FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Predicting Football Match Results through Data Analysis

João Lírio

Mestrado em Engenharia Informática e Computação

Supervisor: Prof. Luís Paulo Reis

July 27, 2023

Predicting Football Match Results through Data Analysis

João Lírio

Mestrado em Engenharia Informática e Computação

Approved in oral examination by the committee:

President: Prof. Henrique Lopes Cardoso
Referee: Prof. Paulo Novais
Referee: Prof. Luís Paulo Reis

July 27, 2023

Resumo

Esta tese explora a aplicação de algoritmos de machine learning e análise de dados na previsão de resultados do futebol para a temporada 2022/2023. O objetivo principal é avaliar o desempenho de vários algoritmos, incluindo K-Nearest Neighbors (KNN), XGBoost, Regressão Logística, Support Vector Machines (SVM), Análise Discriminante Linear (LDA), Naive Bayes e Decision Trees, na previsão precisa dos resultados dos jogos. O estudo também investiga o impacto de técnicas de balanceamento de conjuntos de dados, especificamente over-sampling, na precisão das previsões. Além disso, a tese examina a geração de odds com base nas previsões dos modelos de aprendizagem automática e compara-as com as odds de conhecidas casas de apostas para identificar oportunidades de apostas favoráveis.

Através de análises abrangentes de dados e modelação algorítmica, a pesquisa alcançou uma precisão acima de 60% em todos os algoritmos, destacando o seu potencial na previsão de resultados de futebol. No entanto, a exploração das técnicas de oversampling para balanceamento do conjunto de dados revelou resultados mistos, sugerindo que nem sempre ajuda a aumentar a precisão das previsões.

Além disso, a análise das odds geradas e a sua comparação com as odds das casas de apostas proporcionaram descobertas interessantes. Para além de atingir taxas de acerto bastante elevadas, a capacidade dos modelos em identificar oportunidades de apostas favoráveis demonstrou o seu potencial como ferramentas de apoio à tomada de decisão para entusiastas das apostas desportivas.

Os resultados desta tese contribuem para o campo da análise desportiva e informática, oferecendo insights sobre o desempenho de algoritmos de machine learning na previsão de jogos de futebol e o impacto de técnicas de balanceamento de conjuntos de dados. Os resultados destacam a importância de previsões precisas no mundo dinâmico do futebol e fornecem perspetivas valiosas para stakeholders, incluindo entusiastas do futebol, analistas e entusiastas das apostas. Pesquisas futuras podem aproveitar esses insights para avançar ainda mais a precisão e a confiabilidade das previsões do futebol, abrindo caminho para desenvolvimentos na interseção entre o machine learning e o desporto que todos gostamos.

Abstract

This thesis explores the application of machine learning algorithms and data analytics in predicting football results for the 2022/2023 season. The primary objective is to evaluate the performance of various algorithms, including K-Nearest Neighbors (KNN), XGBoost, Logistic Regression, Support Vector Machines (SVM), Linear Discriminant Analysis (LDA), Naive Bayes, and Decision Trees, in accurately forecasting match outcomes. The study also investigates the impact of dataset balancing techniques, specifically oversampling, on prediction accuracy. Additionally, the thesis examines the generation of odds based on the machine learning models' predictions and compares them with bookmakers' odds to identify favorable betting opportunities.

Through comprehensive data analysis and algorithmic modeling, the research achieved above 60% accuracy across all algorithms, highlighting their potential in predicting football results. However, the exploration of oversampling techniques for dataset balancing revealed mixed results, suggesting that it may not always enhance prediction accuracy significantly.

Furthermore, the analysis of generated odds and their comparison with bookmakers' odds yielded interesting findings. In addition to achieving very high accuracy rates, the models' ability to identify favorable betting opportunities demonstrated their potential value as decision support tools for sports betting enthusiasts.

The findings of this thesis contribute to the field of sports analytics and informatics, offering insights into the performance of machine learning algorithms for football prediction and the impact of dataset balancing techniques. The results underscore the importance of accurate predictions in the dynamic world of football and provide valuable perspectives for stakeholders, including football enthusiasts, analysts, and betting enthusiasts. Future research endeavors can build upon these insights to further advance the accuracy and reliability of football predictions, paving the way for developments in the intersection of machine learning and the beautiful game.

Acknowledgements

When I arrived at FEUP, I was a 18 year old kid that didn't no at all where was going and only decided to enter the course because of the love of playing video games and technology in general. Now, after all these six years, I feel a very different person with much more knowledge about life and informatics. This was only possible, because of every person that supported me and shared experiences with me through the years and that I must thank.

First and foremost, I would like to thank my supervisor, professor Luís Paulo Reis, for his continuous support throughout the entire research process. His valuable insights, feedback, and challenging ideas and perspectives were very important to always push me forward.

Furthermore, I am grateful to all my friends and colleagues who have been there all the time, helping me with all the the difficulties I had through the course whether in projects, tests or exams of the different subjects. Also for all the memories and the good times that I will carry with me forever.

Last but not least, I would like to thank my family for always supporting and never doubting me despite all the difficult times that I came through. You have given me everything and I really hope that I can repay everything you gave me.

Thank you.

João Lírio

"All of humanity's problems stem from man's inability to sit quietly in a room alone."

Blaise Pascal

Contents

1 Introduction **1**
- 1.1 Goals . . . 1
- 1.2 Document Structure . . . 2

2 Football Datasets **3**
- 2.1 Overview . . . 3
- 2.2 National and international football websites . . . 4
- 2.3 Different datasets in football . . . 5

3 Football match result prediction systems **7**
- 3.1 Overview . . . 7
- 3.2 Machine Learning . . . 8
- 3.3 Related works . . . 9

4 Data and Exploratory Analysis **12**
- 4.1 Data . . . 12
- 4.2 Exploratory Analisys . . . 13
  - 4.2.1 Predictability across leagues . . . 15

5 Football results prediction exploration **17**
- 5.1 Goal based prediction model . . . 17
  - 5.1.1 Results analysis . . . 18
- 5.2 Adding recent performance statistics . . . 19
- 5.3 Results from recent matches . . . 20
- 5.4 Exploring Home Advantage . . . 22
- 5.5 Results prediction in the main European Leagues . . . 25

6 Datasets Over-sampling and Under-sampling **28**
- 6.1 Over-sampling Techniques . . . 28
  - 6.1.1 Random Over-sampling . . . 28
  - 6.1.2 Synthetic Minority Over-sampling Technique (SMOTE) . . . 30
- 6.2 Discussion and Practical Implications . . . 32

7 Generating betting odds with machine learning algorithms **33**
- 7.1 Methodology . . . 33
  - 7.1.1 Odds Generation . . . 33
- 7.2 Comparing Generated Odds . . . 35
  - 7.2.1 Comparing Generating Odds From Different Algorithms . . . 35
  - 7.2.2 Comparison With Bookmakers Betting Odds . . . 36

8 Conclusions **39**
8.1 Future Work 40

References **42**

List of Figures

4.1 Average goals scored per game each season 13
4.2 Number of home wins vs home losses each season 14
4.3 Teams with most home wins 15
4.4 Leagues predictability 16

5.1 EPL teams home attack strength 2022/2023 season 18
5.2 Accuracy of Training set 1 - HAS, HDS, AAS, ADS 19
5.3 Comparison between Training set 1 and Training set 2 20
5.4 Comparison between Training set 2 and Training set 3 22
5.5 Comparison between Training set 3 and Training set 4 24
5.6 Comparison between Training set 1 and Training set 4 25
5.7 Comparison of Training set 4 between main European Football Leagues 26

6.1 Comparison between Original Dataset and Resampled Dataset with RandomOver-Sampler 29
6.2 Comparison between Original Dataset and Resampled Dataset with SMOTE . . . 31

7.1 2022/2023 Premier League Season's Round 31 Matches Outcome Probabilities . 34
7.2 2022/2023 Premier League Season's Round 31 Matches Generated Odds 34
7.3 2022/2023 Premier League Match Odds 35
7.4 2022/2023 Premier League Match Normalized Odds 35

List of Tables

2.1 Table comparing two type of football platforms 5
2.2 Table comparing two football datasets 6

3.1 Comparing approaches [9] 11

7.1 Odds Generated by Machine Learning Algorithms 36
7.2 Comparison of Bet365 Odds with KNN Odds 37
7.3 Comparison of Bet365 Odds with XGBoost Odds 38
7.4 Comparison of Bet365 Odds with Decision Tree Odds 38

Abbreviations

FIFA	Fédération Internationale de Football Association
EA	Electronic Arts
CSV	Comma-separated Values
ANNs	Artificial Neural Networks
RNNs	Recurrent Neural Networks
LSTM	Long Short-Term Memory
NBA	National Basketball Association
UEFA	Union of European Football Associations
BLR	Binomial Logistic Regression
EPL	English Premier League
KNN	K-Nearest Neighbour
SVC	Support Vector Classifier
HAS	Home Attacking Strength
HDS	Home Defensive Strength
AAS	Away Attacking Strength
ADS	Away Defensive Strength
FTR	Full Time Result
SMOTE	Synthetic Minority Oversampling Technique

Chapter 1

Introduction

Football, as one of the most popular sports worldwide, captivates the hearts of millions of fans with its unpredictability and excitement. The ever-increasing competition and unpredictability of match outcomes have sparked a growing interest in developing accurate prediction models. The popularity and global reach of football have made it a fertile ground for data-driven analysis, opening up new avenues for understanding the intricate dynamics of the game. From player statistics and team formations to match history and recent performances, a wealth of data is readily available to unravel the patterns that influence match outcomes. The ability to accurately predict match outcomes has long been a subject of interest for enthusiasts, analysts, and even bookmakers. In recent years, advancements in machine learning algorithms and the availability of vast historical data have opened up new avenues for predicting football results with greater precision.

The traditional methods of football prediction relied heavily on human intuition, subjective analysis, and limited statistical data. However, with the emergence of powerful machine learning algorithms and the availability of extensive historical data, we now have the tools to extract meaningful patterns, identify key influencing factors, and make data-driven predictions. This paradigm shift has enabled us to explore and exploit the wealth of information contained within the vast repositories of football match data.

1.1 Goals

This thesis embarks on a compelling journey into the realm of sports prediction, with a specific focus on forecasting football match results in the top European Leagues. The pursuit of accurate match predictions is not merely an academic exercise; it has practical applications for bettors, bookmakers, sports analysts, and football fans alike. The ability to predict match outcomes with high accuracy holds immense value in sports betting markets, enabling bettors to make more informed decisions and bookmakers to set odds that reflect the probabilities of different outcomes. For sports analysts and enthusiasts, accurate predictions offer valuable insights into team performances, player dynamics, and strategic nuances that influence the beautiful game's trajectory.

Against this backdrop, the primary objective of this thesis is twofold. First and foremost, it aims to explore the efficacy of machine learning algorithms in predicting football results for the highly anticipated 2022/2023 season. By systematically analyzing historical data from the top European Leagues, including England, Spain, Italy, Germany, and Portugal, this study seeks to identify underlying patterns, trends, and interdependencies that contribute to match results and with that provide valuable insights into the potential of these techniques in enhancing the accuracy of football predictions.

Secondly, this thesis delves into the impact of dataset balancing techniques on the performance of machine learning models. Specifically, it investigates the effectiveness of oversampling methods in addressing class imbalances within the dataset. By systematically analyzing the effect of oversampling on model performance, the research aims to shed light on the role of data preprocessing techniques in improving the accuracy and reliability of football predictions.

Furthermore, this study ventures into the realm of generating odds based on the predictions of machine learning models. The thesis will compare these generated odds with the odds offered by bookmakers to evaluate the models' ability to identify favorable betting opportunities. This analysis holds potential implications for individuals interested in sports betting and underscores the value of machine learning algorithms as decision support tools in this domain.

By pursuing these goals, this thesis aims to contribute to the field of sports analytics and informatics by advancing our understanding of the intricacies underlying football results. The research endeavors to leverage the power of machine learning algorithms and historical data to uncover patterns, trends, and factors influencing match outcomes. Through the convergence of data-driven methodologies and the captivating world of football, this study seeks to push the boundaries of prediction accuracy and unearth valuable insights for football enthusiasts, analysts, and betting enthusiasts alike.

1.2 Document Structure

The rest of the document is structured as follows: in Chapter 2, we explore the football websites existent in Portugal and worldwide and compare them, also comparisons between two football datasets were made. Chapter 3 presents some background on some different machine learning algorithms and a comparison between related works of predicting results not only from football but also other sports. Chapter 4 makes some analysis about the dataset used in this research in order to find some patterns in the different leagues and teams. In Chapter 5 we perform the exploration of our predictions using different parameters and see which ones provide better results. And also, compare how different leagues performed we those same training sets. Chapter 6 was about trying to balance the dataset used using over-sampling techniques and trying at the same time to maintain the accuracy levels or at least not to get much worse. In Chapter 7 it was generated betting odds through the machine learning algorithms used in the previous chapters and then was made a comparison with betting odds from known bookmakers. Finally, a reflection on the work developed and possible future expansions, in Chapter 8, concludes the document.

Chapter 2

Football Datasets

2.1 Overview

Sport is something that has been part of humanity since the begging of it, and lived a continuous process of development until reaching the state in which we see it today. Talking more particularly about football, the biggest and the world's most popular sport [4]. At the turn of the twenty-first century, FIFA estimated that there were approximately 250 million football players in over 200 countries, and over 1.3 billion football fans [14]. Its growth has always gone hand in hand with the growth of sport in general. Nowadays, football is more connected with technology than ever, and that connection will only continue to grow. Football it's more and more a sport in which every detail counts to reach the victory and everything is analyzed to the smallest detail, this importance of the detail makes statistics gain much relevance and importance to football. All these factors together led to a considerably growth of online sports platforms. These platforms, like mobile applications or websites, dispose every type of information. From the final result of a game to individual statistics of every player that participated in that specific game and a final note for every player based on his exhibition.

Nonetheless, there is another market that has grown a lot and has increased the use of the sports platforms. I am talking about the sports betting market. The global sports gambling market is estimated to worth up to 3 trillion dollars, with football betting representing 65 percent of this figure [11]. Sport result prediction is an interesting and challenging problem due to the inherently unpredictable nature of sport. Despite its difficulty, predicting the results of sports matches is of significant interest to many different stakeholders, including bookmakers, bettors and fans. And these two last groups of people use a lot the sports applications, to check the statistics of the last games and make their bets based on that or just to simply verify the results of the games they bet on.

2.2 National and international football websites

All over the world football websites and applications show the information in a very similar way. Currently, football has five major leagues (the English, the Spanish, the German, the Italian and the French league) and every football platform in the world presents all the information they have about these five leagues, the sixth being the national league of the country in which the site or application operates.

About the websites, the information they have and how they present it, we have sports news online platforms (which in Portugal are for example *MaisFutebol*, *ojogo.pt* and *zerozero.pt*) that show in the foreground the latest news related to national and international football, but which usually have a section where it is presented the standings of the national championships and the major international championships, the latest results and some key statistics like the goals each team scored and conceded, the best scorers and the players with most yellow and red cards.

Other type of platforms are the websites or mobile applications that are one hundred percent dedicated to providing to the users statistics of their favourite sport (for example *flashscore.com* and *sofascore.com*). As football is the biggest sport in the world, it turns out to be the sport with the most information. This websites dispose obviously the type of information that the sports news online platforms mentioned before dispose, but beside that it is possible to find statistics about the major leagues of almost every country in the world and some countries secondary leagues. It is also disposed standings of the games played at home and the games played away in order to understand which team are most successful in these conditions. Other type of statistics presented are the individual ones, with standings of the players with more goals, mores assists, more successful dribbles, more chances created, more tackles, more interceptions, percentage of accurate passes and many more other statistics of this type. Another very important feature this platforms present is that in addition to having detailed statistics of all games, they also present them live of the games that are happening in the moment. The type of information presented in this section is obviously the current score of the game, but also the minutes the goals were scored, the minutes of the cards and substitutions, the lineups of each team, statistics of each team like ball possession, total shots and more specific ones like duels won or long balls completed. In addition, it is also possible to monitor the individual performance of each player on the field in a detailed way, seeing their range of action, the touches, the shots, the fouls they made and an overall score of their performance so far in order to see who are the players doing it better on the pitch.

This live feature these platforms present is what brings the majority of the users. Being great part of them bettors that want to follow the games up to the minute and even bet live. The other part are just users that like the sport and like to be aware of the games that are happening or want to see how their favourite team is performing.

The following table compares the two different types of football online platforms that were analyzed in this section.

Different types of football platforms		
	Sports news platforms	Sports statistics platforms
Latest football news	Yes	No
Top five leagues standings	Yes	Yes
Top five leagues results	Yes	Yes
Other leagues statistics	No	Yes
Players statistics	No	Yes
Live scores	Yes	Yes
Live detailed statistics	No	Yes

Table 2.1: Table comparing two type of football platforms

2.3 Different datasets in football

When talking about datasets in the world of football, probably we are talking about very large datasets with a lot of information. It is possible to find databases with information from almost 20 seasons about the major leagues of football in Europe, with detailed statistics about thousands of different games. To better understand the topic, a comparison was made between two different datasets that are available in Google Dataset Search.

The most famous dataset about football is the *European Soccer Database* implemented by Hugo Mathien. This database has already been used many times in many projects of data analysis and machine learning. The data presented in this data set was sourced from `http://football-data.mx-api.enetscores.com/` for scores, lineup, team formation and events, `http://www.football-data.co.uk/` which captures the published market odds offered by a number of bookmakers over many leagues. The odds are recorded on Friday afternoons for weekend games and on Tuesday afternoons for midweek games. And finally `http://sofifa.com/` for players and teams attributes from EA Sports FIFA games. Currently, it has more than 25,000 matches, information about more than 10,000 players like name, birth date, height, weight... A very interesting type of data this database has is the players attributes that are sourced from EA Sports FIFA video games series, including weekly updates of this attributes. It also has information about the major 11 leagues of 11 european countries and those statistics go from season of 2008 to 2016. More than 10,000 of the games in the database have detailed match events (goal types, possession, corners, crosses, fouls, cards, ...) and team line ups have a squad formation using X and Y coordinates. Most of the matches also have winner betting odds from up to 10 different providers.

The other dataset analyzed was another dataset available on *Kaggle* developed by Joseph Mohr and called *European Club Football Dataset*. This database has data for over 23,000 matches from various top football leagues in Europe. In this dataset the data start years depend on the league and on the type of information. For example, the Ligue 1 (the major French football league) has data about the classification tables from 2002, data from aggregate stats also from 2002, but data about the games only from 20018. There are six leagues in this dataset, being them

the six major leagues in Europe (English, Spanish, German, Italian, Dutch, French). The data presented in here is scraped from https://www.espn.com/soccer/. The dataset has the usual classification tables with the total points, the wins, losses and draws, goals scored and goals conceded. For each game is available various different information like the date and time of the match, the attendance, the local and then statistics of the game like ball possession, shots made, fouls committed, offsides and many other statistical information. There is also a "events" file that contains the substitutions, the cards and the goals with the minute that each one happened. Then there is tables with best scorers classifications, classifications of players with most assists in every season and classifications of the teams in a season ordered by the number of cards each one had.

Table 2.2 compare the two datasets analyzed to be more easy to understand what each dataset has and which one can possibly be more advantageous.

Comparison between the two datasets		
	European Soccer Database	European Club Football Dataset
Number of leagues	11	6
Total matches	+25,000	+23,000
Total players	+10,000	N/A
Seasons	2008 to 2016	2000 to present (*)
Players and teams attributes	Yes	No
Detailed match events	Yes	Yes
Betting odds	Yes	No
Type of files	SQLITE	CSV

Table 2.2: Table comparing two football datasets

(*) The start season depends on the league and on the type of data

Chapter 3

Football match result prediction systems

3.1 Overview

Prediction systems have been developed for many purposes in the last years, including areas such as the weather, student performance, and stock market fluctuations [7]. Football is another area in which predictive systems are extensively explored, more specifically with regard to the prediction of match results.

Football is a sport that, following the example of basketball, baseball, American football and others, has become very supported by statistics in recent years. Which means that there are more and more different types of data regarding teams and players performances, from the most specific and detailed like the number of missed passes in high risk zones of every player, to the most general like the ball possession of a team. These factors allied to the popularity of sports made many organizations invest to obtain better results in predicting football matches; accordingly, the prediction of game results has become an area of interest. Which led to the growth of bookmakers, since bookmakers rely on computers that calculate the winning odds (and other types of events that happen in a match) of each team in a given match, the development of predictive models and algorithms grew exponentially.

In football, match result prediction systems have been developed with techniques such as artificial neural networks, naive Bayesian system, k-nearest neighbor algorithms (k-nn), and others [7]. Obviously, the choice of which method to use depends on the type of data as well as the feature sets and usually the main goal is to achieve the highest prediction accuracy as possible.

This chapter will consider different methods of predicting football match results, investigate them, make comparisons between them and try to understand which ones are more accurate, more efficient and with that discover the best method to approach the goal of this project.

3.2 Machine Learning

Machine learning is a branch of artificial intelligence that is concerned with building systems that require minimal human intervention in order to learn from data and make accurate predictions [6]. This subfield of artificial intelligence is growing in the programming fields at a very high rate. Machine learning is composed of two phases, namely, a learning phase and a prediction phase. The learning phase involves the following: 1) preprocessing (normalization, reduction, data cleansing); 2) learning (supervised, unsupervised and reinforcement); 3) error analysis (precision/recall, over fitting, test/cross validation etc.); and 4) model building [7]. The prediction phase takes the output of the learning phase, which is the model to predict new data sets. The predicted data helps management or decision makers make informed decisions that are further used to build a knowledge discovery database [7].

Machine learning usually can be of three different types supervised, unsupervised, and reinforcement learning. In supervised learning, the system searches for patterns and analyses the data to train itself for different scenarios according to the data provided. When the system starts giving consistent and accurate results, it is deployed on real-life data. In unsupervised learning, there are no labels given to the learning algorithm or the trainer, leaving it on its own to find structure in its input in an unstructured manner. Unsupervised learning can be a goal in itself in terms of ascertaining hidden patterns in data or a possible means towards an end, which is in the prediction proper [8]. Reinforcement learning is when the model is learned in a simulated environment.

Neural Networks Artificial Neural Networks (ANNs) usually contains interconnected components (neurons) that transform a set of inputs into a desired output, trying to mimic a biological neural network [5]. In neural networks each node executes a simple task to help the system learn something from the input fed to the system. This model can adapt to changing input; so the network generates the best possible result without needing to redesign the output criteria. [9]

Deep Learning Deep learning is a branch of learning that is based on endowing large neural networks with multiple hidden layers in order to learn features that have strong predictive abilities [6]. It is used when we have to face complex tasks, which often require dealing with lots of unstructured data, such as image classification, natural language processing, or speech recognition, among others. Deep learning takes longer times to train the system, but to compensate it gives more accurate results when the training process is finished and consolidated. [9]

Recurrent Neural Networks RNNs are very similar to Artificial Neural Networks with the improvement of allowing loops between different nodes to allow the system to remember specific

outputs related to some particular features. These loops bring to the system the capability of memorizing, basically remembering the data for some time in the system. [9]

Long Short Term Memory Long Short-Term Memory (LSTM) networks are a type of recurrent neural network capable of learning order dependence in sequence prediction problems, in other terms they are designed to remember the values for a long or a short period of time. If there is a feature that has a considerable weight in deciding the output, it will remember the feature for a longer period of time. This is a behavior required in complex problem domains like machine translation, speech recognition, and more. [9]

XGBoost XGBoost, short for Extreme Gradient Boosting, is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. It is widely used for both classification and regression tasks and has gained popularity in various domains, including finance, healthcare, and, of course, sports analytics. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks. However, when it comes to small-to-medium structured/tabular data, decision tree based algorithms are considered best-in-class right now. [12]

Logistic Regression Logistic regression, despite its name, is a classification model rather than regression model. Logistic regression is a simple and more efficient method for binary and linear classification problems. It is a classification model, which is very easy to realize and achieves very good performance with linearly separable classes. It is an extensively employed algorithm for classification in industry. The logistic regression model can be generalized to multiclass classification. Scikit-learn has a highly optimized version of logistic regression implementation, which supports multiclass classification task. [16]

3.3 Related works

Considering now works related to the prediction of the winner in a sports game, more specifically football matches, it will be analyzed different techniques that were used in the past and finally compare them and try to understand which ones had more accurate results.

Zdravevski and Kulakov employed classification techniques that can be found in WEKA to predict the winner of a game. The parameters used were the number of players injured in the team, winning streak of the team before playing the game, the fatigue of the team before playing the game, win percentage of the teams and offensive and defensive ratings of the team. In this work was used data of two back to back seasons of NBA League from the basketball reference website [17].

In 2018, Lam in their research has used a Bayesian linear regression model, Gaussian process regression model and Sparse Spectrum Gaussian Process Regression model to calculate the winning probabilities of a match. In that research were used 11 attributes in total to calculate each team's strength and every player's ability. For these solution they used the data of the 2014/2015 season of the NBA League from the database available on the NBA website [10].

This next research is specifically about football matches and the system is based in Support vector machine. Igiri used Gaussian combinational kernel type and generated a total of 79 support vectors at a count of 100,000 iterations. For this research they used the data available for all the seasons of the English Premier League, more specifically their dataset to work upon [7].

Rotshtein in their research propose a model which predicts the result of an upcoming match using the method of identifying non-linear dependencies in the knowledge base. Five different types of a match result were defined to construct the prediction system based on fuzzy logic, namely loss with a high score, loss with a low score, draw, win with a low score and finally a win with a high score. To apply the fuzzy knowledge base a fuzzy knowledge approximator was used. Here was used the data of the championship of Finland from seasons 1994 to 2001 with results of 1056 matches [15].

Pettersson and Nyquist used Long Short Term Memories in recurrent neural networks to predict the outcome of an ongoing football match. They considered all the players playing in the team to be at the same advantage and of the same capabilities. Their dataset included matches from leagues and tournaments across the globe from 54 different countries that come under UEFA as well as the countries of American and Asian origin [13].

Cui in its research used 25 different variables related to a football match as terminal sets, and the fitness measure is set as the total number of the wrong prediction of the games. It was used the official site of the English Premier League [3].

Maurizio Carpita, Enrico Ciavolino and Paola Pasca made a study exploring the Kaggle European Soccer Database that was mentioned in section 2.3. The main goal was to estimate the win probability of the home team with the binomial logistic regression (BLR). The predictive power of the BLR model and its extensions has been compared with the one of other statistical modelling approaches (Random Forest, Neural Network, k-NN, Naïve Bayes). Results showed that role-based indicators substantially improved the performance of all the models used in both this work and in previous works available on Kaggle. The base BLR model increased prediction accuracy by 10 percentage points, and showed the importance of defence performances, especially in the last seasons [1].

Table 3.1 shows in a summarized way the information provided in this section, becoming easier to compare results.

Author	Dataset	Accuracy	Remarks
Zdravevski and Kulakov	Data of two back to back seasons of NBA League. It contains detailed statistics of each game played during a season.	The accuracy of reference classifier was comparatively low when compared to any other classifier (around 5-10% low)	The data was collected from the official NBA source, and hence, the data was more accurate. Low prediction accuracy
Lam	Data of the 2014/2015 season of the NBA League. It contains box score statistics, which can be converted to the player performance statistics	TLGProb had an accuracy of 85.28% in NBA 2014/2015 season	TLGProb performs well when compared to the existing NBA predictive models. A small dataset was used
Igiri	Players' performance and manager indices were gathered from the 2014-2015 season of the English Premier League	Prediction accuracy was 53.3%	Can handle nominal data. Low prediction accuracy. Does not support large datasets
Rotshtein et al.	Data of the championship of Finland from seasons 1994 to 2001 with results of 1056 matches	85% accuracy in genetic tuning and 84% accuracy in neural tuning	High accuracy. Due to insufficient learning samples, the tuning of this kind of system is a very difficult task
Pettersson and Nyquist	It contains 35,234 matches for two years (2015-2017) from 54 different countries. The dataset also includes the information for each player in the teams	The accuracy was 96.63% for many to one approach and 88.68% for the many-to-many approach	High accuracy. The dataset comprised of only 2–3 seasons of the leagues and tournaments
Cui et al.	25 different features of every footbal game from the English Premier League season of 2011/2012 (380 games) and a part of the 2012/2013 season	Testing accuracy was 56%, and the overall accuracy of the system was near to 70%	Low accuracy. The accuracy rate of each function is not the same
Maurizio Carpita, Enrico Ciavolino and Paola Pasca	Kaggle European Soccer Database described in section 2.3. More than 25,000 matches and 10,000 players from 11 leagues of 11 European countries	The final results show an accuracy of about 54% of the home team WINs and about 73% of the home team NOWINs	The k-NN model showed a slightly lower accuracy than the other models. NB model proved to be more balanced in predicting both WINs and NOWINs

Table 3.1: Comparing approaches [9]

Chapter 4

Data and Exploratory Analysis

We now present the process of selecting and exploring datasets related to the season 22/23 in the top European leagues. Choosing the most appropriate datasets requires evaluating data sources for relevance and quality, and make sure that the data is official so we can do comparisons with what actually were the last seasons in those leagues.

After choosing the datasets that would be used, it was necessary to start using exploratory data analysis techniques to gain insights into the seasons. The goal was to create descriptive statistics, visualizations and summary measures in order to uncover patterns and characteristics in some of the leagues and teams present in the datasets. Those visualization techniques aid in interpreting the quite extensive data due to then number of games contained in it.

4.1 Data

After exploring several databases, it was decided to use data obtained from the Football-Data website (https://www.football-data.co.uk/data.php). This website contains data from 21 first European divisions in addition to the secondary divisions of some of these countries. Additionally, Football-Data provides data for 16 other worldwide premier divisions. The datasets from this source contained CSV files that had detail about each match and in game features (corners, shots on target, faults, etc.) as separate observations. Another interesting feature of these datasets is the betting odds data from up to 10 major online bookmakers for every match of the last few years, that feature was used to make some comparisons that were covered in a later chapter.

For this study, it was used the top four European Leagues (England, Spain, Italy, German) and the Portuguese top league. The data used only refer to this last season of 2022/2023, taking into account that the history of the teams in previous years ends up not having much weight in the events of this last season. However, to make some studies about these five leagues we used data from the last eight seasons (from 2015/2016 to the 2022/2023 season) in order to understand better the play style of a league and any tendency that each may have.

4.2 Exploratory Analisys

To start off, it was analyzed the average number of goals scored per game by league in each of the eight seasons. At first it was thought to analyze the total number of goals scored, but since the number of matches played in the German and Portuguese leagues is considerably smaller than in the other leagues it came to the conclusion that would not make much sense.

In the graph bellow is possible to see an overview of the data collected about the subject, which helps to better understand the reality of those numbers. Looking at the graph is possible to see that Bundesliga (German league) is the one with the highest average of goals per game in more seasons and has an average greater than three goals per game since the 2018/2019 season. On the other hand, Liga Portugal and La Liga are the ones with lower average of goals per game. Another relevant data is the goals per game in Liga Portugal in the 2016/1017 season being significantly lower than 2.5 goals. Premier League (English league) is the one with more consistent results being between 2.7 and 2.85 goals per game in every season.

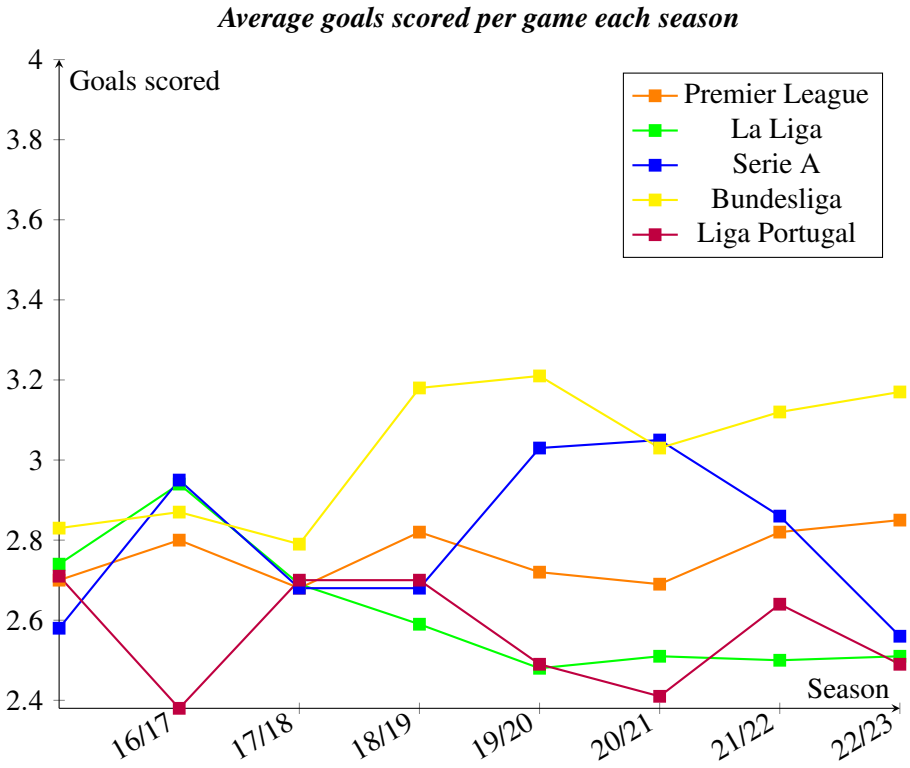

Figure 4.1: Average goals scored per game each season

The second question to be analyzed was the veracity of the 'Home Team Advantage' since this is discuss a lot in football and sports in general, and has a lot of influence when predicting results about any match that is not on neutral ground. In the graph below we can see the Home wins/Away wins ratio and once all the values it is possible to observe that every league in every

season had more home wins than away wins except Premier League in the 2020/2021 season. In fact, 2020/2021 season in general was the one that in which there was less relevance in the Home Team Advantage. Maybe this is explained by this years being the most affected by COVID-19 what led to the matches played during that season being played behind close doors with no support from the stands, and the truth is that the fans are what turns playing at home so advantageous. Looking league by league, it appears that La Liga is the one where the home advantage has more weight.

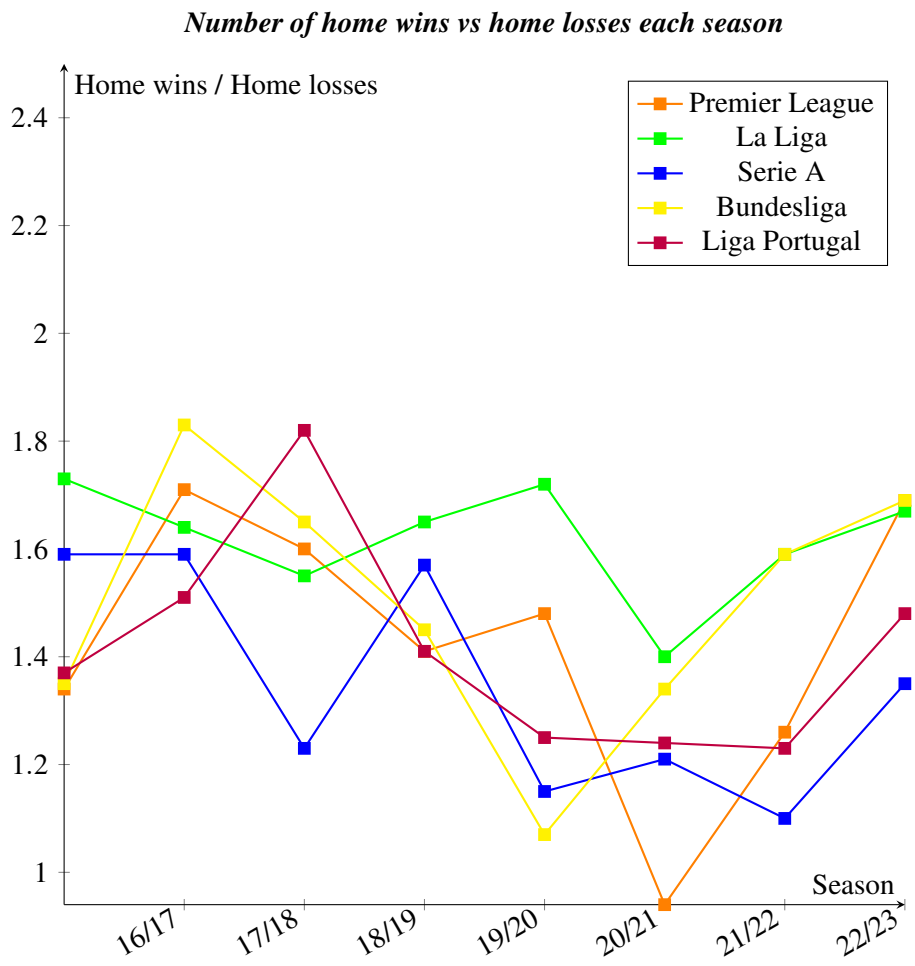

Figure 4.2: Number of home wins vs home losses each season

To evaluate which team were more dominant for the last years in these leagues, it was plotted the teams with most home wins in the last 8 years. The Spanish home advantage seen in the last graph can also be checked here with three of the first six teams with most wins being Spanish. Also the Portuguese league has three teams with lots of dominance playing home, even more taking into account that they have two fewer home games per season (16 games in the 8 seasons). German league is the less represented in the graph with only two teams and one of the them being tied for the last place in the graph.

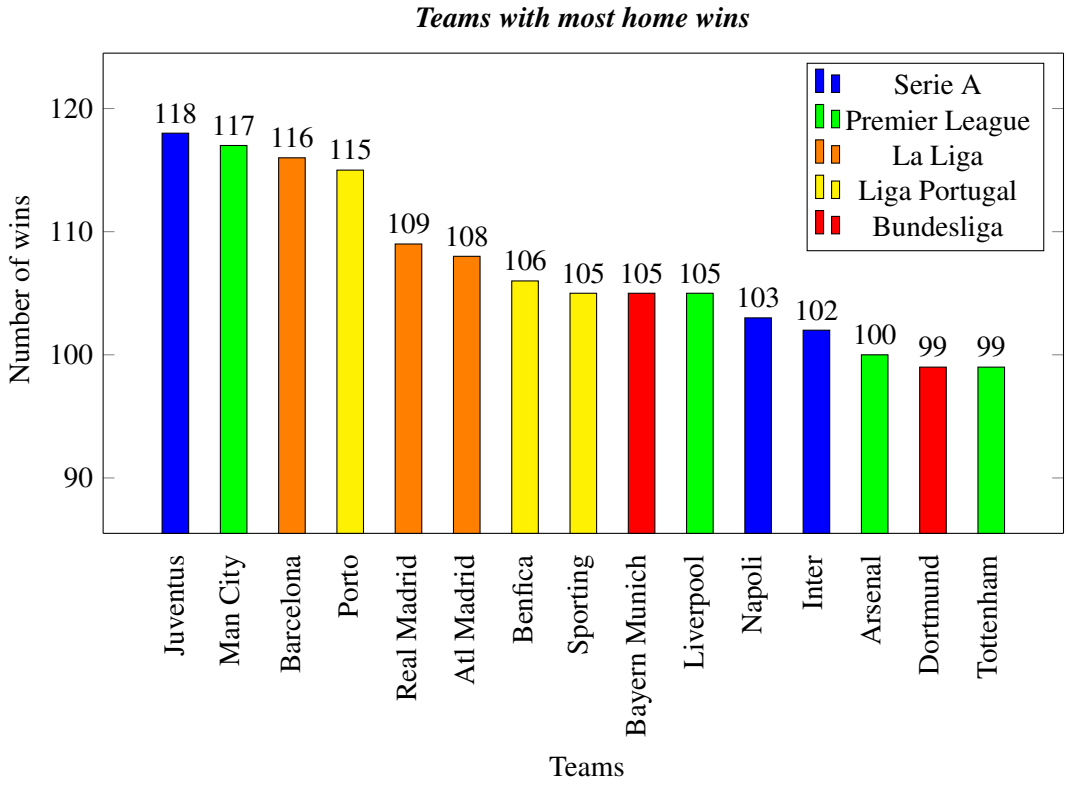

Figure 4.3: Teams with most home wins

4.2.1 Predictability across leagues

The final study made about the dataset was about the predictability of each league in the last years. To measure something that seems a bit abstract it was decided to use the odds of the three possible results of a match present in the dataset given by Bet365, one of the most popular bookmakers. These odds have been converted to probabilities and with the probabilities it was calculated the entropy with base *e*. Entropy, in *Information Theory*, is capable of measuring the "uncertainty" or "randomness" of a variable's output [2].

Observing the graph, it can be inferred that the Spanish La Liga was the most "unpredictable" in the last four seasons with values of entropy very very close to 1. Another interesting pattern is value of entropy being higher this last season than seven seasons ago in every league except the English Premier League.

Analysing the entropy of some teams it was possible to conclude that top teams of each league like Barcelona, Bayern Munich, Manchester City, Benfica, Napoli have an entropy around 0.75 becoming much more predictable than the others.

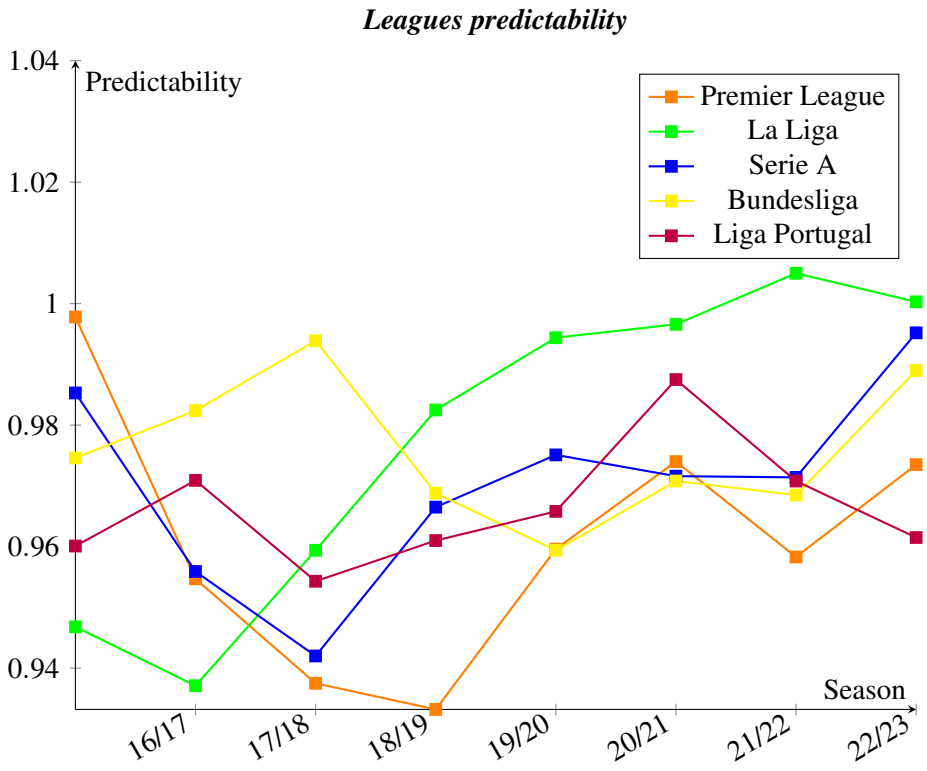

Figure 4.4: Leagues predictability

Chapter 5

Football results prediction exploration

After an exploratory analysis of the dataset, it is possible to say that we are ready to start exploring different approaches to predict football results. That is the focus of this chapter and it will cover a range of topics, including the utilization of various algorithms, training parameter optimization, database balancing techniques, and comparison of generated odds with bookmakers betting odds.

It was highlighted the importance of training parameter tuning and model evaluation to improve prediction performance. Additionally, it addresses the challenges posed by class imbalances in the dataset and explores techniques to balance the database. Finally, comparing the generated odds with betting odds allows us to assess the predictive power of the models developed and, in other situations, potentially identify valuable betting opportunities.

5.1 Goal based prediction model

Since the main objective in a football match is to score goals (and not concede) and that is what define the winner of a match. It seems that looking to historic data, goals scored and conceded were the most valuable statistic that it could be used. Therefore, a few statistics based in goals were calculated for each team. That statistics were attacking strength and away strength at home and away.

Using the attack strength at home example. To calculate it, it was used the goals scored at home of the the team in question, the number of games at home played until the day and the average number of goals scored by home teams in every match. For defensive strength it is the same but with the goals conceded:

Attacking strength at home = Goals scored at home / Number of games at home / Average number of goals scored at home

Defensive strength at home = Goals conceded at home / Number of games at home / Average number of goals conceded at home

The following graph shows the veracity of the Home Attack Strength since that Man City and Arsenal were in fact the teams with more offensive power in the all season.

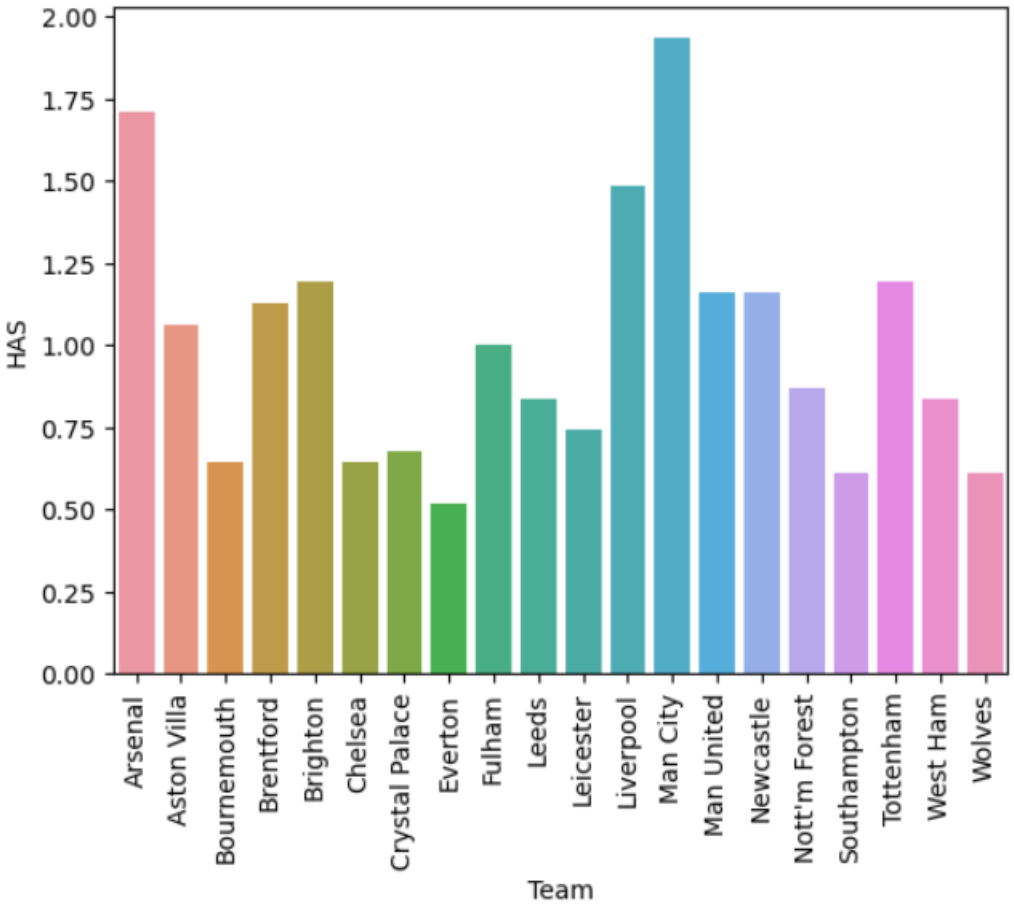

Figure 5.1: EPL teams home attack strength 2022/2023 season

5.1.1 Results analysis

The goal was trying to correctly predict the final outcome of maximum games (home win, draw or home loss). This turns now into a multi-class classification problem and with that the idea is to try out different machine learning methods to model the dataset available. In a multi-class classification problem, various machine learning algorithms can be applied to make accurate predictions. Among these algorithms, the KNN classifier, XGB classifier, Logistic Regression, Linear SVC, Linear Discriminant Analysis, Bernoulli Naive Bayes and Decision Tree Classifier presented better results to the initial experiments carried out and, therefore, were the algorithms used throughout all the experiments.

Since there are three possible results to any match, with a model that works randomly picking any of the three outcomes as the final result, we would end up with an overall accuracy of 33%. The main goal is to obtain clearly better results than what a random model would give us.

Training features and training target datasets used in this situation were about the first three hundred games of the 2022/2023 English Premier League season (until round 30). Consequently, the test features and test target datasets were about the remaining 80 games of the same championship (last eight rounds). As already mentioned, this first try is only using teams' attacking

strength and defensive strength.

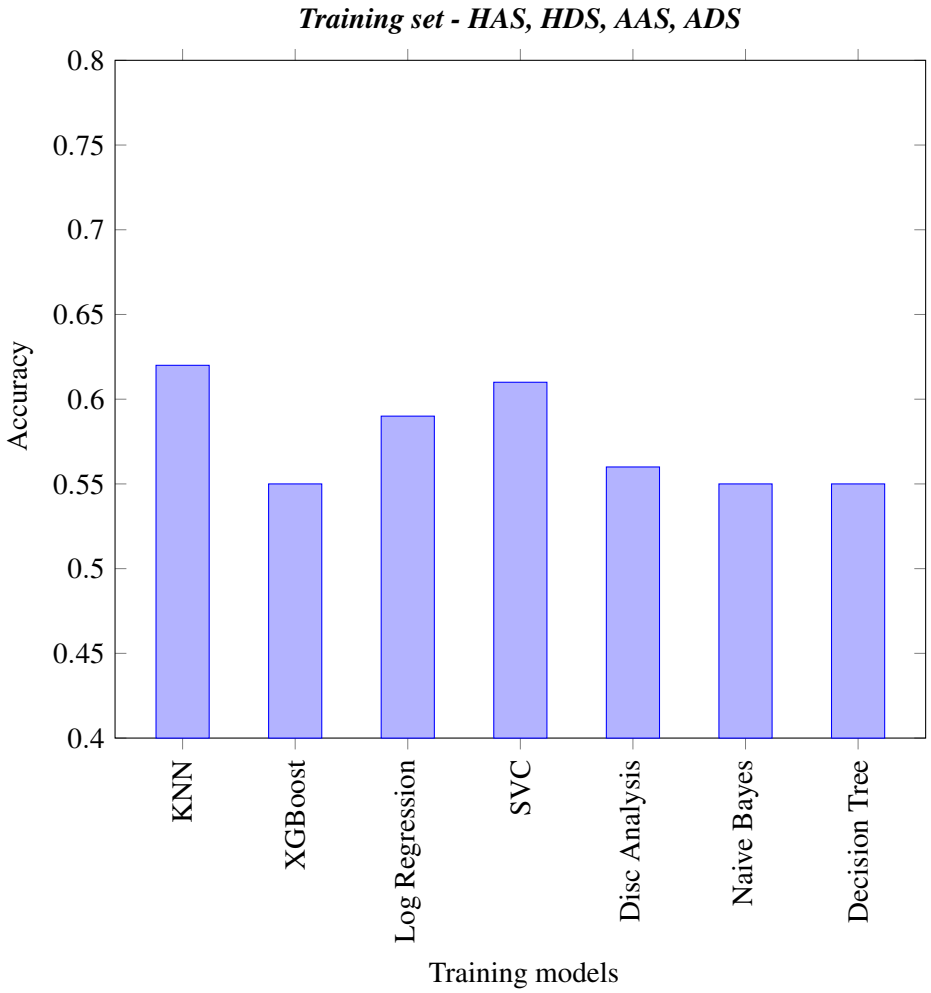

Figure 5.2: Accuracy of Training set 1 - HAS, HDS, AAS, ADS

5.2 Adding recent performance statistics

For this second test, it was decided to give more relevance to the recent form of the teams. With that, it was used the average number of shots on target, the average number of corners and the average number of goals scored in the last five matches of both teams in relation to the match in question. Using these statistics, it was made the difference of both teams in each parameter:

pastGoalsDiff = Average Goals Home Team - Average Goals Away Team

pastShotsDiff = Average Shots Home Team - Average Shots Away Team

pastCornerDiff = Average Corners Home Team - Average Corners Away Team

The two graphs below show the results of the first training set analysed in the previous section and the other with the results of this second training set with the additional statistics.

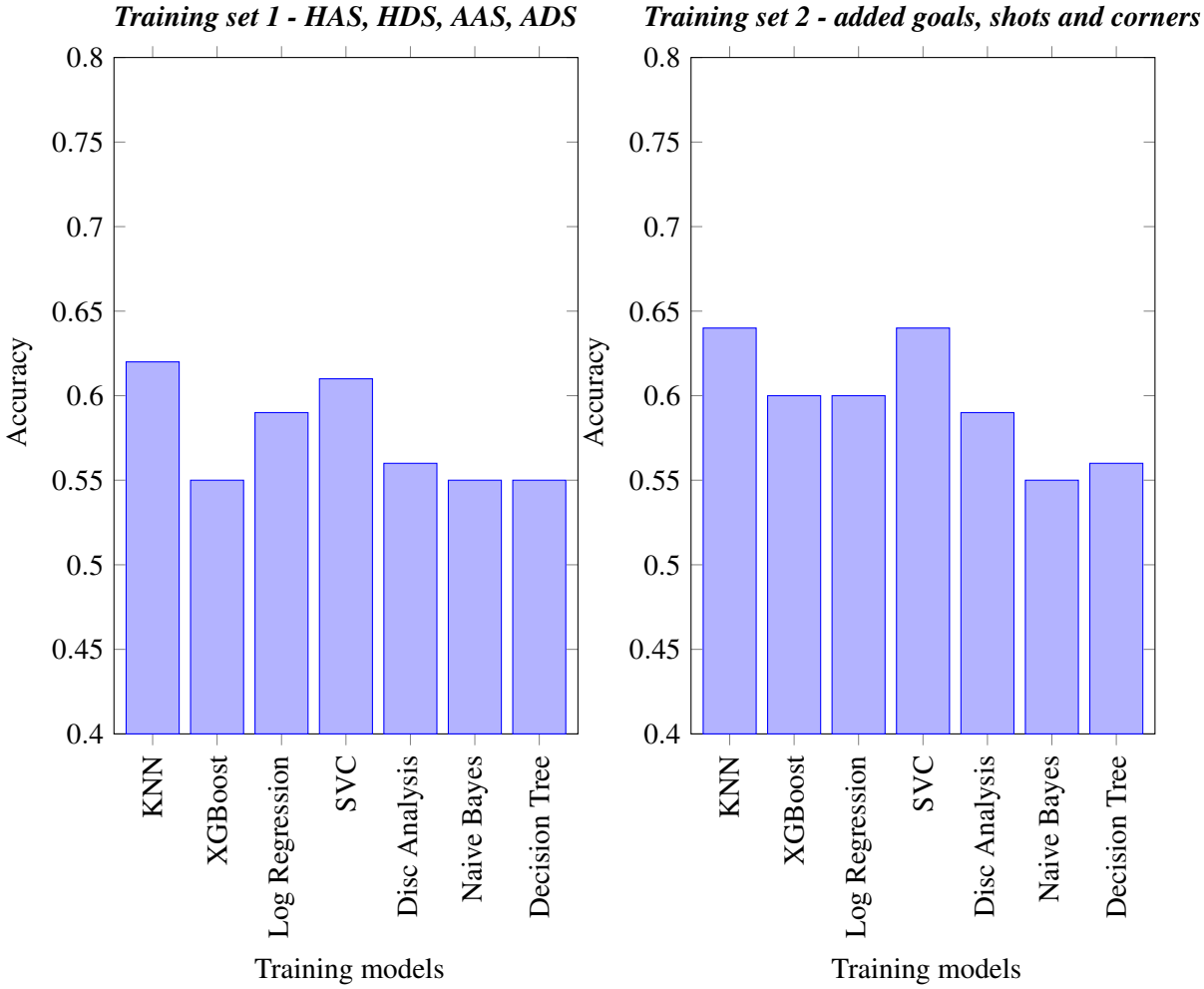

Figure 5.3: Comparison between Training set 1 and Training set 2

In a first analysis, it would be expected that the results would be better in training set 2, since that usually who has more shots is closer from the win and the corners are synonymous of possible good chances to score.

The reality is that comparing both graphics it is possible to see a little improvement from training set 1 to training set 2, specially on XGboost, SVC and Disc Analysys algorithms, but not the expected and the not the required for a final result. Also a good sign is that no algorithm had worse results in relation to the first experiment.

5.3 Results from recent matches

The next attempt to improve accuracy in our football results prediction system will be based in teams recent form. It is know by all that momentum is a really important factor in football and

any other sport. And every athlete or sports team that comes from good results, even if it doesn't come from such good performances, is always closer from winning the next match. So what was done here was try to transfer the theory referred to what is being studied here in practice.

In a first approach, what was done for the situation presented here, was to take the last five matches the home and the away team played before the game in question. With for every match in the training set were created two arrays, one for the home team and other for the away team. Each array had the outcome of the last five matches of the respective team. If the team had won the match, it was added the value 1 to the array. If it was a draw, the value 0 was added and if it was a loss, the value -1 was the one added. Looking for the example of Arsenal in Round 38 (last the game of the season) vs Wolves, the array that refers to Arsenal would be something like **[-1, 1, 1, -1, -1]**, meaning that in the last five matches, Arsenal won two, loss three and didn't have any draw. Once it is not possible to use arrays in training sets, it was created another two variables that corresponded to the sum of the values of each array. Looking back at the array of the Arsenal example, this new variable would be -1 **(- 1 + 1 + 1 - 1 - 1 = -1)**.

So, for this third training set were used the variables used in the last two training sets (*HAS, HDS, AAS, ADS, pastGoalDiff, pastShotsDiff, pastCornerDiff)* plus this two new variables ***sumResH*** and ***sumResA***:

sumResH = Sum Last 5 Results Home Team

sumResA = Sum Last 5 Results Away Team

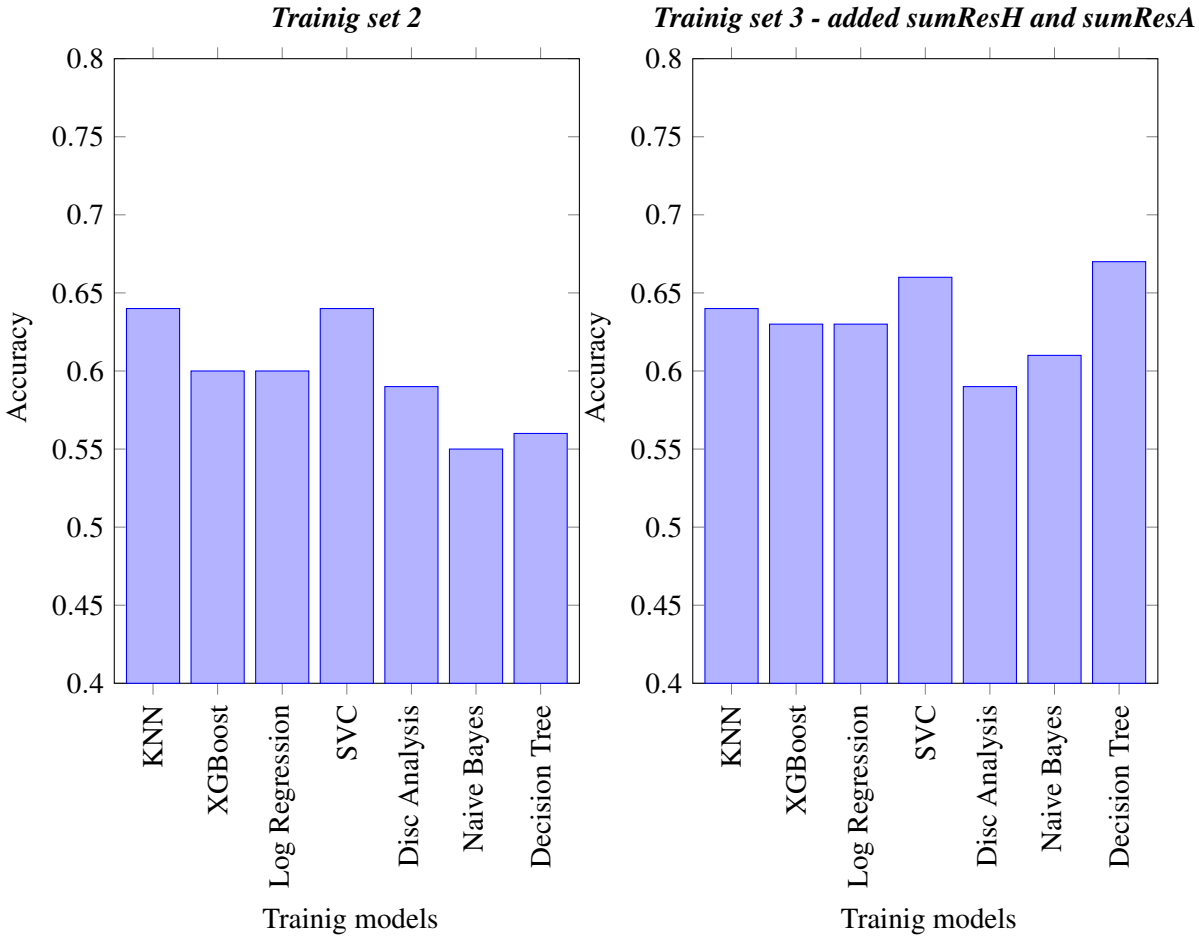

Figure 5.4: Comparison between Training set 2 and Training set 3

The two graphs above show the results from training set 2 and training set 3 in order to easier observe the evolution of the accuracy results from the second experiment to the third experiment. Looking to both of them, it is possible to notice an improvement in the results of every algorithm again altough minimal in some cases (except *Linear Discriminant Analysis* that stayed exactly with same accuracy). Also in this third training set, every algorithm reached values of 60% at least which is a very good indicator. Highlight also for the decision tree algorithm, which acieved an accuracy of 67

5.4 Exploring Home Advantage

Despite the good results in the last experiment with good improvements and an overall accuracy of above 60%. There is still a very important factor to explore in the experiments. Has it is known by everyone, in every sports, the Home Advantage factor is something that can make big difference in the outcome of a game. Every season in every championship, there are teams that perform much more positively in home games than away. With all this in mind, it was decided to explore this Home Advantage present in football adding some more parameters to our training set.

The approach taken in this section was to use the same logic used in the last training set, i.e. take the last five games each team made and build an array both of them. The main difference here is that, in the case of the home team, it would be the games played at home until that round, and in the case of the away team, it would be all the games played away in that season also until that round. This would emphasize both teams form in the last games playing in similar conditions of the next game and in case of a club using the home factor quite efficiently, that would also be taken into account.

Using again the same example used before, Arsenal at round 38 vs Wolves, this time the array would be **[1, 1, 1, 1, 1, 1, 1, 0, 1, 0, -1, 1, 1, 1, 1, 0, 1, -1]**, meaning that in all games play at home in that season, Arsenal only lost two and drew three, winning all remaining games which is a very good indicator for them to win the next one. As in the previous section, two other variables corresponding to the sum of the array were created. The ***sumResH_HM*** and ***sumResA_AM***. The ***sumResH_HM*** in this example would be 2 (**1 + 1 + 1 + 1 + 1 + 1 + 1 + 0 + 1 + 0 - 1 + 1 + 1 + 1 + 1 + 0 + 1 - 1 = 11**).

sumResH_HM = Sum Results Home Team Played at Home

sumResA_AM = Sum Results Away Team Played Away

Next two graphs compare the results of the training set 3 and this new training set with the home advantage added.

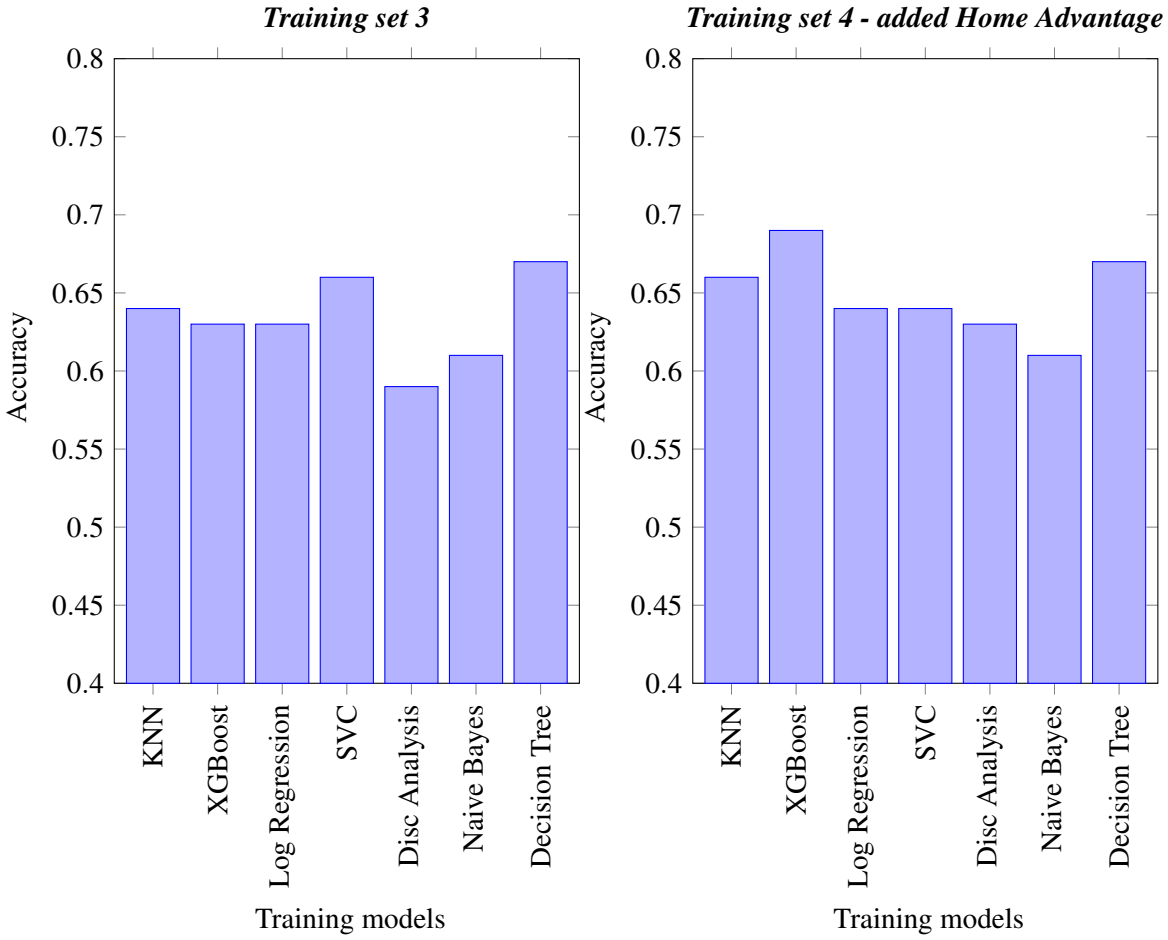

Figure 5.5: Comparison between Training set 3 and Training set 4

Now with the results of this last training set that emphasises more the home advantage for teams we were able to improve even more the results, now with every algorithm definitely above 60%. In SVC algorithm there is a slight decrease in accuracy but in every other the results got better or stayed the same. Highlight for XGBoost algorithm that managed to reach almost 70% accuracy which a very good number when talking about predicting football results.

Since these are the best results achieved with the algorithms used, it was decided to compare the initial graph of the initial training set (training set 1) composed only of goals statistics with the final training set (training set 4) that has every parameters from the *Home Attack Strength* to the the *SumResH_HM*. The results have improved significantly from one experiment to the other, specially in some cases like the XGBoost algorithm that went from a 55% accuracy to 69% and Decision Tree that went from 55% to 67% accuracy which are really good improvements.

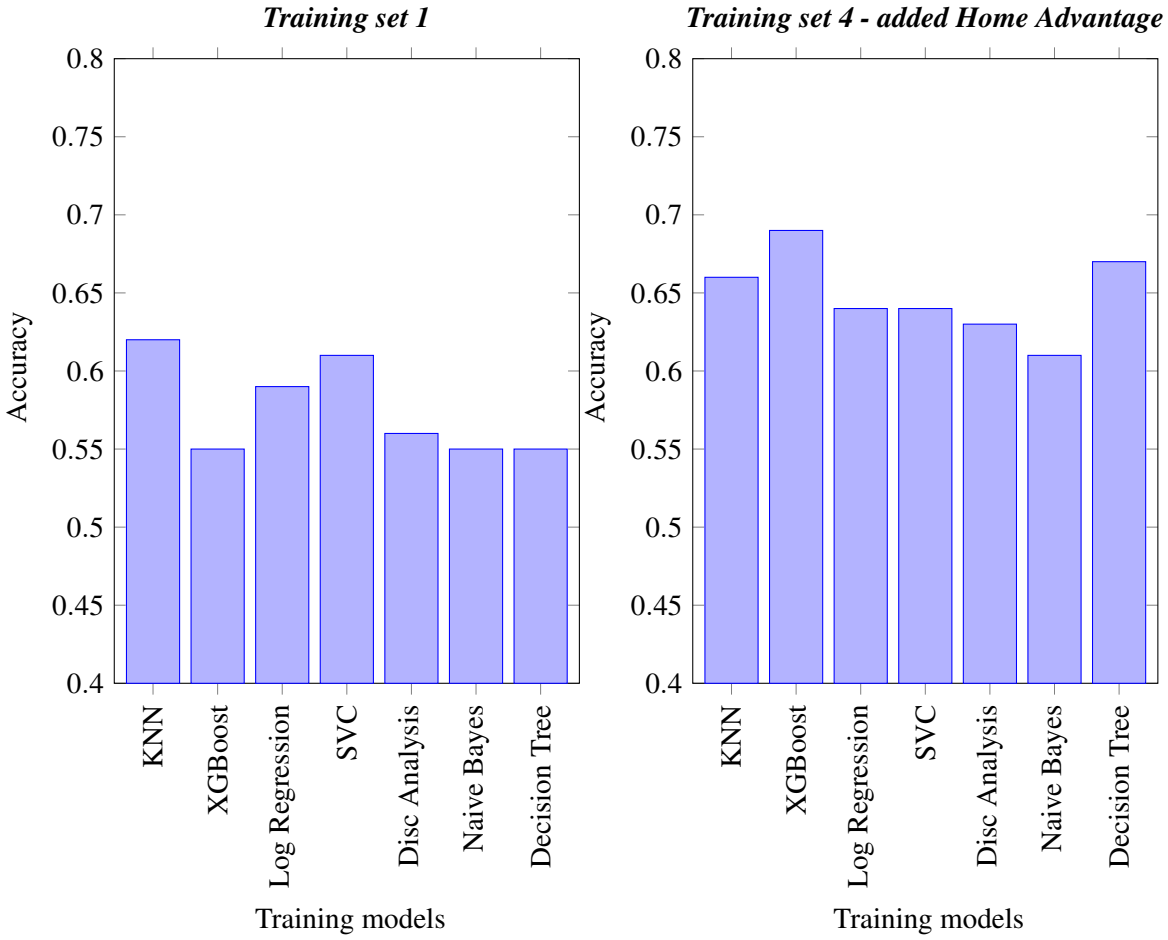

Figure 5.6: Comparison between Training set 1 and Training set 4

5.5 Results prediction in the main European Leagues

After analysing in a more deeply way the top division of English Football, it's time to look to other top leagues in Europe and see how the training algorithms behave with them. For that, the training algorithms were run with all the parameters as in training set 4. And after that, results will be compared between all leagues used.

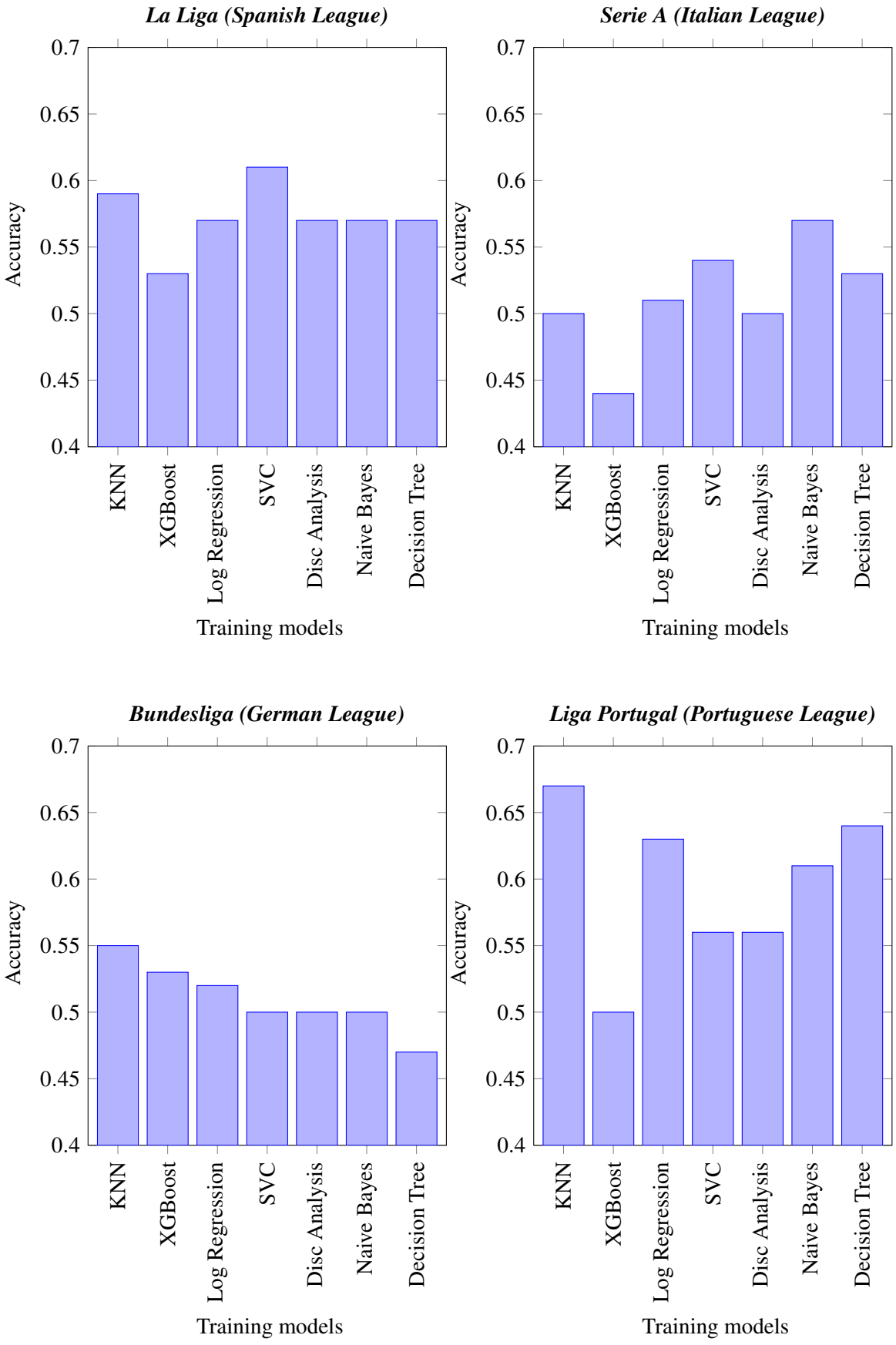

Figure 5.7: Comparison of Training set 4 between main European Football Leagues

With the experiments made for all the leagues and with all the graphs together, it is possible to say that any of them could reach the results obtained with the English Premier League. This could be justified maybe because this algorithms, the arguments used in them and the parameters used in the training sets were all tested and chosen after several tests and attempts to improve the effectiveness by looking only at the English League. Taking this into account, lower results would be expected. However, average values of accuracy were good, being above 50% in every league. It should be also noted that of the four leagues mentioned, Spanish and Portuguese are the ones with the highest levels of accuracy. Mainly the Portuguese that reached values above 60% more frequently.

Now analysing the main purpose of this section, it is noticeable that KNN algorithm is probably the most suitable for football matches prediction. Since it was the one with better accuracy in German and Portuguese Leagues and the second one in Spanish League. Only in the Italian League was it not so accurate but still kept at 50% accuracy.

Chapter 6

Datasets Over-sampling and Under-sampling

The imbalance of data in football match prediction datasets poses challenges for accurate model training and prediction. In many cases, football match prediction datasets suffer from class imbalance, where one class (e.g., home team win, draw, or away team win) is significantly over represented compared to others. This class imbalance can negatively impact the performance of prediction models. Therefore, effective data balancing techniques such as over-sampling and under-sampling are crucial to address this issue and improve the accuracy and reliability of soccer match prediction models.

The primary objective of this chapter is to investigate and analyze the application of over-sampling and under-sampling techniques in the context of football match prediction datasets. It aims to evaluate the impact of data balancing on the performance of football match prediction models and identify the most effective strategies for achieving accurate predictions.

6.1 Over-sampling Techniques

Class imbalance occurs when one class is significantly more prevalent than others, posing challenges for accurate model training and prediction. Over-sampling techniques aim to increase the representation of the minority class by generating synthetic samples or duplicating existing instances. By examining and evaluating popular over-sampling methods such as random over-sampling and the Synthetic Minority Over-sampling Technique (SMOTE), this section aims to provide insights into their implementation, benefits, and potential considerations when applied to football match prediction datasets. The effectiveness of these techniques in improving the performance of prediction models will be assessed using relevant evaluation metrics.

6.1.1 Random Over-sampling

This section delves into the random over-sampling technique, which involves randomly duplicating instances from the minority class until a balanced dataset is achieved. The chapter discusses

the implementation process, potential challenges, and the impact of random over-sampling on the performance of football match prediction models.

In the original dataset, the counter for the number of Home Wins, Draws and Away Wins was ***Home Wins: 144, Away Wins: 93, Draws: 74***. After resampling the dataset with the ***RandomOverSampler()*** function, the counter went to ***Home Wins: 144, Away Wins: 144, Draws: 144*** as it was expected.

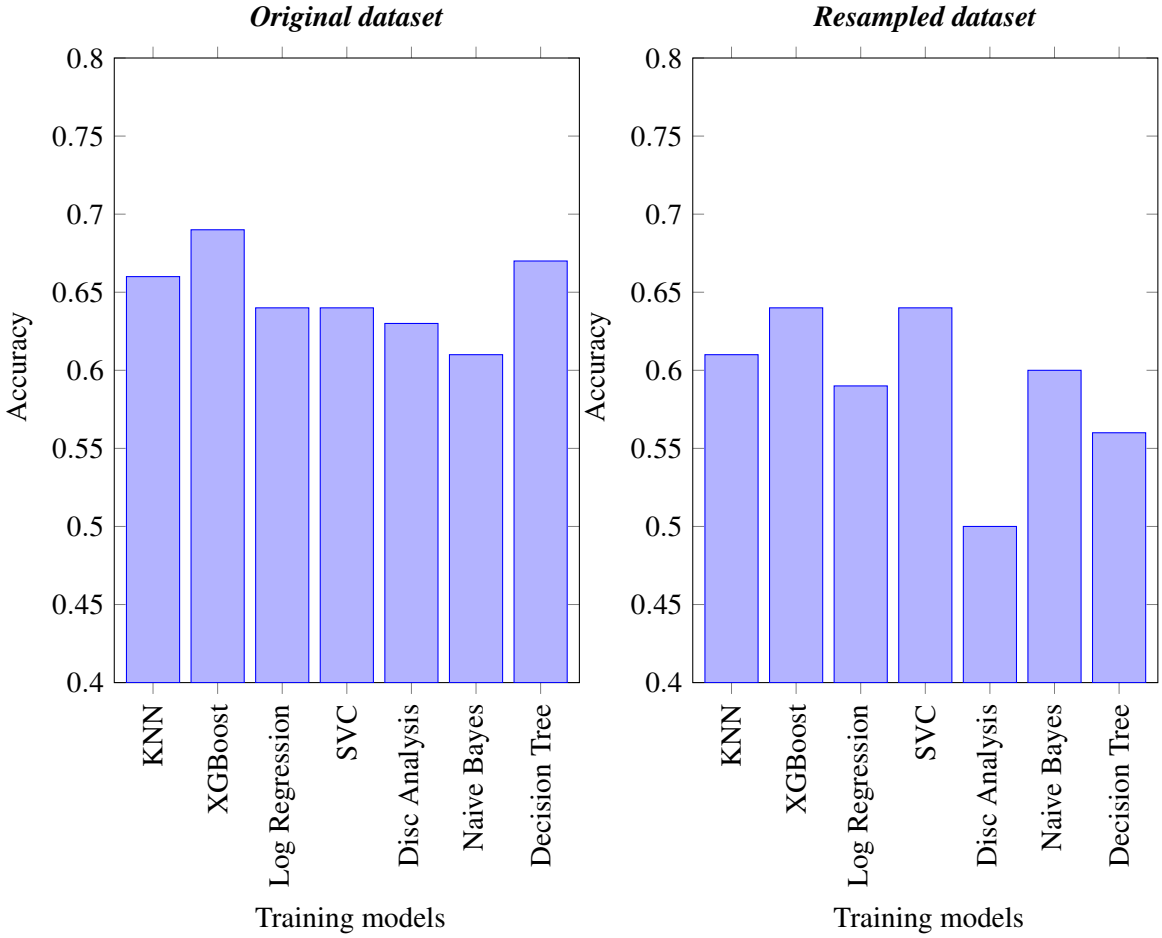

Figure 6.1: Comparison between Original Dataset and Resampled Dataset with RandomOverSampler

Upon analyzing the accuracy values, some interesting patterns emerge. The results indicate that the random over-sampling technique has varying effects on different algorithms. For instance, KNN and XGBoost show a slight decrease in accuracy from 0.66 to 0.61 and 0.69 to 0.64 respectively. The same happened with Logistic Regression dropping from 0.64 to 0.59. Furthermore, Linear Discriminant Analysis and Decision Tree classifiers exhibit significant decreases in accuracy, with Linear Discriminant Analysis dropping from 0.63 to 0.50 and Decision Tree dropping from 0.67 to 0.56. On the other hand, Bernoulli Naive Bayes and Linear SVC showcase consistent accuracy values between the normal dataset and the dataset with random over-sampling.

The findings from this experiment suggest that the impact of random over-sampling on accuracy varies depending on the algorithm employed. While some algorithms demonstrate stable or slightly decreased accuracy values, others experience more substantial decreases. It is important to consider these results and assess the overall performance of the algorithms based on additional evaluation metrics. Moreover, the suitability of random over-sampling as a data balancing technique may depend on the specific characteristics of the soccer match prediction dataset and the algorithm being used.

6.1.2 Synthetic Minority Over-sampling Technique (SMOTE)

Exploring now the SMOTE algorithm, SMOTE (Synthetic Minority Over-sampling Technique) is a popular algorithm used to address class imbalance in machine learning datasets. It focuses on the minority class by generating synthetic instances to balance the class distribution. The key idea behind SMOTE is to create synthetic samples by interpolating between existing minority class instances.

As in the Random Over-Sampling, in the original dataset, the counter for the number of Home Wins, Draws and Away Wins was ***Home Wins: 144, Away Wins: 93, Draws: 74*** and after resampling the dataset with the ***SMOTE()*** function, the counter went to ***Home Wins: 144, Away Wins: 144, Draws: 144*** as it was expected.

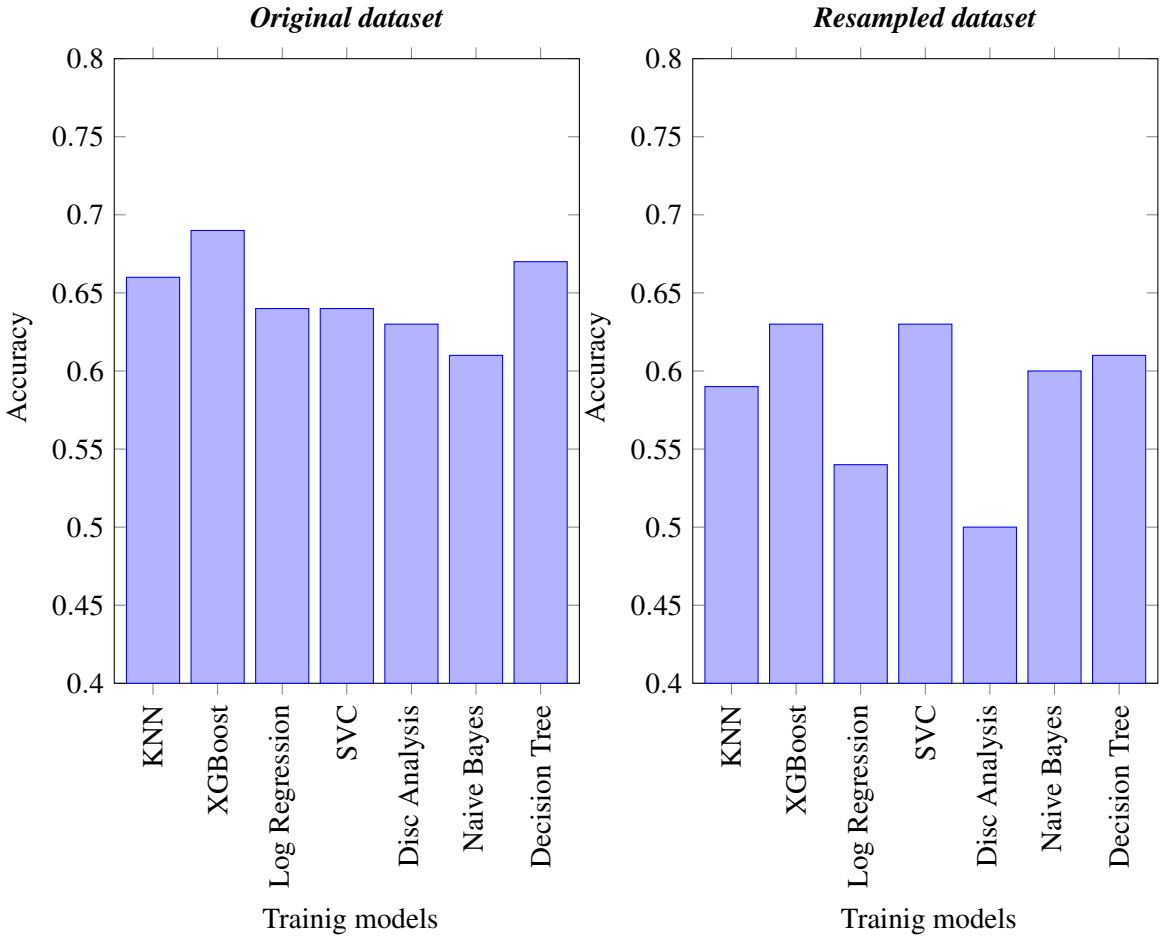

Figure 6.2: Comparison between Original Dataset and Resampled Dataset with SMOTE

Upon analyzing the results, we observe that the application of SMOTE has varying effects on the accuracy of different algorithms. While some algorithms, such as KNN and XGBoost, experience a decrease in accuracy, others, like Linear Discriminant Analysis, show a more substantial drop. Logistic Regression also demonstrates a significant decrease in accuracy.

The evaluation of the soccer match prediction dataset with and without the application of SMOTE reveals interesting insights. The accuracy of different algorithms varied when SMOTE was applied, with some algorithms experiencing a decrease in accuracy while others showed more significant drops.

These findings emphasize the importance of carefully considering the impact of SMOTE on the performance of prediction models for soccer match prediction tasks. While SMOTE can help address class imbalance and improve the accuracy of certain algorithms, its effectiveness may vary depending on the specific dataset and the algorithms being employed.

6.2 Discussion and Practical Implications

The results revealed valuable insights regarding the trade-offs between class imbalance correction and accuracy optimization. When evaluating the dataset with random over-sampling, we observed mixed outcomes. While some algorithms have shown to maintain accuracy, others experienced a decrease. This indicates that random over-sampling may not be the most effective approach for maximizing accuracy in soccer match prediction.

Similarly, when applying SMOTE to address class imbalance, the results showed a mixed impact on the accuracy of different algorithms. Some algorithms experienced a decrease in accuracy, suggesting that SMOTE may not always lead to improved performance.

Furthermore, it is important to consider the specific characteristics of soccer match prediction datasets. The input of away wins may possibly biase the home advantage factor, since the parameter used in the training set to measure the strength of a team at home or away is based on the number of victories obtained under the same conditions. These factors may contribute to the decrease in accuracy observed with certain algorithms.

Considering the main goal of achieving the highest accuracy possible in soccer match prediction, the findings suggest that blindly applying over-sampling techniques, whether random or SMOTE, may not be the most effective strategy. Instead, a more nuanced approach is required, taking into account the specific dataset characteristics and carefully selecting the algorithms and sampling techniques that align with the problem domain.

Future research should focus on exploring alternative methods that address class imbalance while minimizing the potential decrease in accuracy.

Chapter 7

Generating betting odds with machine learning algorithms

The world of sports betting is bigger then ever and it is growing more than every other type of gambling. Specially in football with the bookmakers being the main sponsors of so many teams in the major leagues and beyond, almost every person that watches football frequently have made at least a bet in some moment of their lives.

With that, it was decided that would be interesting to delve into the process of generating odds using the machine learning algorithms discussed in previous chapters for football match prediction.

We aim to explore how these algorithms can be leveraged to generate odds that reflect the likelihood of specific outcomes in soccer matches. Additionally, we compare the generated odds with those provided by bookmakers to assess the effectiveness and accuracy of our models in predicting match outcomes.

7.1 Methodology

To generate odds, we utilize the preprocessed and feature-engineered dataset obtained in previous chapters. The goal here was to approximate the most to the odds made by the bookmakers, so we used the ***Training Set 4*** since it is the one in which the best results were obtained and therefore, logically, the generated odds will be more in line with what is supposed. The algorithms used were the same used earlier, however, only the ones that reached accuracy values above 65% will be refered (***KNN, XGBoost, Decision Tree***).

7.1.1 Odds Generation

Before obtain the decimal odds (the most commonly used by sports bettors in Europe), the goal is to get the probabilities of each event to happen in every match. The process is very similar to the one that took us to the accuracy values. The algorithms are trained using the training set selected, but then it comes what differentiates this from what was done in Chapter 5. In Chapter

5 the function that generates the results (Home Win, Draw or Away Win) from every match was the *predict()* function. Here, what was intended was not the expected result but the probabilities of each of the three results happening. For that, it was used the *predict_proba()* function that generates these same probabilities for the games we select.

For better understanding of what *predict_proba()* returns, an example is shown bellow using the Round 31 of the Premier League 2022/2023 season. The function returns an array in which line represents a match. First element of the line is probability of Home Team winning, second is for the Draw and the last element of each line is the probability for the Away Team win.

```
1 print(XGB_probabilities)
```

```
[[0.16848354 0.4674202  0.36409622]
 [0.44248223 0.4508319  0.10668585]
 [0.21702504 0.44895485 0.33402014]
 [0.27523723 0.6728246  0.05193809]
 [0.11327855 0.19459468 0.6921268 ]
 [0.14524974 0.5255827  0.3291676 ]
 [0.28604624 0.38129276 0.332661  ]
 [0.23598082 0.17916416 0.584855  ]
 [0.13371843 0.5970034  0.26927814]
 [0.23794828 0.24343193 0.51861984]]
```

Figure 7.1: 2022/2023 Premier League Season's Round 31 Matches Outcome Probabilities

After getting the probabilities, it was necessary to convert it into the format of decimal betting odds. For that it was used the conversion formula $Odd = 1//Probability$. After converting every value, it was obtained a table like this.

```
1 print(XGB_odds)
```

```
[[ 5.935298   2.1394026  2.7465267]
 [ 2.2599778  2.2181218  9.373315 ]
 [ 4.6077633  2.2273955  2.9938314]
 [ 3.6332295  1.4862714 19.253693 ]
 [ 8.827796   5.1388865  1.444822 ]
 [ 6.884694   1.9026502  3.037966 ]
 [ 3.4959383  2.6226566  3.0060632]
 [ 4.2376323  5.5814734  1.7098254]
 [ 7.4784007  1.6750324  3.7136323]
 [ 4.202594   4.1079245  1.9281946]]
```

Figure 7.2: 2022/2023 Premier League Season's Round 31 Matches Generated Odds

7.2 Comparing Generated Odds

To assess the accuracy of our generated odds, we compare them with the odds provided by bookmakers. We collect odds data from one of the most famous bookmakers in the world (BET365) for the same set of matches in our test dataset. We then calculate the discrepancy between our generated odds and the bookmaker odds for each match outcome.

Before make all the comparisons it was necessary to adjust the odds from the bookmaker. It is known that in every bookmaker the probabilities assigned to different outcomes do not add up to 100%. . This phenomenon occurs due to the inclusion of the bookmaker's margin or overround. The bookmaker sets odds in such a way that their revenue is guaranteed regardless of the outcome of the event. By incorporating a margin, the bookmaker ensures a profit margin while still offering odds that entice bettors. Therefore, when examining odds from bookmakers, it is important to recognize that the probabilities implied by the odds will not sum up to 100% and that the margin plays a significant role in shaping the odds landscape. With that in mind, it was made an adjustement in every betting odd used for the comparisons.

The following image shows an example of an odd of a match from the last Premier League season.

	HomeTeam	AwayTeam	B365H	B365D	B365A
220	Arsenal	Man City	2.9	3.5	2.35

Figure 7.3: 2022/2023 Premier League Match Odds

Normalizing the odds for this game they would naturally increase in every possible outcome and we would get something like the values in the image below.

	HomeTeam	AwayTeam	H_odd	D_odd	A_odd
220	Arsenal	Man City	3.06	3.7	2.48

Figure 7.4: 2022/2023 Premier League Match Normalized Odds

Now that we have the generated odds and that the odds from the bookmakers have been normalized, we are able to compare them and see if there are big discrepancies between them.

7.2.1 Comparing Generating Odds From Different Algorithms

To start, it was decided to generate the odds from Round 32 of the Premier League 2022/2023 season, with the three machine learning algorithms mentioned already (***KNN, XGBoost, Decision Tree***) and it will be compared the odds each of them generated.

Looking for table with the odds generated, the first situation that calls attention is in the Decision Tree algorithm will *null* values and 1.00 odds. Supposedly, that means that the team with

Generated Odds			
Match	KNN	XGBoost	Decision Tree
Man City vs Arsenal	1.42 - 5.67 - 8.50	1.40 - 6.17 - 8.05	1.35 - 7.46 - 8.08
Everton vs Newcastle	8.50 - 2.43 - 2.13	13.15 - 5.20 - 1.37	null - null - 1.00
Southampton vs Bournemouth	4.25 - 5.67 - 1.70	3.44 - 3.78 - 2.25	2.47 - 4.37 - 2.73
Tottenham vs Man United	2.43 - 4.25 - 2.83	7.55 - 3.18 - 1.81	3.00 - null - 1.50
Crystal Palace vs West Ham	1.42 - 5.67 - 8.50	5.14 - 2.07 - 3.09	2.47 - 4.37 - 2.73
Brentford vs Nott'm Forest	1.89 - 3.40 - 5.67	2.27 - 2.40 - 6.92	2.07 - 2.64 - 7.25
Brighton vs Wolves	1.54 - 4.25 - 8.50	2.17 - 4.19 - 3.33	1.35 - 7.46 - 8.08
Bournemouth vs Leeds	2.43 - 3.40 - 3.40	1.48 - 11.07 - 4.30	2.47 - 4.37 - 2.73
Fulham vs Man City	8.50 - 2.83 - 1.89	9.26 - 1.48 - 4.65	null - 1.00 - null
Man United vs Aston Villa	1.54 - 4.25 - 8.50	1.43 - 6.37 - 6.82	1.35 - 7.46 - 8.08

Table 7.1: Odds Generated by Machine Learning Algorithms

the 1.00 odd is clearly favorite and the chances of the outcome with the *null* value to happen are practically zero.

In some matches the three algorithms generate similar odds like in the first match of the table (Man City vs Arsenal), each algorithm gives favoritism to the draw with odds from 1.35 to 1.42 and the least favorite to win is Arsenal with odds going from 8.05 to 8.50.

In other cases, it happens exactly the opposite like in the penultimate match (Fulham vs Man City). In this case, for the KNN algorithm the favorite to win was Man City with a 1.89 odd, but for the XGBoost algorithm the Fulham win is the outcome most likely to happen with an odd of 1.43. The Decision Tree algorithm in this same game give clearly favoritism to Fulham with null values for the other two possible outcomes.

7.2.2 Comparison With Bookmakers Betting Odds

Now looking for Bet365 betting odds from this same games, will be made comparisons with the values of table in the earlier subsection and try to see which algorithm has closer values from the Bet365 values.

7.2.2.1 KNN vs Bet365

Starting with the KNN algorithm, the table bellow shows the values from the KNN algorithm and Bet365 from the same Round 32 of the Premier League 2022/2023 season. Also, was added FTR column that shows the outcome of each game to better compare which odds are more suitable.

Analysing the table, something that is possible to see is that like in bookmakers betting odds, KNN very rarely puts the lower odd (most probable outcome) in the Draw. In the case of bookmakers, the Draw is never the outcome with the lowest odd. More often happens the opposite, in games that are expected to be very balanced, the Draw is always the highest draw. With KNN, this also never happened in this ten games.

Generated Odds			
Match	FTR	Bet365	KNN
Man City vs Arsenal	H	1.64 -4.76- 5.55	1.42 - 5.67 - 8.50
Everton vs Newcastle	A	5.07 -3.84- 1.84	8.50 - 2.43 - 2.13
Southampton vs Bournemouth	A	2.36- 3.57 -3.36	4.25 - 5.67 - 1.70
Tottenham vs Man United	D	2.94 -3.79- 2.52	2.43 - 4.25 - 2.83
Crystal Palace vs West Ham	H	2.77- 3.42- 2.89	1.42 - 5.67 - 8.50
Brentford vs Nott'm Forest	H	1.73 -3.99- 5.78	1.89 - 3.40 - 5.67
Brighton vs Wolves	H	1.58- 4.55 -6.83	1.54 - 4.25 - 8.50
Bournemouth vs Leeds	H	2.67- 3.67 -2.83	2.43 - 3.40 - 3.40
Fulham vs Man City	A	12.60- 6.30- 1.31	8.50 - 2.83 - 1.89
Man United vs Aston Villa	H	1.82- 4.10 -4.84	1.54 - 4.25 - 8.50

Table 7.2: Comparison of Bet365 Odds with KNN Odds

Now analysing game for game, there is some cases where the values differ a lot from one source to other. Like in Southampton vs Bournemouth (3rd row of the table), Bet365 previewed a very tight game with small advantage for Southampton and the highest odd for the Draw. For KNN, Bournemouth was the favourite with a 1.70 odd for winning the match, which is a quite favorable odd for the away team. The truth, is that Bournemouth ended winning the match, so KNN got the upper hand this time. Also, is possible to find games with very similar odds between the two sources as the Brentford vs Nott'm Forest and Brighton vs Wolves matches. In the case of the first match, the biggest variance is 0.59 in the Draw outcome.

Looking for the games each source predicted correctly, Bet365 gave the lowest odd for the right outcome in 8 of 10 matches and KNN algorithm had an accuracy of 9 out of 10 which is quite impressive.

7.2.2.2 XGBoost vs Bet365

Now with the XGBoost algorithm, the table bellow shows the values from the XGBoost algorithm and Bet365 from the same matches used for the KNN algorithm. After analysing the table, some curiosities present in the values were presented below.

In the case of XGBoost, this algorithm also gives favoritism to the Draw in 2 out of 10 games, which deviates a little from the values obtained at Bet365.

In general, the odds from XGBoost are not so close to the Bet365 ones as the KNN odds are. In such a way that there is no game to highlight in that aspect. What is possible to observe is that XGBoost has more difficult to give the favoritism to one of the teams, as the odds generated indicate in the majority of the cases a minimally balanced match.

XGBoost predicted the lowest odd correctly in 7 of the 10 games, which also a very good value but obviously not so great as the KNN algorithm and the Bet365 odds.

Generated Odds			
Match	FTR	Bet365	XGBoost
Man City vs Arsenal	H	1.64 -4.76- 5.55	1.40 - 6.17 - 8.05
Everton vs Newcastle	A	5.07 -3.84- 1.84	13.15 - 5.20 - 1.37
Southampton vs Bournemouth	A	2.36- 3.57 -3.36	3.44 - 3.78 - 2.25
Tottenham vs Man United	D	2.94 -3.79- 2.52	7.55 - 3.18 - 1.81
Crystal Palace vs West Ham	H	2.77- 3.42- 2.89	5.14 - 2.07 - 3.09
Brentford vs Nott'm Forest	H	1.73 -3.99- 5.78	2.27 - 2.40 - 6.92
Brighton vs Wolves	H	1.58- 4.55 -6.83	2.17 - 4.19 - 3.33
Bournemouth vs Leeds	H	2.67- 3.67 -2.83	1.48 - 11.07 - 4.30
Fulham vs Man City	A	12.60- 6.30- 1.31	9.26 - 1.48 - 4.65
Man United vs Aston Villa	H	1.82- 4.10 -4.84	1.43 - 6.37 - 6.82

Table 7.3: Comparison of Bet365 Odds with XGBoost Odds

7.2.2.3 Decision Tree vs Bet365

For the last the Decision Tree algorithm, the table bellow as in the other cases shows also the values from the Decision Tree algorithm and Bet365 from the same matches used for the other two tables.

Generated Odds			
Match	FTR	Bet365	Decision Tree
Man City vs Arsenal	H	1.64 -4.76- 5.55	1.35 - 7.46 - 8.08
Everton vs Newcastle	A	5.07 -3.84- 1.84	null - null - 1.00
Southampton vs Bournemouth	A	2.36- 3.57 -3.36	2.47 - 4.37 - 2.73
Tottenham vs Man United	D	2.94 -3.79- 2.52	3.00 - null - 1.50
Crystal Palace vs West Ham	H	2.77- 3.42- 2.89	2.47 - 4.37 - 2.73
Brentford vs Nott'm Forest	H	1.73 -3.99- 5.78	2.07 - 2.64 - 7.25
Brighton vs Wolves	H	1.58- 4.55 -6.83	1.35 - 7.46 - 8.08
Bournemouth vs Leeds	H	2.67- 3.67 -2.83	2.47 - 4.37 - 2.73
Fulham vs Man City	A	12.60- 6.30- 1.31	null - 1.00 - null
Man United vs Aston Villa	H	1.82- 4.10 -4.84	1.35 - 7.46 - 8.08

Table 7.4: Comparison of Bet365 Odds with Decision Tree Odds

Despite Decision Tree being the algorithm with the weirdest values of the three algorithms, it was able to keep accuracy in good levels as it managed to get the lowest odd correctly in 7 of the 10 games like the XGBoost algorithm. But with the games that it didn't choose correctly being different ones.

A curiosity present here is that in the Southampton vs Bournemouth (3rd row of the table), this algorithm was the only one of the three to give favoritism to Southampton like Bet365 did, despite the winner of the match being Bournemouth.

Decision Tree was also able to get some very similar odds to the Bet365 ones, being better than XGBoost in this topic. The main example is the Bournemouth vs Leeds game, with the less closer value being in the Draw outcome with a difference of 0.7 (4.37 - 3.67).

Chapter 8

Conclusions

In this work, we targeted two distinct areas that, nonetheless, quite intertwined. The main focus, was to explore the prediction of football results in the main European Leagues in the 2022/2023 season, but for that it was also necessary to do Data Exploration. Exploration allows for deeper understanding of a dataset, making it easier to navigate and use the data later. The better we know the data we are working with, the better the analysis will be. Successful exploration reveals new ways for the goal, and helps to identify future analytics questions and problems. With the dataset we had in hands, it was used Jupyter Notebook to explore the data. Jupyter Notebook is an interactive computing environment that supports Python, so we used a powerful python library called Pandas, and with it we are able to create tables and graphs in order to begin exploring and determining what patterns and trends are found in the dataset. The objective of this data exploration was to find patterns in the leagues that could help us to predict the outcomes of the matches, so the challenge here was to take statistics of the matches and the teams throughout the season and turn them into parameters of the training sets that would train the machine learning algorithms.

After all that process, and several attempts to improve the accuracy of the algorithms used, testing different parameters and different statistics of the matches and the teams, we were able to achieve accuracy rates of over 60% across all the algorithms tested and almost 70% with XGBoost and Decision Tree algorithms. These results seems to be quite interesting, since it was only used historical data, and in football, as any other sport, the outcome of any match is quite unpredictable and there are several other factors that come into play during a game. Player's form, strategic nuances in formation, players injuries, fatigue level and even the responsibility of playing against big rivals and teams constitute a major part of a team's performance in a match. Unfortunately, we weren't able to find data about such characteristics freely. Looking for data we used in the work, adding teams' recent form in the last matches (results from last five games in this case) and adding home advantage in the case of teams that perform much better at home turned to be quite important as the results suffered nice improvements after that.

Additionally, in this thesis we explored the impact of dataset balancing through over-sampling techniques. Two different over-sampling methods were used, RandomOverSampler and SMOTE,

and then results were compared with the original dataset. While initial expectations were in a optimistic way, the results demonstrated that oversampling did not significantly improve the accuracy of the models. This finding suggests that the inherent biases in the dataset were not effectively addressed through oversampling alone. Further investigation into alternative balancing methods or the incorporation of additional features to enhance model performance may be warranted in future search.

Another noteworthy aspect of this research was the generation of odds based on the predictions of the machine learning models. These odds were then compared with the odds offered by bookmakers known worldwide. The results obtained in this chapter were quite interesting, in some cases with really goods levels of accuracy based in the real outcome of the matches. The comparison yielded intriguing findings, in a quite considerable number of matches we were able to get odds similar to the betting odds by the bookmakers, but also in many cases revealing disparities between them. Some disparities were not so good for the machine learning algorithm side, as in those cases the values obtained were a little silly. Those situations are explained by being matches with particular characteristics, like injured players or teams fatigue, that were not taken into account in our study. On the other side, some disparities presented captivating opportunities for further exploration, indicating instances where the machine learning models outperformed bookmakers or identified potential value in betting markets. That may be also explained by not taking into account the individual quality of a team, which is something that has quite importance for the bookmakers, but that does not always translate into practice specially when a team comes from a bad sequence of games.

Overall, this thesis contributes to the field of sports prediction by highlighting the potential of machine learning algorithms in accurately forecasting football results. The attainment of above 60% accuracy across all algorithms showcases the robustness of the models in capturing intricate patterns and trends in historical data. Moreover, the comparison of generated odds with bookmaker odds underscores the potential value in utilizing machine learning models as decision support tools for identifying favorable betting opportunities.

8.1 Future Work

This research could be continued in different ways. As already mentioned, including features like players injuries, suspended players, fatigue level, strategic nuances in formation, would help us to get a better understanding of how to more accurately model the data available and also which factors contribute more to a team's victory.

Further research and analysis could be carried on by applying these models more deeply in other leagues and seasons, since the Premier League 2022/2023 season was the only one that was explored in every chapter of the work.

Moving forward, future research should also focus on refining dataset balancing techniques in order to at least don't drop the accuracy values so significantly. This would be made by emphasizing the need for a holistic and contextualized approach to dataset balancing in football prediction,

considering the unique dynamics of the sport and the multitude of factors that contribute to match outcomes.

References

[1] Maurizio Carpita, Enrico Ciavolino, and Paola Pasca. Exploring and modelling team performances of the kaggle european soccer database. *Statistical Modelling*, 19(1):74–101, 2019.

[2] Tom Carter. An introduction to information theory and entropy. *Complex systems summer school, Santa Fe*, 2007.

[3] Tianxiang Cui, Jingpeng Li, John R. Woodward, and Andrew J. Parkes. An ensemble based genetic programming system to predict english football premier league games. In *2013 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS)*, pages 138–143, 2013.

[4] Eric Dunning. *The development of soccer as a world game*. In Sports Matters: Sociological Studies of Sport Violence and Civilisation. Routledge, 1999.

[5] Gabriel Fialho, Aline Manhães, and João Paulo Teixeira. Predicting sports results with artificial intelligence – a proposal framework for soccer games. *Procedia Computer Science*, 164:131–136, 2019.

[6] Patrick Hall, Jared Dean, Ilknur Kaynar Kabul, and Jorge Silva. An overview of machine learning with sas® enterprise miner™. *SAS Institute Inc*, 2, 2014.

[7] Chinwe Peace Igiri. Support vector machine—based prediction system for a football match result. *IOSR Journal of Computer Engineering (IOSR-JCE)*, 17(3):21–26, 2015.

[8] Chinwe Peace Igiri, Oscar Uzoma Anyama, Abbasiama Ita Silas, and Iibi Sam. A comparative analysis of k-nn and ann techniques in machine learning. 2015.

[9] Sarika Jain, Ekansh Tiwari, and Prasanjit Sardar. Soccer result prediction using deep learning and neural networks. In Jude Hemanth, Robert Bestak, and Joy Iong-Zong Chen, editors, *Intelligent Data Communication Technologies and Internet of Things*, pages 697–707, Singapore, 2021. Springer Singapore.

[10] M. W. Y. Lam. One-match-ahead forecasting in two-team sports with stacked bayesian regressions. *Journal of Artificial Intelligence and Soft Computing Research*, Vol. 8, No. 3:159–172, 2018.

[11] Daily Mail. Global sports gambling worth 'up to 3 trillion dollars'. Available at http://www.dailymail.co.uk/wires/afp/article-3040540/Global-sports-gambling-worth-3-trillion.html, Accessed August 9, 2022, 2015.

[12] Vishal Morde and Venkat Anurag Setty. Xgboost algorithm: Long may she reign. *Towards Data Science*, 8, 2019.

[13] Nyquist R Pettersson D. Football match prediction using deep learning. Master's thesis, Doctoral dissertation, Master's thesis, Chalmers University of Technology, 2017.

[14] Bernard Joy Jack Rollin Eric Weil Richard C. Giulianotti, Peter Christopher Alegi. Football (association football, soccer). Available at https://www.britannica.com/sports/football-soccer, Accessed August 5, 2022, 2022.

[15] Alexander P Rotshtein, Morton Posner, and AB Rakityanskaya. Football predictions based on a fuzzy model with genetic and neural tuning. *Cybernetics and Systems Analysis*, 41(4):619–630, 2005.

[16] Abdulhamit Subasi. Chapter 3 - machine learning techniques. In Abdulhamit Subasi, editor, *Practical Machine Learning for Data Analysis Using Python*, pages 91–202. Academic Press, 2020.

[17] Eftim Zdravevski and Andrea Kulakov. System for prediction of the winner in a sports game. In Danco Davcev and Jorge Marx Gómez, editors, *ICT Innovations 2009*, pages 55–63, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.