



INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Football match result prediction: a business analytics approach

Diogo Miguel Filipe Romão

Master in Management

Supervisor:

Doctor Raul Manuel Silva Laureano, Associate Professor, ISCTE Business School

September 2023

Department of Marketing, Strategy and Operations

Football match result prediction: a business analytics approach

Diogo Miguel Filipe Romão

Master in Management

Supervisor:

Doctor Raul Manuel Silva Laureano, Associate Professor, ISCTE Business School

September 2023

ACKNOWLEDGEMENTS

To my parents, I want to express my gratitude for your constant support throughout the years and for always providing the best conditions for my development. Your dedication and encouragement were essential to my journey. To my sister, I thank you for always being there and for the unconditional support you have given me. Your presence was a constant source of motivation.

To my girlfriend, I lack words to express how much your support meant to me throughout this entire process. From the beginning, you closely followed every step of this thesis. You stood by my side during moments of doubt, during the most imaginative ideas the thesis brought, during times of stress and joy. Your tireless support was the foundation that allowed me to reach this point.

To my grandparents, even though you may not fully comprehend what a thesis is, I want to thank you for your constant encouragement and for frequently asking, "So, have you finished your thesis yet?" Today, with pride, I can answer yes, thanks to your support.

To Professor Doctor Raul Laureano, I want to express my gratitude for showing interest in my research topic from the very beginning. Your support and guidance were crucial in many of the decisions I had to make over this past year. My sincere thanks!

To Diogo Morgado, I appreciate your efforts and the interest you have shown in my work. Your collaboration has opened new horizons that we hope will bear fruit in the coming years.

To Professor Nuno Santos, I want to thank you for being readily available to help me in a quick meeting when everything seemed to be going wrong. It is gestures like these that make a difference, and I deeply appreciate them.

I want to thank all those who have been by my side, supported me, and believed in me throughout this journey. My most sincere thanks to each one of you.

Finally, but not least, I want to thank ChatGPT for its assistance in translating this thesis and for its 24 hours per day availability.

ABSTRACT

Currently, there is a gap in the market for predicting football match outcomes before a match takes place. To address this challenge, a predictive model based on pre-match variables was developed, focusing on analytical techniques, including machine learning (ML). This study focuses on football match outcome predictions and their relevance in the fields of sports, investments, and entertainment. The goal is to identify the most effective algorithms for predicting outcomes.

This study focus on five European leagues: the English Premier League (EPL), Spanish La Liga, Italian Serie A, German Bundesliga, and Portuguese Liga Bwin, for the seasons 2019-2020, 2020-2021, and 2021-2022. The predictive model proved promising in anticipating football match results. The C5 algorithm showed the best results, with an 85.40% accuracy on test data. Additionally, the model was validated in a new sports season, demonstrating its effectiveness in predicting future outcomes. This is essential to ensure the model reliability in future contexts and reinforces its utility for investors, football clubs, and fans. Through financial analysis, profitable strategies were identified, and a monitoring dashboard was built to track the financial performance of the predictive model and to evaluate the effectiveness of the developed investment strategies.

In summary, this study contributed to the understanding of football predictions, benefiting investors, football clubs, betting houses, and fans by providing a robust approach to anticipate match results.

Keywords: Football predictions; Machine Learning; Investments

JEL Classification System: C53, C63

RESUMO

Atualmente existe uma lacuna no mercado das previsões de resultados de futebol antes de um jogo ocorrer. Para resolver este desafio, desenvolve-se um modelo preditivo baseado em variáveis pré-jogo, com foco em técnicas analíticas, incluindo *machine learning* (ML). Este estudo foca-se assim nas previsões de resultados de jogos de futebol e na sua relevância nas áreas de desporto, investimentos e entretenimento. O objetivo é identificar os algoritmos mais eficazes para prever resultados.

Este estudo foca-se em cinco ligas europeias: a Premier League Inglesa, La Liga Espanhola, Serie A Italiana, Bundesliga Alemã e a Liga Bwin Portuguesa, para as épocas 2019-2020, 2020-2021 e 2021-2022. O modelo preditivo revelou-se promissor na antecipação de resultados de jogos de futebol. O algoritmo C5 foi o que apresentou melhores resultados, com uma *accuracy* de 85,40% em dados de teste. Adicionalmente, o modelo foi validado numa época desportiva nova, o que demonstra sua eficácia em prever resultados futuros. Isto é essencial para garantir a confiabilidade do modelo em contextos futuros e reforça a sua utilidade para investidores, clubes de futebol e adeptos. Foram identificadas, através de uma análise financeira, algumas estratégias lucrativas, tendo sido construído um *dashboard* de monitorização para acompanhar o desempenho contínuo das previsões e das estratégias de investimento.

Em síntese, este estudo contribuiu para o entendimento das previsões de futebol, beneficiando investidores, clubes de futebol, casas de apostas e adeptos ao fornecer uma abordagem sólida para antecipar os resultados dos jogos.

Palavras-chave: Previsões de futebol, Machine Learning, Investimentos

JEL Classification System: C53, C63

LIST OF ACRONYMS AND ABBREVIATIONS

A – Away

ANN – Artificial Neural Network

BN – Bayesian Network

CSL – Chinese Football Super League

CHAID – Chi-squared Automatic Interaction Detection

CLPZ – Conference League Playoff Zone

CPZ – Champions Playoff Zone

CRISP-DM – Cross-Industry Standard Process for Data Mining

CZ – Champions Zone

D – Draw

DB – Database

DF – Decision Forest

DM – Data Mining

DPZ – Relegation Playoff Zone

e.g., – For example

ELPZ – Europa League Playoff Zone

EPL – English Premier League

FT – Full time result

GB – Gradient Boosting

H – Home

HNN – Hybrid Neural Networks

HT – Half time result

KDD – Knowledge Discovery in Database

KNN – K nearest neighbours

LEZ – Europa League Zone

LR – Logistic Regression

LSTM – Long Short-Term Memory

MAE – Mean Absolute Error

MAPE – Mean Absolute Percentage Error

ML – Machine Learning

MSE – Mean Square Error

MZ – Maintenance Zone

N.A. – Not available

NB – Naive Bayes
NN – Neural Networks
OCL – Original Challenge learning set
PL – Profit-loss
RMSE – Root-Mean-Square Error
RNN – Recurrent Neural Networks
ROI – Return on investment
RPS – Rank Probability Score
RT – Random Trees
RZ – Relegation Zone
SD – Standard deviation
SLR – Systematic literature review
SVM – Support Vector Machine
UI – Unit Investment
UK – United Kingdom
XGBoost – Extreme Gradient Boosting

INDEX

ACKNOWLEDGEMENTS	i
ABSTRACT	iii
RESUMO	v
LIST OF ACRONYMS AND ABBREVIATIONS.....	vii
INDEX.....	ix
INDEX OF TABLE	xi
INDEX OF FIGURES.....	xiii
1. INTRODUCTION	1
1.1. Theme and its importance	1
1.2. Research problem	1
1.3. Objectives and contributions	2
1.4. Methodological approach	2
1.5. Dissertation structure.....	3
2. SYSTEMATIC LITERATURE REVIEW	5
2.1. Protocol for the Systematic Literature Review	5
2.2. Critical analysis	8
2.2.1. Contexts of model application	8
2.2.2. Methodology	12
2.2.3. Results obtained from the models	17
2.2.4. Main contributions, limitations, and future recommendations	20
2.3. Articles evaluation.....	25
3. CRISP-DM	27
3.1. Business understanding.....	27
3.2. Data understanding.....	28
3.3. Data preparation	30
3.3.1. Dataset	30
3.4. Modeling	32

3.5.	Evaluation	35
3.6.	Deployment	36
4.	RESULTS AND DISCUSSIONS	39
4.1.	Predictive model for match outcome	39
4.2.	Most important predictors	40
4.3.	Game profiles with a high propensity for a result	42
4.4.	Other assessments of the predictive model	44
4.4.1.	Generalization to new data	44
4.4.2.	Financial perspective	47
4.5.	Discussion and practical implications	52
5.	CONCLUSIONS	55
5.1.	Summary	55
5.2.	Contributions	56
5.3.	Limitations	58
5.4.	Future research recommendations	58
	REFERENCES	61
	APPENDIX	65
	A: Variables extracted and excluded for analysis	65
	B: Data dictionary	66

INDEX OF TABLE

Table 2.1: Inclusion and exclusion criteria for articles in the SLR	6
Table 2.2: Articles under study.....	7
Table 2.3: Evaluation questions	8
Table 2.4: Contexts of the articles.....	9
Table 2.5: Description of the algorithm	11
Table 2.6: Methodology of the articles	12
Table 2.7: Dependent variable	14
Table 2.8: Common variables.....	15
Table 2.9: Other independent variables.....	15
Table 2.10: Algorithm, type of validation and evaluation metrics.....	16
Table 2.11: Metrics definition	17
Table 2.12: Results of the best models	18
Table 2.13: Contributions and limitations of the articles.....	21
Table 2.14: Future recommendations.....	22
Table 2.15: Critical analysis of contributions and limitations	24
Table 2.16: Articles evaluation	25
Table 3.1: Variables used in preparation and modeling	29
Table 3.2: Description of each algorithm used in the research	32
Table 3.3: Parameterizations of the best models for each algorithm (part 1).....	34
Table 3.4: Parameterizations of the best models for each algorithm (part 2).....	34
Table 3.5: Dimensions of the best models for each algorithm	34
Table 3.6: Confusion matrix	35
Table 3.7: Metrics for evaluation	35
Table 4.1: Results of the algorithms in the test sample (part 1).....	39
Table 4.2: Results of the algorithms in the test sample (part 2).....	39
Table 4.3: Predictions of the model by championship (2022- 2023)	44
Table 4.4: Actual results of the matches by championship (2022-2023).....	44
Table 4.5: Confusion matrix in new data (2022-2023).....	45
Table 4.6: Profit-Loss by championship	47
Table 4.7: Top 5 most lucrative teams playing at home (2022-2023).....	47
Table 4.8: Profit-loss per matchday (2022-2023).....	48
Table 4.9: Predictions for the matchday 6 (2022-2023)	49
Table 4.10: Questionnaire responses.....	51

Table 4.11: Best algorithm	52
----------------------------------	----

INDEX OF FIGURES

Figure 2.1: Article selection process.....	6
Figure 2.2: Cloud word	7
Figure 3.1: Data sources	29
Figure 3.2: Sample selection process auxiliary dataset.....	31
Figure 3.3: Model training method	33
Figure 4.1: Top 10 - Variables importance	40
Figure 4.2: Average goals conceded by the away team per prediction	41
Figure 4.3: Average ball possession by the home team per prediction.....	41
Figure 4.4: Predicted outcome by home team position in the standings	41
Figure 4.5: Relative standings by the home team per prediction	41
Figure 4.6: Average accuracy vs accuracy per matchday	45
Figure 4.7: Home team precision vs home team precision per matchday	46
Figure 4.8: Draw precision vs draw precision per matchday	46
Figure 4.9: Away team precision vs away team precision per matchday	47
Figure 4.10: Dashboard for tracking the profit and loss (2022-2023).....	50
Figure 4.11: What is the most profitable prediction? (2022-2023)	50
Figure 4.12: Which is the most profitable league? (2022-2023).....	50
Figure 4.13: Which is the most profitable confidence interval? (2022-2023)	51

1. INTRODUCTION

Predictions in football have garnered increasing interest among researchers and sports enthusiasts. The ability to predict match results can provide competitors an advantage, including clubs, coaches, betting companies, investors, and even the fans as well. This introduction aims to explain the interest in football predictions and highlight how important they have become in recent years. The problem and the goals of this investigation are exposed, as well as, a brief discussion of the methodology employed, and the dissertation structure is covered.

1.1. Theme and its importance

Football is one of the most famous and far-reaching sports in the world. Consequently, anticipating the outcome of a football match represents a captivating challenge.

Predicting the outcome of a football match can also aid administrators and clubs in making better decisions to defeat opponents and, consequently, win competitions (Nivetha et al., 2022). Therefore, predicting football match results has attracted numerous individuals.

The vast quantity and diversity of factors that can influence a football match result, such as teamwork, players skills, weather conditions, and home-field advantage, have posed significant problems and difficulties in accurately predicting the outcomes of each match. As a result, in recent years, the subject of match result prediction in football has gained popularity, and various models have been developed to predict the attributes that lead a team to win, draw, or lose a match (Razali et al., 2017).

1.2. Research problem

In the context of football predictions, variables related to in-game actions, such as shots, fouls, and cards, which are dynamically updated throughout a match and have a direct impact on the result, are frequently used. Since these variables depend on the course of the match, making predictions becomes quite complex. Therefore, to make pre-game predictions, it is necessary to use static variables, which do not depend on any live events and are only dependent on past results.

The research problem is the difficulty of predicting the outcome before a match takes place, and it is common (e.g., Guan & Wang, 2022; Nivetha et al., 2022) to incorporate post-game results into variable models that can only be accessed after the match has ended, preventing, for instance, coaches and investors from adjusting their strategies based on these predictions.

As a result, the primary goal of the current study is to develop a predictive model that uses pre-game variables (variables known before the match has started) to predict the outcome of a football match. This aims to respond the following research question: Is it possible to predict the outcomes of football matches?

Based on the research problem, this study explores the effectiveness of various analytical techniques, including ML, and their practical utility that is, the ability to predict the results of future matches. To achieve this, it is necessary to analyse ML algorithms, their trends, historical football matches data, and to examine appropriate evaluation metrics to measure the accuracy of predictions.

1.3. Objectives and contributions

Considering the research question formulated, this investigation has the following main objectives: O1) To create a predictive model for predicting match outcomes (Home team wins, draw, away team wins); O2) To identify the most important predictors of a match result; O3) To identify profiles of matches where home team wins, draw, or away team wins; and O4) To develop a dashboard to track the financial performance of the predictive model and to evaluate the effectiveness of the developed investment strategies.

Accomplishing these objectives and, in turn, addressing the research question, offers several contributions. Firstly, to the literature, as there are still few articles addressing the topic of football predictions, and none of them reach a consensus on the best algorithm and main predictor variables. Secondly, the present model will also be relevant for football clubs and betting companies, as having access to the variables and to the developed model it can lead to better management of their business resources.

1.4. Methodological approach

This research, in addition to analysing the existing literature related to the research topic, adopts the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology.

This methodology outlines the process of converting data into information through data mining (DM) (Chapman et al., 1999) and is the most widely used and comprehensive when compared to other methodologies in data analysis and predictive model construction (Laureano et al., 2014).

In this research, the adopted methodology is based on secondary data, with a focus on quantitative data. The data is obtained from five European leagues: the EPL, Spanish La Liga, Italian Serie A, German Bundesliga, and Portuguese Liga Bwin, for the seasons 2019-2020, 2020-2021, and 2021-2022, resulting in a sample of 5254 football matches. To classify the study as a success in these

predictions, four success criteria are set, namely, a minimum accuracy, precision, recall, and F1 Score of 80%. Given the complexity of the problem and to fulfil the purpose of this study, analytical techniques are employed, including the C5, Chi-squared Automatic Interaction Detection (CHAID), Random Trees (RT), Support Vector Machine (SVM), and Neural Networks (NN) models.

Finally, there is a goal to create a dashboard, which purpose is to track the financial performance of the predictive model and to evaluate the effectiveness of the developed investment strategies. This dashboard will be evaluated by potential users, and the objective is deemed successful if the average evaluation score is higher than 8 on a scale from 1 (very dissatisfied) to 10 (very satisfied).

1.5. Dissertation structure

In order to predict the outcome of a football match, this work is organized into five chapters that reflect the different phases of the research.

In the first chapter, the topic and its importance are discussed, along with the research question, objectives, main contributions of the study, and the methodology adopted. The second chapter reflects the state-of-the-art on predicting football match results using analytical techniques, including ML. This chapter presents the protocol for the systematic literature review (SLR) and the four main dimensions analyzed (context of model application, methodology, results obtained from the models, and contributions, limitations, and future recommendations).

Regarding the methodology used, that is presented in the third chapter, where the entire process of data collection and processing is described, as well as the analytical techniques used in data analysis and the metrics for model evaluation. The fourth chapter presents and discusses the results obtained, and finally, in the fifth chapter, conclusions are presented, addressing contributions, limitations, and recommendations for future research.

2. SYSTEMATIC LITERATURE REVIEW

This SLR chapter focuses on identifying the models that have already been developed and their contributions to the advancement of football predictions. Therefore, studies that predict outcomes in football are described, highlighting their key characteristics such as the study purpose, methods and variables employed, key findings, and, when applicable, the authors recommendations for additional research.

Therefore, the goal is to analyse the work that has already been done about football predictions. To achieve this goal, this SLR aims to provide an answer to the following question: How can the outcome of a football match be predicted? Consequently, this study aims to respond as well to these three questions: (1) What contexts are models applicable to? (2) What methodologies are used? (3) How are the results obtained evaluated?

This chapter foundation is an SLR, which emphasizes this study reproducibility by characterizing it as a research methodology centred on certain protocol requirements (Caldas et al., 2017). The method for selecting the articles, including the criteria for including and excluding them, as well as an assessment of their quality, must be presented in the protocol in an explicit manner (Vilelas, 2017).

2.1. Protocol for the Systematic Literature Review

The articles included in this SLR underwent a preliminary selection process that took into account key words, previous research sources, and inclusion and exclusion criteria. The scientific data source used was Scopus (www.scopus.com), which was chosen due to its amplitude, practical and academic relevance. This data source was chosen because it contains publications pertinent to the study scientific fields and is widely used in SLR (Vilelas, 2017). The automatic research of articles published in journals or conference proceedings in the database (DB) was carried out by means of a query applied to the articles title.

Only the field title was searched for words or expressions. In the literature were identified key words used such as synonyms for football, outcomes, and machine learning. The key words that were used in the inquiry were football, results, machine learning, and estimation. The query used was: (football or "Premier League" or soccer) and (result* or outcome* or score or match or game*) and ("machine learning" or ML or "deep learning" or "data mining" or ai or "artificial intelligence" or "neural network*" or ann) and (estima* or prediction* or forecast* or predictive). This was approved by two experts in the field: a scholar and a professional for the football team "Os Belenenses". This way, it was possible to locate twenty-seven articles related to the topic. One of these articles was not included because of the impossibility to found it. This was not exhaustive of all algorithms, therefore the ad hoc

article talked about Bayesian Networks (BN), was not taken into consideration in the query related to algorithms. As a result, it was found by an ad hoc research project. Later, criteria for inclusion and exclusion were applied to this collection of articles (Table 2.1).

Table 2.1: Inclusion and exclusion criteria for articles in the SLR

Inclusion Criteria	Articles that address football predictions;
	Articles that address predictive models;
	Articles that discuss the use of Machine Learning techniques in football;
Exclusion Criteria	Book chapters, corrections and other documents;
	Articles whose objective is not to predict the final outcome of a football match;
	Articles that are not fully available;
	Articles with unclear methodology.

By applying the criteria for inclusion and exclusion criteria as well as the reading of the summary, introduction, and methodology, 12 documents that did not meet the entry requirements were removed, as shown by the Figure 2.1.

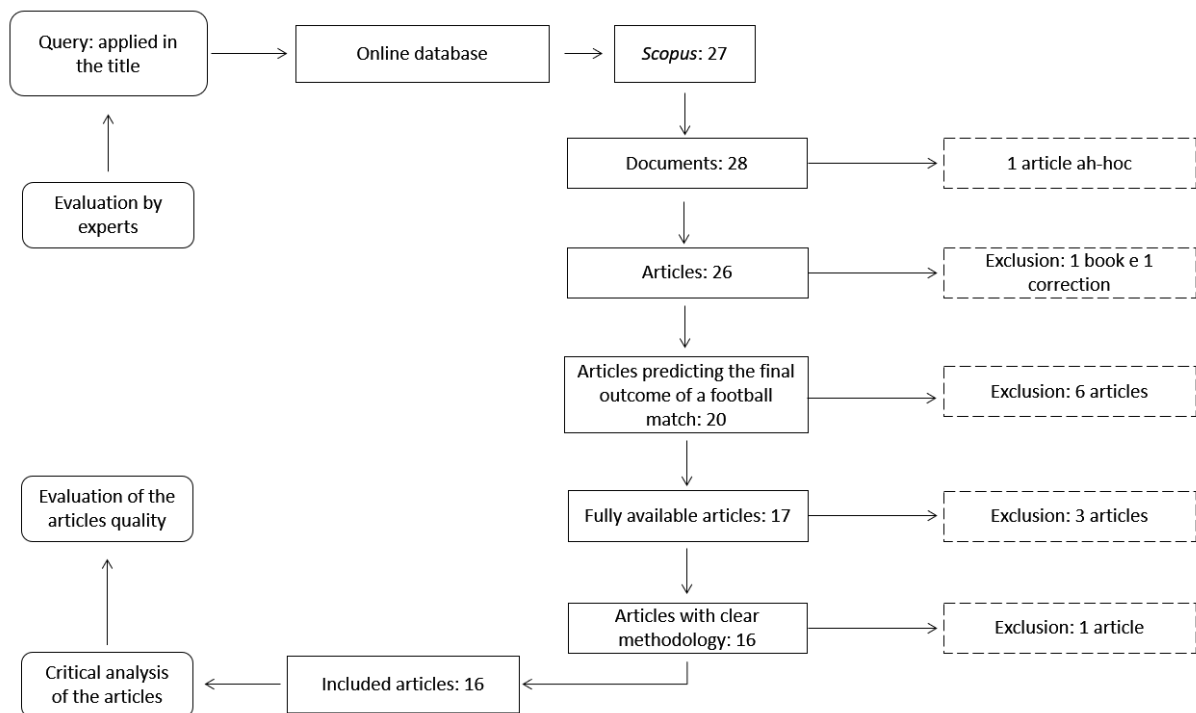


Figure 2.1: Article selection process

Following that, it is intended to discuss contemporary literature that falls within the parameters of the search, given that 90% of the literature found for this study is quite recent (roughly four years) in date. The 16 articles in Table 2.2, were also included, with a note that nine of them are published in proceedings of scientific conferences.

ID	Year	Title	Authors	Publication Place	Query on ah-hoc
1	2022	Optimization Analysis Of Football Match Prediction Model Based On Neural Network	Guan, S. & Wang, X.	Neural Computing and Applications	Query
2	2022	Football Result Prediction Based On Machine Learning	Yu, W.	IOS Press	Query
3	2022	A Deep Learning Framework For Football Match Prediction	Nivetha, S. K., Geetha, M., Suganthe, R. C., Prabakaran, R., Manoj, M. S. & Sameer, A. M.	Proceeding: International Conference on Computer Communication and Informatics	Query
4	2021	Prediction Of Football Matches Results: Decision Forest Against Neural Networks	Azeman, A. A., Mustapha, A., Razali, N., Nanthaaomomphong, A. & Abd Wahab, M. H.	Proceeding: 18th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology	Query
5	2021	A Machine Learning Approach To Football Match Result Prediction	Carloni, L., De Angelis, A., Sansonetti, G. & Micarelli, A.	Communications in Computer and Information Science	Query
6	2020	Football Match Result Prediction Using Neural Networks And Deep Learning	Tiwari, E., Sardar, P. & Jain, S.	Proceeding: 8th International Conference on Reliability, Infocom Technologies and Optimization	Query
7	2020	Football Match Outcome Prediction By Applying Three Machine Learning Algorithms	Azeman, A. A., Mustapha, A., Mostafa, S. A., AbuSalim, S. W.G., Jubair, M. A. & Hassan, M. H.	International Journal of Emerging Trends in Engineering Research	Query
8	2019	Predictive Analysis And Modelling Football Results Using Machine Learning Approach For English Premier League	Baboota, R. & Kaur, H.	Internacional Journal of Forecasting	Query
9	2019	Match Outcomes Prediction Of Six Top English Premier League Clubs Via Machine Learning Technique	Musa, R. M., P. P. Abdul Majeed, A., Mohd Razman, M. A. & Mohd Razman, M. A.	Communications in Computer and Information Science	Query
10	2018	Incorporating Domain Knowledge In Machine Learning For Soccer Outcome Prediction	Berrar, D., Lopes, P. & Dubitzky, W.	Machine Learning	Query
11	2018	Applying Neural Networks To Football Matches Results Forecasting	Korotyyeva, T., Tushnyskyyy, R. & Kulyk, V.	Proceeding: IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies	Query
12	2017	A Comparative Study Of Data Mining Techniques On Football Match Prediction	Rosli, C. M. F. C. M., Saringat, M. Z., Razali, N. & Mustapha, A.	Proceeding: 1st International Conference on Computing, Technology, Science and Management in Sports	Query
13	2017	Predicting Football Matches Results Using Bayesian Networks For English Premier League (EPL)	Razali, N., Mustapha, A., Yatim, F. A. & Ab Aziz, R.	Proceeding: International Research and Innovation Summit	Ah-hoc
14	2010	A Neural Network Method For Prediction Of 2006 World Cup Football Game	Huang, Kou-Yuan. & Chang, Wen-Lung.	Proceeding: 6th IEEE World Congress on Computational Intelligence, International Joint Conference on Neural Networks	Query
15	2008	A Comparative Study On Neural Network Based Soccer Result Prediction	Aslan, B. G. & Inceoglu, M. M.	Proceeding: Seventh International Conference on Intelligent Systems Design and Applications	Query
16	2003	A New Model To Forecast The Results Of Matches Based On Hybrid Neural Networks In The Soccer Rating System	Cheng, T., Cui, D., Fan, Z., Zhou, J. & Lu, S.	Proceeding: 5th International Conference on Computational Intelligence and Multimedia Applications	Query

Prediction

Football Match

Neural

Result

Network

Machine Learning

Based

English

Soccer

Forecasting

League

Comparative

Knowledge

Cup

Mining

System

Premier

Outcome

Deep

Study

Applying

Using

Algorithms

Three

Game

Method

Approach

Techniques

Decision

Modelling

Predictive

incorporating

Forecast

via

Model

Technique

Analysis

World

Six

Domain

Data

Clubs

Framework

New

Top

optimization

Rating

predicting

Following the review of the chosen articles with the aim of addressing the specific research questions, a quality assessment of these articles was performed, considering their contribution to

achieving the objective and, particularly, addressing the three specific questions posed. Therefore, for each research question, evaluation criteria were established and operationalized into questions. A value of 0 was assigned if the criteria were not met, 0.5 if partially met, and 1 if fully met (Table 2.3).

Table 2.3: Evaluation questions

In what contexts can the models be used?	Q1: Does the article describe and justify the objective of applying the model?
	Q2: Does the article clearly describe the scope of the study?
	Q3: Does the article clearly identify the targeted stakeholders of the study?
What methodologies are used?	Q4: Is the model applied to a relevant and contextualized sample?
	Q5: Does the article provide a detailed description of the data understanding phase?
	Q6: Does the article provide a detailed description of the data preparation phase?
	Q7: Does the article describe and justify the predictor variables used?
	Q8: Does the article describe and justify the algorithm used?
	Q9: Does the article compare different algorithms?
	Q10: Does the article clearly apply validation methods and various evaluation metrics?
How are the study results evaluated?	Q11: Does the article clearly present the results of the evaluation metrics?
	Q12: Does the article clearly describe the contributions of the study?
	Q13: Does the article clearly describe the limitations of the study?
	Q14: Does the article provide and justified the future recommendations?

2.2. Critical analysis

2.2.1. Contexts of model application

Over the past decades, there has been a growing trend in the attention given to statistical methodologies, and football is no exception. While data analysis in football remains an evolving field, data is beginning to play an increasingly important role in the sport (Yu, 2022).

All the models discussed in this study fall under the category of supervised learning. In simple terms, this means that we know our independent variables (x) and the desired output (y), creating a relationship of dependency between the variables. Applied to football, this occurs when we are aware of the variables x (e.g., goals, cards) and we know our output (e.g., match winner H, D, and A).

The analysis of Table 2.4, allows to conclude that 81.25% of the target of the studied articles aim to predict the outcome of a match in H, D, and A, with only three articles deviating from this objective (Yu, 2022; Carloni et al., 2021; Huang & Chang, 2010). In addition to this, there is a general agreement among the scientific community about the country chosen for their studies, with more than half of the articles (n=9) using England in their analysis (e.g., Yu, 2022; Nivetha et al., 2022).

Regarding the stakeholders with impact on the results, it is important to note that 37.5% (n=6) are related to the academic community (e.g., Guan & Wang, 2022; Yu, 2022; Azeman et al., 2021). It is relevant to mention that 18.8% of the articles (n=3) have betting companies as their primary stakeholders (Carloni et al., 2021; Baboota & Kaur, 2019; Korot'yeyeva et al., 2018). As a result, readers

perspectives on the articles may differ because they may be written with the intention of support betting houses with their predictions or assist investors in making money from betting houses.

Table 2.4: Contexts of the articles

ID	Goal	Target	Country	Stakeholders
1	Predicting the final outcome of a match based on three NN models	Predict the final result in H, D and A	China	Academic
2	Binary prediction of the final outcome	Predict whether the H team wins or does not win	England	Academic
3	Predicting the final outcome of a match based on an RNN and LSTM model	Predict the final result in H, D and A	England	Business
4	Predicting the final outcome of a match based on two ML models: DF and NN	Predict the final result in H, D and A	England	Academic
5	Developing a prediction system focused on generating profit in the betting markets using ML techniques	Predict the final outcome	N.A.	Betting house
6	Building a prediction model with RNN and LSTM	Predict the final result in H, D and A	England	Business
7	Implementing three ML algorithms: DF, ANN, and SVM to predict the outcome of a game	Predict the final result in H, D and A	England	N.A.
8	Predicting the final outcome of a match by evaluating four ML methods	Predict the final result in H, D and A	England	Betting house
9	Predicting the final results of the top 6 clubs in England using six different SVM techniques	Predict the final result in H, D and A	England	Academic
10	Development of two new models: recency feature extraction and rating feature learning	Predict the final result in H, D and A, and predict the number of goals scored by each team	Group 2	Academic
11	Predicting the final outcome of a match using the NN model	Predict the final result in H, D and A	N.A.	Betting house
12	Finding the model with the best predictive ability through data mining techniques and analyzing the defined metrics	Predict the final result in H, D and A	England	Business
13	Predicting the final outcome of a match using a BN model	Predict the final result in H, D and A	England	N.A.
14	Predicting the probability of a team victory	Predict the probability of the final outcome	World Cup	N.A.
15	Predicting the final outcome of a match using two vector-based ML models	Predict the final result in H, D and A	Italy	Academic
16	Building a result prediction model in a rating system that can predict the final outcomes of a football game based on HNN	Predict the final result in H, D and A	Italy	N.A.

Notes: H: Home; D: Draw; A: Away; NN: Neural Networks; RNN: Recurrent Neural Networks; LSTM: Long Short-Term Memory; ML: Machine Learning; DF: Decision Forest; ANN: Artificial Neural Network; SVM: Support Vector Machine; BN: Bayesian Network; HNN: Hybrid Neural Networks; Group 2: Austria, Belgium, Switzerland, Chile, China, Ecuador, England, France, Germany, Holland, Israel, Italy, Japan, Republic of Korea, Morocco, Mexico, Portugal, Russia, Scotland, Spain, Tunisia, United States of America, Venezuela, South Africa; N.A.: not available.

Given that ML models are more sophisticated than conventional statistical models, there is interest in understanding whether these techniques can accurately predict the outcomes of football matches. Considering this, Guan and Wang (2022) conducted the most current study in which they attempted to predict the outcome of a meeting using three different NN models: the Grey Fuzzy Model, the Grey Extreme Learning Machine Model, and the Combination Model. On the other hand, Yu (2022) study aims to evaluate the binary prediction ability of ML algorithms including Logistic Regression (LR), SVM, and Extreme Gradient Boosting (XGBoost) on a binary level, specifically whether the home team wins or not.

Nivetha et al. (2022) and Tiwari et al. (2020) argue that the Recurrent Neural Networks (RNN) with Long Short-Term Memory (LSTM) model appears to be the most effective method for predicting football match outcomes. The authors claim that the human factor has crucial characteristics that enable the prediction of sports results, as it can qualitatively analyze emotional factors that may impact a match outcome. Consequently, this human factor could be advantageous and facilitate result prediction or it can also present disadvantages and complicate the prediction process. Accordingly, it is by using ML techniques that is possible to decrease human error and increase the predictive capability. This is because, according to numerous studies, even the color of a team equipment can influence the referee decisions (Nivetha et al., 2022).

Yu (2022), Carloni et al. (2021), Azeman et al. (2020), Baboota and Kaur (2019), and Musa et al. (2019) employed the SVM algorithm in their studies. The first four studies focus on predicting results for a specific league, while Musa et al. (2019) shift the perspective of the literature by aiming to develop a model capable of predicting outcomes for the top-6 teams of the EPL, namely Manchester United, Manchester City, Liverpool, Arsenal, Chelsea, and Tottenham Hotspur. In this context, this study explores six different variations of the SVM model: linear, quadratic, cubic, fine radial basis function (RBF), medium RBF, and coarse RBF. It is also noteworthy that the Hybrid Neural Networks (HNN)¹ and Vector Quantization (LVQ)² algorithms are defined using ChatGPT. In Table 2.5 the definitions of the studied algorithms are presented.

¹ Prompt sent to ChatGPT to get the definition of HNN: Based on the following query applied at scopus: (football or "Premier League" or soccer) and (result* or outcome* or score or match or game*) and ("machine learning" or ML or "deep learning" or "data mining" or ai or "artificial intelligence" or "neural network*" or ann) and (estima* or prediction* or forecast* or predictive). Give me the definition of HNN and list any authors that you find."

² Prompt sent to ChatGPT to get the definition of LVQ: Based on the following query applied at scopus: (football or "Premier League" or soccer) and (result* or outcome* or score or match or game*) and ("machine learning" or ML or "deep learning" or "data mining" or ai or "artificial intelligence" or "neural network*" or ann) and (estima* or prediction* or forecast* or predictive). Give me the definition of LVQ and list any authors that you find."

Table 2.5: Description of the algorithm

Algorithm	Description	ID of the articles
DT	It is structured in the form of a tree to make decisions for predictions. This hierarchical structure consists of nodes representing inputs that generate an output (Quinlan, 1986).	13
DF	It works by creating multiple decision trees and then voting on the most popular outputs (Azeman et al., 2020; Baboota & Kaur, 2019).	4; 7
RF	It is a specific variant of DF that combines multiple DT to perform classification and regression tasks. It uses an ensemble learning strategy to achieve more accurate and robust results. It consists of a set of individual DT, where each tree is trained with a random sample of data and a random selection of inputs (Breiman, 2001).	5; 8
NN	It is a model inspired by the functioning of the human brain. It is composed of artificial neurons that are connected in layers and work together to process information (LeCun et al., 2015).	1; 4; 11; 13; 14
ANN	It is a variant of NN. During the training process, the weights of the connections between neurons are iteratively adjusted to minimize the error between the predicted outputs and the desired outputs (LeCun et al., 2015).	5; 7
HNN	It is an algorithm that combines different types of NN (e.g., Convolutional NN (CNN), Recurrent NN (RNN), Deep NN (DNN)) (ChatGPT, 2023).	16
LSTM	Unlike traditional RNNs, LSTM has internal memory cells that can store and access long-term information, allowing relevant information to be remembered for long periods of time (Nivetha et al., 2022).	3; 6
LR	Classification algorithm used to predict sports outcomes and with the ability to provide empirical insights through regression coefficients (Prasetio, 2016).	2; 5
SVM	Variant of support vector machine that parameterises the number of support vectors (West & Bhattacharya, 2016)	2; 5; 7; 8; 9
GB	Creates multiple sets of decision trees based on the principle of boosting. The variables are sequentially trained on the entire dataset, which improves the model accuracy after each iteration (Baboota & Kaur, 2019; Friedman, 2001).	8
XGBoost	It is a variant of the GB algorithm. It uses an ensemble learning approach, combining multiple weak decision trees to create a stronger model (Chen & Guestrin, 2016).	2; 10
KNN	The main difficulty when using this algorithm is to discover the optimal value of k that will optimize the model. If the value of k is too low, the result can be very sensitive to the number of nearest neighbors. On the other hand, if the value of k is too high, the model may include too many irrelevant cases and potentially perform poorly (Berrar et al., 2019; Che Mohd Rosli et al., 2018).	5; 10; 13
BN	It is a graphical algorithm where nodes represent discrete or continuous variables, and the arcs represent direct connections between them (Pearl & Russel, 2003).	12; 13
NB	It is a purely probabilistic algorithm that uses the principles of conditional probability to calculate the posterior probabilities of an event belonging to a specific class. The reason it is called "naive" is due to its strong assumption of independence, meaning all features are considered independent of each other (Baboota & Kaur, 2019).	5; 8
LVQ	It belongs to the family of vector quantization models. It is used for classification and pattern recognition (ChatGPT, 2023).	15

Notes: i) DT: Decision Tree; DF: Decision Forest; RF: Random Forest; NN: Neural Networks; ANN: Artificial Neural Network; HNN: Hybrid Neural Networks; LSTM: Long Short-Term Memory; LR: Logistic Regression; SVM: Support Vector Machine; GB: Gradient Boosting; XGBoost: Extreme Gradient Boosting; KNN: K nearest neighbours; BN: Bayesian Network; NB: Naive Bayes; LVQ: Vector Quantization; RT: Random Trees; CHAID: Chi-squared Automatic Interaction Detection; RNN: Recurrent Neural Networks; ii) the references in the descriptions of the algorithms are from unidentified articles from the SLR

In addition to the SVM algorithm, the NN algorithm has also been widely used in the literature, with 31.25% (n=5) of the studies employing the traditional NN algorithm (Guan & Wang, 2022; Azeman et al., 2021; Korotyeyeva et al., 2018; Rosli et al., 2018; Huang & Chang, 2010). Given that the Artificial Neural Network (ANN) (used by Carloni et al., 2021 e Azeman et al., 2020) and the HNN (used by Cheng et al., 2003) are two variations of the NN algorithm, respectively, then half of the studies (n=8) use NN algorithms or one of its variants.

2.2.2. Methodology

Knowing the algorithms used in the various studies, it is important to consider other methodological points of view. There are studies that support the handling and analysis of data using the Knowledge Discovery in Database (KDD) methodology. This implies that initially, DB are identified for the study, followed by a process of data cleaning and data selection and transformation, where the most relevant information is filtered and transformed into the required format for the algorithm. Finally, a data mining process converts this into knowledge (Azeman et al., 2021; Azeman et al., 2020).

Taking into consideration the analyzed literature, it can be observed from Table 2.6, the authors such as Yu (2022), Nivetha et al. (2022), Azeman et al. (2021), Tiwari et al. (2020), Azeman et al. (2020), Baboota and Kaur (2019), Musa et al. (2019), Rosli et al. (2018), and Razali et al. (2017) focus their studies on the EPL and utilize data from the United Kingdom (UK) football betting website (www.football-data.co.uk), where football data with match statistics could be obtained. Accordingly, over half of the articles under study (56,25%, n=9) are covered by these two criteria.

Table 2.6: Methodology of the articles

ID	Season	Sample	Sample Size	Data Base	Software	Data Processing	Are comparisons made between algorithms?
1	N.A.	CFSL	N.A.	N.A.	N.A.	N.A.	Yes
2	2016-2019	EPL	1050	FD e SF	Sklearn	The data from the first three matchdays of each season has been removed.	No
3	2009-2015	EPL	2280	FD	N.A.	N.A.	No
4	2005-2006	EPL	380	FD	Azure Machine Learning Studio.	The database is clean and there were no missing values.	Yes
5	1 ^a division: 2008-2020; 2 ^a division: 2010-2019	N.A.	36461	N.A.	N.A.	The data cleaning operation reduced the sample, but it does not explain the process of the beginning of the season.	Yes
6	2010-2018	EPL	3040	FD	Tensorflow library in python.	N.A.	No
7	2010-2013	EPL	1140	FD	N.A.	N.A.	Yes
8	2005-2016	EPL	3630	FD e FI	Data pre-processing: Python; Modelos de ML: Scikit-Learn.	It starts from matchday 6 until matchday 38.	Yes
9	2013-2016	Group 1	456	FD	MATLAB 2016a.	All data that does not include the selected teams has been excluded.	Yes
10	2013-2018	Group 2	31318	OCL	N.A.	A continuous league solves the issue of the beginning of the season. A superleague solves the issue of promotion and relegation.	Yes
11	N.A.	N.A.	N.A.	N.A.	Python, JavaScript, Node.JS, Chrome, Redis Network, Open Source Tools.	N.A.	No
12	2013-2016	EPL	1140	FD	Weka 3.8 and GMDH Shell DS.	N.A.	Yes
13	2010-2013	EPL	1140	FD	WEKA.	N.A.	No
14	2006	World Cup	64	FIFA	N.A.	It only starts after the group stage.	No
15	2001-2002	Italian Serie A	252	SW	N.A.	It starts from matchday 6.	Yes
16	2001-2002	Italian Serie A	252	N.A.	N.A.	It starts from matchday 6.	Yes

Note: CFSL: Chinese Football Super League; EPL: English Premier League; ML: Machine Learning; Group 1: games of Manchester United, Manchester City, Liverpool, Chelsea, Arsenal e Tottenham; Group 2: Austria, Belgium, Switzerland, Chile, China, Ecuador, England, France, Germany, Holland, Israel, Italy, Japan, Republic of Korea, Morocco, Mexico, Portugal, Russia, Scotland, Spain, Tunisia, United States of America, Venezuela, South Africa; FD: www.football-data.com.uk; SF:

While there is an agreement among the authors regarding the analyzed league, the same cannot be said for the seasons under study. One of the main factors that might explain the high variability in seasons and years studied is the period during which the research was conducted. For instance, if a study is conducted in 2022, it is more relevant to use data from more recent seasons. So, only Yu (2022), Azeman et al. (2020), Musa et al. (2019), Razali et al. (2017), and Rosli et al. (2018), approximately 33% of the articles, demonstrate consensus that three seasons is the maximum time during which past data remains relevant for predicting future events, aligning with the findings of Rebelo (2012). Consequently, data older than three years becomes less relevant over time as it fails to reflect the team real value in each season.

It is noteworthy that only five articles remove data from the initial matchdays (Yu, 2022; Baboota & Kaur, 2019; Huang & Chang, 2010; Aslan & Inceoglu, 2008; Cheng et al., 2003), as there are various random factors at the start of a season that play a significant role and can distort the training process (Aslan & Inceoglu, 2008).

One of the best solutions for data preparation at the beginning of the seasons is provided by Berrar et al. (2019). This presupposes that to carry out this model, the continuous visualization of the seasons must be carried out. As a result, the difficulty of how to visualize the first matches is solved, and these matches will now just be evaluated as the next matches from the previous season. With this approach, it becomes possible to maintain the performance trajectory from the previous season and avoid resetting each team at the start of each season.

From this principle, another issue emerges, which is the existence of multiple divisions. In every league, there are typically first and second divisions, and in each season, at least two teams are relegated from the first to the second division while two teams are promoted from the second to the first division. Many of them lack first-league registrations, or the data from the last time they competed in the league is irrelevant because the players are no longer the same or the time spent outside the league was too long. Therefore, considering a league as a continuous season does not fully resolve the issue of teams moving between divisions. To address this, Berrar et al. (2019) suggest including all clubs from both the first and second divisions in a super league. This approach helps mitigate the problem of teams shifting between divisions due to promotions and relegations (Berrar et al., 2019).

There is no consensus on the best or recommended software to be used. Weka is being used by Rosli et al. (2018) and Razali et al. (2017) for their models, but they are using different versions of the program. The programming language Python is used by Tiwari et al. (2020), Baboota e Kaur (2019), and Korotyeyeva et al. (2018), but each time with a different library.

As defended by Shin and Gasparyan (2014), the addition of a DB with player ratings from the video game "FIFA" may improve the quality of the forecast. Therefore, Yu (2022), Baboota and Kaur (2019), and Berrar et al. (2019) decided to incorporate a DB with team ratings into their models. Along with that, it is still observed that 62.5% of the articles compare various algorithms. This is a significant aspect, as only through such comparisons it can be discern which model exhibits superior performance.

In relation to the variables used in the literature, it can be shown from Table 2.7 that a significant portion of studies (81.25%, n=13) attempt to predict the outcome of a football match based on three results (H, D, and A). On the other hand, 56.25% (n=9) does not specify how the independent variables from their studies are calculated. With this as a foundation, it becomes difficult for future studies to use the same methodology as these.

Table 2.7: Dependent variable

ID	Dependent Variable	Measure used in the independent variables
1	Predict the outcome of the match (H, D, A)	N.A.
2	Predict if the home team wins or does not win	Average
3	Predict the outcome of the match (H, D, A)	Sum
4	Predict the outcome of the match (H, D, A)	N.A.
5	Predict the number of goals in the match	N.A.
6	Predict the outcome of the match (H, D, A)	Sum
7	Predict the outcome of the match (H, D, A)	N.A.
8	Predict the outcome of the match (H, D, A)	Average
9	Predict the outcome of the match (H, D, A)	N.A.
10	Recency Feature Extraction: Predict the outcome of the match (H, D, A) Rating Feature Learning: Predict the number of goals scored by each team.	Recency Feature Extraction: average Rating Feature Learning: sum
11	Predict the outcome of the match (H, D, A)	N.A.
12	Predict the outcome of the match (H, D, A)	N.A.
13	Predict the outcome of the match (H, D, A)	N.A.
14	Predict the probability of a team winning a game	Average
15	Predict the outcome of the match (H, D, A)	N.A.
16	Predict the outcome of the match (H, D, A)	Average

Note: H: Home Team; D: Draw; A: Away Team; N.A.: Not available

One of the biggest challenges in developing a predictive model is choosing the variables. As a result, it is necessary to analyze which seem to be the most pertinent variable (Baboota & Kaur, 2019). It is vital to remember that there is not a general agreement on which independent variables are most significant to predict the outcome of a football match. Therefore, the Table 2.8 presents the most common variables in the literature along with the corresponding study that utilizes them.

Table 2.8: Common variables

Variables	ID
HandATeam	1; 3; 4; 7; 13; 16
HandATeamGoalScoredAtFT	1; 3; 4; 6; 7; 9; 12; 13; 14; 16
HandATeamGoalScoredAtHT	4; 7; 13
HandATeamShots	1; 4; 7; 9; 12; 13; 14
HandATeamShotsOnTarget	1; 4; 7; 9; 12; 13; 14
HandATeamCorners	1; 4; 7; 9; 11; 12; 13; 14
HandATeamOffside	1; 14
HandATeamBallPossession	1; 14
HandATeamGoalDifference	2; 3; 8
HandATeamPoints	2; 3
HandATeamFoulsCommitted	4; 7; 9; 11; 13; 14
HandATeamYellowCards	4; 7; 9; 13; 14
HandATeamRedCards	4; 7; 9; 13; 14
HandATeamOverallRating	2; 8; 15
HandATeamAttackRating	8; 10
HandATeamDefenseRating	8; 10
HandATeamConsecutiveWin(3)	3; 6
HandATeamConsecutiveWin(5)	3; 6
StatisticsOfMeetingsBetweenTeams	1; 11
OutcomeMatch	9; 16
PenaltyStatistics	11; 14

Note: H: Home Team A: Away Team; FT: Full time; HT: Halt time

It can be concluded that the variable related to goals scored by each team is the most frequently used (n=10), followed by corners (n=9), shots (n=8), and shots on target (n=8). The study by Carloni et al. (2021) stands out for adopting a different approach to variable selection. In their work, they use odds from betting houses as inputs, including maximum, minimum, average, standard deviation (SD), and variance, as illustrated in Table 2.9.

Table 2.9: Other independent variables

ID	Variables
1	HandATeamPasses, HandATeamSteals, HandATeamCompetitiveStatus, WinValue;
2	HandATeamLastResults(3);
3	HandATeamGoalConcededAtFT, HandATeamConsecutiveLoss(3), HandATeamConsecutiveLoss(5);
4	Referee;
5	MaxOfAwayWin, MaxOfDraw, MaxOfGoal, MaxOfOver, MaxOfUnder, MeanOfAwayWin, MeanOfDraw, MeanOfGoal, MeanOfOver, MeanOfUnder, MinOfAwayWin, MinOfGoal, MinOfNoGoal, MinOfOver, MinOfUnder, OverUnderOutcome, StdAwayWin, StdOfOver, StdOfDraw, StdOfGoal, StdOfOver, StdOfUnder, VarHomeWin, VarOfAwayWin, VarOfDraw, VarOfGoal, VarOfOver;
8	HandATeamForm, HandATeamStreak, HandATeamPastKShotsOnTarget, HandATeamPastKGoals, HandATeamPastKCorners, HandATeamMidfieldRating, HandATeamWeightedStreak, FormDifferential, StreakDifferential, PastKShotsOnTargetDifferential, PastKGoalsDifferential, PastKCornersDifferential, AttackRatingDifferential, MidfieldRatingDifferential, DefenseRatingDifferential, OverallRatingDifferential, GoalDifferenceDifferential, WeightedStreakDifferential;
10	AttackingPerformance, DefensivePerformance, RecentPerformance, StrengthOpposition, HomeTeamAdvantage;
11	HandAFormLastMatches(5), HandAGameStatistics, Meteorology, MomentOfDay, HandATeamGeneralStandings;
14	DirectFreeKicksToGoal, IndirectFreeKicksToGoal, OwnGoals, Cautions, Expulsions;
16	WinRatio, DrawRatio, LoseRatio, AverageLose, Morale.

Note: H: Home Team; A: Away Team; FT: Full time

As shown in Table 2.10, 87.5% of studies use accuracy as their primary method to evaluate the results.

Table 2.10: Algorithm, type of validation and evaluation metrics

ID	Type of validation	Evaluation Metrics
1	N.A.	Accuracy, MAE, MSE, RMSE and MAPE
2	N.A.	Accuracy and F1 score
3	N.A.	Accuracy, F1 score, precision, recall and support
4	N.A.	Accuracy, precision and recall
5	Hold out: 80/20	Accuracy
6	N.A.	Accuracy
7	10-folds cross-validation	Accuracy, precision and recall
8	6-folds cross-validation	RPS, F1 score, precision and recall
9	5-folds cross-validation	Accuracy and cohen's kappa
10	K-folds cross-validation	RPS
11	Inverse error propagation	Accuracy
12	10-folds cross-validation	Accuracy
13	10-folds cross-validation	Accuracy, F1 score, precision and recall
14	N.A.	Accuracy
15	N.A.	Accuracy
16	N.A.	Accuracy

Note: MAE: Mean Absolute Error; MSE: Mean Square Error; RMSE: Root-Mean-Square Error; MAPE: Mean Absolute Percentage Error; RPS: Rank Probability Score; N.A.: Not available

From the remaining 12.5%, Baboota and Kaur (2019) and Berrar et al. (2019) use the Rank Probability Score (RPS) as their primary evaluation metric. A lower RPS indicates better predictive capability of the model (Baboota & Kaur, 2019). It can be observed that there is substantial agreement within the literature that the accuracy is the preferred evaluation metric, yet it should be complemented by others (e.g., precision and recall). In this context, Nivetha et al. (2022), Azeman et al. (2021), Azeman et al. (2020), and Razali et al. (2017) also utilize, in addition to accuracy, metrics such as F1 score, precision, recall, and support. However, Azeman does not use the F1 score in their studies (as defined in Table 2.11). In addition to this, it is essential to consider the confusion matrix in the evaluation process (e.g., Nivetha et al., 2022; Baboota & Kaur, 2019).

Table 2.11: Metrics definition

Metrics	Formula	Explanation
Accuracy	$\frac{\text{True Positive}}{\text{Total Predictions}}$	Number of correct predictions divided to the total number of predictions made by the model.
Precision	$\frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$	Evaluates the proportion of correct predictions in relation to the total number of positive predictions (in column).
Recall	$\frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$	Evaluates the proportion of correct predictions in relation to the total number of actual positive cases (in row).
F1 Score	$\frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$	It combines precision and recall into a single measure. Essentially, it is the harmonic mean between the two metrics, providing a balanced measure of the model performance.
Support	$\frac{\text{Total Predictions of the item}}{\text{Total Predictions}}$	Ratio of samples containing a specific item. That is, the percentage of times the model predicts H, D, or A.
RPS	$\frac{1}{r-1} \sum_{i=1}^{r-1} \left(\sum_{j=1}^i (p_j - e_j) \right)^2$	It represents the difference between the acumulative distributions of the forecasts and the actual outcomes over the total number of predictions (Jolliffe & Stephenson, 2003; Murphy, 1970).

Note: H: Home; D: Draw; A: Away

Source: West and Bhattacharya (2016, p. 4)

Regarding validation methods, around 38% use the k-folds cross-validation method, with six articles not specifying the method they use (also approximately 38%). Excluding the articles that do not specify the validation method, about 66% (6 out of 9 in the sample) of the studies use the k-folds cross-validation method. This consensus indicates that k-folds cross-validation is considered the preferred method for model validation, as it helps mitigate the issue of overfitting, which is the tendency of a model to overly adapt to the training data (Musa et al., 2019). Additionally, Seni and Elder (2010) assert that k-folds cross-validation is a powerful technique for evaluating model prediction performance. However, it is noteworthy that only Carloni et al. (2021) utilize the hold-out 80-20 method, where 80% of the data is used for training and the remaining 20% for testing purpose.

Considering this, Azeman et al. (2020), Baboota and Kaur (2019), Musa et al. (2019), Berrar et al. (2018), Rosli et al. (2018), and Razali et al. (2017) employ cross-validation for model validation, with variations in the choice of k (number of folds). As a result, the first issue that arises while using this method is locating a value that optimizes the models results. It is necessary to consider the fact that the higher the value of k, the longer the computation takes. After doing some tests, Baboota and Kaur (2019) concluded that due to a uniformity issue for all the algorithms used in the study, the k that optimizes the same is k=6. Musa et al. (2019) found that k=3 was optimal for their research, while other authors used k=10. This variability in the choice of k underscores the need for consideration and experimentation to find the most suitable value for the modeling task.

2.2.3. Results obtained from the models

Regarding the obtained results, it is important to note that they depend on the algorithm and variables used. Therefore, referring to Table 2.12, it is evident that Musa et al. (2019) model, which focused

exclusively on the top-6 clubs in England, achieved an accuracy of 100% and a perfect agreement level within their model. Concentrating on predicting outcomes for specific clubs can indeed enhance the model performance, but it does restrict predictions to the defined teams.

Table 2.12: Results of the best models

ID	Algorithm	Accuracy	Precision	Recall	F1 Score	RPS
1	Gray fuzzy extreme learning machine prediction combination model	73.00%	N.A.	N.A.	N.A.	N.A.
2	LG	67.30%	N.A.	N.A.	66.01%	N.A.
3	LSTM Model	86.00%	86.00%	86.00%	86.00%	N.A.
4	Multiclass Decision Forest	88.95%	87.57%	85.61%	N.A.	N.A.
5	ANN	59.00%	N.A.	N.A.	N.A.	N.A.
6	LSTM Model	80.75%	N.A.	N.A.	N.A.	N.A.
7	DF	88.94%	88.94%	88.94%	N.A.	N.A.
8	GB	58.80%	53.00%	51.67%	50.67%	21.56%
9	SVM - Linear	100.00%	100.00%	100.00%	100.00%	N.A.
10	XGBoost	N.A.	N.A.	N.A.	N.A.	20.23%
11	NN - 16 input signals	75.00%	N.A.	N.A.	N.A.	N.A.
12	Decision Trees	99.56%	N.A.	N.A.	N.A.	N.A.
13	BN	75.09%	75.23%	75.10%	75.10%	N.A.
14	NN - 8 input signals	76.90%	N.A.	N.A.	N.A.	N.A.
15	LVQ _B	53.25%	N.A.	N.A.	N.A.	N.A.
16	LVQ	52.29%	N.A.	N.A.	N.A.	N.A.
Average		75.66%	81.79%	81.22%	75.56%	20.90%

Note: ANN: Artificial Neural Network; BN: Bayesian Network; DF: Decision Forest; GB: Gradient Boost; HNN: Hybrid Neural Networks; LSTM: Long Short-Term Memory; NN: Neural Networks; RPS: Rank Probability Score; SVM: Support Vector Machine; N.A.: Not available

The model presented by Rosli et al. (2018) achieved results that were not far behind those presented by Musa et al. (2019), as it shows an accuracy of 99.56% for the decision tree (DT) algorithm across the three seasons. This was followed by NN with an accuracy of approximately 96.83%, BN with 76.41%, and finally, k nearest neighbours (KNN) with around 77.54%. Through the utilization of the employed variables and algorithm, the authors were able to predict EPL results with a remarkable accuracy of 99.56% over three seasons. This stands as the best result in the literature based on the accuracy metric, excluding the Musa et al. (2019) model, which predicts outcomes solely for specific teams.

The LG model presented by Yu (2022) achieved an accuracy of approximately 67%. This value is not optimal and leaves room for improvement. On the other hand, the LSTM model discussed by Nivetha et al. (2022) and Tiwari et al. (2020), despite differences in the variables used, they display intriguing results, showing an accuracy of 86% and 80.75%, respectively. It is worth noting that both Nivetha et al. (2022) and Azeman et al. (2020) present their overall results. In other words, the

precision, recall, and F1 Score values shown in Table 2.12 represent the averages across the three classes (H, D, and A).

To predict the number of goals in a match, Carloni et al. (2021) used an ANN model. The accuracy level of the resulting prediction was 59%. The authors evaluated the system performance on upcoming matches, by making the testing phase on matches already played. Through this analysis, the authors managed to achieve a return on investment (ROI) of 26.54%. This outcome highlights the potential for profitability in the sports betting market through the development of a model based on ML techniques. This demonstrates that a well-constructed ML model can not only provide accurate predictions but also yield positive financial returns, demonstrating its practical utility beyond the academic realm.

Azeman et al., in both their 2021 and 2020 studies, achieved similar results. In both cases, they concluded that the decision forest (DF) model exhibited the highest accuracy of 88.94% and 89%, respectively, when compared to the NN, ANN, and SVM algorithms. Nonetheless, the second-best performing algorithm was the ANN with 72%, followed by NN with 71.58%, and SVM with 70%. As a result, the authors claim that the DF model performs better than the NN because it is simpler, requires less computational effort, and works best with smaller and simpler DB (Azeman et al., 2020; Azeman et al., 2021).

Only Berrar et al. (2018) utilized the RPS as the primary evaluation metric for their model. In this case, the authors obtained RPS values of 20.54% for RPS_{KNN} and 20.23% for $RPS_{XGBoost}$ in the rating feature learning model. These results demonstrate that not only does the model containing rating variables yield better outcomes, but also the XGBoost algorithm outperforms KNN. Baboota and Kaur (2019) also present the gradient boosting (GB) model with the best RPS result at 21.56%.

Abording the NN results, Korotyeyeva et al. (2018) report an accuracy for the NN model with 8 and 10 hidden layers of 65% and 68%, respectively. When testing the same model with two different algorithms, the accuracy dropped to 33% for Naive Bayes (NB) and 47% for the clustering method. Based on the findings, the authors further attempted to improve the model by removing the less significant variables, leaving just the most pertinent ones. This resulted in accuracy of 75% when the system uses 16 input signals and 70% when it uses 8 input signals.

In comparison to Huang and Chang (2010) study, this one has an accuracy rate of 76.90%, however, because the World Cup sample size is just 64 matches, the results in this case may not be considered significant. Finally, Aslan and Inceoglu (2008) report the LVQ_B model as having obtained a better accuracy result (53.25%) compared to the LVQ_A (51.29%). In comparison to the previously mentioned results, Aslan and Inceoglu (2008) model performs significantly worse than the outcomes of the models of the most current literary authors. On the other hand, when compared to Cheng et al. (2003) results, these are marginally superior, perhaps because Cheng et al. (2003) present the worst results in the literature (accuracy = 52.29%).

From the point of view of investors, the models put out by Musa et al. (2019) and Rosli et al. (2018) show very promising results, with accuracy rates of 100% and 99.56%, respectively. But it is vital to note that the Musa et al. (2019) model only provides predictions for the six main EPL teams. Additionally, both the model developed by Musa et al. (2019) and the model developed by Rosli et al. (2018) were based on variables that were only made available after the match was completed. Consequently, these models do not provide meaningful information to predict the outcome of a match.

2.2.4. Main contributions, limitations, and future recommendations

All studies should take into account their contributions, limitations, and future research recommendations. This state-of-the-art presented many approaches and techniques used to predict the outcomes of football matches, with a summary of the contributions and limitations of the examined studies in Table 2.13.

Table 2.13: Contributions and limitations of the articles

ID	Contributions	Limitations
1	The combination of two models improves predictive capability.	N.A.
2	It shows that three seasons is the optimal number for predicting future events with past data.	N.A.
3	It demonstrates that an algorithm with RNN and LSTM has a clear advantage over all other ML models.	N.A.
4	It shows that decision trees (DF) perform better than neural networks (NN) on simpler databases.	N.A.
5	It shows that it is possible to be profitable in betting markets through the creation of a machine learning model.	N.A.
6	The more information the system is fed, the more accurate the outcome of the RNN system will be.	They do not address the most impactful variables of their model, which limits future studies on these topics.
7	Decision trees work well with discrete variables, while artificial neural networks (ANN) and support vector machines (SVM) work better with continuous variables.	There is no data available on shot types and passes.
8	Using ratings enhances the predictive capability of the models.	Unable to predict results in the first 5 matchdays.
9	The variables used are considered highly valuable for predictions and can assist coaches and managers in assessing the probabilities of winning.	N.A.
10	The key to success appears to be domain knowledge in the data processing process and the inclusion of rating variables.	N.A.
11	The fundamental principle of NN training is essentially to adjust the weights in the synapses to minimize the error, which is the difference between the output obtained and the actual value.	N.A.
12	Club management can use this prediction method to forecast the future of their team.	The techniques used in this study are static over time because they rely on data from previous games, so for analysis on upcoming games, real-time data acquisition is necessary (in a dynamic manner).
13	The results can be used as a starting point for future research.	N.A.
14	It shows that it is possible to make football predictions in an international football team competition (World Cup).	Unable to predict the probability of a draw.
15	The chosen variables are a critical success factor; the study supports the idea that neural networks can be used to predict football outcomes with better capability than statistical approaches.	N.A.
16	The chosen variables are a critical success factor; with modifications to the model, it is possible to use it in different sports such as tennis and badminton.	N.A.

Note: DT: Decision Tree; ML: Machine Learning; DF: Decision Forest; NN: Neural Networks; ANN: Artificial Neural Network; SVM: Support Vector Machine; RNN: Recurrent Neural Networks; LSTM: Long Short-Term Memory; N.A.: Not available

In terms of contributions, Baboota and Kaur (2019) highlight the enhanced predictive capacity of models when incorporating player ratings. On the other hand, Musa et al. (2019) work highlights the significance of variable selection, specifically utilizing metrics such as team shots, shots on target, and corners. The integration of player ratings, as suggested by Baboota and Kaur (2019), reflects the growing recognition of the importance of individual player performance in match predictions. Meanwhile, Musa et al. (2019) emphasis on accessible variables brings practicality to the forefront, making their model more actionable for real-world applications in football coaching and decision-making. This most recent study gives coaches a valuable perspective because the variables used are simple to access and can be helpful in predicting the likelihood that a team will win, tie, or lose a particular match.

In addition to the authors already mentioned, Razali et al. (2017) argue that the methodology used in the study, which was based on the BN algorithm and an analysis of data from the EPL over three seasons, adds a lot of value for the scientific community and leaves the expansion of the methodology to other tournaments as a future recommendation. For Berrar et al. (2019) the key to success is the domain knowledge of the topic. The authors established that successful model performance is intertwined with a deep understanding of the domain and effective data handling highlights the importance of the data pre-processing. Based on their contributions, the authors concluded that rating variables are a key component of an ML model for football predictions and that knowledge of these variables is also important.

In terms of the models limitations, only three authors explicitly specify them in their studies. On one hand, it may be beneficial to show that the models appear to be robust, but on the other hand, it may be difficult for the authors to admit any flaws in front of the scientific community. It is noteworthy that Azeman et al. (2020) claim in their study that taking into account information about the type of passes made during a match may have an impact on the model results because, while one team may have more passes than the other, if those passes are in its own half of the field or are not dangerous, that team is not likely to put the other team in danger.

Regarding recommendations for future research (Table 2.14), approximately 37.50% (n=6) of the studies suggest adding new variables to their analyses to enhance their results.

Table 2.14: Future recommendations

ID	Future Recommendations
1	N.A.
2	N.A.
3	Continue the model by adding more seasons or including individual player statistics.
4	The study can be expanded to other leagues.
5	Replace ANNs with RNNs, as the latter is more efficient in measuring different time intervals or using variations in odds over time before the start of the game.
6	Increase the number of variables, include player statistics, or try to predict the number of goals or cards.
7	Improve the DF model to increase accuracy.
8	Include data on injuries, presence of key players, psychological factors, or the use of social media. It is possible to have the model predict fair betting odds.
9	Extend the methodology used to other leagues.
10	Add more variables such as yellow and red cards, fouls, ball possession, passing or running rates, salary, age, fitness condition, average height and attack rates.
11	Add more hidden layers to improve accuracy.
12	Use new variables such as travel distance for each team, current team performance, or the number of yellow and red cards.
13	N.A.
14	N.A.
15	It is possible to experiment with new leagues, new variables, and different types of models. Future studies should focus on developing a well-balanced prediction performance model rather than solely evaluating overall performance.
16	Carefully select the appropriate input structure and test under different conditions.

Note: DF: Decision Forest; ANN: Artificial Neural Network; RNN: Recurrent Neural Networks; N.A.: Not available

Nivetha et al. (2022), Tiwari et al. (2020), and Baboota and Kaur (2019) recommend the inclusion of player statistics as variables in their studies. Incorporating such data could prove advantageous for the model, as understanding which players are in better form or dealing with injuries can lead to more accurate predictions. If the team key player is in good form and on the field, the likelihood of winning is higher, while, if a top player is injured, the probability for winning may decrease. However, Yu (2022), Berrar et al. (2019), as well as Aslan and Inceoglu (2008) claim that including many different types of data is potentially useful with the main challenge being the data availability. For some of the major leagues, the basic match statistics (e.g., shots, corners, cards, or fouls) are immediately available. However, these data may not be available to predict the outcome of matches in lower leagues. In addition to the variables mentioned, Rosli et al. (2018) suggests incorporating the away team travel distance and recent team performance as potential new variables. It is also crucial to obtain data indicating whether a player is between the ball and the goal, because that position drastically reduces the likelihood of a shot resulting in a goal (Azeman et al., 2020).

According to the earlier explanation, each line of data stands for a different match. Therefore, the values of variables such as goals scored, shots, or corners refer specifically to that match. If authors use this data to predict the outcome of the same match, it becomes impossible to forecast future results, as they would be using real-time variables and values. This is a critical point to consider, as accurate values for these statistics are only available at the conclusion of the match. Consequently, if these variables were used, they would merely be estimations of what could be expected from each team, rather than actual values. Such an approach could potentially compromise the accuracy of predictions. These types of statistics are valuable for explaining (rather than predicting) outcomes of completed matches. Additionally, if someone attempts to predict the outcome of a match and already knows which goals each team scored, it is quite probable that the model will have excellent predictive ability (for previous events). Therefore, if the goal is to predict future outcomes, none of the models that are based on real values add any value. As a result, it is essential that the variables used to construct predictive models always be historical variables, that is, variables that we have access prior to the start of the match. The conclusions regarding the contributions and limitations of the examined articles are presented in Table 2.15.

In terms of contributions, it is important to highlight one made by Berrar et al. (2019), which involves seeing each season as a single competition with the aim of fixing the issue with the opening matches. In addition, it suggests the creation of a super league to address issues related to the rise and fall of divisions. From my perspective, this seems to be a very novel solution that, on a practical level, fixes the issues raised in certain studies. On the other hand, Yu (2022) notes that in the EPL, the home team winning percentage is 47.43%, demonstrating that playing at home has advantages. Additionally, Baboota and Kaur (2019), found that the home factor is an important variable in predictive models.

Regarding limitations, it is important to note the article by Musa et al. (2019), which, despite having a model with 100% accuracy, can only predict outcomes for specific teams. This restriction directly relates to the study goal. If the goal is to obtain a model with accuracy close to 100%, it may be simpler to use a collection of pre-selected teams (e.g., the major European teams), as their results tend to be more accurate. However, using a model with this limitation is not ideal if the goal is to predict the outcome of football matches for all teams.

Table 2.15: Critical analysis of contributions and limitations

ID	Contributions	Limitations
1	N.A.	It does not specify the season and it does not mention the specific database used for the analysis.
2	It shows that there is a strong correlation between the average and the goal scored difference; It demonstrates that the EPL has an advantage when playing at home.	Inability to predict the outcomes of the first three matchdays and to predict whether a game will end in a draw.
3	It shows that an algorithm with RNN and LSTM has a clear advantage over all other ML models; They do not utilize in-game variables.	N.A.
4	N.A.	N.A.
5	N.A.	Lack of explanation regarding the countries included in the study.
6	N.A.	Despite the results, the authors do not address the most impactful variables in their model, which limits future studies on these topics.
7	N.A.	N.A.
8	Good ability to define the most important variables, and the variable "form" of the teams is also highly significant compared to others.	N.A.
9	N.A.	It only predicts results for specific teams.
10	The initial matchdays should be viewed as a continuation of the season, meaning that there is only one continuous season rather than a succession between seasons; The creation of a super league (combining the first and second divisions) should be considered to minimize the issue of relegation and promotion between divisions.	N.A.
11	N.A.	The country of the sample, its size, the databases used, and the metrics to evaluate the model are not disclosed in the study.
12	N.A.	This analysis is only applicable to games that have already taken place.
13	N.A.	N.A.
14	N.A.	It only predicts the results after the group stage.
15	One of the early articles on NN in football.	The sample size is very small and not very relevant as it only predicts results from the second half of the championship.
16	N.A.	It does not predict the first 6 matchdays; It only presents the results of one season, so we cannot draw many conclusions from it.

Note: EPL: English Premier League; LSTM: Long Short-Term Memory; RNN: Recurrent Neural Networks; ML: Machine Learning; NN: Neural Networks; N.A.: Not available

In conclusion, after reading these articles, it is difficult to understand how to replicate the majority of the methodologies used. It is possible to understand that the models that are currently used in the literature are only capable of predicting results of matches that have already occurred (e.g., Guan & Wang, 2022; Azeman et al., 2021; Azeman et al., 2020), not future outcomes. The authors generally do not provide explanations regarding the calculation of independent variables. This limitation primarily arises from using variables that are not accessible before a match takes place.

2.3. Articles evaluation

Given that this study is based on a SLR, it is important to facilitate the work for new researchers and professionals. Therefore, the relevance of the articles was assessed for each of the three specific research questions (Table 2.3). To achieve this, criteria were defined and operationalized as questions, with a score assigned to each article: 0 if it does not address the question; 0.5 if it partially addresses the question; and 1 if it fully addresses the question. In the Table 2.16 it is possible to see the 16 articles analysed in this review, and their evaluation.

Table 2.16: Articles evaluation

In what contexts can the models be used?					What methodologies are used?								How are the obtained results evaluated?					Total ₁
ID	Q1	Q2	Q3	Sub-Total ₁	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Sub-Total ₂	Q11	Q12	Q13	Q14	Sub-Total ₃	-
1	1	1	0.5	2.5	0	0.5	0.5	0.5	1	1	0.5	4.0	0.5	1	0	0	1.5	8.0
2	1	0.5	1	2.5	1	1	1	1	0.5	0	1	5.5	1	0	0	0	1	9.0
3	1	1	1	3	0.5	1	0.5	1	1	0	0.5	4.5	1	0	0	1	2	9.5
4	1	1	1	3	0	1	1	0.5	1	1	1	5.5	1	1	0	1	3	11.5
5	1	0.5	1	2.5	0	0.5	0.5	0.5	1	1	1	4.5	1	1	0	1	3	10.0
6	1	1	1	3	0	1	0	0.5	1	0	0.5	3.0	1	0	0	1	2	8.0
7	1	1	0	2	0	1	0	0.5	1	1	1	4.5	1	1	1	1	4	10.5
8	1	1	1	3	1	1	1	1	1	1	1	7.0	1	1	1	1	4	14.0
9	1	1	1	3	0	1	0.5	1	1	0.5	1	5.0	1	1	0	1	3	11.0
10	1	1	0.5	2.5	0.5	1	1	1	1	1	1	6.5	1	1	0	1	3	12.0
11	1	0.5	1	2.5	0	0	0	0.5	1	0	1	2.5	1	1	0	1	3	8.0
12	1	1	1	3	0	1	0	0.5	1	1	1	4.5	1	1	1	1	4	11.5
13	1	1	0	2	0	1	0	0.5	1	0	1	3.5	1	1	0	0	2	7.5
14	1	1	0	2	0	1	1	0.5	1	0	0.5	4.0	1	1	1	0	3	9.0
15	1	1	0.5	2.5	1	1	1	0.5	1	1	0.5	6.0	1	1	0	1	3	11.5
16	1	1	0	2	0	0.5	1	1	1	1	0.5	5.0	1	1	0	1	3	10.0
Total ₂	16.0	14.5	10.5	-	4.0	13.5	9.0	11.0	15.5	9.5	13.0	-	15.5	13.0	4.0	12.0	-	-

Note: Total₁: Total score for each article; Total₂: Total score for each quality criterion; Sub-Total: Total score of quality criteria related to the methodology.

From the articles analysed in this state-of-the-art review, it is possible to conclude that only one article achieved the highest possible score in terms of quality, with a score of 14/14 (Baboota & Kaur, 2019). Therefore, this article stands out and should be considered by any researcher or professional with an interest in the topic of predicting football match outcomes.

Regarding the specific research questions, the one that stands out with the highest evaluation is the question: "What methodologies are used?" This question evaluates how the studies were conducted. Additionally, within this dimension, the second-best criterion (Q8) focuses on whether the articles explain the models they are using. However, based on Table 2.7, it is possible to identify certain gaps in the authors justification and calculation of the independent variables. Only 43.75% (n=7) of the studies provide a clear explanation of how these variables are calculated. Additionally, it is noted that only 25% (n=4) (Yu, 2022; Baboota & Kaur, 2019; Huang & Chang, 2010; Cheng et al., 2003) of the studies use averages without explaining the origin of these calculations. It is important to understand if these averages are derived from the latest matches, for example. If so, how many matches are considered for these calculations? Consequently, it becomes quite challenging to comprehend how authors utilize these variables to develop their models.

Therefore, in response to the previously posed questions, it is possible to analyse the results of the different dimensions in the context of the studied articles. In the first dimension, which addresses the application contexts of the models, an average quality score per criterion of $41/3 = 13.7$ points was achieved. In this dimension, the quality criterion with the best performance (Q1) stands out, indicating that all analysed articles clearly explain their objectives. In this dimension, it is recommended to read Nivetha et al. (2022), Azeman et al. (2021), and Tiwari et al. (2020).

In the second dimension, which deals with the methodologies used, both the second-best quality criterion (Q8) and the worst quality criterion (Q4) were identified. When it comes to this last point, many articles fall short of adequately outlining the reasons behind the historical selections, the competition, and the data base that were used (e.g., Guan & Wang, 2022; Azeman et al., 2021; Carloni et al., 2021). In this dimension, it is recommended to read Baboota and Kaur (2019), Berrar et al. (2019), and Aslan and Inceoglu (2008).

Finally, in the third dimension, which addresses the evaluation of obtained results, the third best quality criterion score is found (Q11). This indicates that articles present their general results (e.g., best model and evaluation metrics). However, the quality criterion (Q13) related to explaining the limitations of the studies has one of the lowest scores. In this dimension, it is recommended to read Azeman et al. (2020), Baboota and Kaur (2019), and Rosli et al. (2018).

3. CRISP-DM

This chapter explains the procedures used in the empirical part of the study, which the goal is to create a model that can predict the outcome of a football match. As a result, the methodology to be used is crucial since it must be simple to understand to make it easier for future researchers to replicate it.

So, the CRISP-DM methodology is used to carry out this model. This methodology involves the conversion of data into knowledge or information that is useful to the stakeholders (Chapman et al., 1999). The first phase is business understanding, where objectives are defined and the problem to be solved is outlined. The second phase is data comprehension, which is based on the gathering and analysis of the available data. The third phase is data preparation, during which the data cleaning and treatment processes are carried out. The fourth phase is modeling, where predictive models are constructed. The fifth phase is evaluation, which involves analyzing the performance of developed models. Finally, the sixth phase is deployment, where the project is implemented, and conclusions are drawn.

3.1. Business understanding

This initial phase focuses on understanding the business, seeking to clarify the problem at hand and define the objectives for the project, both business and analytics-related (Chapman et al., 1999; Schröder et al., 2021).

Concerning the business objectives, it is necessary to set a specific goal for each primary stakeholder. Consequently, the first objective centers on investors, the second on the public, and the third on clubs. The objectives consist of:

- Provide forecast advisory services based on predictions to attract and retain clients.
- Utilize match result predictions to manage fan expectations.
- Employ predictions to generate strategic insights and recommendations for clubs, coaches, and players.

In analytical terms, the goal of this study is to predict the outcomes of matches from five European leagues: the EPL, Spanish LaLiga, Italian Serie A, German Bundesliga, and Portuguese Liga Bwin. The data will span the seasons of 2019-2020, 2020-2021, and 2021-2022. This selection aligns with the points of view of Yu (2022), Azeman et al. (2020), Musa et al. (2019), Razali et al. (2017), Rosli et al. (2018), and Rebelo (2012), who emphasize that data older than three years loses relevance over time, as it does not accurately reflect the team current strength. To classify the study as successful in these predictions, the predictive model performance is expected to meet four success criteria: achieving a

minimum accuracy, precision, recall, and F1 Score of 80% on test data. The best model will be the one that provides the best results in terms of accuracy and precision in test data.

For a second analytical objective, which involves creating a dashboard, it is essential to assess the satisfaction of potential users of this dashboard. This objective is deemed successful if the average evaluation score is higher than 8 on a scale from 1 (very dissatisfied) to 10 (very satisfied). The goal of this dashboard is to track the financial performance of the predictive model and to evaluate the effectiveness of the developed investment strategies. So, the dashboard aims to answer the following questions: Q1) What is the most profitable prediction?; Q2) Which is the most profitable league?; Q3) Which is the most profitable confidence interval? It is important to consider that the confidence interval mentioned in Q3 is focused on the confidence level of each model prediction.

For the execution of this research, simple tools are necessary, such as a computer, to prepare the data and develop the model and the dashboard. All the necessary data is processed using Excel and Power Query. Excel is used for the preparation of all the necessary variables, and furthermore, the development of the dashboard for tracking the financial performance of the predictive model and to evaluate the effectiveness of the developed investment strategies. Power Query is employed to automate the entire data cleaning. Finally, to train and test the model, SPSS Modeler (v. 18.3) is used. The project has a duration of 12 months, with the aim of applying the findings to the 2023/2024 season.

3.2. Data understanding

A second phase involves collecting, identifying, and exploring the necessary data for the model development. In this stage, the data quality issues are identified, and preliminary insights are acquired from the data (Chapman et al., 1999).

In the context of this study, secondary data sources are utilized. One of the data sources used were the Football-Data.co.uk website to acquire relevant information. Additionally, to increase the study value and extract more beneficial information for predicting football match outcomes, a paid tool called CgmBet System is also used. This tool helps to extract specific match related details such as team names, goals, shots, and cards. Furthermore, data from the SoFifa website, containing FIFA game player ratings is incorporated. This inclusion is proven to improve the predictive quality by researchers such as Yu (2022), Baboota and Kaur (2019), Berrar et al. (2019), and Shin and Gasparyan (2014). Consequently, the inclusion of FIFA player ratings can potentially improve the model predictive capacity. Therefore, the data for this study originates from these three complementary data sources, as exemplified in Figure 3.1.

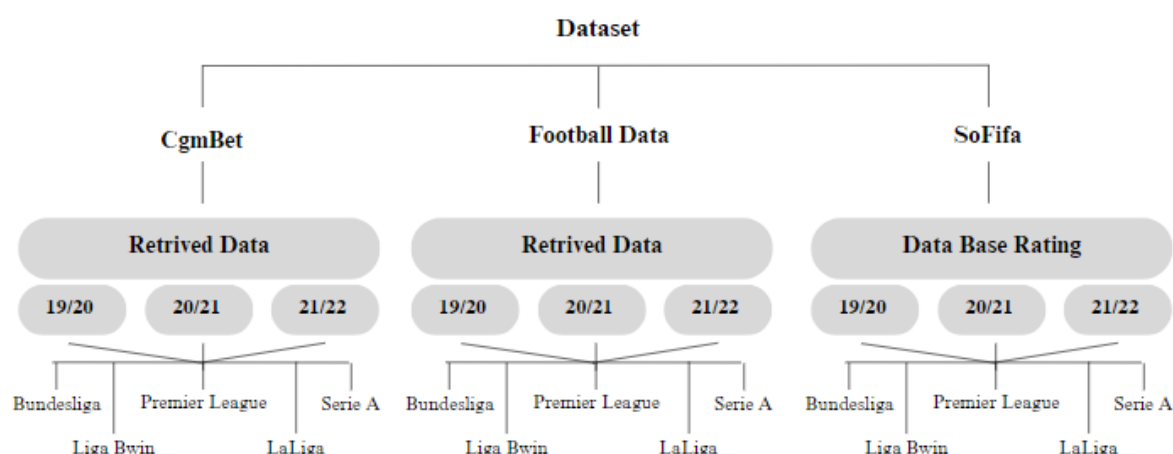


Figure 3.1: Data sources

Once the data is extracted, the selected sample consists of 5254 football matches and 191 variables. These reflect all the extracted DB columns. After a thorough analysis, it becomes evident that some variables are not suitable to use as predictors and are consequently removed.³ Remarkably, it is observed that 65% of these variables are in-game statistics of the match, while the remaining variables are betting odds sourced from bookmakers.

The Table 3.1 describes the 23 auxiliary variables (utilized for creating new variables) and the 22 variables directly employed in modeling. The auxiliary variables indicate the overall match statistics that are used to calculate the history for the subsequent matches (always with values that can be obtained before the match)⁴, making them auxiliary variables to predict the outcome of the match. Although these auxiliary variables are not employed in model construction, they were considered relevant for result prediction and show team performance in a football match. It is important to highlight that variable selection was based on statistical criteria (e.g., utilization of common and easily accessible statistical data such as goals, shots, or fouls), theoretical considerations, and the author expertise in the subject.

Table 3.1: Variables used in preparation and modeling

Role	Football Data	CgmBet	SoFifa
Preparation	HTeamTotalRedCards, ATeamTotalRedCards	Season, Date, Matchday, FTHTeamGoalsScored, FTATeamGoalsScored, HTHTeamGoalsScored, HTATeamGoalsScored, HTeamTotalShots, ATeamTotalShots, HTeamTotalShotsOnTarget, ATeamTotalShotsOnTarget, HTeamTotalCorners, ATeamTotalCorners, HTeamTotalFaults, ATeamTotalFaults, HTeamTotalYellowCards, ATeamTotalYellowCards, HTeamTotalBallPossession, ATeamTotalBallPossession, HTeamFirstGoalMoment, ATeamFirstGoalMoment	-
Modeling	-	Championship, HTeam, ATeam, Hour, AveHTeamGoalScore, AveHTeamGoalConceded, AveATeamGoalScore, AveATeamGoalConceded, HTeamGeneralStandings, HTeamGeneralStandingsAtH, ATeamGeneralStandings, ATeamGeneralStandingsAtA	HTeamOverallRating, HTeamAttackRating, HTeamMidfieldRating, HTeamDefenceRating, ATeamOverallRating, ATeamAttackRating, ATeamMidfieldRating, ATeamDefenceRating

Note: H: Home; A: Away; FT: Full time; HT: Halt time; N.A.: Not available

³ See appendix A: Variables extracted and excluded for analysis, page 65.

⁴ These variables are described in appendix referring to the sub-section 3.3.

3.3. Data preparation

The third phase refers to data preparation, which refers to the entire data cleaning and analysis procedure used to handle the initial three DB up until the creation of a final dataset (Chapman et al., 1999). In this regard, several variables were computed, all of which are considered relevant for modeling the problem.

The variable dependent used in this study is the full-time (FT) outcome. This variable distinguishes itself as being nominal because the outcome could be H, D, or A. Therefore, "H" indicates that the home team won, "D" that there was a tie in the match, and "A" that the away team won the meeting. In addition, four more variables were created, and they are as follows: half-time result (HT), which follows the same logic as the FT column but for the halftime interval of the match; the month, which returns the month of the match (January, February, March, April, May, June, July, August, September, October, November, December); the day of the week, which returns the day of the week the match was played (Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday); and the time of day, a variable dependent on the match kick-off time. If the time is earlier or equal to 1:00 PM, it is considered "morning." If the time is between 1:00 PM and 8:00 PM, it is labeled as "afternoon," and finally, if the time is later or equal to 8:00 PM, it corresponds to "night."

In addition, 62 more variables were created, categorized as averages, sd, standings, and team characteristics. It is evident that 20 variables correspond to the moving average of the last 5 matches, and another 20 variables refer to the sd of the last 5 matches, both for the home team and the away team. In cases where the last 5 matches are not available, the average is calculated based on the number of matches already played, solving the issue of initial rounds in the leagues (only by the 10th round are there 5 matches of historical data).⁵

3.3.1. Dataset

The dataset is a union of three DB, and it is essential that there be one elo between them to make the union of the DB easier. As a result, it is crucial to standardize the data (club names and seasons).

All data processing was conducted using Power Query, a tool within Microsoft Office Excel. After harmonizing the DB, only columns containing relevant information were kept. The dataset is ordered in a descending manner, meaning from the oldest to the most recent date match, to facilitate the creation of new variables.

⁵ The data dictionary for all variables used, divided into six dimensions (past behavior, standings variables, game variables, team characteristics, rating variables e odds variables), is presented in appendix B: Data dictionary, page 66.

To address the challenge presented by the first two rounds of the initial season and the issue of teams being promoted to a higher division, the employed solution is to utilize the team average and respective sd from the previous season. Consequently, Figure 3.2 illustrates the acquisition of the required data to resolve this problem.

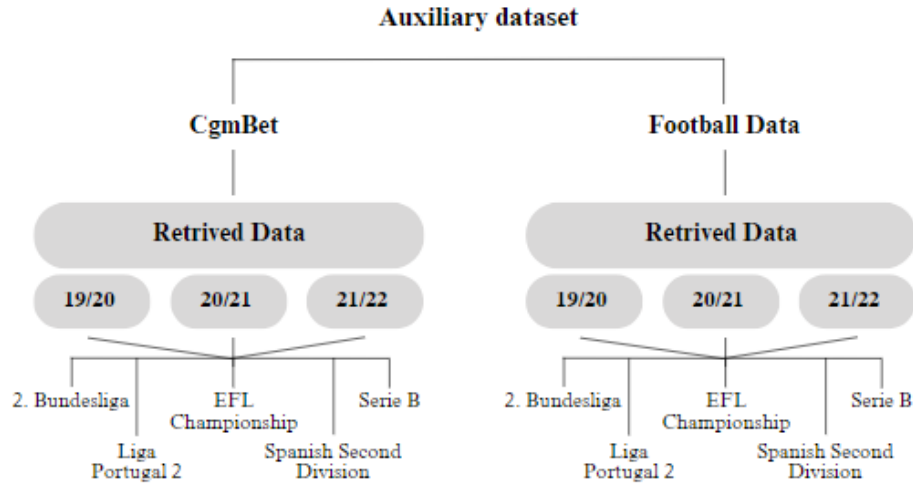


Figure 3.2: Sample selection process auxiliary dataset

In this case, it was not possible to find data concerning red cards for the Portuguese second division. As a result, the errors arising from teams promoted to a higher division had to be corrected using an alternative approach (assuming the mode value, which in this case is 0).

It is important to highlight a change in the dataset. The data related to teams that were promoted to the top division by decision of the secretariat were removed, as historical data for these teams could not be obtained (in this case, only two matches involving Gil Vicente were removed). Another intriguing and crucial point to be clarified is that for teams that were relegated in one season and promoted back to the higher division in the following season, the data for the last 5 matches is related to the season they were relegated, not the previous season. This is justified because of the characteristics of the top division compared differ significantly to the second division. Therefore, it is preferable to have data from the analyzed division rather than the lower division.

Finally, the set of variables included in the modeling can be divided into five dimensions: past behavior, standings variables, game variables, team characteristics, and rating variables. The only difference between these dimensions and the ones mentioned earlier is in the game and rating variables dimensions. Also, the game variables that are not accessible before the match takes place are not used, the ELO rating and the form variables are also not used due to the inability to access the calculation method for these variables.

3.4. Modeling

Supervised learning is based on using data (i.e., independent variables) applied to a function for which we already know the desired output (dependent variable or target) (Deo, 2015; Ray & Chaudhuri, 2021). With that said, supervised learning can be divided into two types: classification or regression. The classification involves categorizing a given dataset into two or more groups (for example: home team wins, draws, or loses). On the other hand, regression estimates the likelihood of an event occurring using a dataset. Therefore, it can be concluded that the scope of this study requires a classification-based supervised learning approach.

Following that, five different prediction algorithms (models) were tested to determine which variables have the most influence on the outcome of a football match. To accomplish this, the SPSS Modeler tool (v. 18.3) was employed to run the C5, CHAID, RT, SVM, and NN algorithms. The Table 3.2 provides a description of each algorithm used in the study.

Table 3.2: Description of each algorithm used in the research

Algorithm	Description
C5.0	It is based on DT, being an improved version of the C4.5 algorithm (Quinlan, 1993). It can handle any combination of nominal and numeric attributes and has a process to reduce overfitting (Mantas & Abellán, 2014; Polat & Gunes, 2009). When the initial DT is constructed, the pruning procedure is initiated to decrease the overall size of the tree and the estimated error rate (Quinlan, 1993);
CHAID	It is a DT based on adjusted significance tests (chi-square test of independence) (Kass, 1980). The chi-square tests of independence compare the squared divergences between observed and expected frequencies. When the chi-square value is high, there is a large difference between the two distributions, indicating dependence. The variable with the highest chi-square value is chosen to split the subset (Valle et al., 2012);
RT	It is based on DT, where multiple DT are constructed randomly. Each tree is trained with a random sample of data, and predictions are obtained by aggregating the individual predictions of each tree. This approach of RT helps reduce overfitting and increases the robustness of the model (Breiman, 2001; Liaw & Wiener, 2002);
SVM	It is widely used to solve classification or regression problems. The goal is to use support vectors to find a hyperplane that optimizes the distance between classes (Vapnik, 2013; Hsu et al., 2010);
NN	It is inspired by the functioning of the human brain. These models are composed by neurons, which are organized in layers. Each neuron receives inputs and then performs an internal processing, producing an output that is transmitted to the neurons in the next layer. The goal is to adjust the synaptic weights of the connections between neurons so that the model can make accurate predictions or classifications (Goodfellow et al., 2016; Bishop, 2006).

Note: DT: Decision Tree; RT: Random Tree

However, to improve the quality of models based on DT and NN, ensemble techniques are employed, namely bagging, and boosting. Bagging (Breiman, 1996) and boosting (Freund & Schapire, 1996) are methods used to train data and generate different models. Bagging involves creating replicas of the training dataset with replacement (bootstrap). It generates multiple models and combines them to achieve more reliable predictions, thereby improving generalization ability and increasing model stability (Quinlan, 1996). On the other hand, boosting utilizes the entire training data in each iteration, assigning weights to each instance in the data. These weights emphasize certain instances, allowing the learner to focus on different cases (Quinlan, 1996).

The development of the models was accomplished through the addition of various dimensions of variables, with accuracy serving as the main criteria for whether to keep the dimension of variables in

the model or not. The process began by including the "game variables" dimension. Subsequently, the variables related to "past behavior" were added. In this case, the initial step involved incorporating only the dimension of averages to determine the impact that average variables have compared to sd variables. At this stage, a comparison between the model using the dimension "game variables" and the model using both the dimensions "game variables" and "past behavior" are conducted to determine which one exhibited better accuracy on the test dataset. If the new accuracy was lower or equal to the accuracy of the previous model, then the new dimension of variables was removed. Otherwise, it was retained if the accuracy improved. Therefore, each addition of a new dimension of variables involved a comparison with the previous model. The purpose of testing different dimensions of variables was to identify the combination that contributed most effectively to predicting match outcomes. The Figure 3.3 provides an illustrative representation of the model training method.

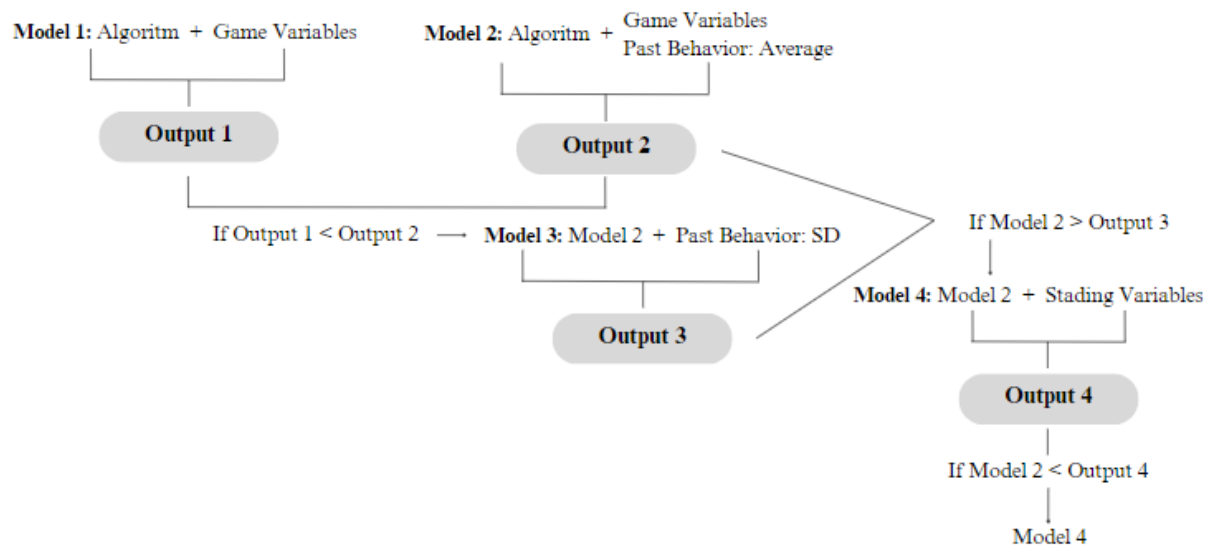


Figure 3.3: Model training method

A final test is conducted when the best model with the combinations of variables that produced the best results in the prediction of the matches results has been identified. This test involved eliminating the variables "Championship," "HTeam," and "ATeam" and creating a model without them. If the new model produced results that were similar or better than the previous model then the new model would be considered the best and, consequently, the final model. This approach allowed for the development of a model that does not incorporate the league and team names, enabling predictions for any match regardless of the league or division. This process was repeated for all five selected algorithms.

In Table 3.3 and Table 3.4, is provide a detailed performance specifications for the top three models of each algorithm. It is important to note that each algorithm listed in Table 3.3 and Table 3.4 corresponds to different dimensions of variables being used as inputs.

Table 3.3: Parameterizations of the best models for each algorithm (part 1)

Models									
	A	B	C	D	E	F	G	H	I
Algorithm	C5	C5	C5	CHAID	CHAID	CHAID	RT	RT	RT
Ensembles	Boosting	Boosting	Boosting	Boosting	Boosting	Boosting	-	-	-
Maximum tree depth	-	-	-	5	5	5	20	20	20
Records in parent branch	-	-	-	2	2	2	-	-	-
Records in child branch	-	-	-	1	1	1	5	5	5
Misclassification costs	No	No	No	No	No	No	No	No	No
Years under review	3 seasons	3 seasons	3 seasons	3 seasons	3 seasons	3 seasons	3 seasons	3 seasons	3 seasons

Note: CHAID: Chi-squared Automatic Interaction Detection; RT: Random Trees

Table 3.4: Parameterizations of the best models for each algorithm (part 2)

Models						
	J	K	L	M	N	O
Algorithm	SVM	SVM	SVM	NN	NN	NN
Ensembles	-	-	-	Boosting	Boosting	Boosting
Maximum tree depth	-	-	-	-	-	-
Records in parent branch	-	-	-	-	-	-
Records in child branch	-	-	-	-	-	-
Misclassification costs	No	No	No	No	No	No
Years under review	3 seasons	3 seasons	3 seasons	3 seasons	3 seasons	3 seasons

Note: SVM: Support Vector Machine; NN: Neural Networks

The Table 3.5 presents the dimensions of variables used in each algorithm. It is important to note that the sub-dimension "average" it is included within the dimension "past behavior." Therefore, the "past behavior" dimension refers to the combined set of "average" and "sd" variables.

Table 3.5: Dimensions of the best models for each algorithm

Models	Dimensions
A	GameVariables, Past behavior: average, Past behavior: SD, Standings, WithoutTeams, WithoutChampionship
B	GameVariables, Past behavior: average, Past behavior: SD, Standings, WithoutChampionship
C	GameVariables, Past behavior: average, Past behavior: SD, Standings
D	GameVariables, Past behavior: average, TeamCaratheristics
E	GameVariables, Past behavior: average, TeamCaratheristics, WithoutChampionship
F	All variables
G	All variables
H	All variables
I	GameVariables, Past behavior: average, Past behavior: SD, Rating, WithoutChampionship
J	All variables
K	All variables
L	All variables
M	GameVariables, Past behavior: average
N	GameVariables, Past behavior: average, Rating
O	GameVariables, Past behavior: average, TeamCaratheristics

Note: i) SD: Standard deviation; ii) Past behavior: average: only four variables represent averages to date.

3.5. Evaluation

In this phase, the results of the models are evaluated by analyzing their performance using the defined metrics. This assessment aims to determine if the achieved performance aligns with the objectives set (Chapman et al., 1999).

The hold-out 70/30 methodology was used to validate the model. This means that 70% of the sample was used to train the model, and the remaining 30% was used to evaluate the model performance. The samples were chosen randomly. To evaluate the models with classification issues the standard metrics were employed: accuracy, precision, recall, F1 score, RPS, and the confusion matrix. These metrics are generally used to assess the performance of classification problems (Sokolova & Lapalme, 2009; Mohandas et al., 2023). The confusion matrix used in the context of this study is detailed in Table 3.6, while the metrics used and their calculation expressions are explained in Table 3.7.

Table 3.6: Confusion matrix

	H	D	A
H	TP	FP	FP
D	FP	TP	FP
A	FP	FP	TP

Note: True Positives: The cases correctly predicted as positive; False Positives: The cases that are wrongly predicted to be positive; Total Predictions: Is the total number of predictions.

Source: Mohandas et al. (2023): page 8-9.

Table 3.7: Metrics for evaluation

Metrics	Formula	Explanation
Accuracy	$\frac{\text{True Positive}}{\text{Total Predictions}}$	Number of correct predictions divided to the total number of predictions made by the model.
Precision	$\frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$	Evaluates the proportion of correct predictions in relation to the total number of positive predictions (in column).
Recall	$\frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$	Evaluates the proportion of correct predictions in relation to the total number of actual positive cases (in row).
F1 Score	$\frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$	It combines precision and recall into a single measure. Essentially, it is the harmonic mean between the two metrics, providing a balanced measure of the model performance.
Support	$\frac{\text{Total Predictions of the item}}{\text{Total Predictions}}$	Ratio of samples containing a specific item. That is, the percentage of times the model predicts H, D, or A.

Note: H: Home; D: Draw; A: Away

Based on these metrics, the best models for each algorithm were evaluated, and the one that demonstrates the highest accuracy and precision on the test dataset will be considered the best performing model. In the context of the study, there is not a specific category that is considered positive. Therefore, the metrics for all three categories are presented in the results chapter. Regarding the model adjustment, it is ensured that it is neither underfitting or overfitted, as both the training and test datasets show high metric values.

In the context of investors, the mentioned metrics can demonstrate the effectiveness of the model. When closely examining the metrics, it becomes evident that precision provides the greatest reassurance to investors. A precision of 80% in the class of the home team winning indicates that out of every 100 model predictions for a home team victory, 80 will be accurate. These metric highlights that it is possible to generate profit by investing in odds higher than 1.20. The odds are the inverse of probability (Angelini and De Angelis., 2012), therefore, in this case, $1/0.80 = 1.20$. So, the higher the odds, the lower the probability of that event happening. On the contrary, the lower the odds of a certain event, the higher the probability of it happening.

It is also crucial to calculate the metrics mentioned earlier for each analysed league because the model might only show promising results in half of them. This approach provides a more comprehensive and accurate evaluation of the model performance, allowing for a better understanding its strengths and weaknesses.

Finally, to obtain an evaluation of the dashboard, it is going to be constructed a questionnaire to be answered by three potential users of the tool (an investor, a fan, and a professional at a football club). The questionnaire consists in six questions, five of which have responses on an ascending scale of satisfaction, ease of use, appropriation, or appealing, ranging the rates from 0 to 10 points. Additionally, there is an open-ended question for an overall evaluation of the dashboard, along with a brief justification for each response. The questions are: Q1) How do you evaluate the usefulness of the dashboard for monitoring the performance of the predictive model and to evaluate the effectiveness of the developed investment strategies, on a scale of 1 (very dissatisfied) to 10 (very satisfied)?; Q2) How would you rate the ease of use of this dashboard on a scale of 1 (not easy) to 10 (extremely easy)?; Q3) Do you think the indicators on this dashboard are adequate for monitoring the model financial perspective? Give your response on a scale of 1 (not appropriate) to 10 (very appropriate)?; Q4) Is the dashboard visual presentation appealing? Give your response on a scale of 1 (not appealing) to 10 (very appealing)?; Q5) What is your overall rating of the dashboard on a scale of 1 (very dissatisfied) to 10 (very satisfied)?; Q6) What is your overall satisfaction of the dashboard?

3.6. Deployment

Finally, the last phase of the CRISP-DM process involves presenting the results obtained by the model, along with an implementation plan, continuous monitoring, and maintenance (Chapman et al., 1999).

In this dissertation, it is crucial to emphasize its significance for investors, football clubs, sports enthusiasts, and the scientific community in the field of football prediction using ML models. The development of a predictive model for football results contributes to the advancement of current literature on the subject. Given that, this field is continually evolving and presenting a successful new solution adds value to the scientific community. However, due to the time constraints imposed on this

research, it was not possible to fully execute this phase. Then, its implementation will take place throughout the 2023-2024 season.

Therefore, at this stage of the project, implementation is manifested in this present dissertation, which provides a detailed explanation of the entire process of creating the model and the dashboard. Additionally, two prototypes are provided, one for the predictive model and another for the dashboard, allowing the generation of insights and the monitoring of the predictive model performance, even though still in a testing capacity.

4. RESULTS AND DISCUSSIONS

This chapter aims to follow up on the methodology discussed earlier, presenting the results of the analyses conducted, interpreting them, and discussing their relevance to each of the defined objectives.

4.1. Predictive model for match outcome

As explained in the methodology, many model combinations were tested, and the results of the respective quality metrics are shown in Table 4.1 and Table 4.2.

To correctly determine the best model from the previous ones, accuracy, and precision greater than 80% in new data (test sample) were used as decision criteria. After analyzing the confusion matrix for each of the models for the target categories (H, D, and A), it can be concluded that the C5 model with three years of historical data is the one that performs best (Table 4.1) as it has the highest accuracy among all. Additionally, it is observed that the precision for category H does not vary much between models. As a result, using the precision of D and the precision of A as decision criteria, it is concluded that the model C5 has the highest values. Furthermore, it is a tree-based model that generates decision rules leading to a specific outcome (prediction of the result). However, by using boosting (combining models), the complexity of the final model has increased.

Table 4.1: Results of the algorithms in the test sample (part 1)

C5				CHAID			RT		
Metrics	H	D	A	H	D	A	H	D	A
Accuracy	85.40%			85.11%			80.85%		
Precision	85.73%	85.60%	84.77%	87.55%	81.77%	84.39%	87.42%	80.21%	74.07%
Recall	91.03%	76.73%	85.27%	88.30%	81.77%	83.40%	79.42%	77.42%	85.84%
F1 Score	88.30%	80.92%	85.01%	87.93%	81.77%	83.89%	83.23%	78.79%	79.52%

Note: i) H: Home; D: Draw; A: Away; CHAID: Chi-squared Automatic Interaction Detection; RT: Random Trees; ii) Accuracy represents the percentage of correct predictions for all classes.

Table 4.2: Results of the algorithms in the test sample (part 2)

SVM				NN		
Metrics	H	D	A	H	D	A
Accuracy	85.18%			84.44%		
Precision	85.39%	85.04%	84.99%	84.58%	85.32%	83.65%
Recall	90.34%	74.88%	86.62%	88.21%	78.37%	84.29%
F1 Score	87.80%	79.64%	85.80%	86.36%	81.70%	83.97%

Note: H: Home; D: Draw; A: Away; SVM: Support Vector Machine; NN: Neural Networks

4.2. Most important predictors

Regarding the most important variables in the model (average goals conceded by the away team in the last 5 matches and the home team position zone in the league table), it can be seen the top 10 in Figure 4.1. It is important to highlight that half of the variables are related to the averages of the last 5 matches. Therefore, the most relevant variable dimension is "Past Behavior: average," followed by the "Standing Variables" dimension, with around 30% (n=3).

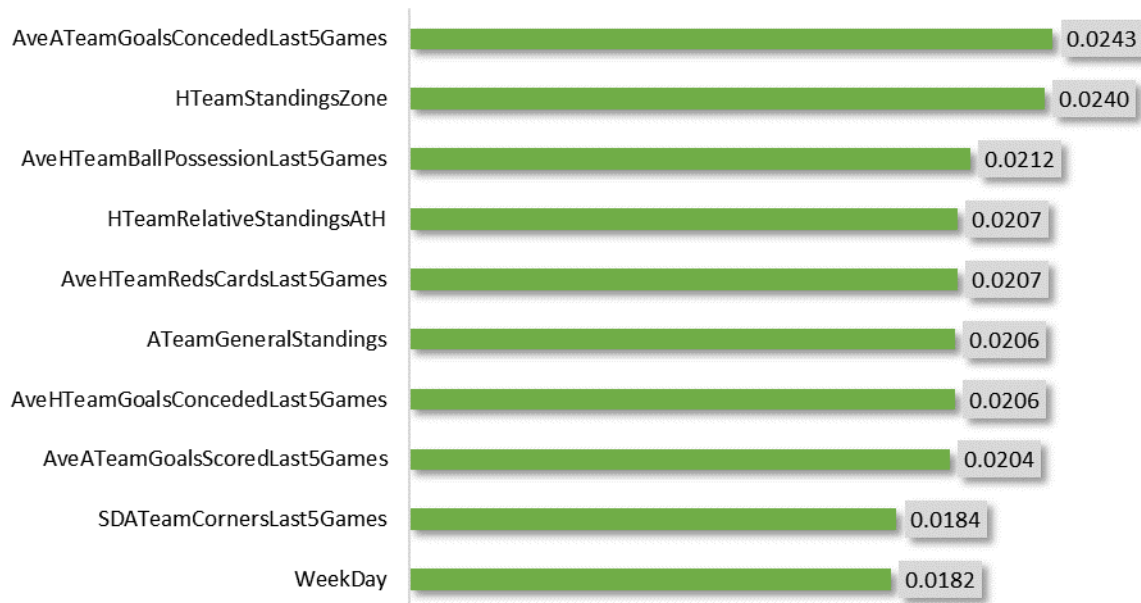


Figure 4.1: Top 10 - Variables importance

Analysing the Figure 4.1, it is clear that five variables associated with the home team have a significant impact on the outcome prediction. Although the most important variable is related to the away team, it is interesting to note that the next four variables are directly associated with the home team. This fact partly explains the relevance of the home team variables in the prediction process, even in a context where the primary variable belongs to the away team, although with very close importance to the variable HTeamStandingsZone, with these two being the most important. Observing the dimensions of "Past Behavior: Average" and "Standing Variables" in a more pormenorized manner, it is concluded that the results have a theoretical and practical meaning.

The Figure 4.2 and Figure 4.3 provide a more detailed analysis of the performance of the most important predictor variable in relation to predictions. It stands out that the prediction of the home team victory is associated with a higher average of goals conceded by the away team. On the other hand, predictions of a draw and the away team victory are associated with a lower number of goals conceded by the away team (Figure 4.2). In other words, as expected, the average number of goals conceded by the away team decreases as the predicted result becomes more favorable to them. In practice, when the away team has a substantial average of goals conceded in their previous matches,

it is reasonable to assume that this team will have a lower probability of winning. This conclusion is relevant in the context of investors.

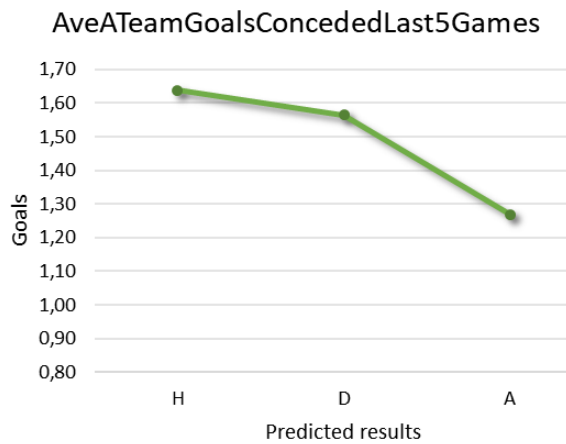


Figure 4.2: Average goals conceded by the away team per prediction

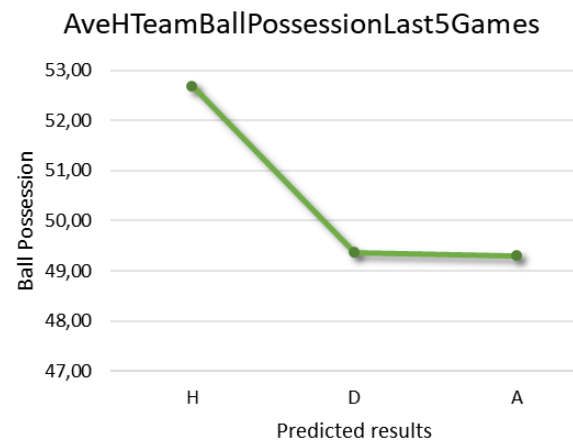


Figure 4.3: Average ball possession by the home team per prediction

The Figure 4.3 shows that the higher the average ball possession of the home team, the more favorable the prediction of the result for them, with not much difference in the averages between the draw and the away team victory. Although there are football matches where the team with less ball possession ends up winning, in general, controlling ball possession often translates into greater influence on the match and, consequently, a higher probability of success for the team with more ball possession.

The Figure 4.4 provides an enlightening visual representation by indicating that teams playing at home in the Champions Playoff Zone (CPZ), Champions Zone (CZ), Europa League Playoff Zone (ELPZ), and Europa League Zone (LEZ) tend to achieve a predicted favorable result (victory) more than 68.26% of the time. On the other hand, in other classification zones, the result tends to be more unpredictable. This information, once again, proves to be quite relevant in the context of investors.

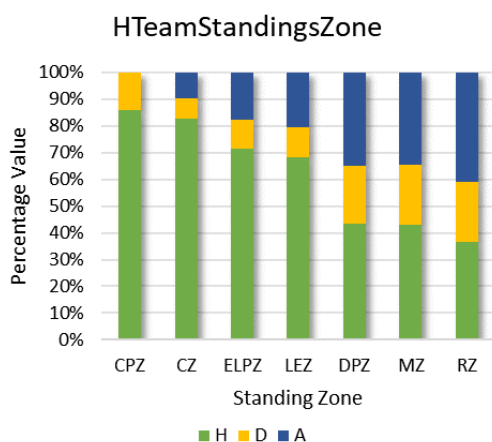


Figure 4.4: Predicted outcome by home team position in the standings

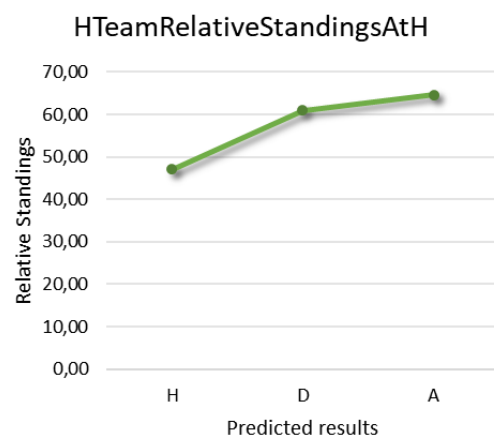


Figure 4.5: Relative standings by the home team per prediction

Furthermore, Figure 4.5 represents the relative ranking of the home team when playing at home. This relative ranking is calculated by dividing the team current position for a particular match by the total number of teams in that league. Therefore, the lower this ratio, better the ranking. The Figure 4.5 shows that a prediction of a home team victory has better rankings (on average) because it has the lowest value among the three possible categories, followed by draws and, finally, the away team.

4.3. Game profiles with a high propensity for a result

In order to identify profiles with predictions for all 3 types of classes, it was used the terminal nodes with high confidence (greater than or equal to 75% for H and A predictions, and greater than or equal to 50% for D. A lower confidence level is considered for D because predicting draws is more challenging, still, its aimed for a confidence level greater than 50%) and some support (10 or more matches) from the C5 model.

Profiles of predictions for the home team victory, translated into propositional rules:

- ATeamRelativeStandings > 26.389, and HTeamGeneralStandings <= 7.500, and AveATeamBallPossessionLast5Games <= 52.305, and HTeamRelativeStandingsAtH > 5.278, and AveHTeamBallPossessionLast5Games > 48.450, and AveATeamYellowCardsLast5Games <= 3.511, and ATeamStadingsZone = "RZ", and AveATeamYellowCardsLast5Games <= 2.700, and MomentOfDay = "Afternoon", and AveATeamGoalsConcededLast5Games <= 2.500, and AveHTeamShotsOnTargetLast5Games > 5.450 and, AveATeamGoalsConcededLast5Games <= 2.300, then H (support: 24 predictions; confidence: 95.80%);
- ATeamRelativeStandings > 26.389, and HTeamGeneralStandings <= 7.500, and AveATeamBallPossessionLast5Games <= 52.305, and HTeamRelativeStandingsAtH > 5.278, and AveHTeamBallPossessionLast5Games > 48.450, and AveATeamYellowCardsLast5Games <= 3.511, and ATeamStadingsZone = "MZ", and AveHTeamCornersLast5Games <= 7.500, and HTeamStadingsZone = "CZ", and AveHTeamYellowCardsLast5Games > 1.450, and AveHTeamFirstGoalMomentLast5Games > 31.075, and SDHTeamGoalScoredLast5Games > 0.561 and, AveATeamBallPossessionLast5Games <= 48.900, then H (support: 43 predictions; confidence: 97.70%).

Profiles of predictions for the draw, translated into propositional rules:

- ATeamRelativeStandings > 26.389, and HTeamGeneralStandings > 7.500, and ATeamGeneralStandings <= 17.500, and AveHTeamGoalsConceded <= 1.839, and HTeamRelativeStandings <= 97.500, and AveATeamShotsOnTargetLast5Games <= 4.708, and ATeamStadingsZone = "MZ", and HTeamRelativeStandingsAtH <= 92.222, and ATeamRelativeStandings <= 55.278, and AveHTeamCornersLast5Games <= 7.059, and

SDATeamFirstGoalMomentLast5Games > 2.250, and WeekDay = "Friday" and, SDHTeamGoalScoredLast5Games > 0.774, then D (support: 11 predictions; confidence: 54.50%);

- ATeamRelativeStandings > 26.389, and HTeamGeneralStandings > 7.500, and ATeamGeneralStandings <= 17.500, and AveHTeamGoalsConceded <= 1.839, and HTeamRelativeStandings <= 97.500, and AveATeamShotsOnTargetLast5Games <= 4.708, and ATeamStadingsZone = "MZ", and HTeamRelativeStandingsAtH <= 92.222, and ATeamRelativeStandings <= 55.278, and AveHTeamCornersLast5Games <= 7.059, and SDATeamFirstGoalMomentLast5Games > 2.250, and WeekDay = "Saturday" and, AveATeamFirstGoalMomentLast5Games <= 65.250, and MomentOfDay = "Afternoon", and SDATeamCornersLast5Games <= 1.777, and AveHTeamBallPossessionLast5Games <= 53.300 and, SDHTeamFirstGoalMomentLast5Games > 12.360, then D (support: 10 predictions; confidence: 100%).

Profiles of predictions for the away team victory, translated into propositional rules:

- ATeamRelativeStandings <= 26.389, and AveHTeamShotsOnTargetLast5Games <= 6.029, and AveATeamShotsLast5Games > 8.700, and AveATeamBallPossessionLast5Games > 47.500, and HTeamGeneralStandings > 4.500, and HTeamRelativeStandings <= 97.500, and AveATeamGoalsConcededLast5Games <= 1.733, and SDHTeamRedsCardsLast5Games <= 0.481, and AveATeamGoalsScoredLast5Games <= 3.225, and Month = "February", and AveHTeamGoalsScored <= 1.867 and, AveHTeamRedsCardsLast5Games <= 0.100, then A (support: 27 predictions; confidence: 81.50%);
- ATeamRelativeStandings <= 26.389, and AveHTeamShotsOnTargetLast5Games <= 6.029, and AveATeamShotsLast5Games > 8.700, and AveATeamBallPossessionLast5Games > 47.500, and HTeamGeneralStandings > 4.500, and HTeamRelativeStandings <= 97.500, and AveATeamGoalsConcededLast5Games <= 1.733, and SDHTeamRedsCardsLast5Games <= 0.481, and AveATeamGoalsScoredLast5Games <= 3.225, and Month = "January" and, SDATeamGoalsConcededLast5Games <= 0.774, then A (support: 30 predictions; confidence: 76.70%).

Analysing the last propositional rule for the prediction of away team victories, this can be translated into a game profile: the better the relative classification of the away team and the average shots on target in the last 5 games of the home team, and the average shots in the last 5 games of the away team, as well as the average possession of the ball in the last 5 games of the away team, are reasonable, and the overall classification of the home team as well as the average of goals conceded in the last 5 games of the away team and the dispersion of red cards in the last 5 games of the home

team is quite good, and the average goals scored in the last 5 games of the away team are relatively reasonable, and the dispersion of goals conceded in the last 5 games of the away team is very low, and finally, the month of the game is January.

4.4. Other assessments of the predictive model

4.4.1. Generalization to new data

A good model should be generalizable to new data, so the results of applying the model to a new sports season, the 2022-2023 season, are going to be presented here. Using the C5 model, predictions were made for the mentioned season, and therefore, all results were based on a new dataset that was not used to train or test the model.

Analysing the results by league, there is a slight tendency for the model to predict a home team win, with the two highest values in the German Bundesliga (57%) and the Liga Bwin (56%) (Table 4.3).

Table 4.3: Predictions of the model by championship (2022-2023)

Predictions			
Championship	H	D	A
Bundesliga	57%	19%	24%
Liga Bwin	56%	18%	26%
Premier League	55%	16%	29%
LaLiga	54%	18%	28%
Serie A	48%	19%	33%

Note: H: Home; D: Draw; A: Away

Table 4.4: Actual results of the matches by championship (2022-2023)

Actual			
Championship	H	D	A
Liga Bwin	49%	18%	33%
Premier League	48%	23%	29%
LaLiga	48%	23%	29%
Bundesliga	47%	25%	28%
Serie A	43%	26%	31%

Note: H: Home; D: Draw; A: Away

Since the 2022-2023 season has already been concluded, Table 4.4 presents the actual results, showing that 49% of the matches in the Liga Bwin ended with a home team win. The comparison between the predicted and actual outcomes demonstrates that the model is in line with the reality of the sports season under analysis. Over a season, it is common to have more home team wins, indicating a slight home advantage.

Regarding the evaluation metrics, these are crucial for understanding the model performance during the matchdays. The accuracy, which represents the percentage of correct predictions compared to the total predictions, has an average of 47%. The Table 4.5 displays the confusion matrix for new data from the 2022-2023 season.

Table 4.5: Confusion matrix in new data (2022-2023)

	H	D	A
H	526	129	167
D	207	82	117
A	212	102	210

Note: H: Home; D: Draw; A: Away

In Figure 4.6, it is possible to observe that the model outperforms the accuracy in 42% (n=16) of the matchdays, illustrating the matchdays where its predictions are more accurate than the overall performance. For instance, on matchday 4, the model achieved an accuracy of 52%, surpassing the overall accuracy by 5 percentage points.

Additionally, it is worth noting that there is not a high variation in accuracy throughout all the matchdays. The model can predict matches in the early matchdays as well as matches in the final matchdays with approximately the same average. The matchday 24 stands out as the one with the worst result (around 30% accuracy). On the other hand, matchdays 6 and 32 stand out as the best ones (around 60% accuracy).

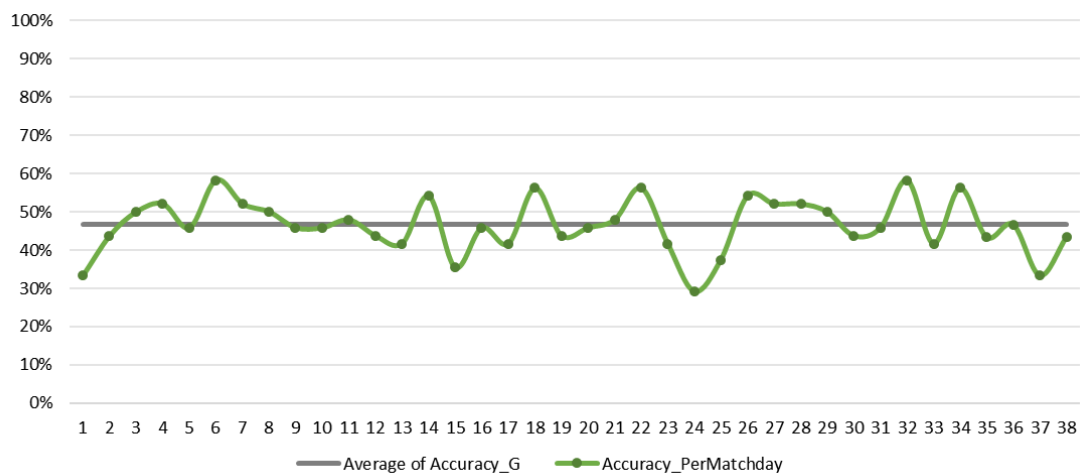


Figure 4.6: Average accuracy vs accuracy per matchday

When the precision of class H is analyzed, it is possible to understand that the model has slightly better performance (average precision of 56%). This metric shows that out of every 100 predictions of the home team winning, the model will get 56 correct, indicating a higher level of precision in this category compared to accuracy. Furthermore, the model overlaps the average precision in 52% of matchdays (n=20) (Figure 4.7). The matchdays 32 and 34 stand out as having the best precision values for H (approximately 75%). On the other hand, the early matchdays have values well below the average (around 40%). Therefore, the precision for H, although showing slightly more variation than accuracy, does not exhibit significant fluctuations.

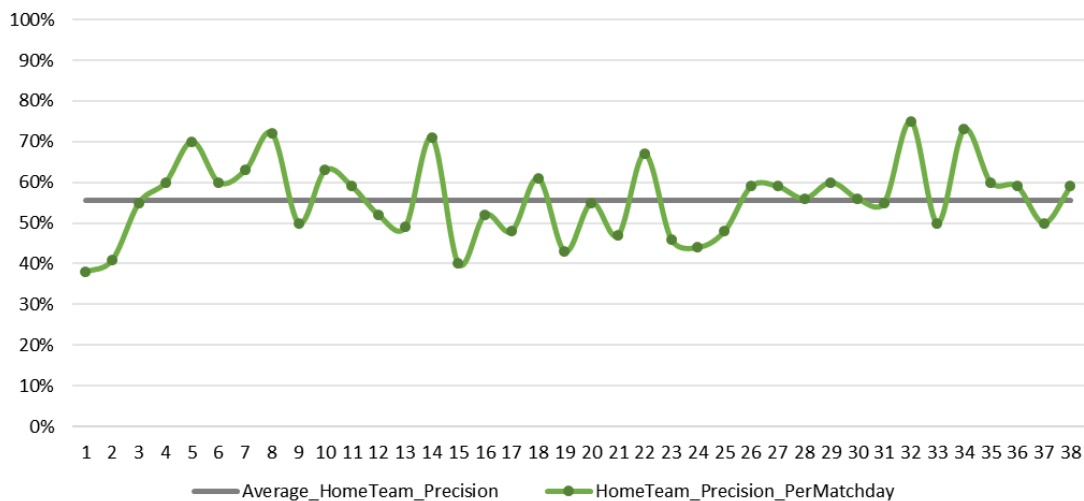


Figure 4.7: Home team precision vs home team precision per matchday

The precision for draws is the metric with the worst model performance (26%). The Figure 4.8 shows the variation across matchdays, which is quite high, including in matchdays 1 and 31, where it did not make any correct predictions. However, it is not all bad in this metric. The best result was, curiously, right after one of the worst matchdays (in matchday 2) with around 55%. Moreover, the model overlaps the average precision for draws in 52% of matchdays (n=20), with a better performance than the average. However, this is a metric seen as an area for model improvement.

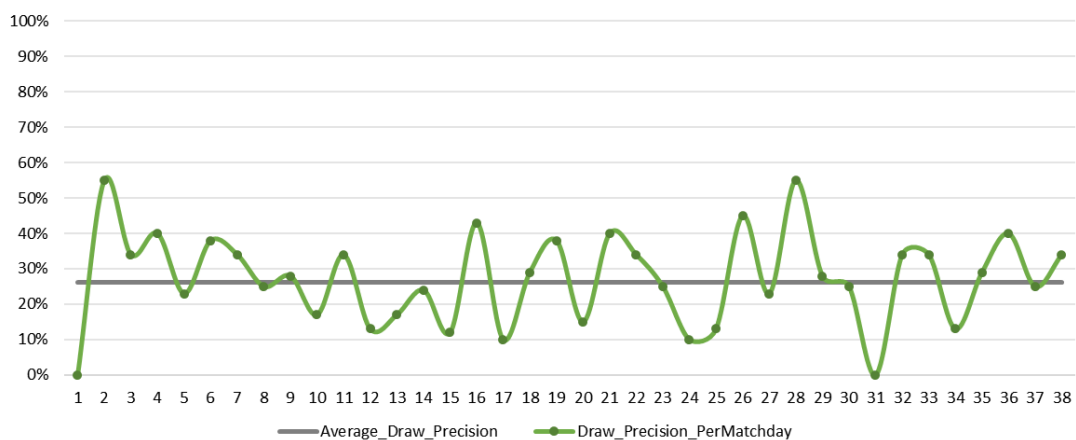


Figure 4.8: Draw precision vs draw precision per matchday

Finally, the precision for the away class is, on average, 43%. The model surpasses this value in 50% of the matchdays (n=19). This result reveals that the model has a reasonable ability to predict the away team victory. The Figure 4.9 shows a high variation in this metric, with a slight decreasing trend over time. Additionally, there is a slight descending trend in the results over the matchdays. In the early matchdays, the model reaches its two peaks (approximately 75% in matchday 1 and 70% in matchday 6). These values were not repeated, and since then, the result has been declining, as matchday 37 presents the worst result in this metric, with around 10% of precision.

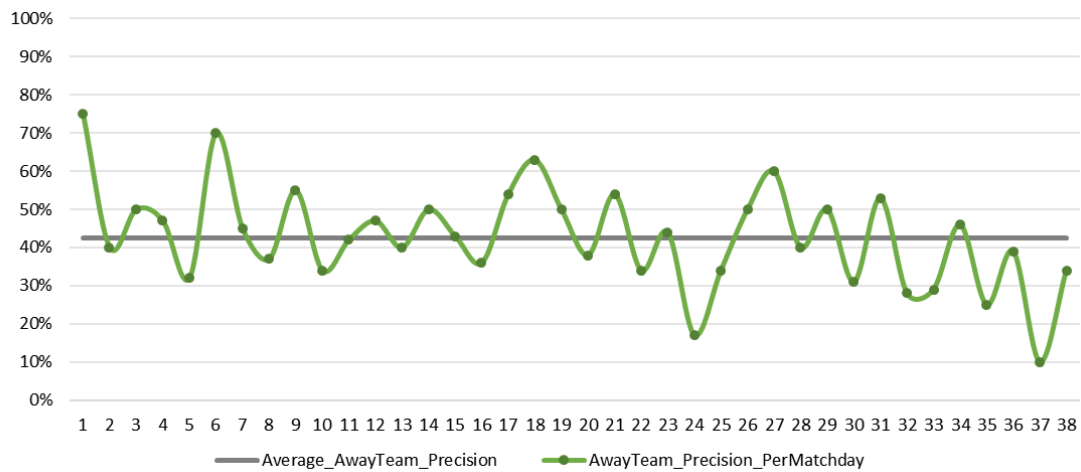


Figure 4.9: Away team precision vs away team precision per matchday

4.4.2. Financial perspective

In a financial approach to the model, a unit of investment (ui)⁶ is considered, meaning that one unit is invested in each prediction of the model. When analyzing the results, it becomes evident that the most profitable league is the EPL, with a return of 31 ui (ROI = 8.27%) over the 2022-2023 season, followed by the Italian Serie A with a profit of 16 ui (ROI = 4.21%). In total, if someone had invested in all the predictions made, would have been achieved an approximate profit of 19 ui (ROI = 1.11%) (Table 4.6).

Table 4.6: Profit-Loss by championship (2022-2023)

Championship	PL
Premier League	31.43
Serie A	16.02
LaLiga	2.44
Bundesliga	-10.95
Liga Bwin	-19.35
Total	19.59

Note: PL: Profit-loss (ui)

Table 4.7: Top 5 most lucrative teams playing at home (2022-2023)

HomeTeam	PL
Getafe	14.87
Bologna	12.40
Everton	11.89
Osasuna	9.94
Milan	9.68
Total	102.20

Note: PL: Profit-loss (ui)

Regarding the most profitable home teams, it is possible to observed that only one corresponds to the most profitable league. The most representative league for the top 5 most profitable home teams are the Italian Serie A and the Spanish La Liga, with two teams each (Table 4.7).

The Table 4.7 shows that, although the EPL is the most profitable league, the most profitable home teams are not exclusively concentrated in this league. On the contrary, the Italian Serie A and Spanish La Liga also have highly profitable teams playing at home. The numbers presented are promising, as they indicate that the model has the potential for positive financial returns over the season.

⁶ Ui: Unit investment. It is considered as an investment unit.

Dividing the 38 matchdays into three parts (1-12; 13-25; 26-38), it is possible to see that the first and third sets of matchdays provide two opportunities for profitable investment. The data show that investing in matches between matchdays 13 and 25 would result in a loss of 56 ui, but the first and third group of matchdays would result in a profit of 41 ui (ROI = 7.18%) and 34 ui (ROI = 6.24%), respectively (Table 4.8).

Table 4.8: Profit-loss per matchday (2022-2023)

Matchday	PL	Matchday	PL	Matchday	PL
1	-16.33	13	-5.90	26	10.28
2	1.66	14	4.39	27	3.56
3	5.63	15	-12.89	28	9.86
4	2.99	16	2.80	29	10.80
5	3.46	17	-13.05	30	-1.04
6	20.80	18	10.28	31	4.41
7	14.63	19	-2.54	32	10.20
8	4.30	20	-1.85	33	-2.67
9	2.47	21	-4.52	34	2.86
10	-2.69	22	8.90	35	-1.93
11	7.74	23	-9.80	36	-1.16
12	-3.31	24	-21.08	37	-8.57
-	-	25	-10.92	38	-2.18
Total	41.35	Total	-56.18	Total	34.42

Note: PL: Profit-loss (ui)

The results indicate that there are specific periods during the season when the model performs more favourably in its predictions. Therefore, investing only in matches from matchdays 1 to 12 and 26 to 38 allows for positive returns, while avoiding the period from matchdays 13 to 25 minimizes potential losses.

The Table 4.9 presents the model predictions for matchday 6 in the five leagues under study. During this matchday, the model made a total of 48 predictions, correctly predicting 28 of them, representing an accuracy rate of 58%. When considering the financial aspect of matchday 6, it can be observed that if one ui were invested in each of the model predictions (i.e., a total of 48 ui), it would have been possible to obtain a profit of 20.80 ui. These results indicate that the model was able to achieve a positive ROI of $20.8/48 = 43.33\%$ for this specific matchday.

Table 4.9: Predictions for the matchday 6 (2022-2023)

HTeam	ATeam	Predicition	OddPredicition	FT	PL
Aston Villa	Man City	A	1.33	D	-1.00
Brentford	Leeds	H	2.17	H	1.17
Chelsea	West Ham	A	5.44	H	-1.00
Everton	Liverpool	D	5.40	D	4.40
Newcastle	Crystal Palace	D	3.47	D	2.47
Nottingham	Bournemouth	D	3.38	A	-1.00
Tottenham	Fulham	H	1.54	H	0.54
Wolves	Southampton	H	2.14	H	1.14
Brighton	Leicester	H	1.94	H	0.94
Man United	Arsenal	A	2.69	H	-1.00
Guimaraes	Santa Clara	H	2.08	H	1.08
Werder Bremen	Augsburg	A	6.25	A	5.25
Bayern Munich	Stuttgart	H	1.22	D	-1.00
Famalicao	Benfica	A	1.42	A	0.42
Frankfurt	Wolfsburg	A	4.13	A	3.13
Hertha	Leverkusen	D	4.14	D	3.14
Hoffenheim	Mainz	A	2.86	H	-1.00
Inter	Torino	H	1.60	H	0.60
Leipzig	Dortmund	H	2.45	H	1.45
Napoli	Spezia	H	1.29	H	0.29
Porto	Chaves	H	1.22	H	0.22
Sampdoria	Milan	D	4.40	A	-1.00
Schalke 04	Bochum	H	2.16	H	1.16
Sporting	Portimonense	H	1.33	H	0.33
Arouca	Boavista	A	2.96	A	1.96
Atalanta	Cremonese	H	1.40	D	-1.00
Bologna	Fiorentina	H	3.57	H	2.57
FC Koln	Union Berlin	A	3.03	A	2.03
Freiburg	Monchengladbach	H	2.20	D	-1.00
Juventus	Salernitana	H	1.60	D	-1.00
Lazio	Verona	H	1.64	H	0.64
Lecce	Monza	H	2.46	D	-1.00
Maritimo	Gil Vicente	H	2.85	A	-1.00
Paços de Ferreira	Casa Pia	A	2.59	A	1.59
Rio Ave	Braga	A	1.58	A	0.58
Sassuolo	Udinese	H	2.72	A	-1.00
Empoli	Roma	A	1.72	A	0.72
Vizela	Estoril	H	2.13	A	-1.00
Valladolid	Cadiz	H	2.06	A	-1.00
Barcelona	Elche	H	1.08	H	0.08
Bilbao	Vallecano	H	1.65	H	0.65
Mallorca	Almeria	D	3.39	H	-1.00
Valencia	Celta Vigo	D	3.03	H	-1.00
Atl Madrid	Real Madrid	A	2.50	A	1.50
Betis	Girona	H	1.75	H	0.75
Osasuna	Getafe	H	1.70	A	-1.00
Real Sociedad	Espanyol	D	3.80	H	-1.00
Villarreal	Sevilla	H	1.76	D	-1.00

Note: H: HomeTeam win; D: Draw; A: AwayTeam win; FT: Full time result; PL: Profit-loss (ui)

The results demonstrate that the model has a promising predictive capability, but it is essential to continue monitoring and evaluating its performance over time to ensure its continuing effectiveness.

The dashboard illustrated in Figure 4.10 allows for tracking the financial performance of the predictive model and the evaluation of the developed investment strategies. In this dashboard, it is possible to make comparisons across various variables, considering the 2022-2023 season.



Figure 4.10: Dashboard for tracking the profit and loss (2022-2023)

To improve the interpretation of the visuals, it is important to consider that each match represents one unit (ui). This means that for each model prediction, one unit is placed (i.e., considering €100 as one unit, if an investor had invested in all EPL predictions, they would have made a profit of €3143 in the 2022-2023 season).

Regarding the questions inquired to the data, the dashboard clearly answers all of them. Therefore, Q1) What is the most profitable prediction? Through the first visualization (Figure 4.11), it is concluded that investing in predictions of the home team winning is the most profitable strategy.



Figure 4.11: What is the most profitable prediction? (2022-2023)

Regarding Q2) Which is the most profitable league? In the second graph (Figure 4.12), it can be observed that the most profitable league is the EPL, with approximately twice the profit compared to the second most profitable one.

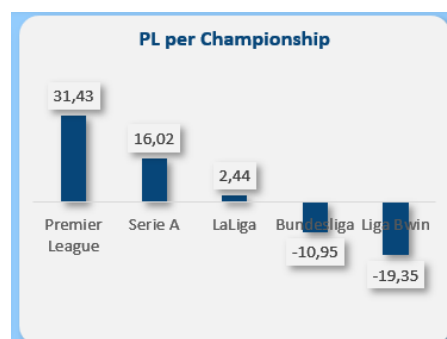


Figure 4.12: Which is the most profitable league? (2022-2023)

Finally, Q3) Which is the most profitable confidence interval? The confidence interval that proves to be the most profitable is the]75% ; 80%] (Figure 4.13). This interval includes all the matches where the prediction result has a confidence between 75% and 80%, with the degree of

confidence provided by the model for each prediction it makes. However, the dashboard presents more visualizations (graphs) that complement the financial analysis by showing, for example, profit or loss per month, the top 10 most profitable matchdays, and the top 10 most profitable home teams, as well as three cards that display the most profitable matches times (Morning, afternoon, and night).

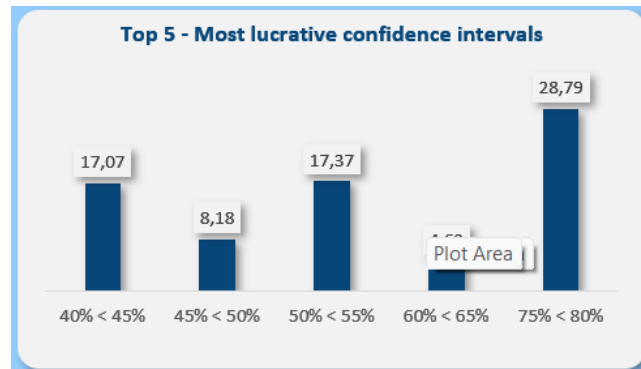


Figure 4.13: Which is the most profitable confidence interval? (2022-2023)

The Table 4.10 presents the responses from the three potential users.

Table 4.10: Questionnaire responses

Q1	How do you evaluate the usefulness of the dashboard for monitoring the performance of the predictive model and to evaluate the effectiveness of the developed investment strategies, on a scale of 1 (very dissatisfied) to 10 (very satisfied)?
R1	9. The dashboard has some small things that make it difficult to understand the model. In some analysis the informations is overlapping. However, it is a great dashboard that allows the investors to make quick desicions.
R2	8. The dashboard unquestionably offers value but has further potential for improvement. It already serves as an excellent tool for users, and it could be further enhanced if it provided more comprehensive information and data.
R3	8. The dashboard provides valuable insights and tools for monitoring the performance of the predictive model and evaluating investment strategies.
Q2	How would you rate the ease of use of this dashboard on a scale of 1 (not easy) to 10 (extremely easy)?
R1	10. Thanks to the slicers, this dashboard is very easy to use and navigate.
R2	9. The dashboard is intuitive, and the slicers significantly contribute to its user-friendliness.
R3	9. It is relatively easy to navigate and use, with an intuitive interface.
Q3	Do you think the indicators on this dashboard are adequate for monitoring the model financial perspective? Give your response on a scale of 1 (not appropriate) to 10 (very appropriate)?
R1	8. Yes but I would also like to see the ROI and maybe the best range of odds.
R2	8. I am not entirely sure if the indicators on this dashboard are adequate for monitoring the financial perspective and I would need more data and information to make an informed assessment.
R3	8. They are quite adequate for monitoring the financial perspective of the model, but there is potential for improvement with additional features and customization options.
Q4	Is the dashboard visual presentation appealing? Give your response on a scale of 1 (not appealing) to 10 (very appealing)?
R1	10. Yes I like it. It is super clean and simple.
R2	9. Certainly, it is an attractive dashboard, and if I were to suggest an improvement, it would involve including brief pop-up descriptions elucidating the specifics of each feature within the dashboard.
R3	9. It is clear and functional.
Q5	What is your overall rating of the dashboard on a scale of 1 (very dissatisfied) to 10 (very satisfied)?
R1	9. This dashboard is simple and easy to use, which is something important to me. The design is also very simple and clean, something I also like.
R2	9. The most important aspect of this dashboard is that able to provide clear and relevant information in a visually appealing manner.
R3	8. The dashboard simplicity of use and visual presentation capabilities are vital, since they make the dashboard accessible and appealing. Overall, it is a valuable tool.
Q6	What is your overall satisfaction of the dashboard?
R1	9. Overall this is a really simple and objective dashboard, wich allows me to make better decisons.
R2	9. In general, this dashboard is very straightforward, enabling me to improve my decision-making capabilities.
R3	9. It provides valuable insights, but I believe it can become even more useful with some refinements and additional features.

Therefore, based on the responses provided by the potential users regarding the dashboard for monitoring the financial performance of the predictive model and evaluating the effectiveness of the developed investment strategies, it is clear that the dashboard is a useful and easy tool to use. However, all potential users have suggested improvements for the dashboard, indicating that there is still room for enhancement.

Finally, it is worth mentioning that the evaluation of the dashboard by three potential users achieved an average rating higher than the previously defined success criterion. Consequently, the objective was successfully met with an overall score of 9 points (exceeding the 8-point goal).

4.5. Discussion and practical implications

The C5 algorithm has proven to be effective in predictions, providing consistent and promising results in terms of accuracy, precision, recall, and F1 Score. It was based on data from the major European leagues (EPL, Spanish La Liga, Italian Serie A, German Bundesliga, and Portuguese Liga Bwin) and utilized 84 predictors classified into dimensions such as past behavior, standings variables, game variables, team characteristics, and rating variables. The most important dimensions were past behavior: average and the standings variables.

As seen in Table 4.11, all four success criteria, specifically, a minimum accuracy, precision, recall, and F1 Score of 80%, were met. Only the recall for draws fell slightly below 80% (approximately 3 percentage points lower).

Table 4.11: Best algorithm

C5			
Metrics	H	D	A
Accuracy		85.40%	
Precision	85.73%	85.60%	84.77%
Recall	91.03%	76.73%	85.27%
F1 Score	88.30%	80.92%	85.01%

Note: H: Home; D: Draw; A: Away

For the second analytical objective of creating a dashboard and achieving an average evaluation score higher than 8 on a scale from 1 (very dissatisfied) to 10 (very satisfied), this goal was also successfully accomplished, with the dashboard receiving a rating of 9 points. The main advantage of this dashboard is its dynamic capability to perform various types of analysis and draw multiple distinct conclusions. This approach makes it easier to track different investment strategies.

Compared to existing literature, the model developed in this study presents remarkable results. In contrast to models in the literature that rely solely on information available before the match, this model stands out by not depending on variables obtained after the match. This distinguishes the

model, as it enables predictions to be made before the matches take place, unlike many models (e.g., Musa et al., 2019; Rosli et al., 2018) that can only make predictions after match have occurred.

The model performs reasonably well in terms of accuracy (47% on new data). Since these results were applied to a new data sample, it is normal for the results to be lower than the training and testing results, as predictions are being made based on a model trained with data from the three previous seasons, consequently, it may be necessary to retrain the model to achieve better results over time. Furthermore, this model is dynamic, meaning that the data from the last three seasons should always be taken into account when predicting the upcoming season.

Therefore, this study demonstrates the viability of using pre-game variables to predict football match outcomes with an accuracy of 85.40% (on test data). Having access to a model with an 85.40% accuracy clearly provides a competitive advantage over competitors. This level of accuracy offers greater confidence in predictions and helps in identifying patterns, trends, and potential surprises in future matches.

One of the major limitations of the literature is the lack of methodological clarity, which makes replicating these methodologies quite challenging. Many variables are not fully detailed (e.g., the variable for goals scored by both teams does not specify if it is calculated as an average of previous matches, a sum, or the goals that occurred in that specific match).

Additionally, the model does not consider teams or leagues as predictor variables, making it applicable to different seasons and leagues. This flexibility differentiates it from much of the literature, as approximately 37.5% (n=6) of models in the literature incorporate teams as predictor variables.

Based on the literature, there does not seem to be a consensus regarding predictor variables. In this research, we aimed to identify various dimensions of possible important variables. As outlined in the methodology, several dimensions of variables were created (past behavior, standings variables, game variables, team characteristics, and rating variables), with the first dimension (past behavior: average) being the most important, followed by the standing variables dimension. This research highlights the importance of historical averages (Past behavior: average) in building effective predictive model.

While there is consensus in the literature regarding the league under consideration, as Yu (2022), Nivetha et al. (2022), Azeman et al. (2021), Tiwari et al. (2020), Azeman et al. (2020), Baboota and Kaur (2019), Musa et al. (2019), Rosli et al. (2018), and Razali et al. (2017) focus their studies on the EPL, this research is more ample and aims to predict results for the EPL, Spanish La Liga, Italian Serie A, German Bundesliga, and Portuguese Liga Bwin.

Berrar et al. (2019) make a valuable contribution by arguing that the key to success lies in domain knowledge in the data pre-processing process and that the inclusion of rating variables improves model performance. Based on these contributions from the authors, it can be concluded that rating

variables are good predictor variables. In the case of this research, rating variables did not prove to be important in the model, likely because the calculation method differs from their study. However, a crucial aspect for the success of this study was the careful selection of predictor variables and the time invested in understanding and pre-processing the data, as suggested by Berrar et al. (2019).

When it comes to validation methods, the literature typically uses only one method, while this study adopts two different methods: the hold-out 70/30 and cross-validation. Using these two methods results in a more robust model construction. In this case, 70% of the sample was used to train the model and 30% for testing, and the C5 model also has the option to apply another validation method, so a 10-fold cross-validation method was also adopted. Regarding evaluation metrics, like 93.75% (n=15) of the literature, it is used the accuracy as the main evaluation metric. Additionally, this study also considers the results of evaluation metrics such as precision, recall, and F1 score (e.g., Nivetha et al., 2022; Baboota & Kaur, 2019).

In this research, there is a need to improve precision in draw predictions because there is high variance (Figure 4.8) and precision below 30% (in new data). This value is low due to the difficulty in predicting drawn matches because they are relatively less common compared to wins or losses, making them more difficult to predict. Besides some of the variables that can influence a draw may not be well represented in the available data used to train the model (e.g., emotional, or psychological factors of players and coaches can be difficult to quantify). So, understanding these metrics is crucial to identify strengths and areas with potential for improvement in the model to improve predictions and provide valuable insights for investors and football clubs.

To determine whether this model could be a beneficial tool to assist football clubs, specifically coaching staff, a phone call was made to a professional at the football club “Os Belenenses”. After the call, it was concluded that, although it is a valuable tool for predictions, it is considered that players can be psychologically influenced and therefore affect their performance. However, it is believed that at the investor level, the model has significant potential and can clearly be an asset for making money.

The strategic approach of investing in specific moments of the season (Table 4.8) can prove advantageous for investors, as it allows them to capitalize their profits during periods of higher model accuracy and minimize risks during periods of lower accuracy. It is also concluded that there are interesting investment opportunities in different leagues (Figure 4.12), and therefore, this diversification of leagues can be an asset from the investors perspective, as it allows them to explore different markets and investment opportunities based on the specific performance of each team and the characteristics of different leagues.

5. CONCLUSIONS

5.1. Summary

This study is focused on predicting football match results using ML techniques and the research question that guided this study was: Is it possible to predict the outcomes of football matches?

To answer this question, four main objectives were established. The first one, O1) To create a predictive model for predicting match outcomes (Home team wins, draw, away team wins), resulted in a successful construction of a predictive model, using the C5 algorithm, that showed promise results in its ability to predict match outcomes (quality metrics exceeding 80%).

For the second objective, O2) To identify the most important predictors of a match result, it was also successfully achieved. Still using the C5 algorithm, it was identified that the most important variable in explaining a result was the average goals conceded by the away team in their last 5 matches, followed by the position zone of the home team in the league table. These variables belong to the sets of past behavior: average and the standings variables, which were also found to be the most important.

Regarding the third objective, O3) To identify profiles of matches where the home team wins, draw, or away team wins it was identified two profiles with predictions for all 3 types of results (six profiles in total), using terminal nodes of the C5 tree with high confidence (greater than or equal to 75% for H and A predictions, and greater than or equal to 50% for D predictions), and with a support of 10 or more matches. This objective was also successfully accomplished.

Finally, the fourth objective, O4) To develop a dashboard to track the financial performance of the predictive model and to evaluate the effectiveness of the developed investment strategies, resulted in an effective tool for monitoring the progress of predictions as well as conducting a financial perspective analysis of football predictions, exploring the profits and losses associated with investment strategies. This objective was successful and received overall satisfaction from potential users of 9 points (in 10 possible).

Having achieved the objectives, it is possible to answer the research question. Therefore, the results show that it is possible to predict the outcomes of football matches using decision trees, specifically the C5 algorithm, which has the advantage of explaining each result prediction based on propositional rules. This prediction is based on a history of 5254 matches from the five European leagues (EPL, Spanish La Liga, Italian Serie A, German Bundesliga, and Portuguese Liga Bwin) and considers as predictor variables the following dimensions: past behavior, standings variables, game variables, team characteristics, and rating variables.

Furthermore, this research identified the most effective analytical techniques, explored the financial aspect of predictions, and provided a monitoring tool for continuous tracking (dashboard).

The results demonstrate that predictions can be made not only in terms of outcomes but also in terms of financial gains associated with investments.

Given these findings, the importance of football predictions in the realms of investment, sports betting, and fan enthusiasm is underscored.

5.2. Contributions

This research represents advancements in the study of football result predictions through analytics techniques, both in the academic dominion and for professionals. In the literature, no studies were identified that employed the CRISP-DM methodology. Therefore, this represents a novel approach in the sports analytics literature, demonstrating that this methodology is also effective within the context of this study.

Having a model that does not consider the league or team names makes it possible to make predictions for any match regardless of the league or division, which is one of the main differentiators compared to the literature. Furthermore, in comparison to some previous studies, this model stands out for its ability to predict matches in the early rounds. These two contributions are significant for scientific knowledge.

For the scientific field of business analytics, this study contributes by illustrating an application of predictive models where the evaluation is not only based on statistical metrics but also on a financial approach, which is of interest to various stakeholders, especially investors.

It also contributes to the analytics field by providing another application of predictive analysis to the sports domain, demonstrating the predictive capability of decision trees and, to some extent, proving that over time, models lose predictive ability and need to be retrained. In the analytics field, it is worth noting that the study currently requires manual weekly updates by the user. Therefore, this model is a prototype, and the goal is to create a pipeline, which means automating the processes. This pipeline will allow for the extraction of data from websites, calculation, and updating of all predictors with new data, and finally, running the model to obtain the predictions. This would enable investors to make their investments each week with the latest forecasts.

From a practical and professional perspective, this study has significant contributions. Investors obtain insights that it is possible to predict football results with high confidence and, therefore, generate profits from their investment activities. Additionally, they can use the model to make more informed investment decisions. By doing so, they can increase their profit potential by identifying odds with higher value (i.e., odds that offer higher returns relative to the assumed risk) which can lead to more favourable financial performance.

With an informed and disciplined approach, the model is a valuable tool for investors seeking a tested method to assist their decision making in order to maximize their returns on football investments. With the proper application of suitable strategies, investors can benefit from more accurate predictions with a higher likelihood of success.

Bookmakers also gain insights from this study, particularly regarding good predictors of results. They can incorporate these predictors into their odds calculation models if they are not already included. They can also use the information about the model predictions to adjust the odds and better manage their risk, which can help avoid substantial losses and maintain the integrity of their operations. Furthermore, by understanding the model predictions, bookmakers can optimize their betting lines, making them more attractive to investors and, consequently, increasing the betting volume.

From the perspective of the public, the model can contribute to managing expectations regarding their favourite team matches. If the model predicts a victory for their team, fans can attend the match with motivation and optimism. However, in the event of a prediction for a loss or a draw, fans should manage their expectations and recognize that their team might not win. This approach enables fans to approach the outcomes more realistically and enjoy the sporting experience to the fullest.

For football clubs, the model can provide valuable insights for better resource management. By obtaining predictions for future matches, the team coach gains an understanding of the odds being in their favour. Therefore, if the team continues with their current approach, there is a higher likelihood of winning. Conversely, if the model predicts a higher probability of defeat, the coach can adjust their strategies and make the most of the limited opportunities they might have in the match. This adaptability can lead to more informed decision making and potentially improved performance on the field.

In the general context, this research makes a significant contribution to the field of sports result prediction through analytical techniques, demonstrating the potential of these approaches to support the decision making. Additionally, the dashboard provides a convenient platform for users to analyse the financial perspective of the predictions. This innovative element adds a practical and useful component to the study, further enriching its applicability.

In summary, the use of pre-game variables and a high accuracy model to predict football match outcomes represents a significant advancement in the field of football predictions. The existing literature predominantly focuses on models using in-game variables. This approach offers valuable insights in various contexts, from clubs and betting houses to investors and fans themselves, providing better expectation management and a solid foundation for decision making for potential investors.

5.3. Limitations

Despite the excellent results, it is important to acknowledge some limitations of the research. It is relatively important to highlight that, although a model with 85.40% accuracy is considered promising, it is crucial to understand that football is full of uncertainties due to numerous uncontrollable variables, making predictions challenging. Some of the factors that can influence results include player injuries, last minute tactical changes, player dismissals, or unexpected events during matches. These kinds of variables were not available and, therefore, the results may be impacted.

So, it is essential to interpret predictions as a complementary tool capable of providing valuable insights but not guaranteeing correct outcomes in all situations. It is relevant to mention that the model mainly considered quantitative variables, leaving out other important factors. For example, factors like the presence (or absence) of key players, the history of coach dismissals in previous matches, or the sale of key players can have a significant impact on results and were not considered in this study due to difficulties in obtaining such data.

While it is possible to adapt this model to different leagues, it is necessary to consider that each league has its unique characteristics, such as playing style or level of competitiveness. Therefore, caution should be exercised when extending the conclusions from this study to leagues outside of Europe (e.g., because matches have significantly different characteristics). In other words, the geographically limited sample considered in the study might partially restrict the generalizability of the results.

5.4. Future research recommendations

In terms of future recommendations, conducting research to predict halftime results would be a significant advancement for the literature. Furthermore, combining this future research with the model used in this study would be a unique contribution to the literature. This approach could bring a financial perspective focused on the first half of the match, allowing investors to adapt their investments (if necessary) for the final prediction at halftime.

Predicting the outcome of a match is already a challenging task, but a new suggestion for future research is to try predicting the result considering the goal difference. From an investor perspective, obtaining a prediction for a match where a team will win by two or more goals provides reassurance to the investor, knowing it will be an easier match and, therefore, placing more confidence in that prediction.

Furthermore, it is recommended that the methodology used in this study be applied to other leagues or that the inclusion of qualitative variables be considered to see if the predictive capability of

the model can be improved. This expansion would allow for a more comprehensive range of influences on match results to be captured, potentially providing more comprehensive and reliable predictions and additional insights.

REFERENCES

- Angelini, G. & De Angelis, L. (2018). Efficiency of online football betting markets. *International Journal of Forecasting*, 35(2), 1-23. <https://doi.org/10.1016/j.ijforecast.2018.07.008>
- Aslan, B., & Inceoglu, M. (2012). *A Comparative Study on Neural Network-Based Soccer Result Prediction*. Seventh International Conference on Intelligent Systems Design and Applications, India. <https://doi.org/10.1109/ISDA.2007.12>
- Azeman, A. A., Mustapha, A., Mostafa, S. A., AbuSalim, S. W.G., Jubair, M. A., & Hassan, M. H. (2020). Football match outcome prediction by applying three machine learning algorithms. *International Journal of Emerging Trends in Engineering Research*, 8(1), 73-77. <https://doi.org/10.30534/ijeter/2020/1181.12020>
- Azeman, A. A., Mustapha, A., Razali, N., Nanthamomphong, A., & Abd Wahab, M. H. (2021, may 19-22). *Prediction of football matches results: Decision forest against neural networks*. ECTI-CON 2021 - 2021 18th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology: Smart Electrical System and Technology, Proceedings, 1032-1035. <https://doi.org/10.1109/ECTI-CON51831.2021.9454789>
- Baboota, R., & Kaur, H. (2019). Predictive analysis and modeling football results using machine learning approach for English Premier League. *International Journal of Forecasting*, 35(2), 741-755. <https://doi.org/10.1016/j.ijforecast.2018.01.003>
- Berrar, D., Lopes, P., & Dubitzky, W. (2019). Incorporating domain knowledge in machine learning for soccer outcome prediction. *Machine Learning*, 108(1), 97-126. <https://doi.org/10.1007/s10994-018-5747-8>
- Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.
- Breiman, L. Bagging Predictors. (1996). *Machine Learning*, 24(2), 123-140.
- Caldas, R., Mundt, M., Potthast W., Neto, F., & Markert, B. (2017). A systematic review of gait analysis methods based on inertial sensors and adaptive algorithms. *Gait & Posture*, 57, 204-210. <https://doi.org/10.1016/j.gaitpost.2017.06.019>
- Carlioni, L., De Angelis, A., Sansonetti, G., & Micarelli, A. (2021). A Machine Learning Approach to Football Match Result Prediction. *Communications in Computer and Information Science*, 473-480. https://doi.org/10.1007/978-3-030-78642-7_63
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (1999). CRISP-DM 1.0 Step-by-step data mining guide. *The CRISP-DM Consortium*.
- Chen, T., & Guestrin, C. (2016, August 13-17). *XGBoost: A Scalable Tree Boosting System*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA.
- Cheng, T., Cui, D., Fan, Z., Zhou, J., & Lu, S. (2003, september 27-30). *A new model to forecast the results of matches based on hybrid neural networks in the soccer rating system*. Proceedings of the 5th International Conference on Computational Intelligence and Multimedia Applications, ICCIMA 2003, 308-313. <https://doi.org/10.1109/ICCIMA.2003.1238143>
- Deo, R. C. (2015). Machine learning in medicine. *Circulation*, 132(20), 1920–1930. <https://doi.org/10.1161/CIRCULATIONAHA.115.001593>
- Freund, Y. & Schapire, R. (1996. July 3-6) *Experiments with a new boosting algorithm*. Machine Learning: Proceedings of the Thirteenth International Conference, Italy.
- Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, 29(5), 1189-1232. <https://doi.org/10.1214/aos/1013203451>
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
- Guan, S., & Wang, X. (2022). Optimization Analysis Of Football Match Prediction Model Based On Neural Network. *Neural Computing and Applications*, 34(4), 2525-2541. <https://doi.org/10.1007/s00521-021-05930-x>
- Hsu, C. W., Chang, C. C., & Lin, C. J. (2010). A practical guide to support vector classification. *National Taiwan University*.

- Huang, K., & Chang, W. (2010, July 18-23). *A neural network method for prediction of 2006 World Cup Football Game*. Proceedings of the International Joint Conference on Neural Networks. doi:10.1109/IJCNN.2010.5596458
- Jolliffe, I. T., & Stephenson, D. B. (2003). *Forecast Verification. A Practitioner's Guide in Atmospheric Science*. West Sussex, England: Wiley.
- Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 29(2), 119-127.
- Koroteyeva, T., Tushnytskyy, R., & Tushnytskyy, V. (2018, september 11-14). *Applying Neural Networks to Football Matches Results Forecasting*. In Proceedings of the International Computer Sciences and Information Technologies. <https://doi.org/10.1109/STC-CSIT.2018.8526756>
- Laureano, R., Caetano, N. & Cortez, P. (2014). Previsão de tempos de internamento num hospital português: Aplicação da metodologia CRISP-DM. *RISTI - Revista Ibérica de Sistemas e Tecnologias de Informação*, 13, 83-98. <https://doi.org/10.4304/risti.13.83-98>
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444.
- Liaw, A., & Wiener, M. (2002). Classification and regression by random Forest. *R news*, 2(3), 18-22.
- Mantas, C. J. & Abellán, J. (2014). Credal-C4.5: Decision tree based on imprecise probabilities to classify noisy data. *Expert Systems with Applications*, 41(10), 4625-4637. <https://doi.org/10.1016/j.eswa.2014.01.017>
- Mohandas, A., Ahsan, M. & Haider, J. (2023). Tactically Maximize Game Advantage by Predicting Football Substitutions Using Machine Learning. *Big data and cognitive computing*, 7(2), 117. <https://doi.org/10.3390/bdcc7020117>
- Murphy, A. (1970). The ranked probability score and the probability score: A comparison. *Monthly Weather Review*, 98, 917-924. [https://doi.org/10.1175/1520-0493\(1970\)098<0917:TRPSAT>2.3.CO;2](https://doi.org/10.1175/1520-0493(1970)098<0917:TRPSAT>2.3.CO;2)
- Musa, R., Majeed, A., Razman, M., & Shaharudin, M. (2019, november 11-19). *Match outcomes prediction of six top English Premier League clubs via machine learning technique*. In Proceedings of the 7th International Conference on Robot Intelligence Technology and Applications, Korea. https://doi.org/10.1007/978-981-13-7780-8_20
- Nivetha, S., Geetha, M., Suganthe, R., Prabakaran, R., Madhuvanan, S., & Sameer, A. (2022, january 25-27). *A Deep Learning Framework for Football Match Prediction*. In Proceedings of the International Conference on Computer Communication and Informatics, India, 1-7. <https://doi.org/10.1109/ICCCI54379.2022.9740760>
- OpenAI. (2023). ChatGPT: A Large-Scale Generative Model for Natural Language Processing [Computer software]. Retrieved from <https://openai.com/research/chatgpt>
- Pearl, J., & Russell, S. (2003). The Handbook of Brain Theory and Neural Networks, 2157-160.
- Polat, K. & Gunes, S. (2009). A novel hybrid intelligent method based on C4.5 decision tree classifier and one-against-all approach for multi-class classification problems. *Expert Systems with Applications*, 36(2), 1587-1592. <https://doi.org/10.1016/j.eswa.2007.11.051>
- Prasetio, D., & Harlili, M. (2016, august 17-18). *Predicting football match results with logistic regression*. In Proceedings of the 4th IGNITE Conference and 2016 International Conference on Advanced Informatics: Concepts, Theory and Application, Indonesia, 1-5. <https://doi.org/10.1109/ICAICTA.2016.7803111>
- Quinlan, J. (1986). Induction of decision trees. *Machine Learning*, 1(1), 81-106.
- Quinlan, J. R. (1993). C4.5: Programs for machine learning. *Machine Learning*, 16, 235-240.
- Quinlan, J. R. (1996, august 4-8). *Bagging, Boosting, and C4.5*. In Proceedings of the thirteenth national conference on Artificial intelligence, Portland Oregon.
- Ray, A., & Chaudhuri, A. (2021). Smart healthcare disease diagnosis and patient management: Innovation, improvement and skill development. *Machine Learning with Applications*, 3, 1-24 <https://doi.org/10.1016/j.mlwa.2020.100011>
- Razali, N., Mustapha, A., Yatim, F., & Aziz, R. (2017, may 6-7). *Predicting football matches results using Bayesian networks for English premier league (EPL)*. International Research and Innovation Summit, Malaysia. <https://doi.org/10.1088/1757-899X/226/1/012099>

- Rebelo, P. (2012). *Ganhar com as apostas desportivas*. Lisboa: Marcador.
- Rosli, C., Saringat, M., Razali, N., & Mustapha, A. (2017, november 25-27). *A Comparative Study of Data Mining Techniques on Football Match Prediction*. In Proceedings of the Journal of Physics: Conference Series, 1st International Conference on Computing, Technology, Science and Management in Sports, Malaysia. <https://doi.org/10.1088/1742-6596/1020/1/012003>
- Schröer, C., Kruse, F., & Marx Gómez, J. (2020). *A Systematic Literature Review on Applying CRISP-DM Process Model*. In Proceedings of the CENTERIS - International Conference on Enterprise Information Systems / Proman - International Conference on Project Management / HCist - International Conference on Health and Social Care Information Systems and Technologies. <https://doi.org/10.1016/j.procs.2021.01.199>
- Seni, G., & Elder, J. (2010). *Synthesis Lectures on Data Mining and Knowledge Discovery*.
- Shin, J., & Gasparyan, R. (2014). *A Novel Way to Soccer Match Prediction*. *Stanford University: Department of Computer Science*, 1-5.
- Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427-437. <https://doi.org/10.1016/j.ipm.2009.03.002>
- Tiwari, E., Sardar, P., & Jain, S. (2020, june 4-5). *Football Match Result Prediction Using Neural Networks and Deep Learning*. In Proceedings of the 8th International Conference on Reliability, Infocom Technologies and Optimization, India. <https://doi.org/10.1109/ICRITO48877.2020.9197811>
- Valle, P. O., Pintassilgo, P., Matias, A. & André, F. (2012). Tourist attitudes towards an accommodation tax earmarked for environmental protection: A survey in the Algarve. *Tourism Management*, 33(6), 1408-1416. <https://doi.org/10.1016/j.tourman.2012.01.003>
- Vapnik, V. (2013). *The nature of statistical learning theory*. Springer Science & Business Media.
- Vilelas, J. (2017). *Investigação: O Processo de Construção do Conhecimento* (2nd ed.). Lisboa: Edições Sílabo.
- West, J., & Bhattacharya, M. (2016). Experimental issues in financial fraud mining. *Procedia Computer Science*, 80, 1734-1744. <https://doi.org/10.1016/j.procs.2016.05.515>
- Yu, W. (2022). Football Result Prediction Based on Machine Learning. *Advances in Transdisciplinary Engineering*, 20, 787-791. <https://doi.org/10.3233/ATDE220080>

APPENDIX

A: Variables extracted and excluded for analysis

Football Data	CgmBet	SoFifa
HTeamOpeningOdd, XTeamOpeningOdd, ATeamOpeningOdd, Over2,5OpeningOdd, Under2,5OpeningOdd, HTeamClosingOdd, XTeamClosingOdd, ATeamClosingOdd, Over2,5ClosingOdd, Under2,5ClosingOdd	HTeamTotalShotsFirstPart, HTeamTotalShotsSecondPart, ATeamTotalShotsFirstPart, ATeamTotalShotsSecondPart, HTeamTotalShotsOnTargetFirstPart, HTeamTotalShotsOnTargetSecondPart, ATeamTotalShotsOnTargetFirstPart, ATeamTotalShotsOnTargetSecondPart, HTeamTotalCornersFirstPart, HTeamTotalCornersSecondPart, ATeamTotalCornersFirstPart, ATeamTotalCornersSecondPart, HTeamTotalFaultsFirstPart, HTeamTotalFaultsSecondPart, ATeamTotalFaultsFirstPart, ATeamTotalFaultsSecondPart, HTeamTotalYellowCardsFirstPart, HTeamTotalYellowCardsSecondPart, ATeamTotalYellowCardsFirstPart, ATeamTotalYellowCardsSecondPart, HTeamTotalBallPossessionFirstPart, HTeamTotalBallPossessionSecondPart, ATeamTotalBallPossessionFirstPart, ATeamTotalBallPossessionSecondPart, HTeamSecondGoalMoment, HTeamThirdGoalMoment, HTeamFourthGoalMoment, HTeamFifthGoalMoment, HTeamSixthGoalMoment, HTeamSeventhGoalMoment, HTeamEighthGoalMoment, HTeamNinthGoalMoment, HTeamTenthGoalMoment, ATeamSecondGoalMoment, ATeamThirdGoalMoment, ATeamFourthGoalMoment, ATeamFifthGoalMoment, HTeamELORating, ATeamELORating, HTeamForm, ATeamForm, HTeamOdd, XTeamOdd, ATeamOdd, Over2,5Odd, Under2,5Odd	HTeamTransferBudget, ATeamTransferBudget, Players

B: Data dictionary

Past behavior variables – A

Variable	Description	Count Distinc	Min	Q1	Q2	Q3	Max	M	SD	Unit of measure	Role
Last 5 games											
AveHTeamGoalsScoredLast5Games	Ave goals scored by the home team in the last 5 games	105	0	1,00	1,40	2,00	5,20	1,51	0,74	Goals	Input
AveHTeamGoalsConcededLast5Games	Ave of goals conceded by the home team in the last 5 games	96	0	0,80	1,20	1,60	5,00	1,27	0,61		
AveATeamGoalsScoredLast5Games	Ave goals scored by the away team in the last 5 games	97	0	0,80	1,20	1,60	5,00	1,28	0,67		
AveATeamGoalsConcededLast5Games	Ave goals conceded by the away team in the last 5 games	93	0	1,00	1,40	1,80	5,00	1,49	0,67		
AveHTeamShotsLast5Games	Ave shots for home team in last 5 games	270	0	11	12,60	14,80	31,00	12,82	3,19	Shots	
AveATeamShotsLast5Games	Ave shots for away team in last 5 games	249	0	9	10,60	12,60	23,00	10,91	2,88		
AveHTeamShotsOnTargetLast5Games	Ave shots on goal for home team in last 5 games	169	0	4	4,60	5,60	14,00	4,75	1,64		
AveATeamShotsOnTargetLast5Games	Ave shots on goal for away team in last 5 games	166	0	3	4,00	5,00	14,00	4,09	1,46		
AveHTeamCornersLast5Games	Ave corners for home team in last 5 games	183	0	4	5,20	6,20	14,00	5,30	1,63	Corners	
AveATeamCornersLast5Games	Ave away team corners in last 5 games	151	0	4	4,40	5,40	10,60	4,54	1,39		
AveHTeamFoultLast5Games	Ave fouls for the home team in the last 5 games	240	0	11	13,00	14,80	25,00	12,99	2,80	Foultts	
AveATeamFoultLast5Games	Ave away team fouls in the last 5 games	237	0	11	13,00	14,80	25,33	13,02	2,74		
AveHTeamYellowCardsLast5Games	Ave of yellow cards for the home team in the last 5 games	117	0	2	2,00	2,60	7,00	2,11	0,79	Cards	
AveATeamYellowCardsLast5Games	Ave of yellow cards for the away team in the last 5 games	115	0	2	2,20	2,80	7,00	2,28	0,82		
AveHTeamRedsCardsLast5Games	Ave of reds cards for the home team in the last 5 games	33	0	0	0,00	0,20	1,00	0,09	0,14		
AveATeamRedsCardsLast5Games	Ave of reds cards for the away team in the last 5 games	33	0	0	0,00	0,20	2,00	0,11	0,17		
AveHTeamBallPossessionLast5Games	Ave possession of the ball by the home team in the last 5 games	435	0	46	50,80	56,00	78,00	51,00	7,72	%	
AveATeamBallPossessionLast5Games	Ave possession of the ball by the away team in the last 5 games	428	0	44	48,60	53,80	77,00	48,84	7,80		
AveHTeamFirstGoalMomentLast5Games	Ave moment of the 1st goal scored by the home team in the last 5 games	716	1	30	38,50	48,00	90,00	39,12	14,20	Minutes	
AveATeamFirstGoalMomentLast5Games	Ave moment of the 1st goal scored by the away team in the last 5 games	739	0	30	40,00	50,40	90,00	40,46	15,79		

Past behavior variables – B

Variable	Description	Count Distinc	Min	Q1	Q2	Q3	Max	M	SD	Unit of measure	Role
Last 5 games											
SDHTeamGoalsScoredLast5Games	SD goals scored by the home team in the last 5 games	192	0	1	0,89	1,20	3,19	0,95	0,45	Goals	Input
SDHTeamGoalsConcededLast5Games	SD of goals conceded by the home team in the last 5 games	183	0	1	0,87	1,17	3,12	0,92	0,44		
SDATeamGoalsScoredLast5Games	SD goals scored by the away team in the last 5 games	182	0	1	0,80	1,17	3,26	0,89	0,42		
SDATeamGoalsConcededLast5Games	SD goals conceded by the away team in the last 5 games	191	0	0,75	0,98	1,20	3,30	0,99	0,46		
SDHTeamShotsLast5Games	SD shots for home team in last 5 games	680	0	2,61	3,66	4,76	9,77	3,72	1,64	Shots	
SDATeamShotsLast5Games	SD shots for away team in last 5 games	634	0	2,42	3,38	4,32	9,52	3,42	1,53		
SDHTeamShotsOnTargetLast5Games	SD shots on goal for home team in last 5 games	336	0	1,33	1,85	2,42	5,64	1,90	0,87		
SDATeamShotsOnTargetLast5Games	SD shots on goal for away team in last 5 games	310	0	1,20	1,67	2,23	5,60	1,76	0,82		
SDHTeamCornersLast5Games	SD corners for home team in last 5 games	389	0	1,60	2,19	2,87	6,50	2,25	0,98	Corners	
SDATeamCornersLast5Games	SD away team corners in last 5 games	359	0	1,47	2,06	2,65	6,34	2,11	0,92		
SDHTeamFoulsLast5Games	SD fouls for the home team in the last 5 games	524	0	2,14	2,93	3,74	8,42	2,96	1,28	Fouls	
SDATeamFoulsLast5Games	SD away team fouls in the last 5 games	541	0	2,23	3,01	3,87	8,00	3,06	1,30		
SDHTeamYellowCardsLast5Games	SD of yellow cards for the home team in the last 5 games	184	0	0,80	1,02	1,36	3,43	1,09	0,47	Cards	
SDATeamYellowCardsLast5Games	SD of yellow cards for the away team in the last 5 games	194	0	0,75	1,10	1,41	3,00	1,10	0,47		
SDHTeamRedsCardsLast5Games	SD of reds cards for the home team in the last 5 games	42	0	0,00	0,00	0,40	1,20	0,15	0,22		
SDATeamRedsCardsLast5Games	SD of reds cards for the away team in the last 5 games	43	0	0,00	0,00	0,40	1,20	0,18	0,23		
SDHTeamBallPossessionLast5Games	SD possession of the ball by the home team in the last 5 games	1 564	0	5,18	7,12	9,37	27,07	7,37	3,21	%	
SDATeamBallPossessionLast5Games	SD possession of the ball by the away team in the last 5 games	1 503	0	5,25	7,08	9,21	28,39	7,27	3,12		
SDHTeamFirstGoalMomentLast5Games	SD moment of the 1st goal scored by the home team in the last 5 games	3 454	0	12,82	19,64	25,68	42,50	18,94	9,10	Goals	
SDATeamFirstGoalMomentLast5Games	SD moment of the 1st goal scored by the away team in the last 5 games	3 008	0	11,53	19,03	25,79	44,00	18,29	9,77		

Past behavior variables – C

Variable	Description	Count Distinc	Min	Q1	Q2	Q3	Max	M	SD	Unit of measure	Role
ToDate											
AveHTeamGoalScored	Ave goals scored by the home team	312	0	1,00	1,35	1,85	8,00	1,43	0,82	Goals	Input
AveHTeamGoalConceded	Ave goals conceded by the home team	247	0	0,88	1,20	1,56	5,00	1,22	0,64		
AveATeamGoalScored	Ave goals scored by away team	269	0	0,80	1,13	1,60	5,50	1,21	0,70		
AveATeamGoalConceded	Ave goals conceded by away team	272	0	1,00	1,43	1,83	8,00	1,42	0,74		

Standings variables – A

Variable	Description	Count Distinct	Min	Q1	Q2	Q3	Max	M	SD	Unit of measure	Role
HTeamGeneralStandings	Home team standings	20	1	5,00	10,00	15,00	20	10,21	5,60	Rank	
HTeamGeneralStandingsAtH	Home team standings at home	20	1	6,00	11,00	15,00	20	10,49	5,52		
ATeamGeneralStandings	Away team standings	20	1	5,00	10,00	15,00	20	10,09	5,59		
ATeamGeneralStandingsAtH	Away team standings at away	20	1	6,00	10,00	15,00	20	10,40	5,51		
HTeamRelativeStandings	Relative standings of the home team (the lower the number, the better classification)	36	5	27,78	55,00	77,78	100	52,91	28,91	%	Input
HTeamRelativeStandingsAtH	Relative standings of home team playing at home (the lower the number, the better classification)	36	5	30,00	55,00	80,00	100	54,39	28,45		
ATeamRelativeStandings	Relative standings of away team (the lower the number, the better classification)	36	5	27,78	50,00	77,78	100	52,26	28,82		
ATeamRelativeStandingsAtA	Relative standings of the away team playing away (the lower the number, the better classification)	36	5	30,00	55,00	80,00	100	53,90	28,41		

Standings variables – B

Variable	Description	Count Distinct	Unit of measure	Role
HTeamStandingsZone	Home team standings zone	CZ - 961 (18%)	Nominal	Input
		LEZ - 458 (9%)		
		MZ - 2627 (50%)		
		RZ - 742 (14%)		
		CPZ - 52 (1%)		
		ELPZ - 109 (2%)		
		CLPZ - 221 (4%)		
		DPZ - 84 (2%)		
ATeamStandingsZone	Away team standings zone	CZ - 991 (19%)	Nominal	Input
		LEZ - 449 (9%)		
		MZ - 2666 (51%)		
		RZ - 698 (13%)		
		CPZ - 50 (1%)		
		ELPZ - 105 (2%)		
		CLPZ - 211 (4%)		
		DPZ - 84 (2%)		

Note: CZ = Champions Zone, LEZ = Europa League Zone, MZ = Maintenance Zone, RZ = Relegation Zone, CPZ = Champions Playoff Zone, ELPZ = Europa League Playoff Zone
CLPZ = Conference League Playoff Zone, DPZ = Relegation Playoff Zone

Game variables – A

Variable	Description	Values	Measure	Unit of measure	Role
Season	Seasons under review	19/20 - 1750 (33%)		Nominal	Preparation
		20/21 - 1752 (33,5%)			
		21/22 - 1752 (33,5%)			
Game	Variable that contains the home team, the away team and the respective season	-		Typeless	None
Championship	Championships under review	Bundesliga - 918 (17%)			Input
		Serie A - 1140 (22%)			
		LaLiga - 1140 (22%)			
		Premier League - 1140 (22%)			
		Liga Bwin - 916 (17%)			
FT	Final result of the game	H - 2205 (42%)			Target
		D - 1332 (25%)			
		A - 1717 (33%)			
HT	Result at half-time	H - 1716 (33%)			None
		D - 2128 (40%)			
		A - 1410 (27%)			
Month	Month of the game	January - 586 (11%)	Qualitative	Nominal	
		February - 612 (12%)			
		March - 371 (7%)			
		April - 463 (9%)			
		May - 422 (8%)			
		June - 195 (4%)			
		July - 258 (5%)			
		August - 285 (5%)			
		September - 458 (9%)			
		October - 516 (10%)			
		November - 466 (9%)			
		December - 622 (11%)			
WeekDay	Week day of the game	Monday - 1947 (37%)			Input
		Tuesday - 1795 (34%)			
		Wednesday - 331 (6%)			
		Thursday - 228 (4%)			
		Friday - 404 (8%)			
		Saturday - 175 (3%)			
		Sunday - 374 (8%)			
MomentOfDay	Time of day when the encounter took place	Morning - 402 (8%)			
		Afternoon - 3752 (71%)			
		Night - 1100 (21%)			

Game variables – B

Variable	Description	Count Distinct	Min	Q1	Q2	Q3	Max	M	SD	Measure	Unit of measure	Role
Date	Date of the game	574	09/08/2019	11/06/2020	29/01/2021	25/10/2021	22/05/2022	19/01/2021	302,90	Quantitative	Day	Preparation
Hour	Time of the game	42	11:00	14:30	17:30	19:45	22:15	17:05	2:38		Hours	Input
Matchday	Championship Matchdays	38	1,00	10,00	19,00	28,00	38,00	18,81	10,62		Ranking	Number
HTeamDaysBetweenGames	Difference of days between the last home team game and the current game	78	2	5	7	8	391	10,19	17,38	Difference	Day	
ATeamDaysBetweenGames	Difference of days between the last away game and the current game	75	2	5	7	8	391	10,08	15,82			
HTeam	Home Team	119	-	-	-	-	-	-	-	Qualitative	Nominal	
ATeam	Away Team	119	-	-	-	-	-	-	-			
FTHTeamGoalsScored	Total goals scored by the home team	10	0,00	1,00	1,00	2,00	9,00	1,50	1,28		Goals	
FTATeamGoalsScored	Total goals scored by the away team	9	0,00	0,00	1,00	2,00	9,00	1,27	1,19			
HTHTeamGoalsScored	Total goals scored by the home team in the 1st half of the match	6	0,00	0,00	0,00	1,00	5,00	0,67	0,83			
HTATeamGoalsScored	Total goals scored by the away team in the 1st half of the match	8	0,00	0,00	0,00	1,00	7,00	0,57	0,78			
HTeamTotalShots	Total shots by home team	39	0,00	9,00	12,00	16,00	46,00	12,85	5,18			
HTeamTotalShotsFirstPart	Total shots by home team in 1st half	21	0,00	0,00	4,00	7,00	22,00	3,99	3,78	Quantitative		Preparation
HTeamTotalShotsSecondPart	Total shots by home team in 2nd half	27	0,00	0,00	4,00	8,00	33,00	4,63	4,36			
ATeamTotalShots	Total shots by away team	34	0,00	8,00	10,00	14,00	36,00	10,96	4,78			
ATeamTotalShotsFirstPart	Total shots by away team in 1st half	18	0,00	0,00	3,00	6,00	17,00	3,37	3,32			
ATeamTotalShotsSecondPart	Total shots by away team in 2nd half	23	0,00	0,00	4,00	7,00	24,00	4,02	3,93			
HTeamTotalShotsOnTarget	Total shots on goal by home team	18	0,00	3,00	4,00	6,00	17,00	4,75	2,65	Shots		
HTeamTotalShotsOnTargetFirstPart	Total shots on goal by home team in 1st half	11	0,00	0,00	1,00	2,00	12,00	1,43	1,66			
HTeamTotalShotsOnTargetSecondPart	Total shots on goal by home team in 2nd half	13	0,00	0,00	1,00	3,00	12,00	1,68	1,89			
ATeamTotalShotsOnTarget	Total shots on goal by away team	18	0,00	2,00	4,00	5,75	20,00	4,09	2,43			
ATeamTotalShotsOnTargetFirstPart	Total shots on goal by away team in 1st half	12	0,00	0,00	1,00	2,00	11,00	1,25	1,55			
ATeamTotalShotsOnTargetSecondPart	Total shots on goal by away team in 2nd half	12	0,00	0,00	1,00	2,00	11,00	1,45	1,70			

Game variables – C

Variable	Description	Count Distin	Min	Q1	Q2	Q3	Max	M	SD	Measure	Unit of measure	Role
HTeamTotalCorners	Total corners by home team	20	0,00	3,00	5,00	7,00	20,00	5,28	2,94	Quantitative	Corners	Preparation
HTeamTotalCornersFirstPart	Total corners by home team in 1st half	11	0,00	0,00	1,00	3,00	10,00	1,61	1,87			
HTeamTotalCornersSecondPart	Total corners by home team in 2nd half	13	0,00	0,00	1,00	3,00	15,00	1,81	2,06			
ATeamTotalCorners	Total corners by away team	18	0,00	3,00	4,00	6,00	17,00	4,54	2,70			
ATeamTotalCornersFirstPart	Total corners by away team in 1st half	11	0,00	0,00	1,00	2,00	10,00	1,40	1,69			
ATeamTotalCornersSecondPart	Total corners by away team in 2nd half	13	0,00	0,00	1,00	3,00	13,00	1,59	1,91			
HTeamTotalFaults	Total faults by home team	33	0,00	10,00	13,00	16,00	33,00	12,96	4,25		Faults	
HTeamTotalFaultsFirstPart	Total faults by home team in 1st half	18	0,00	0,00	4,00	7,00	20,00	4,04	3,52			
HTeamTotalFaultsSecondPart	Total faults by home team in 2nd half	20	0,00	0,00	5,00	8,00	19,00	4,56	3,95			
ATeamTotalFaults	Total faults by away team	30	1,00	10,00	13,00	16,00	30,00	12,99	4,29			
ATeamTotalFaultsFirstPart	Total faults by away team in 1st half	18	0,00	0,00	4,00	7,00	19,00	4,06	3,57			
ATeamTotalFaultsSecondPart	Total faults by away team in 2nd half	20	0,00	0,00	5,00	7,00	23,00	4,50	3,92			
HTeamTotalYellowCards	Total yellow cards by home team	11	0,00	1,00	2,00	3,00	10,00	2,12	1,44	YellowCards		
HTeamTotalYellowCardsFirstPart	Total yellow cards by home team in 1st half	7	0,00	0,00	0,00	1,00	6,00	0,56	0,82			
HTeamTotalYellowCardsSecondPart	Total yellow cards by home team in 2nd half	9	0,00	0,00	0,00	1,00	8,00	0,83	1,15			
ATeamTotalYellowCards	Total yellow cards by away team	9	0,00	1,00	2,00	3,00	8,00	2,27	1,46			
ATeamTotalYellowCardsFirstPart	Total yellow cards by away team in 1st half	6	0,00	0,00	0,00	1,00	5,00	0,59	0,82			
ATeamTotalYellowCardsSecondPart	Total yellow cards by away team in 2nd half	21	0,00	1,00	2,00	5,00	20,00	4,30	5,38			
HTeamTotalBallPossession	Total ball possession by home team	65	18,00	44,00	51,00	59,00	83,00	51,11	10,84	BallPossession		
HTeamTotalBallPossessionFirstPart	Total ball possession by home team in 1st half	70	0,00	0,00	43,00	55,00	86,00	34,06	25,85			
HTeamTotalBallPossessionSecondPart	Total ball possession by home team in 2nd half	78	0,00	0,00	42,00	55,00	95,00	34,05	26,14			
ATeamTotalBallPossession	Total ball possession by away team	65	17,00	41,00	49,00	56,00	82,00	48,89	10,84			
ATeamTotalBallPossessionFirstPart	Total ball possession by away team in 1st half	70	0,00	0,00	41,00	53,00	83,00	32,59	24,89			
ATeamTotalBallPossessionSecondPart	Total ball possession by away team in 2nd half	78	0,00	0,00	40,00	53,00	95,00	32,60	25,20			

Game variables – D

Variable	Description	Count Distinc	Min	Q1	Q2	Q3	Max	M	SD	Measure	Unit of measure	Role
HTeamFirstGoalMoment	Moment of the home team first goal	91	0,00	3,00	24,00	50,00	90,00	29,99	27,67	Quantitative	Minutes	Preparation
HTeamSecondGoalMoment	Moment of the home team second goal	89	0,00	0,00	0,00	52,00	90,00	24,78	32,05			
HTeamThirdGoalMoment	Moment of the home team third goal	77	0,00	0,00	0,00	0,00	90,00	12,27	26,96			
HTeamFourthGoalMoment	Moment of the home team fourth goal	61	0,00	0,00	0,00	0,00	90,00	5,38	19,30			
HTeamFifthGoalMoment	Moment of the home team fifth goal	43	0,00	0,00	0,00	0,00	90,00	2,07	12,53			
HTeamSixthGoalMoment	Moment of the home team sixth goal	26	0,00	0,00	0,00	0,00	90,00	0,53	6,39			
HTeamSeventhGoalMoment	Moment of the home team seventh goal	11	0,00	0,00	0,00	0,00	90,00	0,17	3,62			
HTeamEighthGoalMoment	Moment of the home team eighth goal	5	0,00	0,00	0,00	0,00	90,00	0,07	2,37			
HTeamNinthGoalMoment	Moment of the home team ninth goal	2	0,00	0,00	0,00	0,00	90,00	0,02	1,24			
HTeamTenthGoalMoment	Moment of the tenth goal of the home team	2	0,00	0,00	0,00	0,00	0,00	0,00	0,00			
ATeamFirstGoalMoment	Moment of the away team first goal	91	0,00	0,00	20,00	49,00	90,00	28,16	28,51			
ATeamSecondGoalMoment	Moment of the away team second goal	88	0,00	0,00	0,00	45,00	90,00	20,65	30,86			
ATeamThirdGoalMoment	Moment of the away team third goal	75	0,00	0,00	0,00	0,00	90,00	9,67	24,64			
ATeamFourthGoalMoment	Moment of the away team fourth goal	57	0,00	0,00	0,00	0,00	90,00	3,62	16,11			
ATeamFifthGoalMoment	Moment of the away team fifth goal	37	0,00	0,00	0,00	0,00	0,00	0,00	0,00			
HTeamTotalRedCards	Total red cards for the home team	4	0,00	0,00	0,00	0,00	3,00	0,09	0,30		RedsCards	
ATeamTotalRedCards	Total red cards for the away team	4	0,00	0,00	0,00	0,00	3,00	0,11	0,33			

Team characteristics

Variable	Description	Values	Measure	Unit of measure	Role
HTeamPeriodBetweenGames	Period of days between games from home team	4 days - 1157 (22%) 1 week - 2973 (56%) 2 weeks - 772 (15%) 1 month - 161 (3%) Geater than 1 month - 191 (4%)	Nominal		
ATeamPeriodBetweenGamesATeam	Period of days between games from away team	4 days - 1150 (22%) 1 week - 2975 (56%) 2 weeks - 762 (15%) 1 month - 176 (3%) Geater than 1 month - 191 (4%)			
HTeamMoveUp/HTeamMaintenance	Home team moved up in division or remained in the league	U - 641 (12%) M - 4613 (88%)			
ATeamMoveUp/ATeamMaintenance	Away team moved up in division or remained in the league	U - 641 (12%) M - 4613 (88%)			
HTeam5ConsecutiveWinsAtH	5 consecutive home team wins playing at home	Y - 184 (3,5%) N - 5070 (96,5%)	Qualitative	Input	
ATeam5ConsecutiveWinsAtA	5 consecutive wins by the away team playing away	Y - 88 (1,7%) N - 5166 (98,3%)			
HTeam5ConsecutiveLossessAtH	5 consecutive home team defeats playing at home	Y - 34 (0,6%) N - 5220 (99,4%)			
ATeam5ConsecutiveLossessAtA	5 consecutive away defeats by the away team playing away	Y - 119 (2,3%) N - 5135 (97,7%)			
HTeam5ConsecutiveDrawsAtH	5 consecutive draws by the home team playing at home	Y - 5 (0,09%) N - 5249 (99,1%)	Flag		
ATeam5ConsecutiveDrawsAtA	5 consecutive draws by the away team playing away	Y - 5 (0,09%) N - 5249 (99,1%)			
HTeam3ConsecutiveWinsAtH	3 consecutive home team wins playing at home	Y - 510 (9,7%) N - 4744 (90,3%)			
ATeam3ConsecutiveWinsAtA	3 consecutive wins by the away team playing away	Y - 301 (5,7%) N - 4953 (94,3%)			
HTeam3ConsecutiveLossessAtH	3 consecutive home team defeats playing at home	Y - 223 (4,2%) N - 5031 (95,8%)			
ATeam3ConsecutiveLossessAtA	3 consecutive away defeats by the away team playing away	Y - 464 (8,9%) N - 4790 (91,1%)			
HTeam3ConsecutiveDrawsAtH	3 consecutive draws by the home team playing at home	Y - 88 (1,7%) N - 5166 (98,3%)			
ATeam3ConsecutiveDrawsAtA	3 consecutive draws by the away team playing away	Y - 95 (1,8%) N - 5159 (98,2%)			

Rating variables

Variable	Description	Count Distinct	Min	Q1	Q2	Q3	Max	M	SD	Role
HTeamELORating	Home team ELO rating	720	1472,00	1699,00	1774,00	1881,00	2315,00	1800,54	147,52	None
ATeamELORating	Home team ELO rating	712	1457,00	1697,00	1771,00	1876,00	2315,00	1797,76	147,30	
HTeamForm	Home team form	67	-38,00	-5,00	0,00	4,00	31,00	-0,38	8,70	
ATeamForm	Form of the away team	67	-39,00	-5,00	0,00	4,00	31,00	-0,35	8,76	
HTeamOverallRating	Home team overall rating	20	67,00	74,00	76,00	79,00	86,00	76,24	4,15	Input
HTeamAttackRating	Home team attack rating	24	67,00	74,00	77,00	80,00	92,00	76,99	4,76	
HTeamMidfieldRating	Home team midfield rating	21	67,00	74,00	76,00	79,00	87,00	76,24	4,35	
HTeamDefenceRating	Home team defense rating	20	67,00	73,00	75,00	78,00	86,00	75,63	4,18	
ATeamOverallRating	Away team overall rating	20	67,00	74,00	76,00	79,00	86,00	76,23	4,15	
ATeamAttackRating	Away team attack rating	24	67,00	74,00	77,00	80,00	92,00	76,99	4,76	
ATeamMidfieldRating	Away team midfield rating	21	67,00	74,00	76,00	79,00	87,00	76,24	4,35	
ATeamDefenceRating	Away team defense rating	20	67,00	73,00	75,00	78,00	86,00	75,62	4,18	

Odds variables

Variable	Description	Count Distinct	Min	Q1	Q2	Q3	Max	M	SD	Measure	Role
HTeamOdd	Home team closing odds	779	0,00	1,70	2,33	3,36	39,42	3,01	2,24	Quantitative	None
XTeamOdd	Draw Closing Odds	563	0,00	3,39	3,73	4,45	16,50	4,17	1,34		
ATeamOdd	Away team closing odds	1115	0,00	2,27	3,33	5,30	38,66	4,61	4,01		
Over2,5Odd	Closing odds of over 2,5 goals	211	0,00	1,64	1,87	2,15	3,43	1,92	0,38		
Under2,5Odd	Closing odds of under 2,5 goals	293	0,00	1,76	2,04	2,37	7,31	2,14	0,53		
HTeamOpeningOdd	Home team opening odds	777	0,00	1,70	2,34	3,34	39,42	2,99	2,16		
XTeamOpeningOdd	Draw opening odds	574	0,00	3,41	3,75	4,43	14,35	4,16	1,27		
ATeamOpeningOdd	Away team opening odds	1160	0,00	2,28	3,30	5,22	37,07	4,64	4,07		
Over2,5OpeningOdd	Opening odds of over 2,5 goals	205	0,00	1,64	1,85	2,14	3,46	1,91	0,38		
Under2,5OpeningOdd	Opening odds of under 2,5 goals	273	0,00	1,77	2,04	2,37	6,92	2,12	0,51		
HTeamClosingOdd	Home team closing odds	779	0,00	1,70	2,35	3,41	48,58	3,04	2,37		
XTeamClosingOdd	Tie Closing Odds	588	0,00	3,38	3,72	4,48	21,59	4,20	1,43		
ATeamClosingOdd	Away team closing odds	1157	0,00	2,27	3,36	5,37	41,00	4,70	4,18		
Over2,5ClosingOdd	Closing odds of over 2.5 goals	225	0,00	1,64	1,86	2,16	3,72	1,92	0,40		
Under2,5ClosingOdd	Closing odds of under 2.5 goals	300	0,00	1,76	2,04	2,38	7,31	2,14	0,55		