amount and quality of the input features, specifically when finding repeatable events in soccer. From a computer science point of view, data scientists lack insight on soccer factors, and what is learned from soccer statistics may vary in different leagues, tournaments, and other sports. Extracting relevant opinions from soccer experts in Social Networks by applying Graph Theory could clarify model outcomes, as well, reinforce behavioral prediction techniques through sentiment analysis, and consolidate independence in other sports or fields of study.

1.3 Hypothesis

A model that reinforces sports statistics with factors outside the field such as sentiment, polarity, and fan engagement, from a graph representation of Social Networks, will generate a higher prediction than the baseline of 50% accuracy at minute zero.

- Based on Graph Theory, could users' sentiment be evaluated as a metric to score Fan avidity?
- Is Fan avidity a determinant feature when predicting results previous to a match?
- What percentage of successful predictions are obtained, when considering fan avidity as a score of factors as sentiment, polarity, and fan engagement?
- How is the percentage of correctness compared to a model without considering Fan avidity?

1.4 Objectives

The general objective of this work is to improve the accuracy of soccer predictions before a match. This is done by identifying the key elements of performance for a team and contextualizing the match through Fan avidity, as well, benchmarking the statistics model against the sentiment model.

- Automate search query gathering from Twitter Standard Search API with a rate of 100 tweets and a window of 15 minutes.
- Automate fixture statistics gathering in API Football with a rate of 5000 requests per day.
- Develop a feature engineering mechanism that weights the team's ranking with a constant operation cost; calculating averages and scoring performance.
- Develop a model that fits soccer key indicators for match predictions.
- Develop a model that evaluates centrality on Social Network users applying centrality measures.
- Develop a model that predicts a match result using features like fan relevance, sentiment, and key soccer indicators with an accuracy above the current research on statistics.

Chapter 2

State of the Art

In the early 2000s decade, Sports Analytics has raised as a discipline in Big Data highlighting huge opportunities on Data Science applications. Specifically in Soccer, analytics has being a delayed development, where much of the robust models are found in the 2010s decade. The literature review in this thesis categorizes Soccer Analytics for match prediction under the following items:

- Linear models on performance: oriented on decoding team's performance through regressing some success metric, such as Expected Goals (xG) or goal differential.

- Linear models on betting odds: these models attempt to shorten the error when regressing closing odds to estimate winners.

- Machine learning classifiers: here some variety of models trained with contextual information, for example Sentiment Analysis, and performance team's features are used to classify a match as a win, draw or loss.

Table 2.1 summarizes on how the existing literature can be classify according to the terms designated above, while the last row explains how the work in this thesis is positioned.

2.1 Predicting Match Outcomes Using Statistics

2.1.1 Regression Analysis (RA) on Expected Goals

Herbinet [12] performed RA on shot-based features to generate an expected goals metric, then the difference between the actual number of expected goals and the expected amount of goals was calculated. Herbinet applied RA on the final score and compared those results to classification models for the home win, draw or away win. The data was collected from 5 top European leagues (English Premier League, French Ligue 1, German Bundesliga, Spanish Liga, Italian Serie A) between seasons 2014/15 and 2015/2016 retrieving in a total of 3800 matches. The highest accuracy on the RA model following the Poisson distribution was 44.6%, while the classification model with linear SVC achieved 51.1%.

Table 2.1: Classification on match prediction literature.

	Sport	**Linear Models**	**Machine Learning**	**Team's Performance**	**Betting Odds**	**Contextual Information**
Analysis of the winning probability and the scoring actions in the MLS	Soccer		-	-		
Predicting Football Results Using Machine Learning Techniques	Soccer		-	-		
Quantifying the relation between performance and success in soccer	Soccer	-	-	-		
PlayeRank: Data-driven Performance Evaluation and Player Ranking in Soccer via a Machine Learning	Soccer		-	-		
Who will win it? An In-game Win Probability Model for Football	Soccer		-	-		
The harsh rule of the goals: data-driven performance indicators for football teams	Soccer	-		-		
A profitable model for predicting the over/under market in football	Soccer		-	-	-	
Home-field advantage and biased prediction markets in English soccer	Soccer	-			-	
Exploiting sports-betting market using machine learning	Basketball		-	-	-	
The Economics of Football	Soccer	-			-	
Beating the bookies with their own numbers and how the online sports betting market is rigged	Soccer	-			-	
Mathletics: How Gamblers, Managers and Sports Enthusiasts Use Mathematics in Baseball, Basketball and Football	American Football	-			-	
Predicting wins and spread in the Premier League using a sentiment analysis of twitter	Soccer		-			-
The Wisdom of the Silent Crowd: Predicting the Match Results of World Cup 2018 through Twitter	Soccer		-			-
Analizing Fan Avidity for Soccer Prediction	Soccer		-	-		-

2.1.2 Winning Probability per Minute

Analysis of the winning probability and the scoring actions in the American professional soccer championship [2] finds the conditional probability of winning a match given a minute. It studied 680 matches in the Major League Soccer seasons 2015 and 2016, by applying PCA it found relevant components such as counter-attacks and crosses, and performing Hierarchical Clustering showed the distance between teams. Finally, the conclusion, in this league, was that the last 15 minutes are the least reversible on the final score. Similarly, Robberechts [40] shows a statistical metric to calculate sports team's likelihood of winning at any given point in a game. Some of the features used are the current score and the difference in the Elo Ratings of the teams. The findings for the contextual features are that winning duels affect negatively the win probability and red cards decrease a team's scoring rate. Model's performance improves as the game progresses. However, in pre-game, accuracy is lower than 50%.

2.1.3 Simulating Rankings

Pappalardo and Cintia performed several studies applying Machine Learning in Football, for example, Player Rank [35] uses soccer logs from 18 competitions in a four-season window. The first phase is an exploratory study for weighting aggregated vector features in a match using classification methods, and learning in an unsupervised way to detect roles on players. Later on, the rating phase computes a scalar between weights and the match's features performed by the player, the final ranking computes a role-based performance rating of a player given a match. On The harsh rule of the goals: data-driven performance indicator for football teams [6] the authors tried to simulate four major European Championships (German, English, Spanish and Italian league) using observational data of 1500 football games. They engineer features such as Pezzali score, and relative performance, this is the difference between teams statistics. The results showed that the highest Spearman correlation between the simulated ranking and the real ranking occurred on the German league with 0.89, while the least significance happened on the Spanish League with a correlation of 0.66. These two works are mixed in quantifying the relation between performance and success in soccer [34] taking advantage of Player Rankings at a team level, this research correlates competition's final rankings with its typical performance. The results are evaluated as a classification task on match outcomes, highest accuracy occurs on the french Ligue 1 with 80% but low correlation of 56%, Italian Serie A showed more stable results with a correlation of 74% and accuracy of 70% however those are not simulated evaluations.

A linear model for ranking soccer teams in the Italian soccer 2016-17 championship is proposed by Peretti [37] arguing that the computation of the dominant eigenvalue and its corresponding eigenvector result in the final ordering of the tournament. The matrix from which the vectors are calculated is the square matrix of the coefficients a_{ij}. a_{ij} represents a non-negative number showing the results of the game between team i and team j. The author also synthesized a teams score. The equation 2.1 of the **Score for the ith participant**.

$$s_i = \frac{1}{n_i} \sum_{j=1}^{N} a_{ij} r_j \tag{2.1}$$

Where $r_j$indicates the strength of the team j, and n_i the number of games played by team i.

Predicting the Dutch Football Competition Using Public Data: A Machine Learning Approach [49] uses some interesting candidate features such as team was in a lower league the previous year, number of matches coached by the current coach, the team hired a new coach during the previous month, top-scorer suspended or injured, top-assist suspended or injured, days since the previous match, etc. This research extended an evaluation following a season sequence for Dutch Eredivisie Competition 2000-2013. The highest results were achieved when performing PCA with a Naive Bayes algorithm or Multilayer Perceptron model, however, no more than 54.5% accuracy was reached. A McNemar's test was conducted but did not prove significance in achieving highest accuracy when using a hybrid model that included betting odds features.

2.1.4 Based on Betting Odds

Elaad [10] adapted the Standard Mincer and Zarnowitz Forecast Efficiency Evaluation Framework to prove the incorporation of the home-field advantage in the odds on English football leagues. The author assumes that the prediction error must be equal to the betting market margin. In the end, it is concluded that lower divisions over-predict home wins, and away wins are under-predicted. Beating the bookies with their own numbers [19] studies over 479,440 football matches odds and calculates a payoff given the consensus probability of the match, they adjusted this probability by finding from linear regression the estimated α bias per home, draw, and away win, the strategy was to bet whenever the payoff was greater than zero, they show off an accuracy as high as 44% and 3.5% return. Exploiting sports-betting market using machine learning [17] is so far a novelty that creates its predictions through Machine Learning and correlates those with odds, however, it is applied on basketball only, it also proposes a betting strategy between profit expectation and profit variance. On football, Wheatcroft [53] uses linear regression to predict over/under market in football proposing GAP ratings and maximum odds implied probabilities.

2.2 Measuring Football Conversations

2.2.1 Social Networks From a Graph Perspective

Some research studies, as the one developed by Yan, [55] evaluate the influence of users, represented as nodes, on other entities under the Social Network Analysis around 2017 Champions League. This is performed by calculating a value for each eigenvector by scoring the weight and the importance of the nodes it is connected to. This paper also adds betweenness centrality, which obtains the shortest paths and finds the most repetitive nodes, so that the most influential elements in the network are identified.

Riquelme [39] proposes two new centralization measures for evaluating networks. The model graph is a compound of labels representing the resistance of the actors to be influenced, and the weight of the edges is the power of influence from one actor to another.

The equation 2.2 of the **Node activation** is:

$$\textstyle\sum_{j \in F_t(X)} W_{ij} \geq f(i) \tag{2.2}$$

The activation occurs when the sum of the weight of activated nodes connected to i, in the set of $F_t(X)$ is greater or equal to i's resistance denoted as $f(i)$.

The equation 2.3 of the **Spread of Influence X** is:

$$F(X) = \textstyle\bigcup_{t=0}^{k} F_t(X) = F_0(X) \cup ... \cup F_k(X) \tag{2.3}$$

where t denotes the current spread level of X, and X is an initial activation set.

The first measure considered is called Linear Threshold Centrality and represents how much an actor i can spread his influence within a network, this by convincing his immediate neighbors.

The equation 2.4 of the **Linear Threshold Centrality** is:

$$LTR(i) = \frac{|F(\{i\} \cup neighbors(i))|}{n} \tag{2.4}$$

The second measure is Linear Threshold Centralization, this defines how centralized the network is, by finding a k-core which is the maximal subgraph $C(G)$ such that every vertex has a degree at least k.

The equation 2.5 of the **Linear Threshold Centralization** is:

$$LTC(G) = \frac{|F(C(\hat{G}))|}{n} \tag{2.5}$$

This relation shows that elements outside the core are easier to be influenced.

Kim [21] proposed a formula to address opportunity based on satisfying the fan's requirements. Korean National Football team's comments on the match against Uzbekistan on FIFA World Cup 2018 qualifications were ranked using TF-IDF, which reflects the relevance a word has in the document. After that, a clustering algorithm, such as K-Means, was implemented for topic modeling, once the topic was known, it was assigned a satisfaction value given by the Delphi Method.

The equation 2.6 of the **Delphi satisfaction expression** is:

$$TS_i = \frac{\sum_{j=1}^{j_i} CSi,j}{J_i} \tag{2.6}$$

$CS_{i,j}$ satisfaction level of the *j-th* post in the *i-th* topic, TS_i average satisfaction for the *i-th* topic, and J_i total number of post in the *i-th* topic.

2.2.2 Sentiment Analysis Applications

Dharmarajan [9] applied the Multinomial Naive Bayes Algorithm into two main classifiers. The first one is oriented towards an objective tone, this model is trained with a self-made dataset of well-trusted sources, and the second one is a subjectivity classifier that can either label text as positive or negative. This last one achieved 79,50% accuracy over 32,000 instances, while the first one obtained 77,45% when trained with 86,000 records.

A similar approach is presented by Sajjad [43] who uses the SentiWordNet for acquiring corpus-based and context-based representations for sentiment analysis, where word classification is considered either positive, negative or objective, into a one-dimensional vector. This study compares text feature transformations as unigrams and bigrams by clustering its semantic and statistical similarity. Later on, features are ranked, and those with the lowest scores are eliminated. The author concludes that SVM algorithm obtains better results on the reduced dimensional vector, but when keeping diversity from features the Naive Bayes model has a better performance.

Ljajic [24] proposes a sentiment score by quantifying the logarithmic difference of terms in positive and negative sports comments. Again, sentiment classification is seen as a supervised task that requires creating a domain-specific dictionary and assigning a tag as positive or negative for each of the terms. The author proposed the principle of logarithmic proportion TF-IDF as a labeling mechanism.

The equation 2.7 of the **Polarity compute using TF-IDF** is:

$$tfidf_p = (1 + tf_p) * log_{10}\left(\frac{N_p}{N_{t,p}}\right) \tag{2.7}$$

Where $tfidf_p$ is the polarity of the term in positive comments, tf_p is the term frequency in positive comments, N_p is the number of positive documents, and $N_{t,p}$ is the number of positive documents with term t. The same procedure will be followed for negative ratio $tfidf_n$, where the larger term will be set as a tag.

A methodology, for setting terms as stop words, is also concluded on this research, by finding boundaries due to the logarithmic difference of the terms, on the paper boundaries were set when accuracy stopped improving.

The equation 2.8 of the **Logarithmic difference of term** is:

$$DifLog_t = log_{10}\left(\frac{tfidf_p + 0.001}{tfidf_n + 0.001}\right) \tag{2.8}$$

Jai-Andaloussi [18] aims to summarize highlights in soccer events by analyzing tweets and scoring text sentiment they recommend the deep learning method implemented in Stanford NLP which categorizes comments from 0 being very negative to 4 being very positive. However, as the intention is to obtain the most relevant tweets, the moving-threshold burst detection technique is used.

The equation 2.9 of the **Moving threshold** is:

$$MT_i = \alpha * (mean_i + x * std_i) \tag{2.9}$$

Given l as the length of the sliding window at time t_i, $N(l_1)$ to $N(l_i)$ where N is the number of tweets, the mean and standard deviation at the time i can be calculated. α is the relaxation parameter, and x is a constant between 1.5 and 2.0. A highlight is defined as $N(l_i) > MT_i$.

2.2.3 Sentiment Analysis for Prediction

Schumaker [44] applies sentiment analysis based on a combination of 8 models using either polarity, such as positive, negative, and neutral, and tone such as the objective, subjective, and neutral. This research has an odds-based approach that gathers an odds-maker's match balance sheet on demand of the wagers. The sentiment is calculated by normalizing a specific data model against tweets for a particular club and match.

The equation 2.10 of the **Normalize polarity** is:

$$max(\frac{\sum Tweets|Model_n, Club_1, Match_m}{\sum Club_1, Match_m}, \frac{\sum Tweets|Model_n, Club_2, Match_m}{\sum Club_2, Match_m}) \tag{2.10}$$

When models tested with negative polarity were higher, they could predict a potential loss, whereas models of positive polarity as a possible win.

During World Cup 2018, Talha [48] constructed a database containing 38,371,358 tweets and 7,876,519 unique users, 9 different machine learning models were trained with the 48 matches on the group phase, and tested to predict round 16, and so on. The features considered for this model are detailed information about the user (number of followers, location, likes count, tweets counts, etc) and the tweet (is it a retweet, reply to a user, retweet count, like count, etc), the highest accuracy obtained was 81.25% when using a Multilayer Perceptron algorithm with 30,000 epochs.

2.3 Machine Learning Definitions

2.3.1 Linear Regression

Linear Regression [26] attempts to fit a bi-dimensional space(x, y) in the form of a line by minimizing the summed squared difference between x and y. It traduces the Cartesian plane equation y = mx + b to the expression above. The equation 2.11 of the **Cartesian formulation**.

$$y = \beta X + \beta_0 \tag{2.11}$$

Where X is the matrix of predictors, β is a matrix of coefficients and β_0 is the constant called bias.

R-Square

The coefficient of determination R^2 is the criteria used to know how well is the regression model fit to the data, by assessing the quality of the prediction given a confidence interval, a popular hypothesis test for regression is the upper-tailed F test Analysis of Variance (ANOVA).

The equation 2.12 of the **R-Square**.

$$R^2 = \frac{SS_{reg}}{SS_{total}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{SS_{res}}{SS_{total}} \tag{2.12}$$

Since R^2 represents the proportion of total variation explained by the predictors, where SS_{res} is the residual of the regression. When R^2 is closer to 1, then y_i and $\hat{y}_i$ are very similar and SS_{res} is close to zero, then the regression achieved a good fit.

Mean Squared Error

Feature selection should avoid underfitting, this means to exclude unnecessary features and leading to bias estimation, while discarding overfitting which includes unnecessary features and increases variance. Mean Squared Error is the criteria used to balance this bias-variance tradeoff. The equation 2.13 of the **Mean squared error**.

$$MSE(\hat{y}) = E(\hat{y} - y)^2 \tag{2.13}$$

MSE can also be explained as the sum of the variance and the squared bias of the estimator, smaller MSEs may achieved the balance explained earlier. The equation 2.14 of the **MSE in relation with bias-variance tradeoff**.

$$\begin{aligned} MSE(\hat{y}) &= E\left\{\hat{y} - E(\hat{y})\right\}^2 + \left\{E(\hat{y}) - y\right\}^2 \\ &= Var(\hat{y}) + bias^2(\hat{y}) \end{aligned} \tag{2.14}$$

2.3.2 Support Vector Machines

Linear SVM [25] is a cutting plane algorithm useful in multi-class problems and popular when classifying high dimensional patterns. In a binary classification problem a support line can be defined as: The equation 2.15 of the **Support line on binary classification**.

$$W^t X + b = 1 \tag{2.15}$$

Where values greater or equal than one satisfies the positive class, while values less or equal than minus one the negative class. W decides the orientation of the decision boundary and b is the threshold between the boundary and the center. The objective is to maximize the margin given by $\frac{2}{\|W\|}$ which is the center distance between the planes, equivalent to minimizing $\frac{\|W\|^2}{2}$. The next are some popular metrics for evaluating classification models.

Accuracy

The sum of all predictions divided by the total number of predictions quoted [52].

Precision

Correct positive predictions. From predicted positive, how many were predicted correctly?

Recall

Or True Positive Rate (TPR). From all positive events, how many were predicted correctly?

F1 Score

Low precision relates to high false positives and low recall to high false negatives. F1 Score is also known as the harmonic mean of precision and recall and summarizes the evaluation of the algorithm. The equation 2.16 of the **F1 Score**.

$$F1Score = \frac{2*(precision*recall)}{(precision+recall)} \tag{2.16}$$

2.3.3 Hierarchical Clustering

Hierarchical clustering [25] is part of the Linkage-based clustering which defines the similarity between two objects as the average similarity between objects linked with them. By removing edges connecting the vertices of different communities, it's possible to isolate those communities.

Each vertex is chosen randomly with a probability from vertex i to vertex j, where A is the Adjacency matrix, D is the Diagonal matrix and t is the length of the random walk.

The equation 2.17 of the **Transition probability in a Markov Chain sequence**.

$$P_{ij} = \frac{A_{ij}}{D_{ii}} \tag{2.17}$$

Clusters should be far from each other and closer in its group, the cophenetic distance is the distance between the observation clusters and it should be maximized.

The equation 2.18 of the **Distance between two vertices**.

$$r_{ij} = \left\| D^{-\frac{1}{2}} P_i^t - D^{-\frac{1}{2}} P_j^t \right\| \tag{2.18}$$

The average distance is computed for all points i and j where $\|C_1\|$ and $\|C_2\|$ are cardinalities of clusters C_1 and C_2 respectively, also called the UPGMA algorithm [32].

The equation 2.19 of the **Distance between two communities** C_1 **and** C_2.

$$r_{C_1C_2} = \sum_{ij} \frac{\left\| D^{-\frac{1}{2}} P_i^t - D^{-\frac{1}{2}} P_j^t \right\|}{(\|C_1\| * \|C_2\|)} \tag{2.19}$$

On each step k, the two communities whose mean σ_k between each of the vertex k minimizes the squared distance are chosen. For every pair of adjacent communities with the lowest variation $\Delta\sigma$ are merged.

The equation 2.20 of the **Variation of the squared distances between adjacent communities** C_1 **and** C_2.

$$\Delta\sigma\left(C_1 C_2\right) = \frac{1}{n}\left(\sum_{i \varepsilon C_3} r_{iC_3}^2 - \sum_{i \varepsilon C_1} r_{iC_1}^2 - \sum_{i \varepsilon C_2} r_{iC_2}^2\right) \tag{2.20}$$

Chapter 3

Methodology

During this chapter the aim is to find the key features that could generalize as much as possible the behavior of a given match at pre-game, and propose the Machine Learning models for predicting future matches. First Section 3.1 introduces FootballMLP as the model that gathers and analyzes fixture's statistics. Second Section 3.2 attempts to measure Fan Avidity in Social Networks from a Graph Theory perspective. Further, chapter 4 will show the results of Section 3.1 across different leagues in different periods of time in Section 4.1, while the results of Section 3.2 in Section 4.2 are oriented to evaluate centrality measures as relevant contextual features for the match prediction. Section 4.3 compares these models in terms of classification accuracy while also measuring Ranking Probability Score (RPS) which is commonly used in the literature of match classification.

3.1 FootballMLP

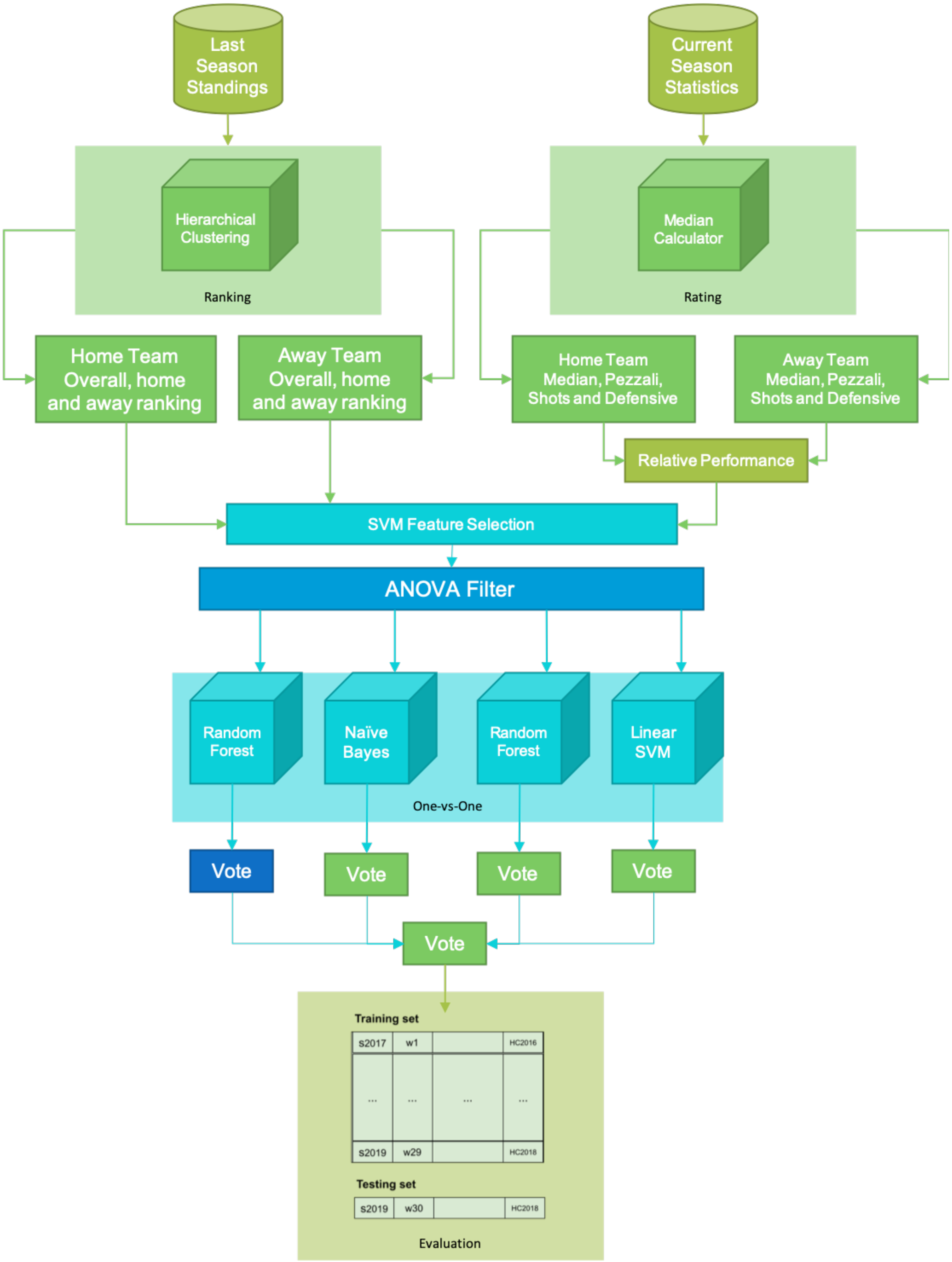
FootballMLP is a framework that lets football users forecast past and future matches [29]. The first phase consists of clustering seasonal team standings, then those clusters are mapped with the median of the current season's statistics on home and away teams. The third phase generates a prediction by majority voting as shown in Figure 3.1.

FootballMLP developed a client library [28] which wraps API-Football that covers 694 football leagues' statistics and standings, detailed coverage is provided on its website [47]. Figure 3.2 shows the relationship between the entities on the library.

First, feature selection is performed on Spain's LaLiga 1,520 fixtures from seasons 2016 to 2019. Second, the model is evaluated on the seven professional leagues listed above on 5,356 fixtures for season 2018-2020.

3.1.1 Ranking

Previous literature has proven the importance of scoring team's strength. Authors as Peña [2] and Cintia [6] had studied team's positions in the league, the first one by showing hierarchy in the MLS, the second one by simulating the LaLiga's final ranks. Given a combination of both, FootballMLP attempts to rate in an unsupervised manner how its a team situated against other teams by using Hierarchical Clustering on the standings taken from API-Football.



Training set

s2017	w1		HC2016
...	...	...	...
s2019	w29		HC2018

Testing set

s2019	w30		HC2018

Figure 3.1: FootballMLP full pipeline.

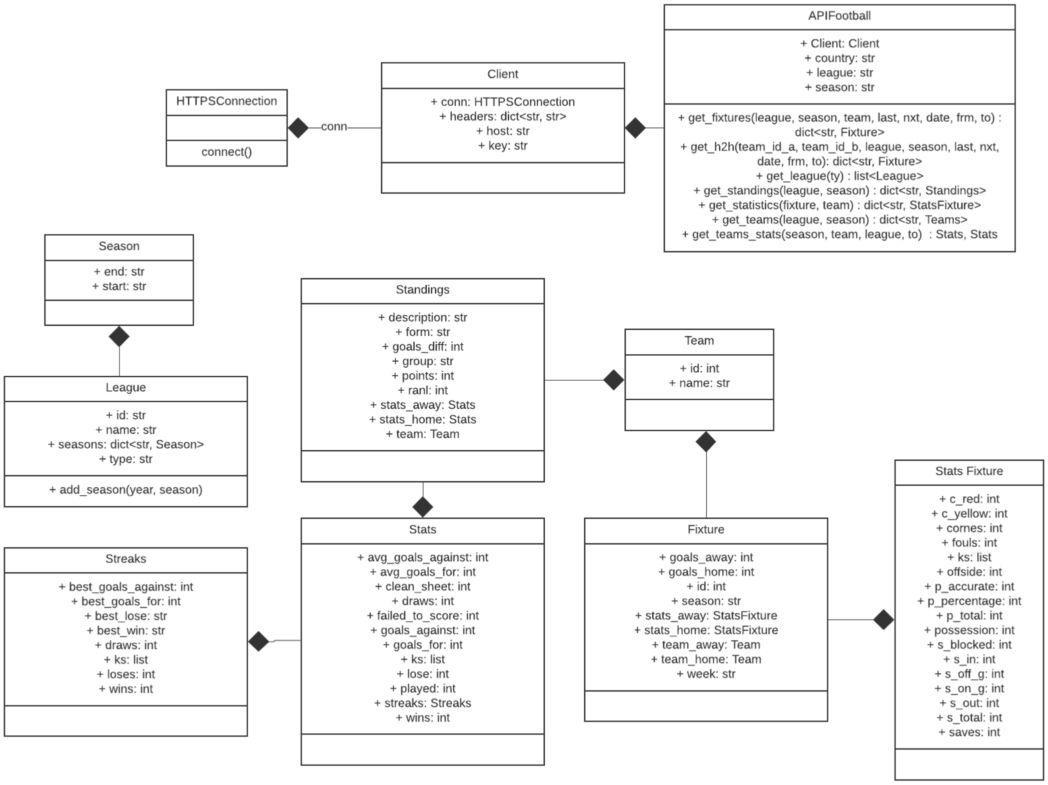


Figure 3.2: APIFootballpy Class Diagram.

This exploratory task uses team's standings at the end of the season, fields such as: rank (position from 1 to 20), points (final points on the season, which strongly correlated to the rank) and description are left out from the set in order to shorten bias. The description feature is the guidance target, this field indicates if the team is promoted to Champions League, Europe League, or relegated. Middle level teams do not get a classification, and sometimes pretty good teams are left out of Champions by goals one or two points, or by goal differences, and that doesn't mean they are below the standards. By performing Agglomerative Clustering, FootballMLP shows the hierarchy of the team's between each other in the league.

Hierarchical Clustering

FootballMLP employs Agglomerative clustering based on random walk to find the similarity measures between vertices. The best linkage method on LaLiga's standing was using average distance.

This clustering algorithm grouped together big teams such as: Barcelona and Real Madrid, as well Atletico de Madrid with Sevilla, Sevilla is the champion of Europe League 2019, the team deserves to be cluster in the highest hierarchy, and Agglomerative clustering is able to detect that. Also, shows how a relegated club as Espanyol had a really bad performance in the last season, and it is an outlier to the rest of the teams. The algorithm shows a really good performance as an unsupervised task, Figure 3.3 uses adjusted Rand Index to quantify both classification methods using Principal Component Analysis for lowering the dimensions in a 2D graph.

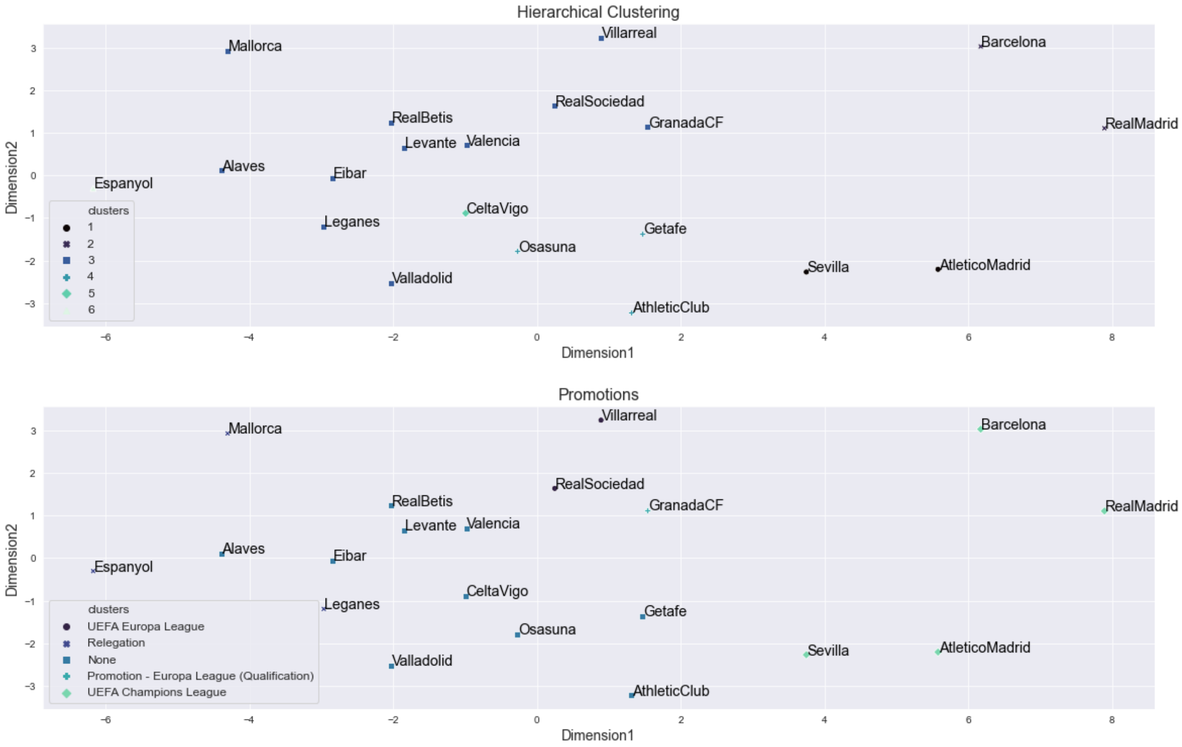


Figure 3.3: Unsupervised clustering vs Real promotions in season 2019.

On Figure 3.4, the Cluster map is a nice visualization to discriminate between top teams and its relation with winning, it helps to detect some interesting facts in the league's teams, for example, the negative correlation that the feature stats_home.streaks.best_lose has on Osasuna and Alaves, those teams were the most hammered when playing at home with a score of 0-5 Osasuna against Barcelona and 0-5 Alaves against Atletico de Madrid.

Another insight here is the offensive strength, when playing home versus playing away. On the failed to score field more negative correlations occurs when playing away rather than playing home, this due to a possible less chances of scoring on away games for top teams. To cover this factor, FootballMLP has automatize the process of translating the encoded Z matrix from the clustering algorithm by finding the desire threshold in the cophenetic distance matrix. The team's performance is different when playing away versus playing at home, this clustering task outputs three new fields: overall, home and away ratings. Those fields showed categorical significance to a match result when a chi-squared test was applied.

3.1.2 Rating

The ranking system is the way FootballMLP separates a team's strength in a league, now, it is important to measure the performance of a team in a given match. This section analyzes

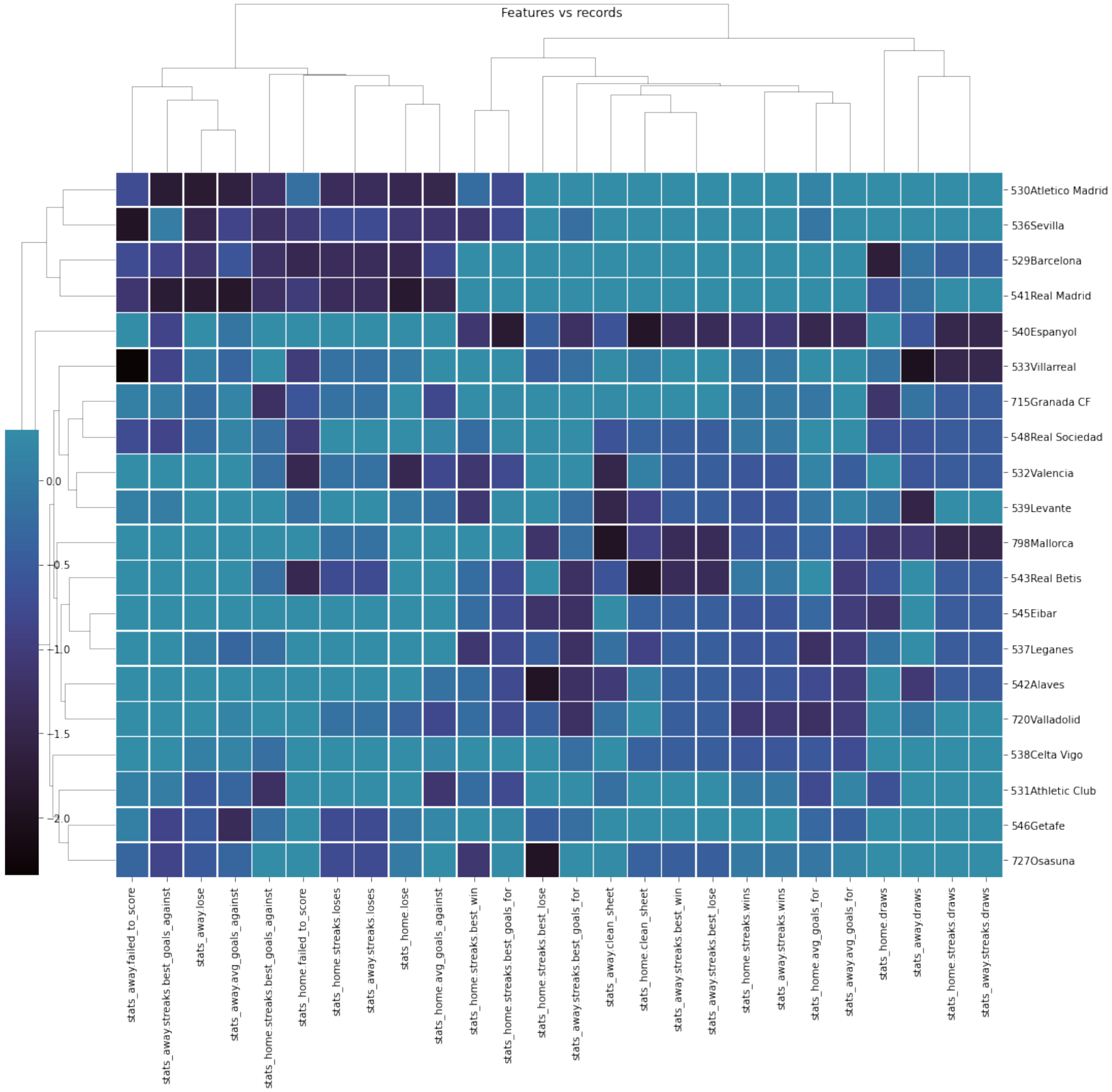


Figure 3.4: LaLiga's dendogram and correlation between teams' standings.

features relevance on team's statistics, as engineers the team's ratings.

Feature Engineering

Based on the relative performance in a game g explained by Pappalardo as the difference between the absolute performances of two opposing teams, he relates a team's behavior to the opponent one quoted [34]. For each of the StatsFixtures' features, FootballMLP calculates its relative performance. The equation 3.1 of the **Typical relative performance** is:

$$\bar{r}_A^{(g)} = [\bar{\delta_1}(A), x_2^{(g)}(A) - x_2^{(g)}(B), ..., \bar{\delta_n}(A)] \tag{3.1}$$

Pezzali The Pezzali score is a defense versus attack efficiency rate, it is based on Pezzali's premise: *It's the harsh rule of the goals: you play a great game but if you don't have a good defense, the opponent scores and then wins* [6]. The equation 3.2 of the **Pezzali score**.

$$PezzaliScore(team) = \frac{|goals(team)|}{|attempts[team]|} * \frac{|attempts(opponent)|}{|goals(opponent)|} \tag{3.2}$$

How many opportunities are you coinciding by attacking? Equation 3.3 is an adaptation of the Pezzali score to the best understanding on FootballMLP fixtures' statistics. The equation 3.3 of the **FootballMLP's Pezzali score**.

$$PezzaliScore(home) = \frac{|goals(home)|}{|s\_off\_g(home)+s\_on\_g(home)|} * \frac{|s\_off\_g(away)+s\_on\_g(away)|}{|goals(away)|} \tag{3.3}$$

The equation 3.3 is calculated for away team as well, and computed its differences.

Shots Fraction This feature attempts to measure the ratio of shots that are inside the penalty area, this means the ability of the team to advance until the last fraction of the field and shoot. The equation 3.4 of the **Shots Fraction** is:

$$ShotsFraction(home) = \frac{|s\_in(home)|}{|s\_in(home)+s\_out(home)|} \tag{3.4}$$

The equation 3.4 is calculated for away team as well, and computed its differences.

Defensive Here FootballMLP tries to rate defensive ability in a team by considering shots blocked by a team over total shots made by the opponent. The equation 3.5 of the **Defensive ability** is:

$$Defensive = \frac{s\_blocked(home)}{s\_total(away)} - \frac{s\_blocked(away)}{s\_total(home)} \tag{3.5}$$

Feature Selection

To identify the best possible features, that could generalize the playing style of any league, Recursive Feature Elimination was employed following the next criteria:

- Fitting a Linear Regression OLS model and leaving out variables with no statistical significance.
- Fitting a Support Vector Machine with a linear kernel and subtracting the features with relevant weights, as well measuring performance on R^2 and MSE.

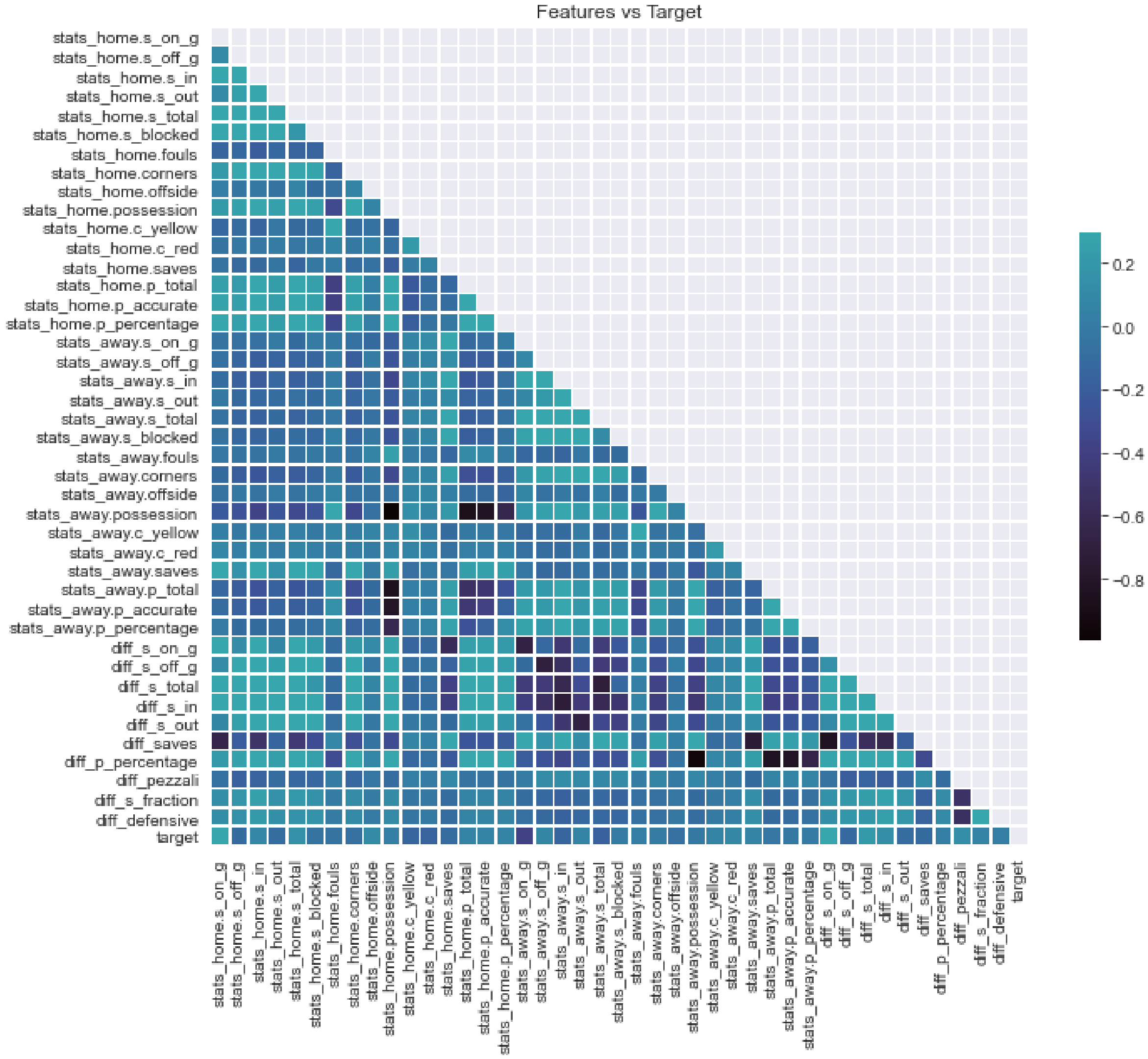


Figure 3.5: LaLiga's correlation between teams' statistics and the winning game.

Linear Regression tries to predict continuous values, but it does not means it cannot work as a first step for classification problems. Appendix 6.1.1 is the model summary obtained from statsmodels [45] when performing regression on score's goals difference, the results indicate a $R^2 = 0.945$ when training over 1,520 matches.

Case of linear regression Perform OLS Regressor on fixture's results will work as a first attempt to create a classification plane. This premise is due on behalf that 0 values represent a goal difference of 0 a draw match, -1 a negative goal difference a lose, and 1 a positive goal difference when home win.

Appendix 6.1.2 shows the model when performing regression on fixture's results a lower $R^2 = 0.724$ is obtained, the model summary of the t-score can be found in Appendix 6.2.. However, when using a classification model as SVM with a linear kernel trained with features accepted by the t-score, metrics such as: R^2, RMSE as seen on Figure 3.7 and Figure 3.8 and F-score in Table 3.1 and Table 3.2 show a slightly better performance.

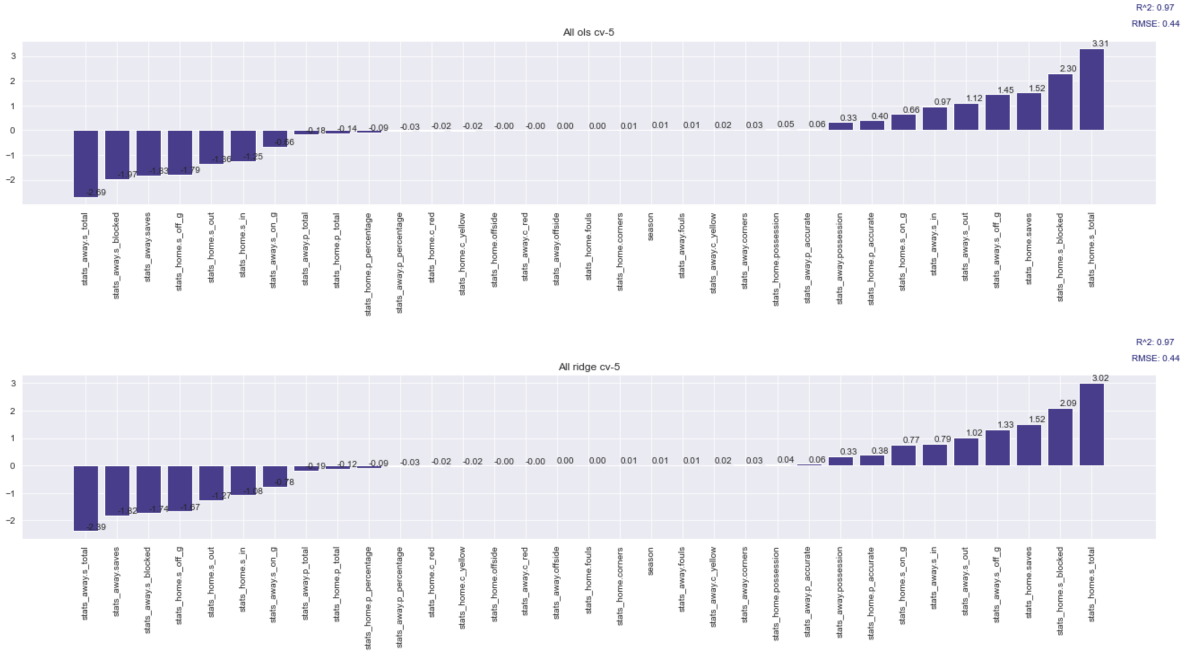


Figure 3.6: LaLiga's linear regression on goals difference using five cross-validation.

Ordinary Least Squares vs Ridge As seen on Figure 3.5 football has highly correlated variables, it is clear that the amount of shots on goal will increase in relation to the shots inside the area. The pass accuracy is the percentage of effective pass frequency, and the number of passes increases in comparison to the possession. When predictors are highly correlated we encounter multicollinearity problems, this can be solved by using Ridge Regression [54] which produces biased estimators with small variances to inflate the singular standard errors matrix. When using Scikit's Learn [36] OLS and Ridge Linear Regressors, we obtained identical values and a high Root Mean Squared Error when predicting goals differences as a result of an overfitted model as indicated on Figure 3.6, which also shows the coefficients learned by the model.

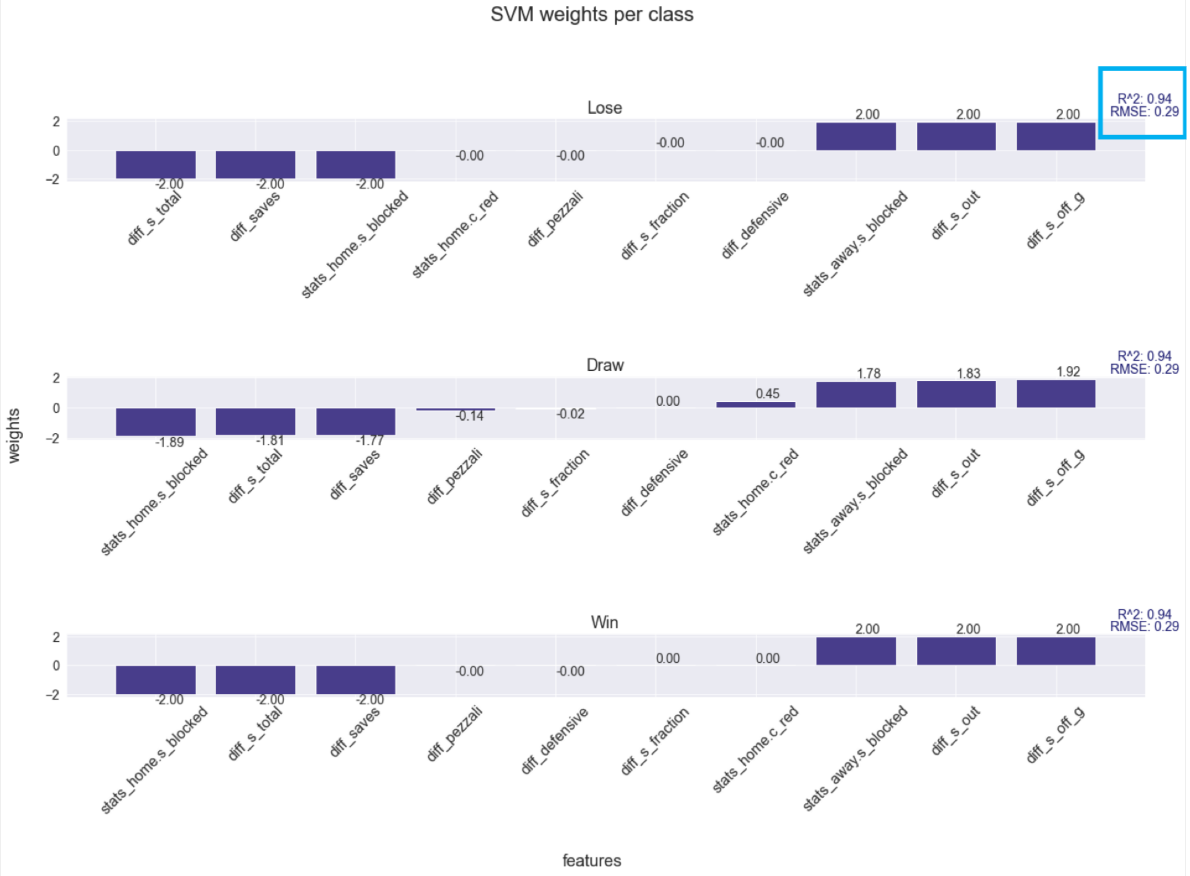


Figure 3.7: LaLiga's SVM classification when selecting features from fixture's results regression using five cross-validation.

Table 3.1: LaLiga's SVM classification report when selecting features from fixture's results regression.

Classification Report				
	precision	**recall**	**f1-score**	**support**
-1	0.93	0.96	0.95	428
0	0.87	0.81	0.84	390
1	0.94	0.95	0.94	702
accuracy avg/total	0.92	0.92	0.92	1520

3.1.3 The Voting Model

ANOVA and feature selection The results from feature selection on Section 3.1.2, adding the Pezzali score and the rankings from past season engineered in Section 3.1.1 (real ranking,

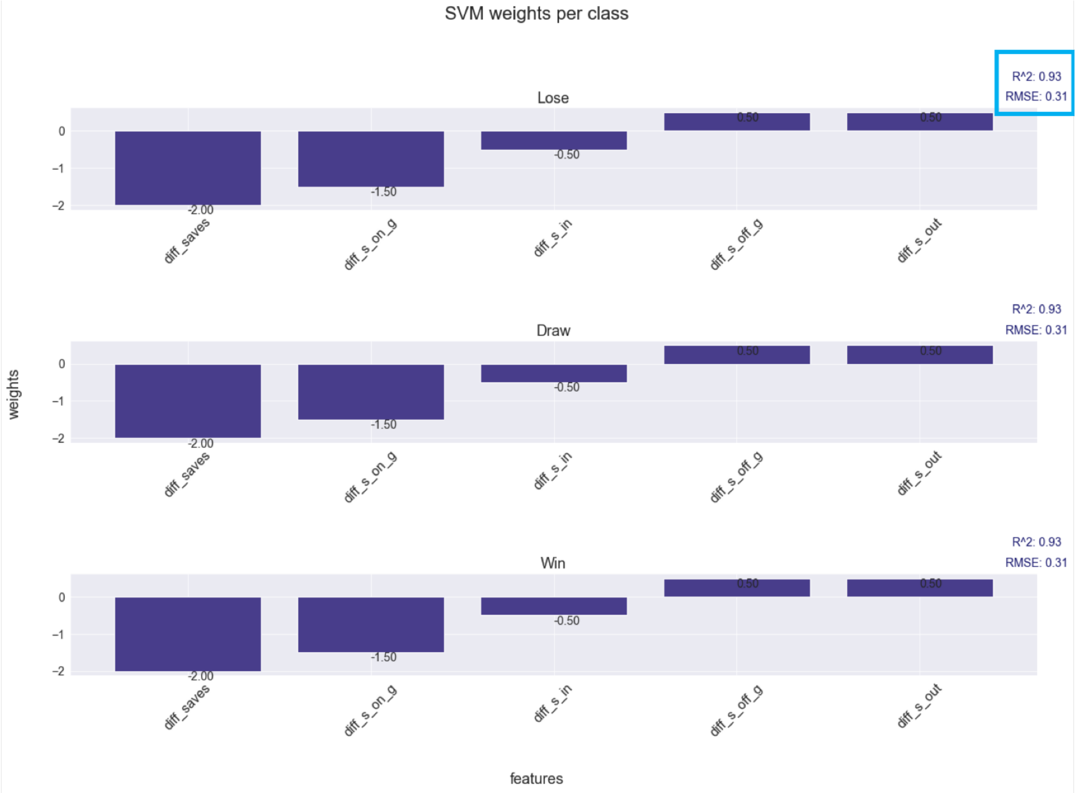


Figure 3.8: LaLiga's SVM classification when selecting features from goals difference regression using five cross-validation.

Table 3.2: LaLiga's SVM classification report when selecting features from goals difference regression.

Classification Report				
	precision	**recall**	**f1-score**	**support**
-1	0.93	0.96	0.94	428
0	0.86	0.81	0.83	390
1	0.94	0.94	0.94	702
accuracy avg/total	0.91	0.91	0.91	1520

home hierarchical ranking, away from hierarchical ranking, and overall hierarchical ranking) were fed into a Random Forest and Linear SVM model. A copy of all statistics with Pezzali and rankings was given to a Random Forest and Naive Bayes algorithm. Later on, in both cases, an ANOVA filter was used to select the 10-th best f-regressor features.

The skewness problem The fixture's statistics don't follow a bell curve in the Normal Distribution, instead, they are skewed, thus the mean is not the best measure of central tendency for the data. The fake test to predict a future match considers the median on the current season up to this match, the first three games add also the last season statistics. The median is calculated on behalf of the performance of the team and its location. When the team will play at home, only past matches at home are considered, the same situation occurs for the opponent team and it's away statistics.

The class-imbalance problem Figure 3.9 shows the frequency per class from seasons 2016 to 2019. To cover this imbalance problem, each of the models is wrapped in a One-vs-One Classifier, from the Sci-kit's library [36] this multiclass approach constructs one classifier per pair of classes. The class with the majority vote is chosen, when a tie occurs, then selects the class with the highest classification confidence by aggregating the computed levels underlying the binary classifiers.

Majority voting While comparing each of the classifiers, it was found that some of them were predicting a win at home versus a loss at home. To solve this disparity problem, it was decided to build a heterogeneous ensemble that performed majority voting.

3.2 Fan Avidity

Soccer is constantly changing over time, and so do its fans. In order to make this a real-time problem, a framework for gathering recent tweets was built. The purpose of this section is to find metrics that could work out as a feature for evaluating social networks into a predictive model. The methodology is summarized in two key components: first tweets are preprocessed for scoring sentiment polarity, and second, they are evaluated as a Social Network problem by applying graph theory.

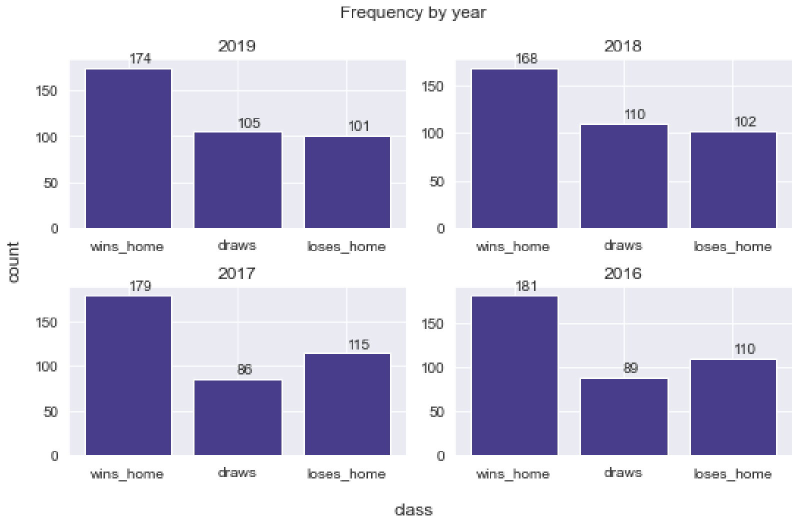


Figure 3.9: LaLiga's results frequencies from season 2016 to 2019.

3.2.1 Classification

The data is obtained through the Twitter's Standard Search API [1]. The queries were performed in the last match week on Premier League's season 19/20 and were limited to the English language. The data was processed into a Dataframe with the next remaining fields as shown in Table 3.3. Twitter's documentation on standard operations shows that appending a string happy face ":)" on the query represents a positive attitude, while ":(" represents a negative attitude. The maximum number of tweets retrieved from a request is 100, to aid the evaluation process, three types of queries were performed: first adding the happy face, second adding the sad face, and finally a neutral request without a face.

Table 3.3: Dataset fields

	Field	**Description**
1	season	A YYYY representation of the match season.
2	weekgame	The number of the current week match.
3	home_team	A three-letter code abbreviating the home team.
4	away_team	A three-letter code abbreviating the away team.
5	favorite_count	The count of favorites in the tweet.
6	lang	A two-letter code abbreviating the language.
7	retweet_count	The count of retweets in the tweet.
8	retweeted	True or false if the tweet is a retweet.
9	text	The text of the tweet.
10	followers_count	The count of followers from the user.
11	verified	True or false if the account is verified.

In the end, a total of 30 JSON requests with a maximum count of 100 tweets were

Table 3.4: Queries

	Match	**Keywords**
1	Arsenal vs Watford	#ARSFC @Arsenal #WatfordFC @WatfordFC #ARSWAT
2	Burnley vs Brighton	#BURFC @BurnleyOfficial #BHAFC @OfficialBHAFC #BURBHA
3	Chelsea vs Wolves	#CFC @ChelseaFC #WWFC @Wolves #CHEWOL
4	Crystal Palace vs Tottenham	#CPFC @CPFC #THFC @SpursOfficial #CRYTOT
5	Everton vs Bournemouth	#EFC @Everton #AFCB @afcbournemouth #EVEBOU
6	Leicester vs Manchester United	#LCFC @LCFC #MUFC @ManUtd #LEIMUN
7	Manchester City vs Norwich	#MCFC @ManCity @NorwichCityFC #MCINOR
8	Newcastle vs Liverpool	#NUFC @NUFC #LFC @LFC #NEWLIV
9	Southampton vs Sheffield Utd	#SaintsFC @SouthamptonFC #SUFC @SheffieldUnited
10	West Ham vs Aston Villa	@WestHam #AVFC @AVFCOfficial #WHUAVL

available for study, Table 3.4 indicates the keywords placed in each fixture query. However, not all fixture requests accomplished these 100 tweets as shown in Table 3.5.

Table 3.5: Requests

	Match	**# of *tweets***
1	Arsenalvs Watford	*247*
2	Burnley vs Brighton	*121*
3	Chelsea vs Wolves	*235*
4	Crystal Palace vs Tottenham	*156*
5	Everton vs Bournemouth	*143*
6	Leicester vs Manchester United	*256*
7	Manchester City vs Norwich	*198*
8	Newcastle vs Liverpool	*268*
9	Southampton vs Sheffield Utd	*128*
10	West Ham vs Aston Villa	*175*

Data cleaning

The data cleaning process performed drop of duplicates with a count of one, and drop off empty tweets, the empty tweets were filtered after removing user mentions and ended up with no text to analyze, this returned a count of 11 empty tweets. At first, the library langdetect was used with a threshold of 50% probability for the English language, however, in practice, the language detection accuracy drops drastically on shorter tweets, so multilingual tweets were kept. The length of the dataset finished with a total of 1,915 tweets.

Data engineering

In order to make the most of the available resources, extra variables were included, two of them relate to text transformations.

- pre_label: integer field that pre classifies the tweet according to the search query, for positive tweets gives 1, negative -1 and 0 if neutral.
- support: integer field that pre represents the support to a given team if it appears on the tweet a mention or hashtag to the home team returns 1 when away team returns -1 and 0 if both appearances happened.
- no_mentions: string field as a version of the tweet without mentions and removing anything that is not plain text.
- with_emojis: string field as a version of the tweet, this sophisticated text transformation keeps mentions, removes links, and uses the emoji library to encode emojis into text, also a regular expression matches happy and sad faces representation and replace happy faces by the word good and sad faces with the word bad.

Classifying polarity was possible by using available resources, such as the Open Source Library Stanza [38], previously named Stanford NLP. Stanza is a language-agnostic processing pipeline that groups together tokenization, lemmatization, part-of-speech tagging, dependency parsing, and named entity recognition. Stanza has a built-in model for Sentiment Analysis [20], this model is trained as a one-layer Convolutional Neural Network using word2vec which are the resulting vectors when applying bag-of-words on 100 billion articles on Google News.

Since it is possible to provide text previously tokenized to Stanza's pipeline, it was preferable to create tokens using NLTK Tweet Tokenizer as it applies regular expressions to maintain mentions. A mention identifies users with the prefix @, while Stanza Tokenizer split those characters underperforming entity recognition.

Figure 1 shows some word clouds comparing the results given on the assumption of the pre_label tag against Stanza's model evaluation.

By applying Stanza's classification it is possible to measure the magnitude of the polarity tags. Now, phrases such as love, good luck, hope, took relevance on the positive tags, while on the negative tags curse words and negations took precedence.

After classifying polarity in tweets from a Machine Learning perspective, two new fields were created a modified support *m_support* and modified sentiment *m_sentiment*. These fields

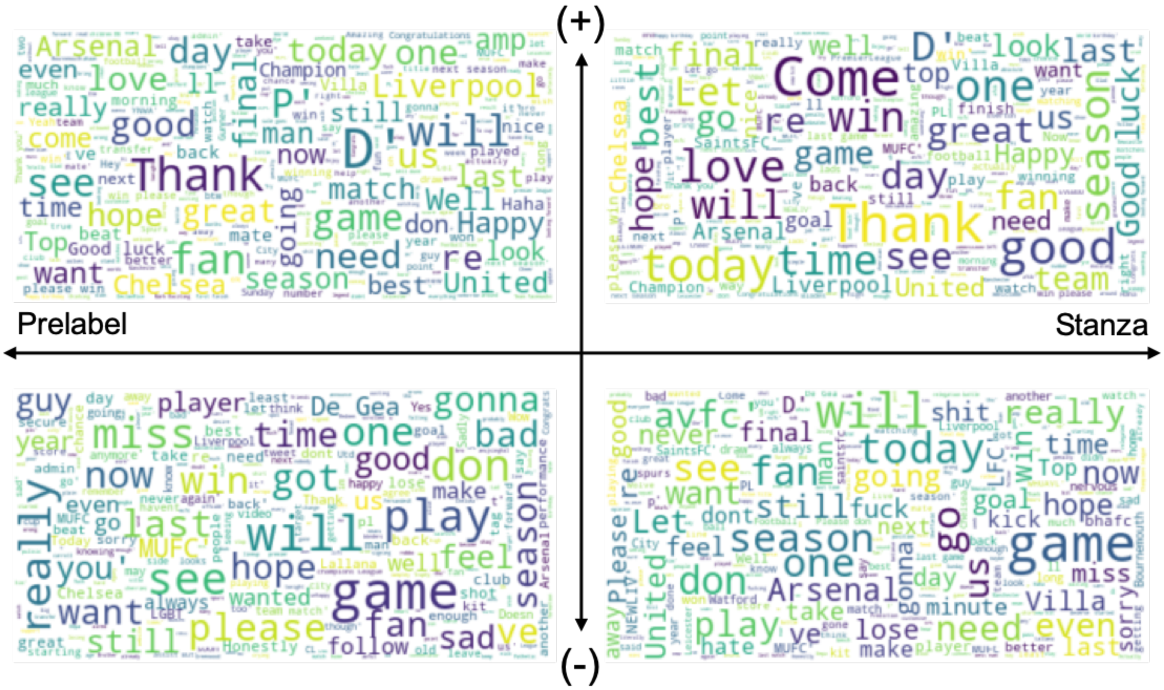


Figure 3.10: Word clouds comparison by polarity and model

were the result of matching a regular expression that identifies suggested scores of form 0:0 or 0-0 since a result with a goal difference greater than zero indicates clear favoritism to one of the adversarial teams.

Figure 3.11 and 3.12 shows relevance on the proposed characteristics, this is measured by the amount of neutral support and sentiment that was able to be classified as positive or negative, and as a side of the home team or away team.

3.2.2 Quantification

Graph theory

Popular teams such as Arsenal, Liverpool, Manchester United, etc., have higher rates of tweets, making it difficult to choose a favorite when comparing against its opponent. This section translates the imbalance of favoritism into a graph analysis.

A simple graph [22] has the form $G = (V, E)$ where V is a set of n vertexes and E is a set of n edges. An edge is a link between two vertexes, so an edge E_k is associated with an unordered pair of the vertex (V_i, V_j).

Here the vertex are the users and each edge represents a tweet to a tagged user, the tagged user is the team which is mentioned on the tweet. The final tuple looks like: $(fan, team, edge_k)$. Two edges were added to the graph when a tweet mentioned both of the teams. Also, it was preferable to choose a multigraph representation, since it is possible to have multiple edges of a fan to the same team.

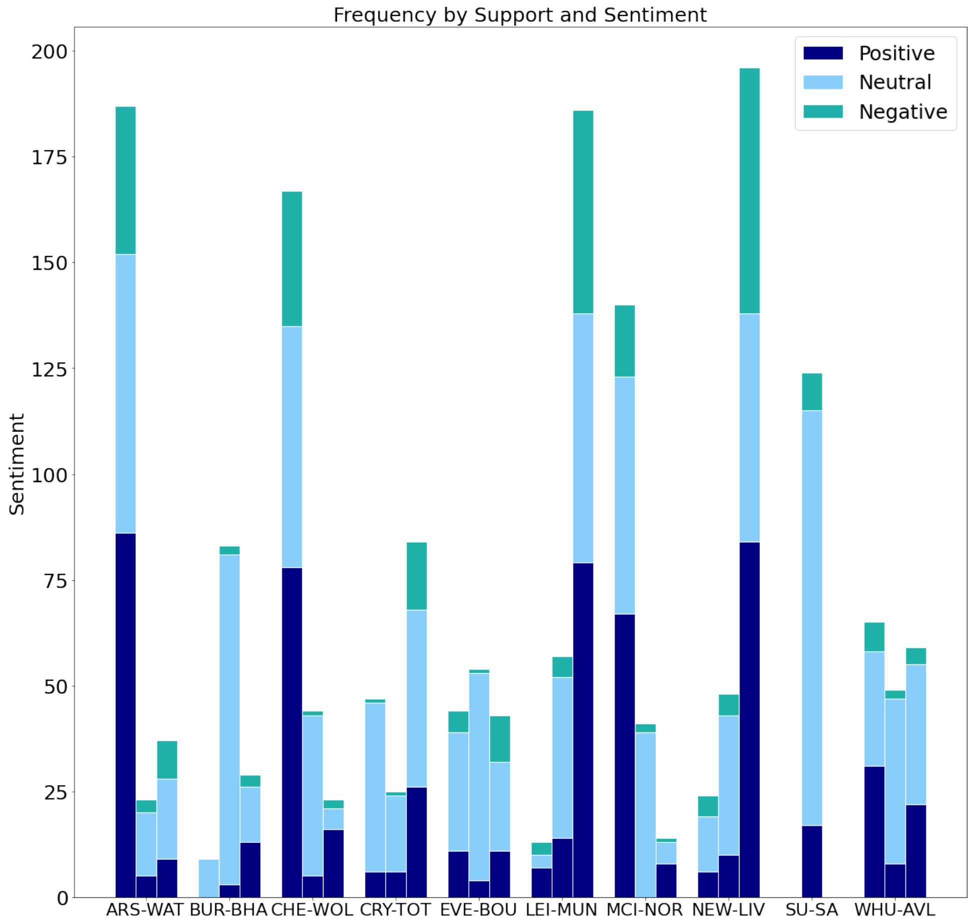


Figure 3.11: The frequency of the pre_label polarity by team support

Edge's weight Setting a singular value for each edge's weight will miss out on tweets having likes or being retweeted, as well, it would not solve the imbalance problem, since the sum of all edges values from the team's vertex to the fans will be equal to the frequency of the fans of a given team.

The equation 3.6 of the **Tweet's weight imbalance** is:

$$E_k(U_i, team) = c - \frac{U_i(likes)+U_i(retweets)}{\sum_{i=0}^{k} support(team) + \frac{\sum_{i=0}^{k} support(match|neutral)}{2}} \tag{3.6}$$

An edge between the user and the team represents the distance the tweet has with the team, whenever a tweet is more retweeted or has a larger amount of likes, it means it is more reachable to the audience of a team, subtracting from a constant c, the relative number of likes and retweets to the support of the team, means reducing the distance between the fan and the team. By calculating the relative influence in a network as the sum of interactions over the frequency of support in a team, a team with fewer followers will represent a greater reach to its network, rather than the reach-in networks with larger amounts of fans, this way the class imbalance problem could be resolved. Neutral support was split in two and added to the frequency of each of the teams as seen in Figure 3.13 where light blue lines are neutral, navy are the positive tweets and green negatives.

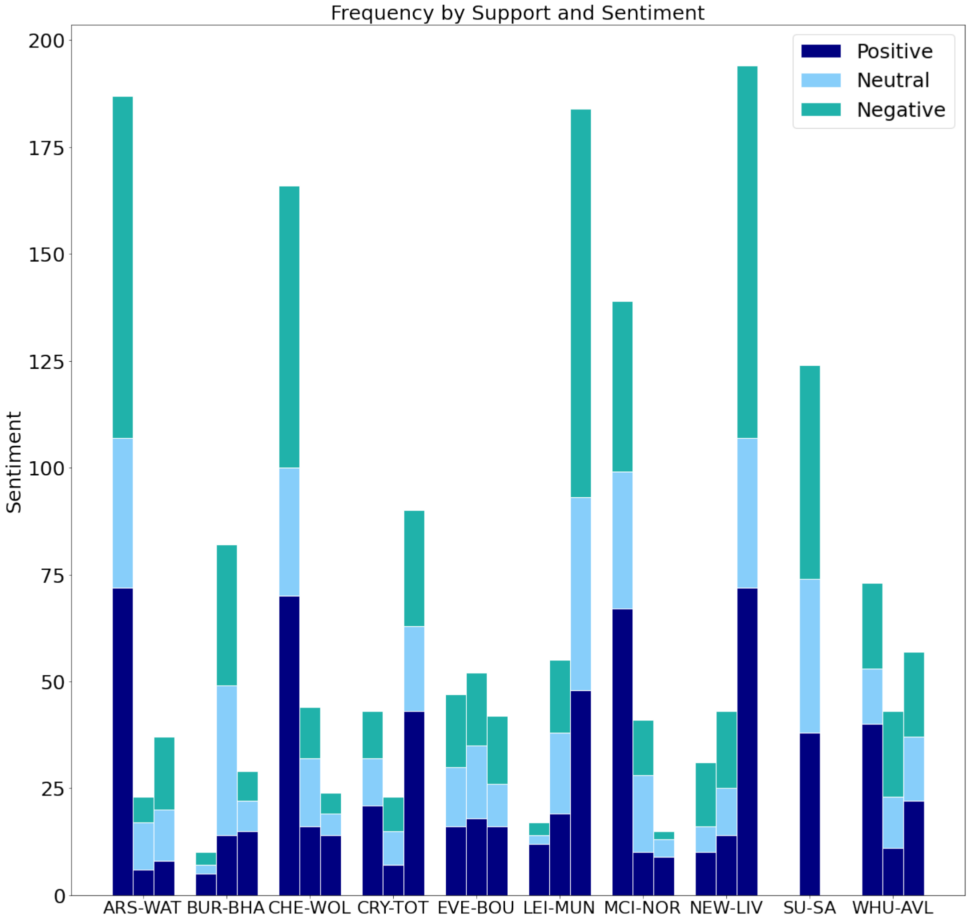


Figure 3.12: The frequency of the modified polarity by team modified support

Inverted polarity This is a counter proposal for solving the imbalance problem, by interchanging support to the adversarial team. This creates a network where negative links to a given team, become positive edges to its opponent and vice versa. Then the network is composed only of positive and neutral polarity represented in Figure 3.14.

Metrics of centrality

A way for evaluating network entities is through indexing centrality [56], this metric indicates the influence of the vertex in the network. Degree centrality is discarded, since it counts the number of links to a node, and as mentioned earlier there is a clear imbalance between the number of fans, so it might present misleading results. Betweenness centrality is not taken into consideration neither, this measure gives precedence to mediation nodes that connect the network, here users that mentioned both of the teams have the highest scores. Closeness centrality is the selected measure for comparing independence and efficiency of communication in an entity [23].

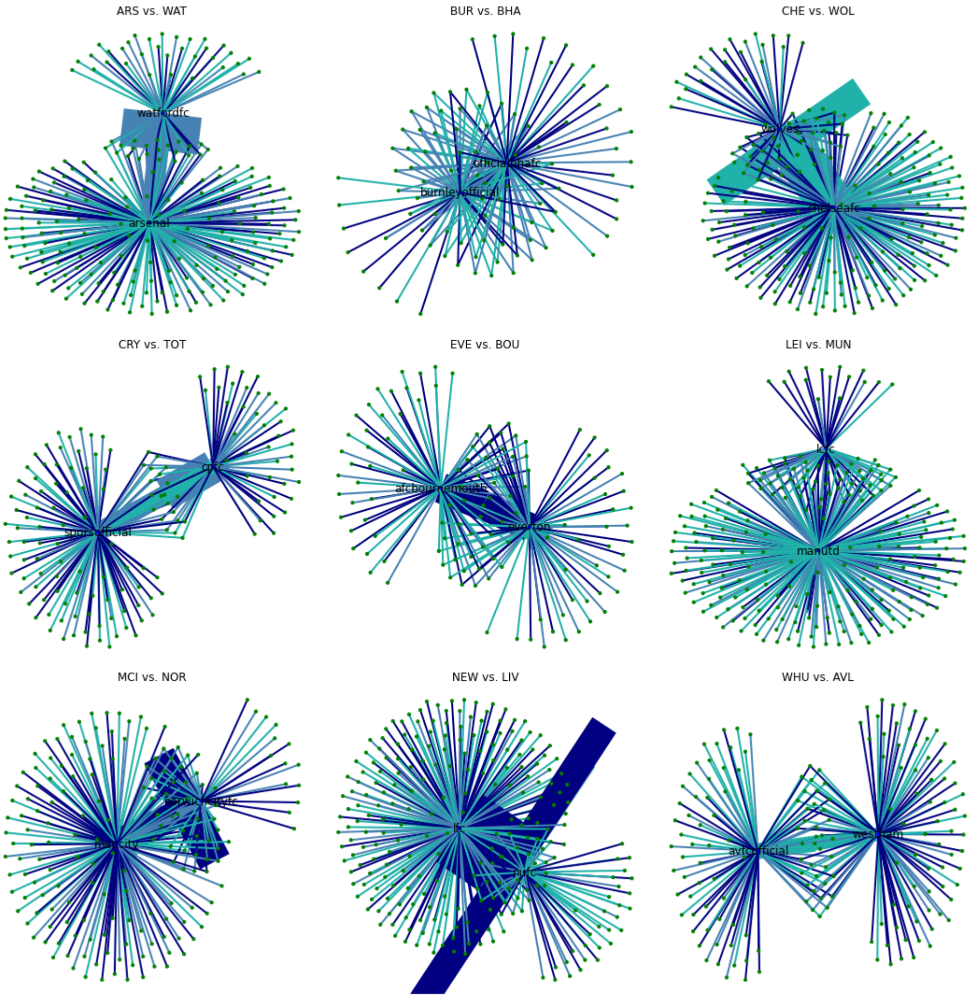


Figure 3.13: Graph using tweet's weight imbalance

Closeness centrality It is computed as the reciprocal of the average of shortest-path distance from an agent A_u to all other agents. The equation 3.7 of the **Closeness centrality** is:

$$C(u) = \frac{n-1}{\sum_{i=1}^{n} d(i,u)} \tag{3.7}$$

Current-flow closeness centrality It is based on information spreading efficiently like an electrical current. Edges are now resistors $r_e = 1/w(e)$ and each vertex has a voltage $v(u)$. The equation 3.8 of the **Current-flow closeness centrality** is:

$$C(u) = \frac{n}{\sum_{i=1}^{n} v(u)-v(i)} \tag{3.8}$$

This represents the ratio n to the sum of effective resistances between u and other vertexes quoted [16]. It is also equivalent to the information centrality which considers all path weights, not only the shortest ones, and instead computes its average from the originated vertex. The information in a path is the inverse of the length of a path [3]. The equation 3.9 of the **Information centrality** is:

$$\bar{I}_u = \frac{n}{\sum_{i=1}^{n} \frac{1}{I_{ui}}} \tag{3.9}$$

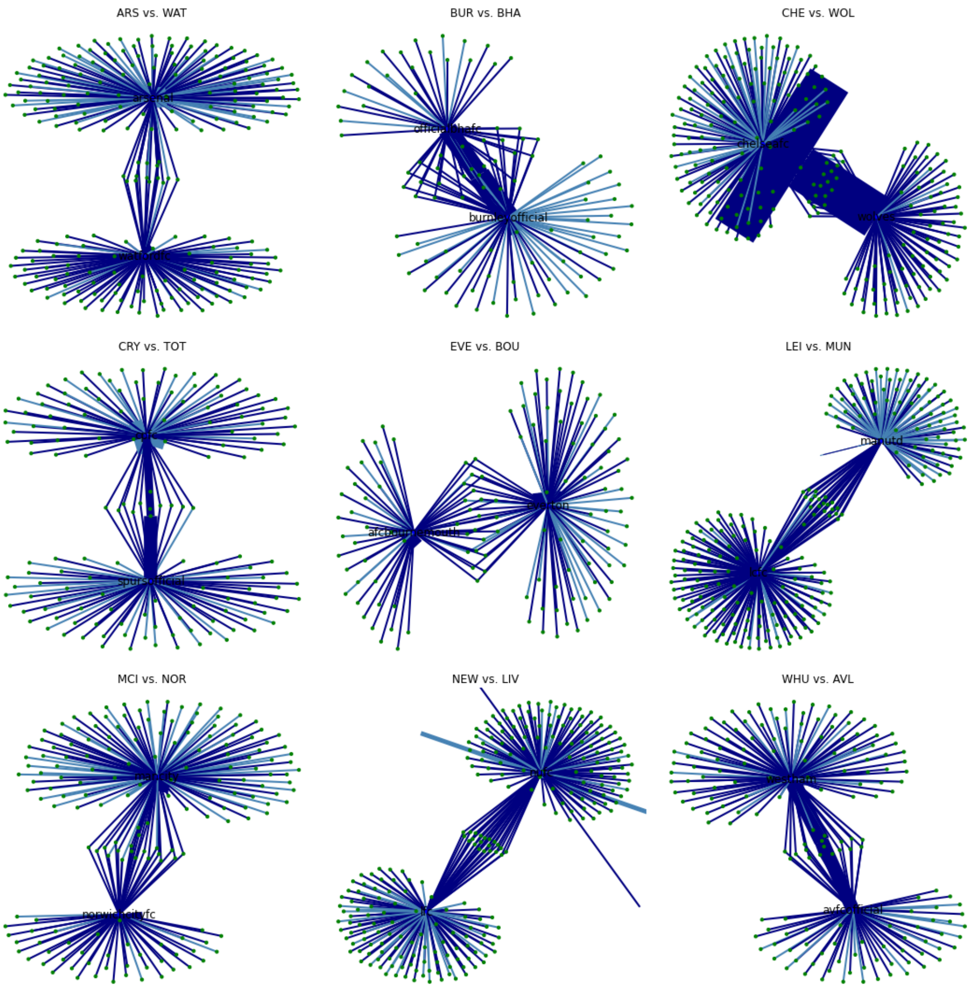


Figure 3.14: Graph using tweet's inverted polarity

Harmonic centrality Applies harmonic mean to overcome outweighs from infinite distances, and it is computed as the sum of the reciprocal of the shortest path distances. The normalized harmonic centrality can reach up to 1 as the maximum connected vertex. Lower values occur when used on an unconnected graph representing the reduced capability of communication in the network [42]. The equation 3.10 of the **Harmonic centrality** is:

$$C(u) = \sum_{i=1}^{n} \frac{1}{d(i,u)} \tag{3.10}$$

Chapter 4

Results and Discussion

This chapter is oriented on the evaluating FootballMLP, the sentiments model and the hybrid model's predictions in a set of chronological matches. As well, it includes a Time Series Analysis to understand periodicity in football results.

4.1 Statistics Model

FootballMLP was evaluated on seasons 2018/19, 2019/20, and partially 2020/21. The sequential order from the match weeks was followed, instead of splitting the fixtures 70% training and 30% testing.

4.1.1 Voting vs Naive Bayes

The novelty of this part of the work is to propose a majority voting model which could help to choose a final result when disparity is met. Table 4.2 shows the classification report for this ensemble model and Table 4.1 shows the classification report for the Naive Bayes classifier in all leagues. Naive Bayes, whose results are displayed on Table 4.1 had the best performance as a single classifier, the assumption of independence of the events is an efficient way to approach the analysis on football, but it is susceptible to overfitting the class with most occurrences, in this case, fixture's wins. Even though both classifiers have an average accuracy of 50%, the majority voting model, as seen on Table 4.2, was able to detect fixtures that ended up as a loss or a draw at home, with higher f1-scores on those labels, therefore a better trade-off between precision and recall.

Table 4.1: Naive Bayes classification report in all leagues.

Classification Report				
	precision	**recall**	**f1-score**	**support**
-1	0.50	0.50	0.50	1660
0	0.30	0.18	0.23	1309
1	0.56	0.68	0.61	2371
accuracy avg/total	0.50	0.50	0.50	5340

Table 4.2: Majority voting classification report in all leagues.

Classification Report				
	precision	**recall**	**f1-score**	**support**
-1	0.49	0.51	0.50	1660
0	0.30	0.22	0.25	1309
1	0.58	0.65	0.61	2371
accuracy avg/total	0.50	0.50	0.50	5340

4.1.2 Leagues' Competitiveness

To compute the true positive rate and false positive rate it is proposed to find the probability estimate of a class in a majority voting system as shown. The equation 4.1 of the **Probability estimate in majority voting classifiers** is:

$$p(C) = \frac{1}{n}\sum_{i \epsilon C}^{n} i \tag{4.1}$$

For n classifiers votes, the probability estimate can be computed as the sum of all votes to a class C over the number of classifiers. Figure 4.2 proved that Draw is the hardest class to be predicted, the baseline AUC (Area Under the Curve) 0.50 was reached on Germany Bundesliga. However, Figure 4.1 and Figure 4.3 coincided, on both, the winning and losing class that the highest areas covered by the algorithm occurred on Eredivisie NL, Premier League GB, and Serie A IT, in those cases values around 0.70 were reached. After conducting this experiment, it could be said that the LaLiga is the less predictable league, when looking at winning games, this same thought infers LaLiga to be an evenly played league with a high competitive level. When compared to Figure 4.4 that shows the accuracy per league, it can be said that Bundesliga got the highest Accuracy for the winning class, however, this decision is not supported in relationship to false positives and true positives as shown in Figure 4.1, meaning that the model is overfitting Bundesliga, this way it can be proposed to keep AUC as a leading metric for measuring performance in a classified league. Figures 4.1 to 4.3 compare leagues predictability per class in form of a ROC (Receiver Operating Characteristic) curve. Figures 4.4 to 4.6 compare leagues accuracy per class. Even though the accuracy in the draws class for Premier League GB is low as seen in Figure 4.5 in this same league it is obtained the third-best AUC = 0.56, this is a result better than random, and due to the imbalance problem, it justifies the model being able to detect draws versus non-draws. In Figure 4.6 an opposite case occurs when Spain La Liga has a high accuracy in the loses class while its AUC is the lowest one, these results showed again that this league is the hardest one to classify correctly. Appendix 6.3 shows the classification report per league.

4.1.3 Time Series Analysis

Time series analysis is the analysis of time-dependent data [33] with two determinant factors which are trend and seasonality. The first one answers, how does the variable tend to rise or fall as time passes? while seasonality tries to detect regular patterns occurring in time repeatedly. This section aims to analyze the results accuracy through time. On top of Figure 4.7 the percentage of predicted correct matches are shown, while the bottom graph displays

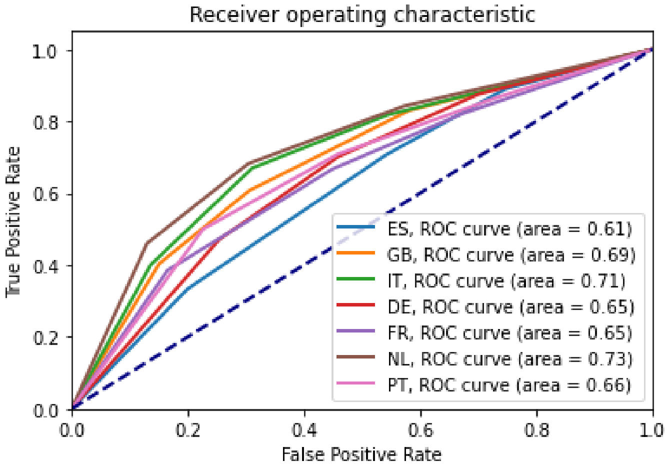


Figure 4.1: ROC curve for the majority voting classifier in the win class per league.

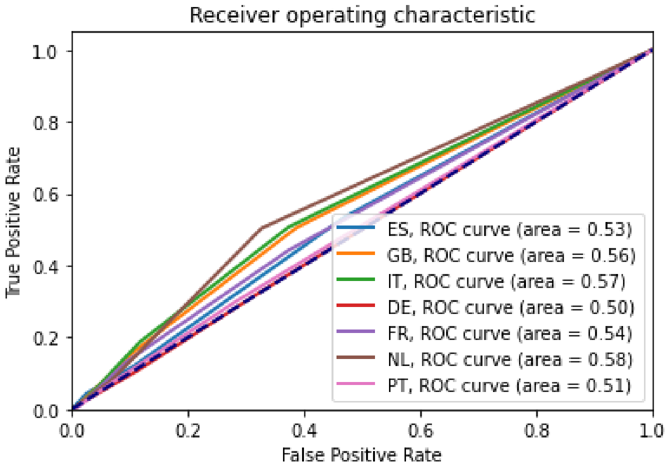


Figure 4.2: ROC curve for the majority voting classifier in the draw class per league.

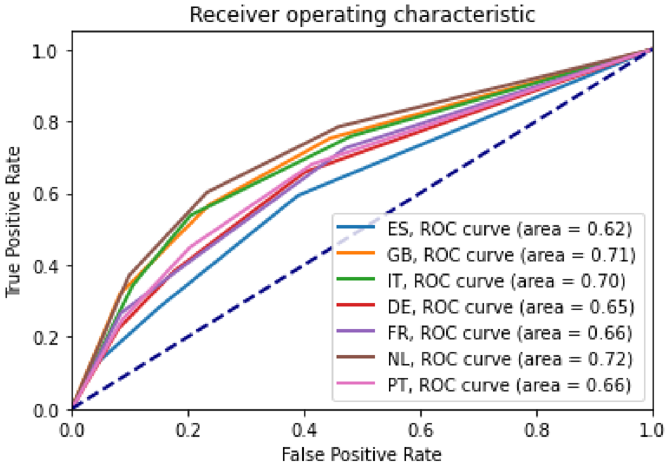


Figure 4.3: ROC curve for the majority voting classifier in the lose class per league.

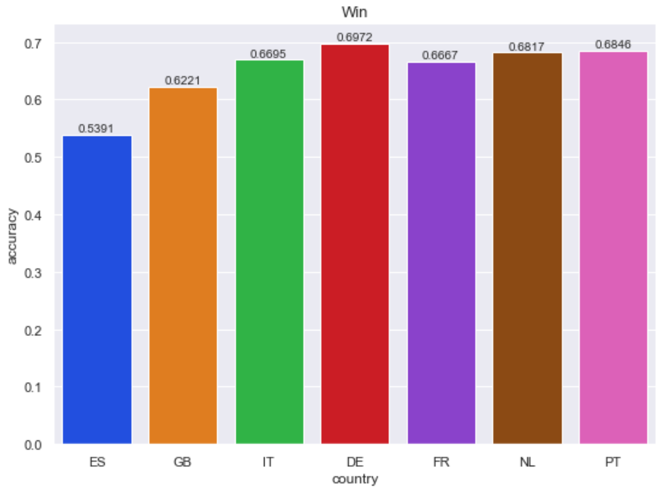


Figure 4.4: Accuracy for the majority voting classifier in the win class per league.

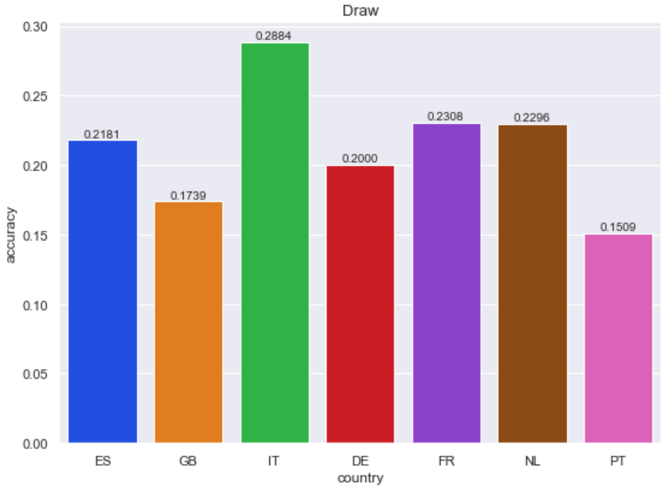


Figure 4.5: Accuracy for the majority voting classifier in the draw class per league.

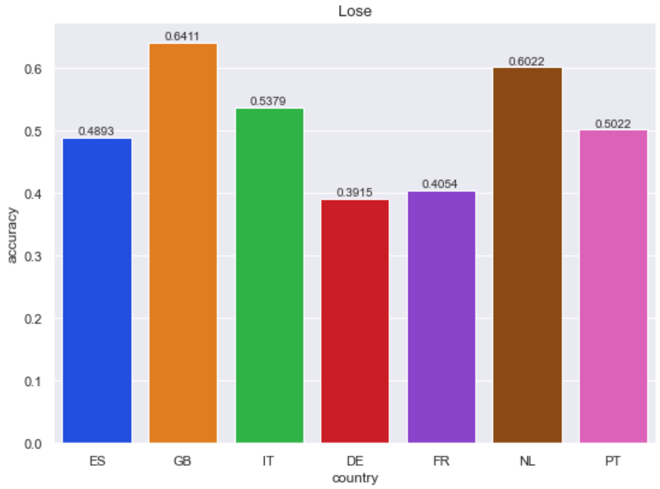


Figure 4.6: Accuracy for the majority voting classifier in the lose class per league.

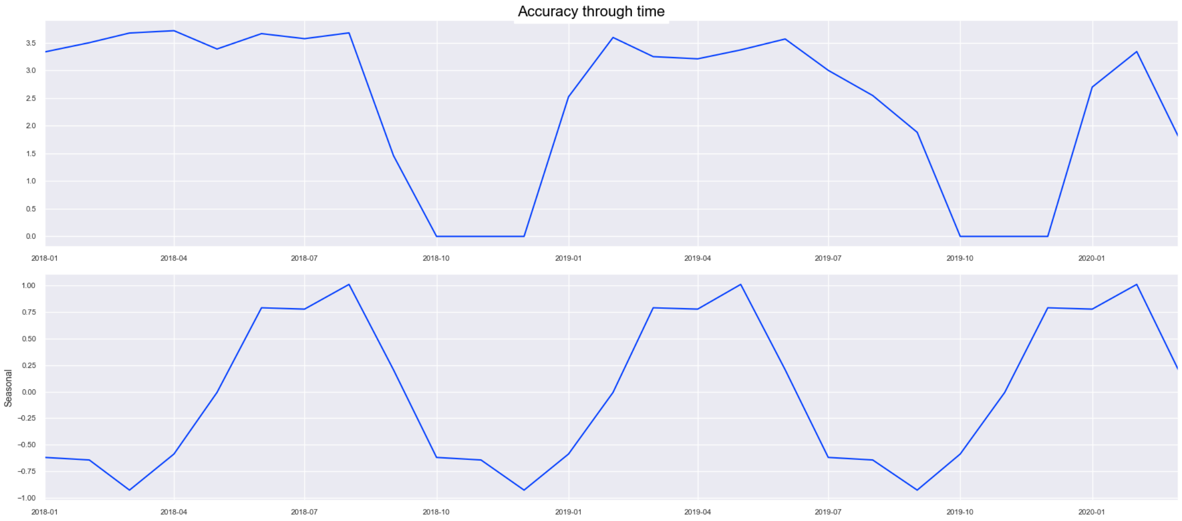


Figure 4.7: Percentage of results accuracy as time series.

seasonality in periods of 9. A maximum of 38 consecutive weeks are display with a break, starting in October, between seasons. As seen in Figure 4.7 apparently between weeks 24 and 32 there are fewer fluctuations in the predictions. Figure 4.8 to 4.10 compares the number of actual occurrences of a class in a week, against the number of predictions of the same class. Metrics such as Root Mean Square Error and R^2 are tested as in linear regression. The number of predicted wins and draws fitted a R^2 around 0.60, however the predicted loses class R^2 was equal to 0.90.

Table 4.3: Seasonality Dickey and Fuller test results.

	Real			**Predicted**		
	-1	0	1	-1	0	1
ADF-Statistic	-0.6143	-6.0792	-2.0471	-2.5082	-2.8429	-2.7848
P-value	0.8677	0.0000	0.2663	0.1135	0.0524	0.06048

To make a strong recommendation about the results of this analysis the idea was to apply the Dickey and Fuller test which delivers evidence or absence for the presence of seasonal unit roots [46]. Table 4.3 displays the result from the Dickey and Fuller test, and it can be said that seasonality is kept while comparing the real and predicted classes. In the win and lose class the null hypothesis is proved with p-values greater than 0.05, meaning that those classes have a time-dependency factor. However, for the Draw class, both cases had p-values less than 0.05, meaning that the number of draws in a week is irrelevant as the competition advances.

4.2 Sentiments Model

For evaluation purposes, the Premier League's study, which consisted of only 10 matches, was extended to a set of 54 matches starting at week 38 from season 2019/2020 on the date of July 26th, 2020. As well, from week one in season 2020/2021 starting in September 12th,

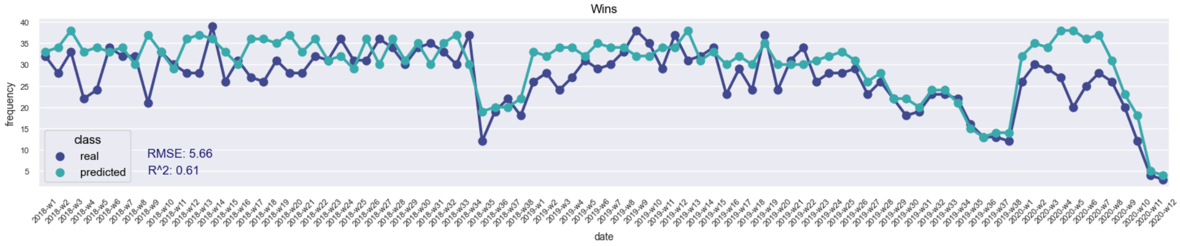


Figure 4.8: Number of wins vs number of predicted wins through time.

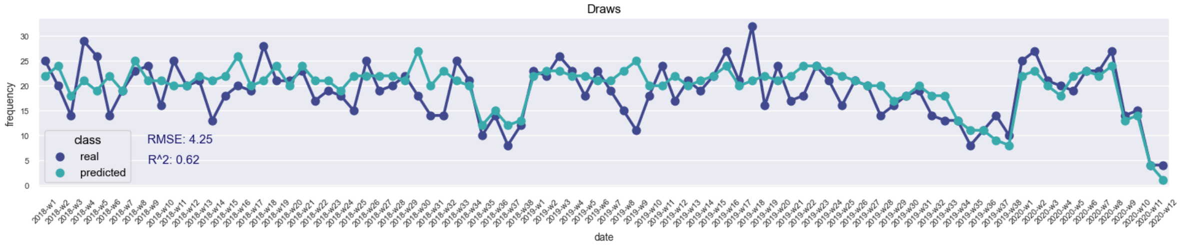


Figure 4.9: Number of draws vs number of predicted draws through time.

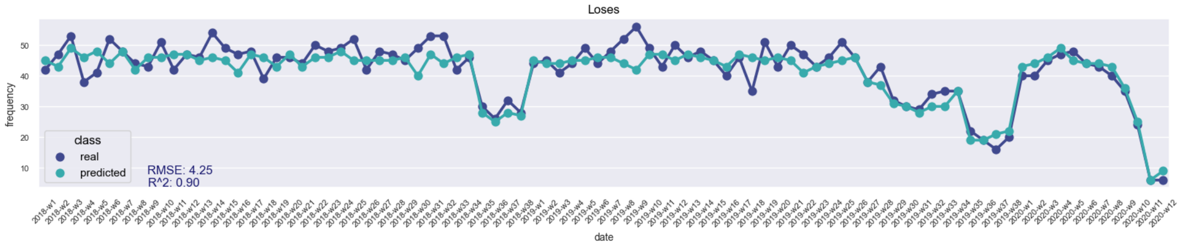


Figure 4.10: Number of loses vs number of predicted loses through time.

2020 till November 8th, 2020 in week 8. In total 7,833 tweets were analyzed. Besides a graph considering the three polarity links, three subgraphs, one for each polarity, were built. Based on centrality measures two cases were considered:

4.2.1 Comparing Networks' Groups

Case 1.

Applying current-flow closeness centrality as a comparison measure inter-team, since low resistance will show efficiency in the way a team communicates to its fans. The difference between the current-flow closeness index on the home team and away team will reflect the favorite team given its communication effectiveness. The equation 4.2 of the **Inter-team closeness** is:

$$diff\_closeness = \|closeness(home) - closeness(away)\| \tag{4.2}$$

Case 2.

Applying harmonic centrality as the leading polarity intra-team, to know which polarity has a better representation of the fan's sentiment towards a team. Communication is more difficult when having fewer connections. For each subgraph, the less fluctuated harmonic centrality given a polarity against the harmonic centrality considering all three polarities will support a good communication capability. The equation 4.3 of the **Intra-team closeness** is:

$$closeness(\frac{team}{polarity}) = \|closeness(team) - closeness(polarity)\| \tag{4.3}$$

Support Vector Machines were used for classifying a match as a win, draw, or lose at home. These models were trained with different centrality indexes (current-flow closeness centrality, harmonic closeness centrality, inter-team closeness and intra-team closeness) and evaluated with five-fold cross-validation. During the pipeline different 10th best features were applied to test ANOVA.

Table 4.4: Classification report when selecting features from tweet's weight imbalance.

Classification Report				
	precision	**recall**	**f1-score**	**support**
-1	0.56	0.61	0.58	23
0	0.00	0.00	0.00	9
1	0.45	0.45	0.45	22
accuracy avg/total	0.44	0.44	0.44	54

Table 4.5 shows higher values on the recall metric than Table 4.4. Although the inverted polarity model has a better recall, the weight imbalance model did not lose precision and gave more diversity to the prediction model by attempting to guess draw matches.

Figure 4.11 plots feature weights, and validates the inter-team centrality as the difference on the normalized harmonic closeness centrality, while the intra-team measure is given by singular polarities of a team. Figure 4.12 confirms the current-flow closeness centrality as

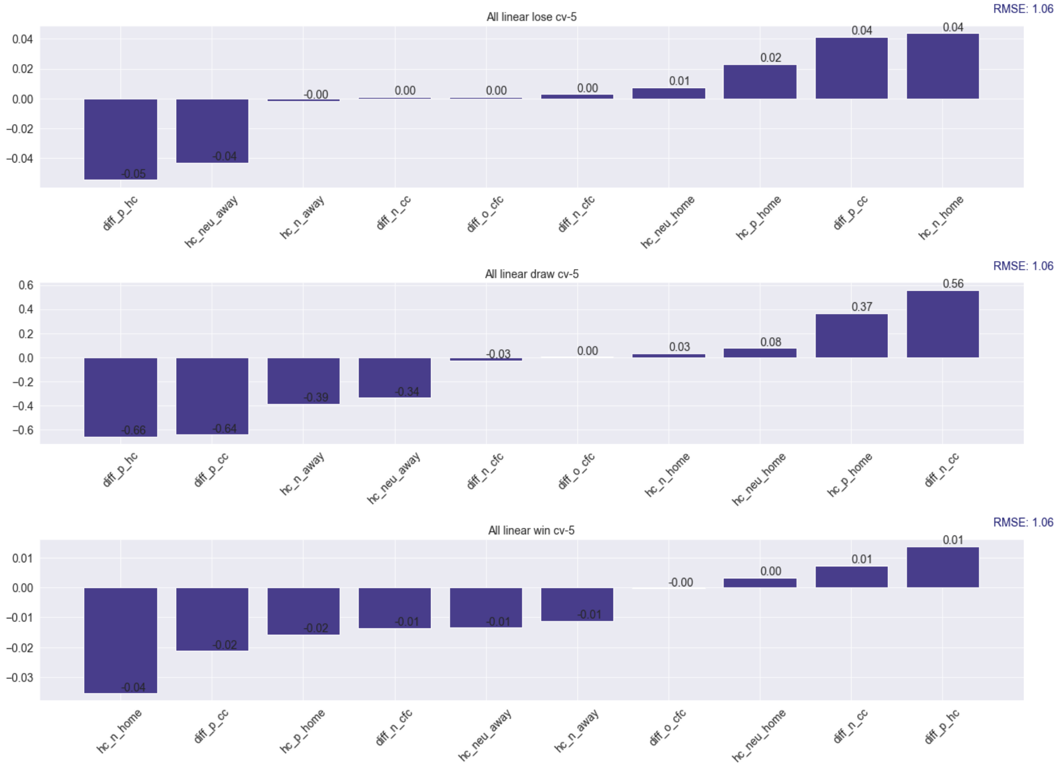


Figure 4.11: SVM's weight bars on tweet's weight imbalance

Table 4.5: Classification report when selecting features from tweet's inverted polarity.

Classification Report				
	precision	**recall**	**f1-score**	**support**
-1	0.52	0.61	0.56	23
0	0.00	0.00	0.00	9
1	0.52	0.64	0.57	22
accuracy avg/total	0.52	0.52	0.52	54

a non-significant measure since this metric does not qualify as one of the best ANOVA test results. It could be presumed that the average path weight from the originated vertex computed by current-flow closeness does not represent the network as good, as selecting the shortest path, as the other closeness variations do.

4.3 The Champion

4.3.1 Sentiments Versus FootballMLP

Figure 4.13 compares the 54 outcomes in Premier Leagues' study by identifying the confusion matrix from FootbalMLP and the Sentiments models, while Table 4.6 shows FootballMLP's classification report in the same set of matches.

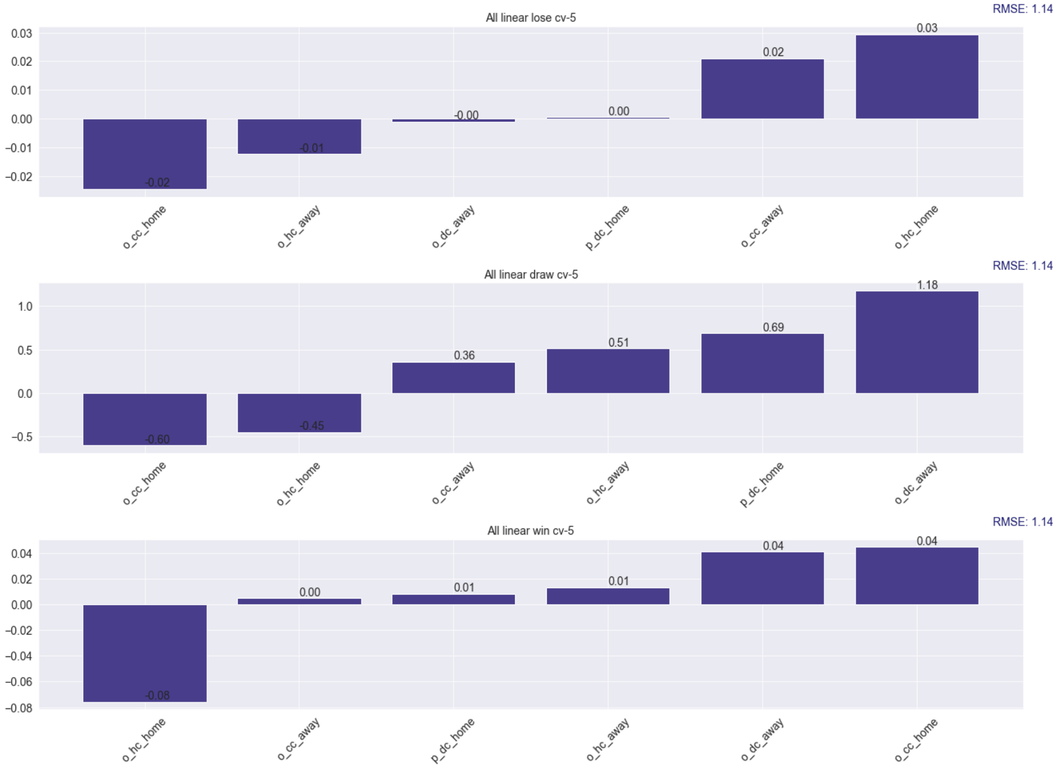


Figure 4.12: SVM's weight bars on the inverted polarity

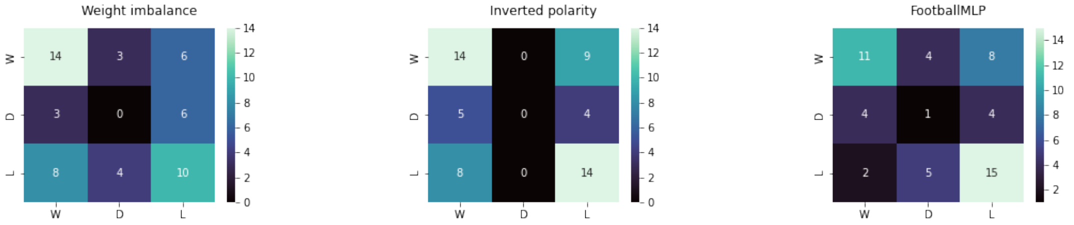


Figure 4.13: Confusion matrix comparing sentiments and statistics.

Table 4.6: Classification report when predicting with FootballMLP.

Classification Report				
	precision	**recall**	**f1-score**	**support**
-1	0.65	0.48	0.55	23
0	0.10	0.11	0.11	9
1	0.56	0.68	0.61	22
accuracy avg/total	0.50	0.50	0.50	54

As Figure 4.13 suggests, both of the sentiments models gave better results when predicting losses at home. The inverted polarity model can discriminate between wins and losses. This model does have better accuracy but is not capable of classifying draws, due to an imbalance problem.

4.3.2 The Hybrid Model

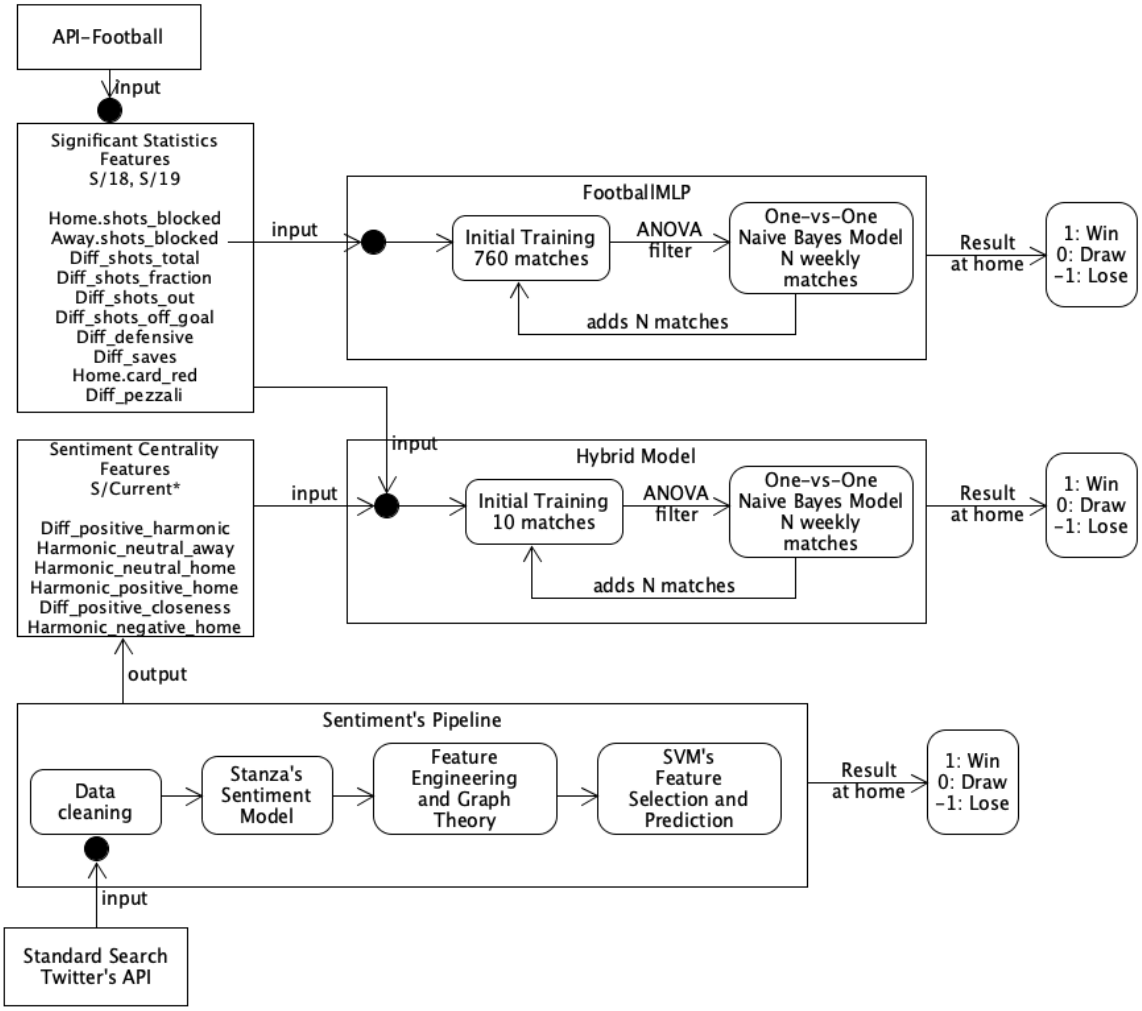


Figure 4.14: Overview of The Hybrid Model

Figure 4.3.2 shows The Hybrid Model as an extension of the significant features found in FootballMLP and the Sentiment Model. For evaluation purposes, the procedure was to keep the sequence of the season. This time 44 matches were predicted with a starting training set of 10 matches. The resulting variables from feature selection on Section **??** were added to the ones on Section 3.1.2 and those were fed into the voting model.

Stats and Weight Imbalance

Stats and Inverted Polarity

The hybrid model in Table 4.7 and Table 4.8 reflected an average accuracy in this 44 matches of 0.55. As well, the SVM coefficients on Figure 4.15 and Figure 4.16 showed the relevance

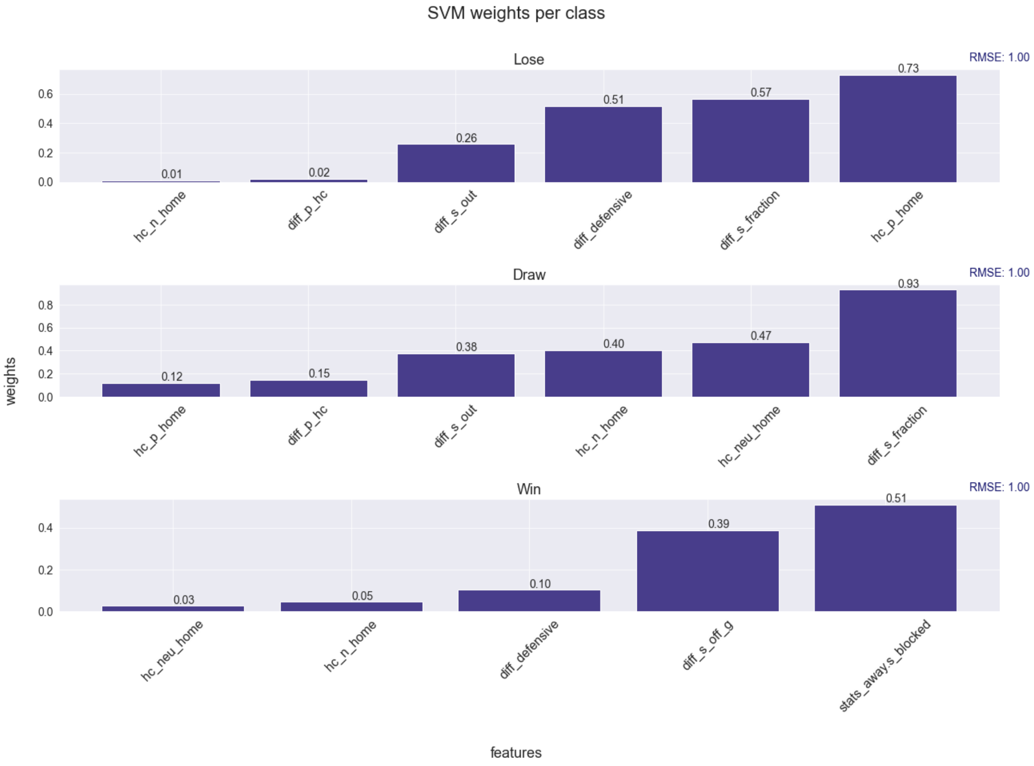


Figure 4.15: SVM's weight bars on the weight imbalance and stats

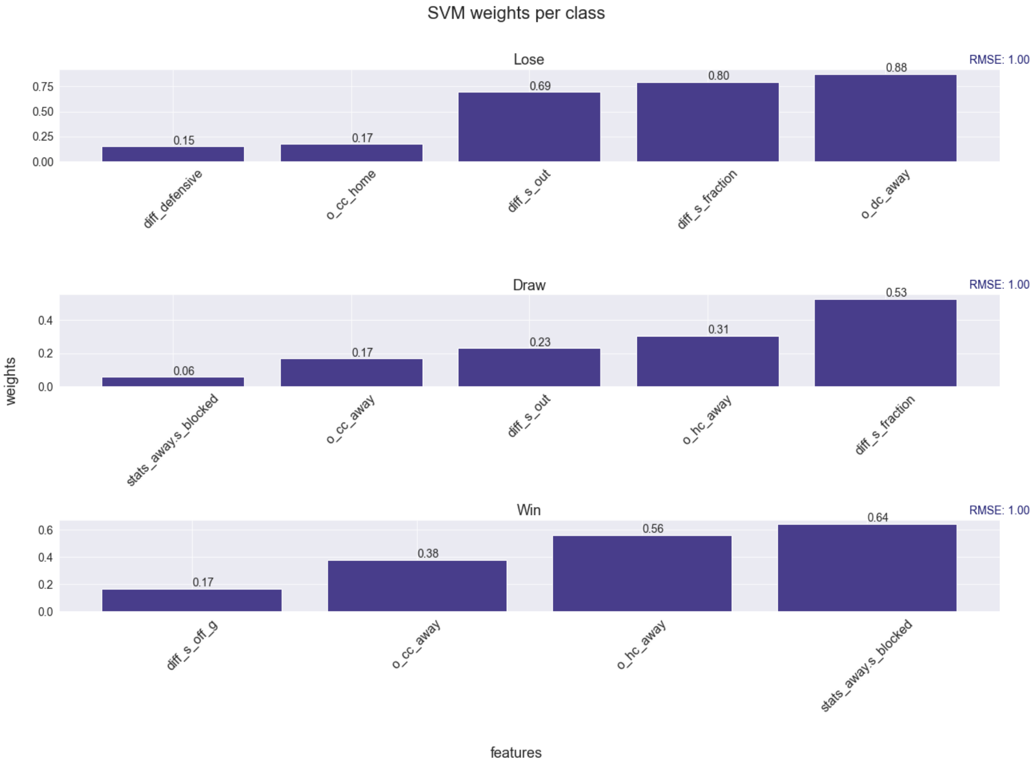


Figure 4.16: SVM's weight bars on the inverted polarity and stats

Table 4.7: Classification report when using Stats and Weight Imbalance Features.

Classification Report				
	precision	**recall**	**f1-score**	**support**
-1	0.52	0.74	0.61	19
0	1.00	0.14	0.25	7
1	0.56	0.50	0.53	18
accuracy avg/total	0.55	0.55	0.55	44

Table 4.8: Classification report when using Stats and Inverted Polarity Features.

Classification Report				
	precision	**recall**	**f1-score**	**support**
-1	0.54	0.68	0.60	19
0	1.00	0.14	0.25	7
1	0.53	0.56	0.54	18
accuracy avg/total	0.55	0.55	0.55	44

when combining sentiment metrics and statistics. The experiments suggest that the inverted polarity characteristics are better to discriminate between wins and loses at home.

4.3.3 Benchmark RPS

According to A Bayesian Approach to In-Game Win Probability in Soccer, it is important to quantify how close is the predicted probability to the class outcome [41], for that the Ranked Probability Score is proposed. This score could be seen as an error score, the larger RPS is, the larger the probability differs from the actual outcome.

The equation 2.6 of the **Ranked Probability Score** is:

$$RPS = \tfrac{1}{2}\textstyle\sum_{i=1}^{2}\left(\sum_{j=1}^{i} p_j - \sum_{j=1}^{i} e_j\right)^2 \tag{4.4}$$

Where p = $[P(Y = win|x), P(Y = draw|x), P(Y = loss|x)]$ are the estimated probabilities in a match, and e encodes the final outcome of a game as a win e = [1, 0, 0], a draw e = [0, 1, 0] and a loss e = [0, 0, 1].

RPS by league

Table 4.9 compares FootballMLP's RPS per classifier and league while using first, the predicted outcome class, and second, the calibrated probabilities. The classifiers tested were Naive Bayes, Random Forest, and Support Vector Machine with linear kernel.

RPS by model

This section pretends to benchmark the Hybrid Model against FootballMLP and Hervert's Bayesian model [15]. Both, the statistical models and hybrid models will use a Naive Bayes classifier to obtain the calibrated probabilities since Table 4.9 proofed a better performance on

Table 4.9: Evaluating FootballMLP using RPS

		Outcome				**Probability**		
League	**# Matches**	**NB**	**RF**	**SVM**	**Vote**	**NB**	**RF**	**SVM**
ES	860	0.3638	0.3643	0.3911	0.3911	0.2230	0.3323	0.3632
GB	860	0.3338	0.3685	0.3567	0.3338	0.2106	0.3534	0.3279
IT	849	0.3174	0.3581	0.3663	0.3039	0.2055	0.3378	0.3467
DE	684	0.3661	0.3872	0.3572	0.36	0.2266	0.3686	0.3447
FR	778	0.3663	0.3817	0.3579	0.3548	0.2188	0.3551	0.3429
NL	627	0.3110	0.3078	0.3524	0.2959	0.1994	0.2917	0.3298
PT	682	0.3519	0.3761	0,3724	0.3519	0.2242	0.3574	0.3528
Total/Average	5340	0.3457	0.3653	0,3659	0.3449	**0.2155**	0.3430	0.3443

this type of classifiers. Figure 4.17 shows the difference in the training data between models and explains how future matches are being predicted.

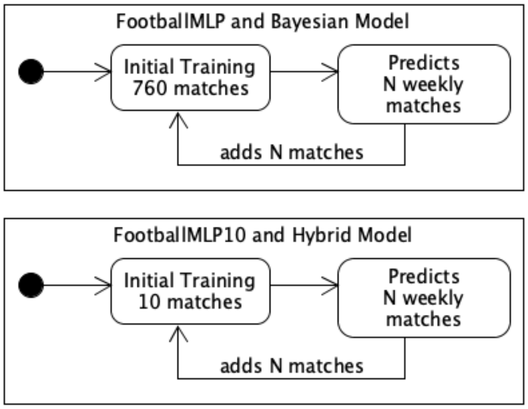

FootballMLP and Bayesian Model
Initial Training 760 matches
Predicts N weekly matches
adds N matches
FootballMLP10 and Hybrid Model
Initial Training 10 matches
Predicts N weekly matches
adds N matches

Figure 4.17: Initial benchmark train set per model

Figure 4.20 compares the models' cumulative rps, while Figure 4.21 contrasts the models' average rps. In total 44 matches were evaluated, Table 4.10 tracks the correlation between the number of total matches and the cumulative rps, the number of matches per week and the average rps, as well the correlation between cumulative and average rps.

Table 4.10: Correlation between # of Matches and RPS

Model	**num_cumrps**	**cum_cumrps**	**cum_avgrps**	**avgrps_cumrps**
FootballMLP	-0.2055	0.9996	-0.6302	-0.6210
Bayes	-0.2106	0.9953	-0.2430	-0.2468
FootballMLP10	-0.2221	0.9970	-0.5669	-0.5307
Hybrid	-0.2177	0.9884	0.1922	0.2014

Each week the number of tested matches varied, so it is important to understand the relationship with its error. This way it can be attempted to determine which model is a better learner through time. The next are the key points found:

- In the four models a negative correlation between the number of matches and its cumulative rps is present, meaning that the error increases with fewer matches predicted in the past, or well decreases with more training matches.

- As expected the number of matches that were predicted is highly correlated to the cumulative rps.

- About the influence of the training set length in each particular predicted match. There is a negative impact for both the statistics and Bayesian model, meaning that as the training set increases the average rps decreases and vice versa. The opposite situation occurred with the hybrid model meaning that as the training set expands, the average rps increases as well. This argument also explains the relationship between average rps and cumulative rps.

As it can be seen in Figure 4.20 over the first 8 weeks the Bayesian Model had a smaller cumulative ranked probability score. However, when looking at Figure 4.21 it's important to remember the initial training set of 10 matches for the hybrid model, and see how its error decreases through each week. After the analysis on Table 4.10 the model that was able to decrease the RPS, at a high rate, while more matches were given was the statistics model from FootballMLP. Figure 4.18 compares the distribution of the predicted classes in a confusion matrix. First, it is compared the models with a large dataset of 760 matches in its training which were Hervert's Bayesian algorithm and FootballMLP. Later, Figure 4.19 compares a narrow version of FootballMLP initially trained with 10 matches, against the extended feature selection on the Hybrid model.

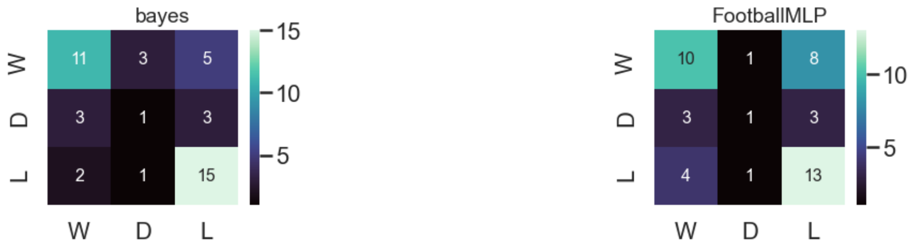


Figure 4.18: Confusion matrix comparing Statistical models with historical data.

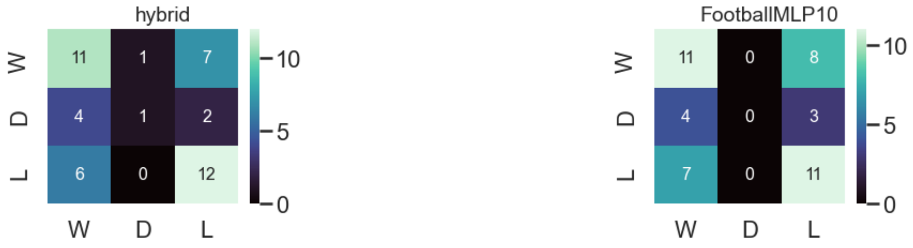


Figure 4.19: Confusion matrix comparing Statistical model and Hybrid model with current data.

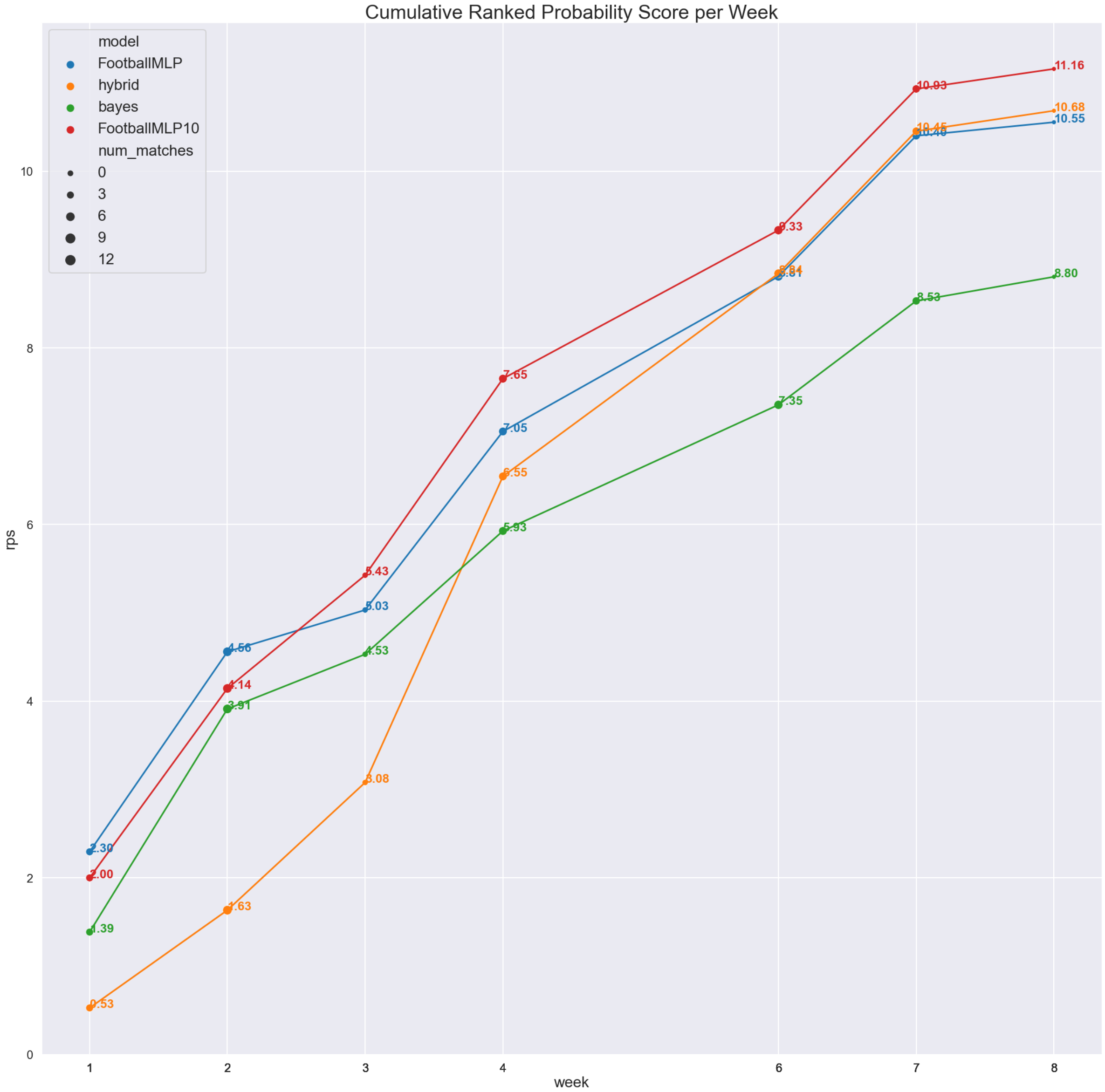


Figure 4.20: Cumulative RPS per week

4.4 Discussion

The purpose of this thesis was to bring up match characteristics that measured Fan Avidity and to see how those features are relevant when compared to other fixture's attributes by correctly predicting a fixture result.

In the first part of the work, FootballMLP tries to narrow the barriers between Machine Learning and Sports Analytics, by studying team's performance per league, and also, by training a predictive pipeline that generalizes well for several leagues, and especially making this pipeline able to interpret the results from several football characteristics.

Section 3.1 beyond showing which characteristics can discriminate match performance, compares several Machine Learning models and concludes Naive Bayes as the best model for

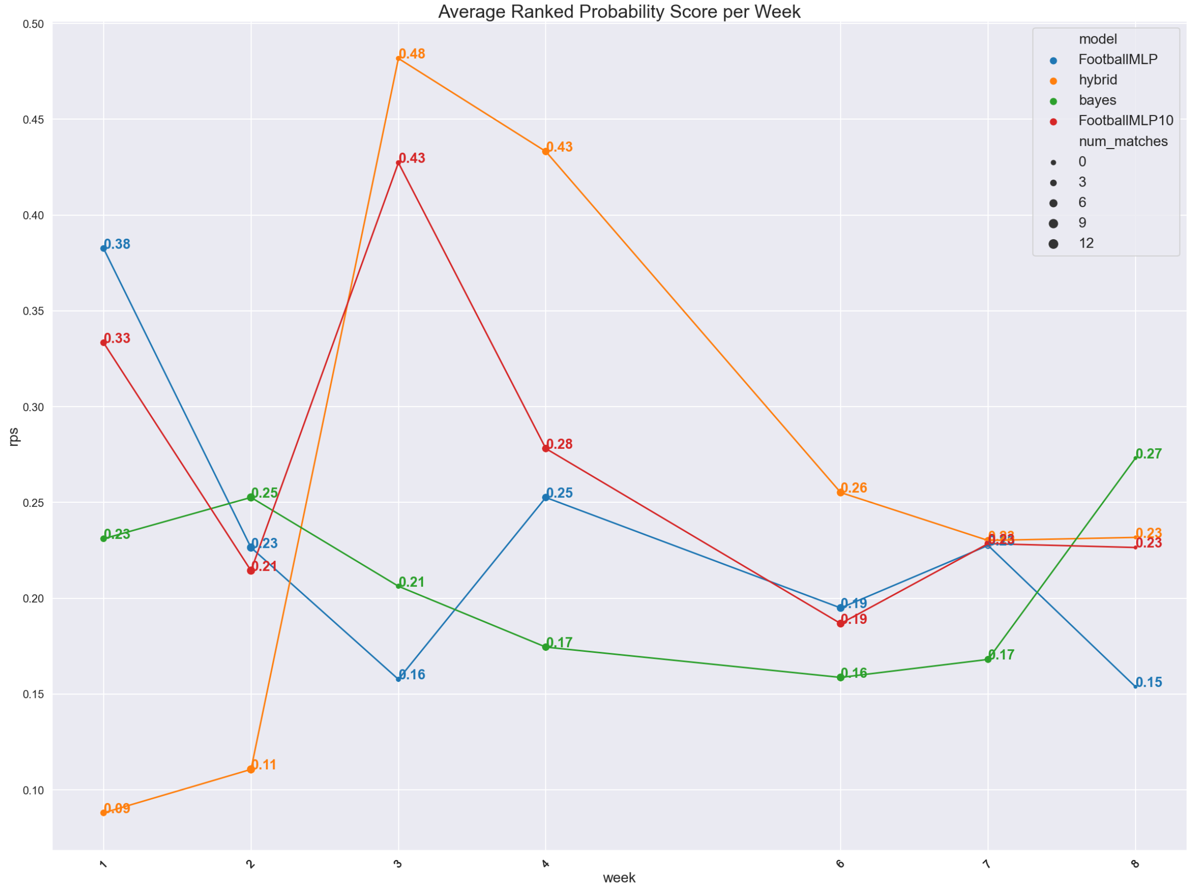


Figure 4.21: Average RPS per week

match classification. Further, this section proposes ensemble methods for solving the disparity problem when the opponent's performance is similar and shows a lift of 0.2 in the F1-score of draws.

Section 3.2 mines Social Networks from football conversations, specifically, it gathers the tweets before a match and computes closeness centrality in a network given a sentiment polarity to measure Fan Avidity in a team and between teams.

The Hybrid Model in Section 4.3.2 aggregates the closeness centrality measures to the performance statistics, when this combination is used on FootballMLP, which pipeline assures to keep the significant features, it is showed that closeness for the away team has a huge weight when classifying the match.

Table 4.11 is a summary of the models presented in this thesis, versus a set of the models found in the literature. The primary condition whether to include or not a model using statistics was to identify its evaluation following the season order, instead of a random train/test split, sentiment's models were not subjected to this condition, since they do not neef to perform aggregates on team's statistics. Also, it is important to emphasize that these models are tested in different competitions and leagues, as well, in different periods. However, the objective was still the same, which was to overcome the baseline 50% accuracy.

The min accuracy column helps to identify when testing multiple models the next two questions: how well the data is a descriptor of the competition? is it constant, or is it luck?

Table 4.11: Match prediction benchmark.

Model	Competition	Seasons	Initial Train Set	Test Set Matches	Min Accuracy	Max Accuracy	Avg Accuracy
Bayes Model	England	20’	760	44	-	-	0.61
Hybrid Model	England	20’	10	44	-	-	0.55
FootballMLP	Spain, England Italy, Germany France, Dutch Portugal	18’, 19’, 20’	10	5340	0.43	0.56	0.50
Sentiment Model	England	20’	-	54	0.44	0.52	0.48
The Wisdom of The Silent Crowd (SA)	World Cup	18’	56	8	0.25	0.87	0.55
Predicting wins and spread (SA)	England	13’	-	122	0.38	0.50	0.45
The harsh rule of the goals	Spain, England Italy, Germany	13’	10	1446	0.49	0.60	0.53
Predicting Football Results	Spain, England Italy, Germany France	14’, 15’	10	3800	0.42	0.51	0.47

For example, in the paper the harsh rule of the goals [6] the Nearest Neighbor model in Germany has the maximum accuracy of 60% and the minimum accuracy of 49% in Spain when using Naive Bayes model. Meanwhile, Predicting Football Results Using Machine Learning Techniques [12] finds the highest accuracy of 51.1% when using expected goals instead of goals in SVC linear classification, and the lowest accuracy of 44.6% when performing the match score linear regression and then doing a Poisson Distribution to generate match probabilities. It should be highlighted that FootballMLP is the larger tested model with over 5,000 matches predicted. Also, if the sentiments model on its own is compared to Schumaker [44] with the highest accuracy of 50.49%, the current model has a slightly better performance with 52% accuracy. However, it is still a limitation gathering and processing this information for extending the evaluation.

In the end, some challenges are found when evaluating a match result. First, when comparing how well labels are classified, and second how well probability is calculated. For the 44 matches benchmarked the distribution is as follows 19 wins, 7 draws, and 18 losses.

For the first part of the evaluation, Figure 4.18 compares models with larger training datasets, which we will refer to as models trained with history, while Figure 4.19 are models with a tight window of the current season. It is interesting how historical models tend to predict more the losing class when compared to the models trained on the latest results. As explained earlier; FootballMLP, FooballMLP10, and the hybrid model follow the same pipeline, but are trained differently, this ensemble mechanism was effective to balance the decision of predicting a draw, when compared to Hervert’s bayesian model. Also, the hybrid model, with a variety of features, showed slightly better performance on draws and losses when compared to FootballMLP10.

The second part of the evaluation deals with the Ranking Probability Score, It is important to point out that larger probabilities in the draws class will minimize the error when misclassifying wins and losses, so overpredicting draws with a large probability will narrow down the cumulative error as in Hervert's bayesian model. Figure 4.21 that plots the average RPS per match in a week, shows how FootballMLP's historical knowledge starts with an accurate score and recognizes how the model is learning through time. While the hybrid model, with stochastic features from sentiment analysis, is able to perform extremely well with no knowledge, and how it is sensitive to noise and larger time frames.

Chapter 5

Conclusion

This thesis' source code is open access and it can be found under the FootballMLP project [29]. The Winning Mining Formula for Football Competitions extends the evaluation of FootballMLP by gathering bookies' odds and calculating the payout of the model. Predicting Soccer Results through Sentiment Analysis [31] is the final deliverable for the Sentiment Analysis study, while Social Networks as Room for Improvement in Football [30] reviews the literature on the way soccer institutions engage fans through social media. And shows how this technology applies artificial intelligence by analyzing sentiment in these massive conversations.

The main research question *Based on Graph Theory, could users' sentiment be evaluated as a metric to score Fan Avidity?* is solved on Section 3.2. Even though there is existing research on sentiment analysis for soccer prediction, this is the first time a social mining approach is used, when engineering characteristics from centrality measures. Also, the idea of solving unbalance problems by computing edges' weights relative to the size of the network, and to the reach of a tweet by encountering the number of likes and shares, is unique.

The second question *Is Fan avidity a determinant feature when predicting results previous to a match?* I will answer yes, the ANOVA pipeline in the Hybrid model proofed that closeness centrality in the away team is a significant feature for the prediction. Questions number three and four try to compare the match prediction accuracy from the statistical model to a model with the context characteristics. It is a surprise that the hybrid model performs extremely well with less training information, however, the statistical model, as seen in FootballMLP, can decrease the error while understanding the team's performance as the season advances.

The challenge that is faced for the hybrid model is capturing conversations before a match, however, this information is interchangeable when needing fewer historical records. When considering sentiment features only, this prediction mechanism can be decoupled from football leagues and even sport, from the fact that it does not consider statistics at all. FootballMLP instead needs a set of repeatable events to work out, the best finding is the ensemble model which can solve the disparity problem when predicting draws.

Future research must try to study the performance behavior of teams' regressions into Time Series Analysis. Also, it is important to extend the results from the mixture of football statistics and sentiment analysis models and to expand this study into several leagues. However, the sentiment analysis field faces some challenges, for example, the lack of multilingual models, since current research is mostly focused on the English language. As mentioned earlier, the advantage of scoring sentiment is to decouple the understanding of statistics that change from sport to sport. The development of a framework that gathers conversation before a game starts on different sports and performs sentiment analysis prediction is something to look forward to. Also, the design of metrics that evaluate the models' accuracy, in order, to penalize in a more accurate way different models classification and probability results is something to work on.

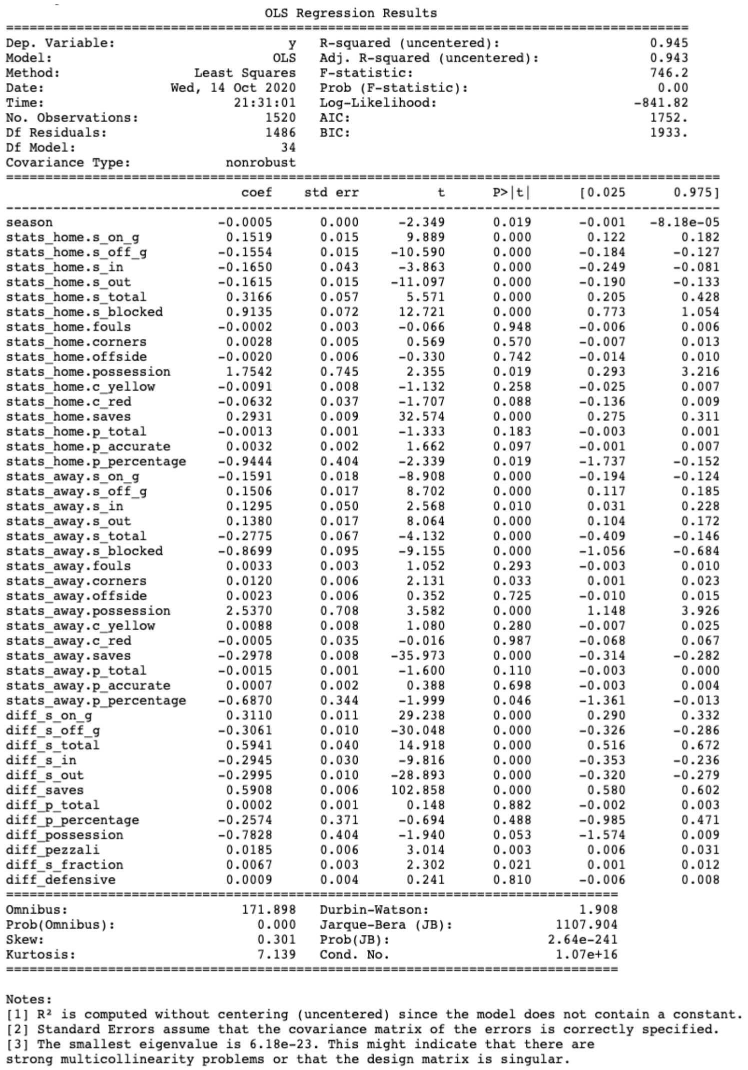
Finally, this thesis is truly inspired by the unpredictability of sports and football pools. The aim was to capture learning from different information systems such as historical statistics and fan's empirical knowledge. Now it is possible to suggest that football leagues are more stable from weeks 24 to 32 and leagues such as Eredivise, Premier League, and Serie A are easier to predict, and draws are phenomena independent from time.

Chapter 6

Appendix

6.1 Model Summary

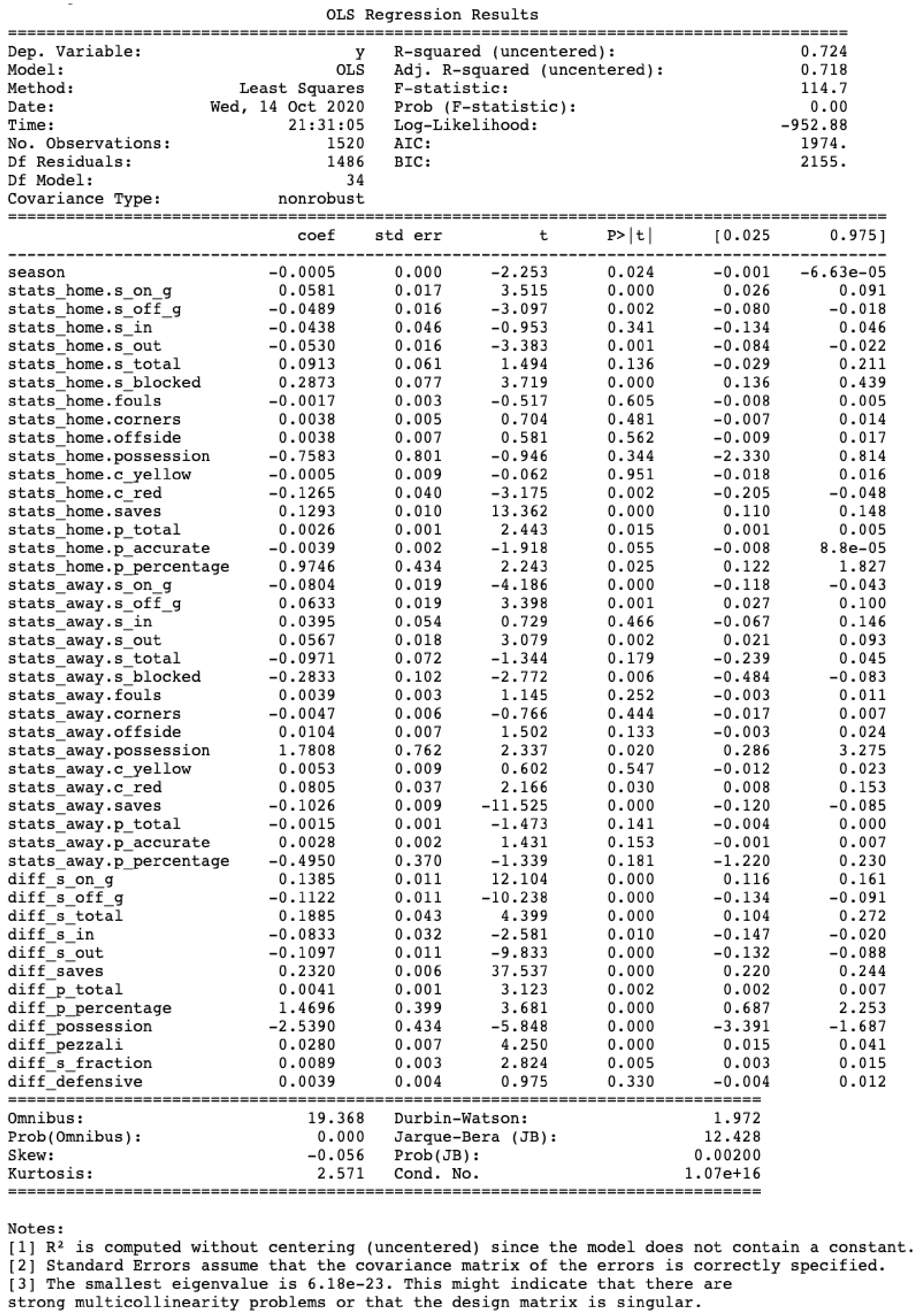
6.1.1 OLS Goals Difference



```
                                 OLS Regression Results
=======================================================================================
Dep. Variable:                      y   R-squared (uncentered):                   0.945
Model:                            OLS   Adj. R-squared (uncentered):              0.943
Method:                 Least Squares   F-statistic:                              746.2
Date:                Wed, 14 Oct 2020   Prob (F-statistic):                        0.00
Time:                        21:31:01   Log-Likelihood:                         -841.82
No. Observations:                1520   AIC:                                      1752.
Df Residuals:                    1486   BIC:                                      1933.
Df Model:                          34
Covariance Type:            nonrobust
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
season                     -0.0005      0.000     -2.349      0.019      -0.001   -8.18e-05
stats_home.s_on_g           0.1519      0.015      9.889      0.000       0.122       0.182
stats_home.s_off_g         -0.1554      0.015    -10.590      0.000      -0.184      -0.127
stats_home.s_in            -0.1650      0.043     -3.863      0.000      -0.249      -0.081
stats_home.s_out           -0.1615      0.015    -11.097      0.000      -0.190      -0.133
stats_home.s_total          0.3166      0.057      5.571      0.000       0.205       0.428
stats_home.s_blocked        0.9135      0.072     12.721      0.000       0.773       1.054
stats_home.fouls           -0.0002      0.003     -0.066      0.948      -0.006       0.006
stats_home.corners          0.0028      0.005      0.569      0.570      -0.007       0.013
stats_home.offside         -0.0020      0.006     -0.330      0.742      -0.014       0.010
stats_home.possession       1.7542      0.745      2.355      0.019       0.293       3.216
stats_home.c_yellow        -0.0091      0.008     -1.132      0.258      -0.025       0.007
stats_home.c_red           -0.0632      0.037     -1.707      0.088      -0.136       0.009
stats_home.saves            0.2931      0.009     32.574      0.000       0.275       0.311
stats_home.p_total         -0.0013      0.001     -1.333      0.183      -0.003       0.001
stats_home.p_accurate       0.0032      0.002      1.662      0.097      -0.001       0.007
stats_home.p_percentage    -0.9444      0.404     -2.339      0.019      -1.737      -0.152
stats_away.s_on_g          -0.1591      0.018     -8.908      0.000      -0.194      -0.124
stats_away.s_off_g          0.1506      0.017      8.702      0.000       0.117       0.185
stats_away.s_in             0.1295      0.050      2.568      0.010       0.031       0.228
stats_away.s_out            0.1380      0.017      8.064      0.000       0.104       0.172
stats_away.s_total         -0.2775      0.067     -4.132      0.000      -0.409      -0.146
stats_away.s_blocked       -0.8699      0.095     -9.155      0.000      -1.056      -0.684
stats_away.fouls            0.0033      0.003      1.052      0.293      -0.003       0.010
stats_away.corners          0.0120      0.006      2.131      0.033       0.001       0.023
stats_away.offside          0.0023      0.006      0.352      0.725      -0.010       0.015
stats_away.possession       2.5370      0.708      3.582      0.000       1.148       3.926
stats_away.c_yellow         0.0088      0.008      1.080      0.280      -0.007       0.025
stats_away.c_red           -0.0005      0.035     -0.016      0.987      -0.068       0.067
stats_away.saves           -0.2978      0.008    -35.973      0.000      -0.314      -0.282
stats_away.p_total         -0.0015      0.001     -1.600      0.110      -0.003       0.000
stats_away.p_accurate       0.0007      0.002      0.388      0.698      -0.003       0.004
stats_away.p_percentage    -0.6870      0.344     -1.999      0.046      -1.361      -0.013
diff_s_on_g                 0.3110      0.011     29.238      0.000       0.290       0.332
diff_s_off_g               -0.3061      0.010    -30.048      0.000      -0.326      -0.286
diff_s_total                0.5941      0.040     14.918      0.000       0.516       0.672
diff_s_in                  -0.2945      0.030     -9.816      0.000      -0.353      -0.236
diff_s_out                 -0.2995      0.010    -28.893      0.000      -0.320      -0.279
diff_saves                  0.5908      0.006    102.858      0.000       0.580       0.602
diff_p_total                0.0002      0.001      0.148      0.882      -0.002       0.003
diff_p_percentage          -0.2574      0.371     -0.694      0.488      -0.985       0.471
diff_possession            -0.7828      0.404     -1.940      0.053      -1.574       0.009
diff_pezzali                0.0185      0.006      3.014      0.003       0.006       0.031
diff_s_fraction             0.0067      0.003      2.302      0.021       0.001       0.012
diff_defensive              0.0009      0.004      0.241      0.810      -0.006       0.008
==============================================================================
Omnibus:                      171.898   Durbin-Watson:                   1.908
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1107.904
Skew:                           0.301   Prob(JB):                    2.64e-241
Kurtosis:                       7.139   Cond. No.                     1.07e+16
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[3] The smallest eigenvalue is 6.18e-23. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

6.1.2 OLS Fixtures' Results

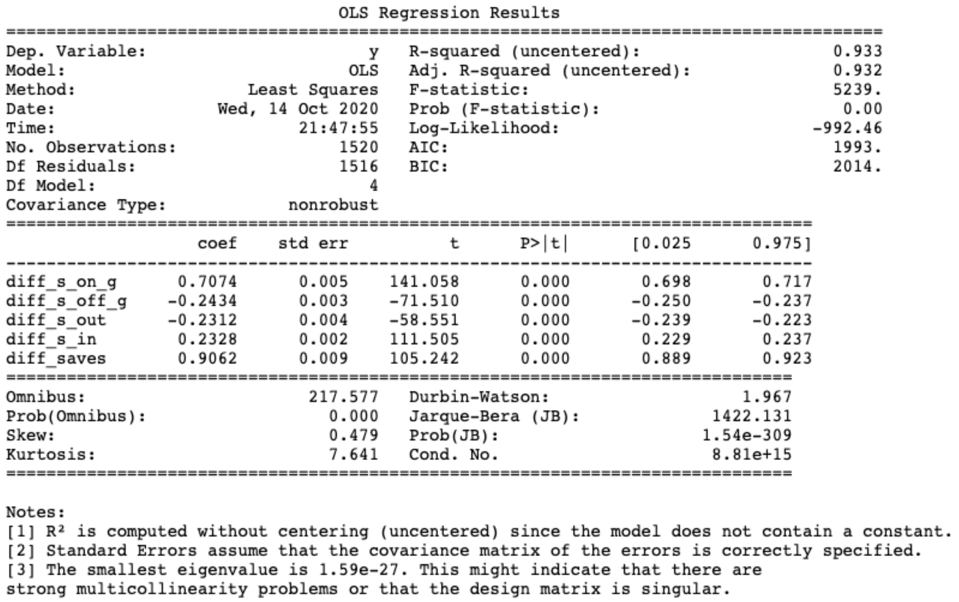


```
                                 OLS Regression Results
=======================================================================================
Dep. Variable:                      y   R-squared (uncentered):                   0.724
Model:                            OLS   Adj. R-squared (uncentered):              0.718
Method:                 Least Squares   F-statistic:                              114.7
Date:                Wed, 14 Oct 2020   Prob (F-statistic):                        0.00
Time:                        21:31:05   Log-Likelihood:                         -952.88
No. Observations:                1520   AIC:                                      1974.
Df Residuals:                    1486   BIC:                                      2155.
Df Model:                          34
Covariance Type:            nonrobust
===========================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
season                    -0.0005      0.000     -2.253      0.024      -0.001   -6.63e-05
stats_home.s_on_g          0.0581      0.017      3.515      0.000       0.026       0.091
stats_home.s_off_g        -0.0489      0.016     -3.097      0.002      -0.080      -0.018
stats_home.s_in           -0.0438      0.046     -0.953      0.341      -0.134       0.046
stats_home.s_out          -0.0530      0.016     -3.383      0.001      -0.084      -0.022
stats_home.s_total         0.0913      0.061      1.494      0.136      -0.029       0.211
stats_home.s_blocked       0.2873      0.077      3.719      0.000       0.136       0.439
stats_home.fouls          -0.0017      0.003     -0.517      0.605      -0.008       0.005
stats_home.corners         0.0038      0.005      0.704      0.481      -0.007       0.014
stats_home.offside         0.0038      0.007      0.581      0.562      -0.009       0.017
stats_home.possession     -0.7583      0.801     -0.946      0.344      -2.330       0.814
stats_home.c_yellow       -0.0005      0.009     -0.062      0.951      -0.018       0.016
stats_home.c_red          -0.1265      0.040     -3.175      0.002      -0.205      -0.048
stats_home.saves           0.1293      0.010     13.362      0.000       0.110       0.148
stats_home.p_total         0.0026      0.001      2.443      0.015       0.001       0.005
stats_home.p_accurate     -0.0039      0.002     -1.918      0.055      -0.008     8.8e-05
stats_home.p_percentage    0.9746      0.434      2.243      0.025       0.122       1.827
stats_away.s_on_g         -0.0804      0.019     -4.186      0.000      -0.118      -0.043
stats_away.s_off_g         0.0633      0.019      3.398      0.001       0.027       0.100
stats_away.s_in            0.0395      0.054      0.729      0.466      -0.067       0.146
stats_away.s_out           0.0567      0.018      3.079      0.002       0.021       0.093
stats_away.s_total        -0.0971      0.072     -1.344      0.179      -0.239       0.045
stats_away.s_blocked      -0.2833      0.102     -2.772      0.006      -0.484      -0.083
stats_away.fouls           0.0039      0.003      1.145      0.252      -0.003       0.011
stats_away.corners        -0.0047      0.006     -0.766      0.444      -0.017       0.007
stats_away.offside         0.0104      0.007      1.502      0.133      -0.003       0.024
stats_away.possession      1.7808      0.762      2.337      0.020       0.286       3.275
stats_away.c_yellow        0.0053      0.009      0.602      0.547      -0.012       0.023
stats_away.c_red           0.0805      0.037      2.166      0.030       0.008       0.153
stats_away.saves          -0.1026      0.009    -11.525      0.000      -0.120      -0.085
stats_away.p_total        -0.0015      0.001     -1.473      0.141      -0.004       0.000
stats_away.p_accurate      0.0028      0.002      1.431      0.153      -0.001       0.007
stats_away.p_percentage   -0.4950      0.370     -1.339      0.181      -1.220       0.230
diff_s_on_g                0.1385      0.011     12.104      0.000       0.116       0.161
diff_s_off_g              -0.1122      0.011    -10.238      0.000      -0.134      -0.091
diff_s_total               0.1885      0.043      4.399      0.000       0.104       0.272
diff_s_in                 -0.0833      0.032     -2.581      0.010      -0.147      -0.020
diff_s_out                -0.1097      0.011     -9.833      0.000      -0.132      -0.088
diff_saves                 0.2320      0.006     37.537      0.000       0.220       0.244
diff_p_total               0.0041      0.001      3.123      0.002       0.002       0.007
diff_p_percentage          1.4696      0.399      3.681      0.000       0.687       2.253
diff_possession           -2.5390      0.434     -5.848      0.000      -3.391      -1.687
diff_pezzali               0.0280      0.007      4.250      0.000       0.015       0.041
diff_s_fraction            0.0089      0.003      2.824      0.005       0.003       0.015
diff_defensive             0.0039      0.004      0.975      0.330      -0.004       0.012
==============================================================================
Omnibus:                       19.368   Durbin-Watson:                   1.972
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               12.428
Skew:                          -0.056   Prob(JB):                      0.00200
Kurtosis:                       2.571   Cond. No.                     1.07e+16
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[3] The smallest eigenvalue is 6.18e-23. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

6.2 Model Summary Feature Selection

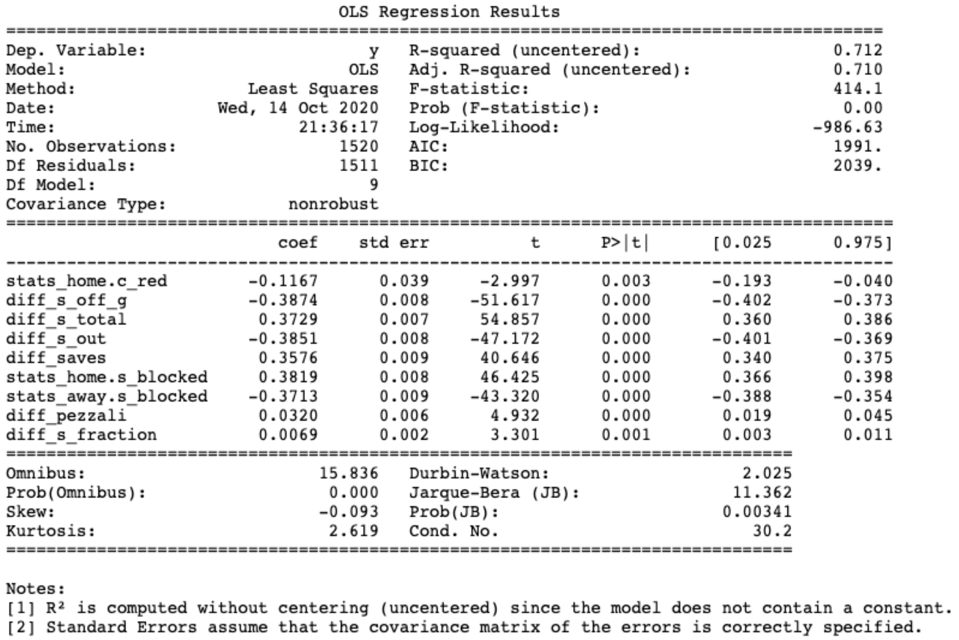
6.2.1 OLS Goals Diffrence



```
                                 OLS Regression Results                                
=======================================================================================
Dep. Variable:                      y   R-squared (uncentered):                   0.933
Model:                            OLS   Adj. R-squared (uncentered):              0.932
Method:                 Least Squares   F-statistic:                              5239.
Date:                Wed, 14 Oct 2020   Prob (F-statistic):                        0.00
Time:                        21:47:55   Log-Likelihood:                         -992.46
No. Observations:                1520   AIC:                                      1993.
Df Residuals:                    1516   BIC:                                      2014.
Df Model:                           4                                                  
Covariance Type:            nonrobust                                                  
================================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
diff_s_on_g     0.7074      0.005    141.058      0.000       0.698       0.717
diff_s_off_g   -0.2434      0.003    -71.510      0.000      -0.250      -0.237
diff_s_out     -0.2312      0.004    -58.551      0.000      -0.239      -0.223
diff_s_in       0.2328      0.002    111.505      0.000       0.229       0.237
diff_saves      0.9062      0.009    105.242      0.000       0.889       0.923
==============================================================================
Omnibus:                      217.577   Durbin-Watson:                   1.967
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1422.131
Skew:                           0.479   Prob(JB):                    1.54e-309
Kurtosis:                       7.641   Cond. No.                     8.81e+15
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[3] The smallest eigenvalue is 1.59e-27. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

6.2.2 OLS Fixtures' Results



```
                                 OLS Regression Results                                
=======================================================================================
Dep. Variable:                      y   R-squared (uncentered):                   0.712
Model:                            OLS   Adj. R-squared (uncentered):              0.710
Method:                 Least Squares   F-statistic:                              414.1
Date:                Wed, 14 Oct 2020   Prob (F-statistic):                        0.00
Time:                        21:36:17   Log-Likelihood:                         -986.63
No. Observations:                1520   AIC:                                      1991.
Df Residuals:                    1511   BIC:                                      2039.
Df Model:                           9                                                  
Covariance Type:            nonrobust                                                  
========================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------
stats_home.c_red       -0.1167      0.039     -2.997      0.003      -0.193      -0.040
diff_s_off_g           -0.3874      0.008    -51.617      0.000      -0.402      -0.373
diff_s_total            0.3729      0.007     54.857      0.000       0.360       0.386
diff_s_out             -0.3851      0.008    -47.172      0.000      -0.401      -0.369
diff_saves              0.3576      0.009     40.646      0.000       0.340       0.375
stats_home.s_blocked    0.3819      0.008     46.425      0.000       0.366       0.398
stats_away.s_blocked   -0.3713      0.009    -43.320      0.000      -0.388      -0.354
diff_pezzali            0.0320      0.006      4.932      0.000       0.019       0.045
diff_s_fraction         0.0069      0.002      3.301      0.001       0.003       0.011
==============================================================================
Omnibus:                       15.836   Durbin-Watson:                   2.025
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               11.362
Skew:                          -0.093   Prob(JB):                      0.00341
Kurtosis:                       2.619   Cond. No.                         30.2
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

6.3 Classification Report per League

Table 6.1: Classification report in Spain's League.

Classification Report				
	precision	**recall**	**f1-score**	**support**
-1	0.35	0.49	0.41	233
0	0.33	0.22	0.26	243
1	0.56	0.54	0.55	384
accuracy avg/total	0.43	0.43	0.43	860

Table 6.2: Classification report in England's League.

Classification Report				
	precision	**recall**	**f1-score**	**support**
-1	0.54	0.64	0.59	287
0	0.27	0.17	0.21	184
1	0.60	0.62	0.61	389
accuracy avg/total	0.53	0.53	0.53	860

Table 6.3: Classification report in Italy's League.

Classification Report				
	precision	**recall**	**f1-score**	**support**
-1	0.56	0.54	0.55	277
0	0.32	0.29	0.31	215
1	0.61	0.67	0.64	357
accuracy avg/total	0.53	0.53	0.53	849

Table 6.4: Classification report in Germany's League.

Classification Report				
	precision	**recall**	**f1-score**	**support**
-1	0.54	0.39	0.45	235
0	0.25	0.20	0.22	165
1	0.52	0.70	0.59	284
accuracy avg/total	0.47	0.47	0.47	684

Table 6.5: Classification report in France's League.

Classification Report				
	precision	**recall**	**f1-score**	**support**
-1	0.45	0.41	0.43	222
0	0.32	0.23	0.27	208
1	0.55	0.67	0.60	348
accuracy avg/total	0.48	0.48	0.48	778

Table 6.6: Classification report in Dutch's League.

Classification Report				
	precision	**recall**	**f1-score**	**support**
-1	0.51	0.60	0.55	181
0	0.29	0.23	0.26	135
1	0.69	0.68	0.68	311
accuracy avg/total	0.56	0.56	0.56	627

Table 6.7: Classification report in Portugal's League.

Classification Report				
	precision	**recall**	**f1-score**	**support**
-1	0.50	0.50	0.50	225
0	0.26	0.15	0.19	159
1	0.56	0.68	0.61	298
accuracy avg/total	0.50	0.50	0.50	682

Bibliography

[1] Standard search api — docs — twitter developer platform.

[2] ALTARRIBA-BARTÉS, A., CALLE, M., SUSÍN, A., GONÇALVES, B., VIVES, M., SAMPAIO, J., AND PEÑA, J. Analysis of the winning probability and the scoring actions in the american professional soccer championship. *RICYDE. Revista internacional de ciencias del deporte 16* (01 2020), 67–84.

[3] AMRIT, C., AND TER MAAT, J. Understanding Information Centrality Metric: A Simulation Approach.

[4] CHASSY, P. Team play in football: How science supports f. c. barcelona's training strategy. *Psychology 04* (01 2013), 7–12.

[5] CHEN, H. Neural network algorithm in predicting football match outcome based on player ability index. *Advances in Physical Education 09* (01 2019), 215–222.

[6] CINTIA, P., GIANNOTTI, F., PAPPALARDO, L., PEDRESCHI, D., AND MALVALDI, M. The harsh rule of the goals: Data-driven performance indicators for football teams. *Proceedings of the 2015 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2015*, December (2015).

[7] CRUZ, X. La generación de una lovemark a través del marketing deportivo. Caso: Selección Peruana de Fútbol. 174.

[8] DESARBO, W. S., AND MADRIGAL, R. Examining the behavioral manifestations of fan avidity in sports marketing. *Journal of Modelling in Management 6*, 1 (2011), 79–99.

[9] DHARMARAJAN, K., ABUTHAHEER, F., AND ABIRAMI, K. Sentiment analysis on social media. 210–217.

[10] ELAAD, G. Home-field advantage and biased prediction markets in English soccer. *Applied Economics Letters 27*, 14 (2020), 1170–1174.

[11] GOUGH, C. Soccer, Apr 2020.

[12] HERBINET, C. Predicting Football Results Using Machine Learning Techniques. *2011 Proceedings of the 34th International Convention MIPRO 48* (2018), 1623–1627.

[13] HERNÁNDEZ GONZÁLEZ, D., AND RECORDER RENTERAL, A. G. *Historia de la Actividad Física y el Deporte*. Impresos Chávez de la Cruz, S.A. de C.V., 2015.

[14] HEROLD, M., GOES, F., NOPP, S., BAUER, P., THOMPSON, C., AND MEYER, T. Machine learning in men's professional football: Current applications and future directions for improving attacking play. *International Journal of Sports Science and Coaching* (10 2019).

[15] HERVERT-ESCOBAR, L., MATIS, T. I., AND HERNANDEZ-GRESS, N. Prediction learning model for soccer matches outcomes. In *2018 Seventeenth Mexican International Conference on Artificial Intelligence (MICAI)* (Oct 2018), pp. 63–69.

[16] HUAN, L., RICHARD, P., LIREN, S., YUHAO, Y., AND ZHONGZHI, Z. Current flow group closeness centrality for complex networks?. 961 – 971.

[17] HUBÁČEK, O., ŠOUREK, G., AND ŽELEZNÝ, F. Exploiting sports-betting market using machine learning. *International Journal of Forecasting 35*, 2 (2019), 783–796.

[18] JAI-ANDALOUSSI, S., EL MOURABIT, I., MADRANE, N., CHAOUNI, S. B., AND SEKKAKI, A. Soccer events summarization by using sentiment analysis. *Proceedings - 2015 International Conference on Computational Science and Computational Intelligence, CSCI 2015*, September 2018 (2016), 398–403.

[19] KAUNITZ, L., ZHONG, S., AND KREINER, J. Beating the bookies with their own numbers - and how the online sports betting market is rigged.

[20] KIM, Y. Convolutional Neural Networks for Sentence Classification. *CoRR abs/1408.5* (2014).

[21] KIM, Y. S., AND KIM, M. 'A wisdom of crowds': Social media mining for soccer match analysis. *IEEE Access 7* (2019), 52634–52639.

[22] KUMAR, R., AND PATTNAIK, P. K. *Graph Theory.* Laxmi Publications Pvt Ltd, 2018.

[23] LIOTTA, G., TAMASSIA, R., AND TOLLIS, I. G. *Graph algorithms and applications 4.* World Scientific, 2006.

[24] LJAJIĆ, A., LJAJIĆ, E., SPALEVIĆ, P., ARSIĆ, B., AND VUČKOVIĆ, D. Sentiment analysis of textual comments in field of sport Sentiment analysis of textual comments in field of sport.

[25] M NARASIMHA, M., AND V SUSHEELA, D. *Introduction To Pattern Recognition And Machine Learning.* No. vol. 5 in IISc Lecture Notes Series. World Scientific, 2015.

[26] MASSARON, L., BOSCHETTI, A., MASSARON, L., AND BOSCHETTI, A. *Regression Analysis with Python.* 2016.

[27] MCMAHON, B. Revenue of 22.8b: Uefa report shows the few teams making money and the many that are not, Jan 2019.

[28] MIRANDA-PEÑA, C. *API-Footballpy*, 2020. Python 3.

[29] MIRANDA-PEÑA, C. *FootballMLP*, 2021. Python 3.

[30] MIRANDA-PEÑA, C. Las redes sociales un área de oportunidad en el fútbol. *Summa Humanitatis 11*, 2 (ago. 2021), 1–15.

[31] MIRANDA-PEÑA, C., CEBALLOS, H. G., HERVERT-ESCOBAR, L., AND GONZALEZ-MENDOZA, M. Predicting soccer results through sentiment analysis: A graph theory approach. In *Computational Science – ICCS 2021* (Cham, 2021), M. Paszynski, D. Kranzlmüller, V. V. Krzhizhanovskaya, J. J. Dongarra, and P. M. A. Sloot, Eds., Springer International Publishing, pp. 422–435.

[32] MÜLLNER, D. Modern hierarchical, agglomerative clustering algorithms, 2011.

[33] NATINGGA, D. *Data Science Algorithms in a Week.* Packt Publishing, 2017.

[34] PAPPALARDO, L., AND CINTIA, P. Quantifying the relation between performance and success in soccer. *Advances in Complex Systems 21*, 3-4 (2018), 1–29.

[35] PAPPALARDO, L., CINTIA, P., FERRAGINA, P., MASSUCCO, E., PEDRESCHI, D., AND GIANNOTTI, F. PlayeRank: Data-driven performance evaluation and player ranking in soccer via a machine learning approach. *ACM Transactions on Intelligent Systems and Technology 10*, 5 (2019).

[36] PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R., DUBOURG, V., VANDERPLAS, J., PASSOS, A., COURNAPEAU, D., BRUCHER, M., PERROT, M., AND DUCHESNAY, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12* (2011), 2825–2830.

[37] PERETTI, A. A linear model for ranking soccer teams. *Journal of Interdisciplinary Mathematics 22*, 3 (2019), 243–263.

[38] QI, P., ZHANG, Y., ZHANG, Y., BOLTON, J., AND MANNING, C. D. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082* (2020).

[39] RIQUELME, F., GONZALEZ-CANTERGIANI, P., MOLINERO, X., AND SERNA, M. Centrality measure in social networks based on linear threshold model. *Knowledge-Based Systems 140*, January (2018), 92–102.

[40] ROBBERECHTS, P., VAN HAAREN, J., AND DAVIS, J. Who Will Win It? An In-game Win Probability Model for Football. 1–13.

[41] ROBBERECHTS, P., VAN HAAREN, J., AND DAVIS, J. A bayesian approach to in-game win probability in soccer. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery amp; Data Mining* (New York, NY, USA, 2021), KDD '21, Association for Computing Machinery, p. 3512–3521.

[42] ROCHAT, Y. Closeness centrality extended to unconnected graphs: the harmonic centrality index.

[43] SAJJAD, T., AND MOSTAFA, F. S. A proposed scheme for sentiment analysis: Effective feature reduction based on statistical information of SentiWordNet. *Kybernetes 47*, 5 (jan 2018), 957–984.

[44] SCHUMAKER, R. P., JARMOSZKO, A. T., AND LABEDZ, C. S. Predicting wins and spread in the Premier League using a sentiment analysis of twitter. *Decision Support Systems 88* (2016), 76–84.

[45] SEABOLD, S., AND PERKTOLD, J. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference* (2010).

[46] SILVA LOPES, A. Deterministic seasonality in dickey–fuller tests: should we care?. *Empirical Economics 31*, 1 (2006), 165 – 182.

[47] SPORTS, A. *API-Football: Coverage*, 2020. API-Football beta version 3.7.2.

[48] TALHA, A., SIMSEK, M., AND BELENLI, I. The Wisdom of the Silent Crowd: Predicting the Match Results of World Cup 2018 through Twitter. *International Journal of Computer Applications 182*, 27 (2018), 40–45.

[49] TAX, N., AND JOUSTRA, Y. Predicting The Dutch Football Competition Using Public Data: A Machine Learning Approach. *Transactions on knowledge and data engineering 10*, SEPTEMBER (2015), 1–13.

[50] TYSON, J. Fan avidity as it relates to proximity of Minor and Major league affiliates.

[51] UEFA. Uefa financial report 2018/19 annex, Jun 2019.

[52] WEI-MENG, L. *Python Machine Learning.* Wiley, 2019.

[53] WHEATCROFT, E. A profitable model for predicting the over/under market in football. *International Journal of Forecasting 36*, 3 (2020), 916–932.

[54] XIN, Y., AND XIAOGANG, S. *Linear Regression Analysis: Theory And Computing.* World Scientific, 2009.

[55] YAN, G., WATANABE, N. M., SHAPIRO, S. L., NARAINE, M. L., AND HULL, K. Unfolding the Twitter scene of the 2017 UEFA Champions League Final: social media networks and power dynamics. *European Sport Management Quarterly 19*, 4 (2019), 419–436.

[56] ZHANG, J., AND LUO, Y. Degree Centrality, Betweenness Centrality, and Closeness Centrality in Social Network. 300–303.

Curriculum Vitae

Master student was born in Aguascalientes, México, on December 11, 1997. She earned the Computer Technology Engineering degree from the Instituto Tecnológico y de Estudios Superiores de Monterrey, Monterrey Campus in December 2019. She was accepted into the graduate program in Computer Sciences in February 2020. During the bachelor program, she was able to study abroad one semester at the University of Texas at Austin, where she found her passion for Artificial Intelligence by taking courses such as Introduction to Data Mining and Machine Learning in Text Analysis and she became interested in the area of Natural Language Processing and Sentiment Analysis. Also, she took the course Integrated Communication for Sports that though her the concept of Fan Avidity. Before finishing undergrad, Clarissa was able to intern at Google as a Developer Programs Engineer for Summer 2019, there she got expertise on back-end services and services in the cloud, her project was to develop a language-agnostic interface built in gRPC that runs regressions' analysis when new changes to Google Cloud Libraries are pushed. Clarissa decided to follow her path on AI by becoming a masters student in Computer Science and to follow her passion in Football by researching the Sports Analytics field. She had successfully published and presented in two international conferences, first Predicting Soccer Results through Sentiment Analysis: A Graph Theory Approach at the International Conference on Computational Science (ICCS'21) hold in Krakow, Poland, as well, The Winning Mining Formula for Football Competitions at the International Conference on Data Science (ICDATA'21) that took place in Las Vegas, USA. She also worked as a Teacher Assistant for the Computer Science Department and the Humanities Department, in this last one, she was invited to publish Social Networks as Room for Improvement in Football as a journal article. Clarissa continues developing herself by taking the specialization in Sports Analytics from the University of Michigan, and studying Computer Vision in her last semester, to boost her career among the sports community. Clarissa achieved the Honorable Mention in Excellence at the undergraduate program. She is an enthusiastic hackathon participant and satisfactorily has won two times, second place at HackMty fall 2019 in Mexico, and first place at Hack Off v3 fall 2020 in India, both of these awards solved challenges related to Data Science applications. In her free time, she enjoys watching football, obviously, and practicing ballet.

This document was typed in using LaTeX 2_{ε}[1] by Ana Clarissa Miranda Peña.

[1] The template `MCCi-DCC-Thesis.cls` used to set up this document was prepared by the Research Group with Strategic Focus in Intelligent Systems of Tecnológico de Monterrey, Monterrey Campus.