UNIVERSITÀ DEGLI STUDI DI BERGAMO

Dipartimento di Scienze Economiche

Corso di Laurea Magistrale in Economics and Data Analysis

Classe n. 56 – Data Science

# A Comparative Framework for Football Match Prediction using Machine Learning Techniques: A Study of Different Approaches for Improved Understanding

Relatore:

Chiar.ma Prof.ssa Michela Cameletti

Tesi di Laurea Magistrale

Enrico Cattaneo

Matricola n. 1051033

ANNO ACCADEMICO 2021 / 2022

# Abstract

This thesis explores using Machine Learning to predict football match outcomes in the top five European leagues from season 2016/2017 to 2021/2022. The study aims to expand on previous literature by analyzing a more extensive range of football-related features and assessing the predictive power of different football knowledge forms to understand better the various components of match information and the overall application domain. Hence, the analysis was conducted on two levels: first, splitting the data into seven subsets, each containing a reduced feature space, and then aggregating them into a comprehensive dataset. Several combinations of classification algorithms and preprocessing techniques, such as categorical encoding, feature scaling, and dimensionality reduction, were tested on each set in a structured way, enabling comparisons.

The models trained on different sub-datasets achieved varying levels of test accuracy, ranging from 48.7% for the Venue set, 53.9% for the Standings set, 54.5% for the Form and Rest set, 52.7% for the Stats set, 60.0% for the Betting Odds set, 54.1% for the Team Attributes set, and 54.2% for the Player Attributes set. In comparison, the model trained on the Comprehensive dataset accomplished a test accuracy of 60.7%. The process of generating such results provides critical insights for further research on practical applications by uncovering the predictive ability of the diverse feature spaces. Moreover, based on the comprehensive dataset, the best-performing model performs satisfactorily, especially compared to prior multiclass analyses. Its potential practical utilization in future studies could be of interest, mainly to investigate its effectiveness in identifying betting strategies.

# Contents

# Chapter 1

# Introduction

Football, sometimes called soccer, is widely recognized as the most popular sport worldwide. According to the global governing body of football, FIFA, there were around five billion football fans and over 250 million football players in 2021 (FIFA, 2021). The popularity of football has resulted in a significant financial impact, with the industry generating billions of dollars in revenue annually. Specifically, the European football market generated €27.6 billion in revenue in the 2020/21 season, a 10% increase from the previous season (Deloitte, 2022), despite the absence of fans from stadiums. The five major European leagues, namely the English Premier League, Germany's Bundesliga, Spain's La Liga, Italy's Serie A, and France's Ligue 1, which represent a 57% share of the European football market, generated €15.6 billion in revenue in the same season, a 3% increase from the previous year (Deloitte, 2022). In addition, the sports gambling industry has been growing at a rapid pace. The global sports betting market size grew from €90.6 billion in 2022 to €98.1 billion in 2023 (The Business Research Company, 2023) and is expected to reach €131 billion by 2028 (Grand View Research, 2021). In particular, as of 2022, football was the most favored sport for betting.

The history of football match result prediction dates back several decades, beginning with human experts analyzing critical factors to generate their predictions. In the 1980s, the first computer-based models were developed using simple algorithms to predict results based on basic information, such as team rankings and historical match results. With the advent of more advanced technology, there has been a growing appeal in using statistical

models and machine learning algorithms to generate more accurate and reliable predictions. These methods permitted using a broader range of features, including player and team statistics, match history, and environmental factors, to create forecasts of varying accuracy. Despite being a relatively new area of research, there have already been several notable successes using machine learning models, developed and deployed in various areas of the football industry outside mere results prediction, such as performance analysis (e.g., player tracking systems), strategy optimization, team dynamics, and fan behavior.

Reliable prediction of football match results using machine learning is a relevant area of research, potentially driving significant improvements in the whole industry, particularly in the sports gambling sector, and having substantial implications for various stakeholders. Bettors could benefit from a more educated decision-making process, potentially leading to greater returns on their investment. Additionally, accurate match prediction could equip managers, analysts, and club management with valuable insights to optimize strategies and player selection, enhancing team performances. Moreover, by detecting suspicious betting patterns, these potential benefits could also extend to the prevention of match-fixing and corruption by regulatory bodies.

The analysis in this thesis considers football match prediction as a multiclass classification problem with three possible classes (home team win, away team win, and draw). Specifically, it uses classification accuracy as the performance metric. This design is preferable since it is the most common among research studies within the specific application domain, allowing comparisons.

The main contributions of this study are twofold: the use of an extensive feature space and the development of a framework that enables the comparison of various forms of football-related knowledge. By doing so, this research attempts to open new avenues for football match prediction by providing a deeper understanding of different aspects of information related to the sport.

First, our approach regards an exceptionally vast and comprehensive feature space of

2,105 attributes, which is larger by almost an order of magnitude compared to sets from earlier analyses. While datasets of larger sizes in terms of the number of collected games (observations) were employed in prior research, the dimensionality of our comprehensive dataset, as the dimension of the feature set, is a novel aspect that clearly distinguishes our study.

Second, the extensiveness of the comprehensive dataset allows the arbitrary partitioning of the feature space into smaller sets, each incorporating knowledge referring to a specific field, differing in the nature and design of the included information or sharing a particular affinity. Such an approach is a fundamental element of our study since it enables comparing the predictive power of different types, characterized by common traits, of football-related information.

More specifically, the structure of our study attempts to address one of the main limitations of the literature on football match prediction, namely, the use of diverse samples due to an open and comprehensive dataset's absence, causing difficulties in comparing approaches and feature sets employed in different studies.

The partition of our comprehensive dataset into lower-dimensional feature spaces serves as a way of comparing the performances of different types of football-related knowledge and various machine-learning techniques. Furthermore, a shared set of procedures was employed throughout this study to ensure comparability across the results obtained when considering different feature sets, including all the sub-datasets (presented in Section 3.3) and the comprehensive dataset (see Section 3.4). In particular, while the data cleaning and feature engineering methods differ across datasets, other preprocessing techniques and classification algorithms were examined in a grid search to identify the optimal preparation methods and classifiers for each individual dataset (described in Section 3.2). Among preprocessing procedures, categorical encoding, feature scaling, and dimensionality reduction methods were included in the grid search. Additionally, six different classifiers and their respective hyperparameter tuning were evaluated within the grid search. This strategy produced a more exhaustive evaluation of each model and aimed to enhance performance. Moreover, it enabled comparisons between the results obtained from different datasets while facilitating the management of the data and the development of new features, generally providing a deeper understanding of the various components of football match information.

In summary, this study represents a significant contribution to the literature on football prediction because it evaluates a more extensive feature space than prior studies, allowing the comparison of various forms of football-related knowledge for more indepth data insights.

The remainder of this thesis is structured as follows. Section 2 reviews the literature on machine learning techniques for predicting football match results and some general reflections on this application domain. Section 3 outlines the data and methods used in our analysis. Specifically, Section 3.1 presents the data sources, Section 3.2 describes the techniques applied in the analysis, Section 3.3 details the preparation of the sub-datasets employed in our study, and Section 3.4 briefly presents the comprehensive dataset resulting from the combination of all sub-datasets.

Section 4 presents the results of the models developed using each sub-dataset and the comprehensive dataset. Section 5 discusses the obtained results, including their implications for the field of football match prediction. Finally, in Section 6, we draw our conclusions and highlight the main contributions of this study.

# Chapter 2

# Literature Review

This section reviews the literature on football match prediction using machine learning techniques. This presentation is not intended as a complete reexamination of the numerous studies conducted in the application domain but instead aspires to provide a valuable summary of the foundations where our analysis was constructed. Bunker & Susnjak (2019) and Horvat & Job (2020) conducted more comprehensive and detailed inspections of the domain, not limiting their analysis to football but also reviewing the research across multiple sports.

While review papers present the literature chronologically, we opted for a different course of action. Since our analysis aims to compare the significance of information forged from diverse natures, the related literature's exposition is conducted by aggregating studies based on the affinity of the assessed feature sets. Nevertheless, papers usually examine knowledge of various genres and are not easily categorizable into definitive classes. Thus, when a clear distinction could not be made, we tried not only to consider the features employed but also to weigh the relative importance and novelty factors brought by that particular analysis.

Among some of the first attempts to predict football results using machine learning, a common approach was not to focus on a single sport domain but to create models valid for multiple sports.

Reed & O'Donoghue (2005) aimed to predict outcomes (home win, draw, or away win) in both the English Premier League (football) and the Premiership Rugby. Sport-specific predictors were not included in the dataset, which contained games from two rugby teams and three football teams across three seasons (498 total game observations). Only seven variables common to both sports were included: position of both teams in the league table, rest period since the previous game, past matches' results (form), distance traveled, and game venue. The prediction accuracy for football contests was 57.9%, achieved with an artificial neural network (ANN), outperforming experts' beliefs.
Similarly, McCabe & Trevathan (2008) considered only shared attributes among multiple sports (Australian National Rugby League, Australian Football League, English Premier League, Super Rugby League) from 2002 to 2007. The 19 features contained information on the league's table, past performance, location, and player availability. Specifically, the best accuracy for football (54.6%) was obtained by employing an ANN.

Another recurrent strategy in the domain has been to investigate past performances, conceived from result-based knowledge of previous matches, such as outcomes and scores, to acquire valuable information for game prediction (analogous grounds of our dataset in Section 3.3.3).
Buursma (2010) employed data from 15 seasons of the Eredivise (Dutch football league) with 4,590 total game instances. The feature set was composed of 11 features depicting result-based past performances. These were averaged or aggregated across the team's previous 20 matches. This best number of past instances to consider was individuated by experimentation. The best accuracy of 55% was obtained using logistic regression.
Zaveri et al. (2018) proposed a more sophisticated approach using conceptually similar features and considering data from the La Liga competition from 2012/2013 to 2016/2017. The authors employed two subsets of features: a match history set including 12 attributes for past game performance and a team versus team database containing records for past encounters for all teams against each other. Different machine learning models were considered in the analysis, and ultimately logistic regression was the bestperforming algorithm with an accuracy of 71.63%.
Hucaljuk & Rakipović (2011) considered a dataset consisting of 96 games in the group stages of the Champions League competition. Data for the tournament's initial stages was

borrowed from previous seasons (with a weight decreasing round after round). Among the 30 features (20 remained after feature selection) inspected were: the team's form intended as the results achieved in the last six games, the outcome of the previous contest between the considered teams that play the game, current rankings position, number of injured players, and the goals scored and conceded per game. Among the considered classifiers, the ANN performed the best, with an accuracy of 68%. At the same time, including additional variables selected by experts did not produce performance improvements.

While form variables consider past performance from an outcome-centered perspective, football statistics-related features appoint in-game events as metrics of previous match performances (as in our dataset in Section 3.3.4).

Huang & Chang (2010) created a model to predict 64 matches from the 2006 World cup omitting the possibility for draws (binary classification). Via a feature selection process based on the domain knowledge of the authors, only eight variables were included, all obtained from past game statistics. Namely, the considered features referred to goals, shots on goal, possession, direct and indirect free kicks, corners, and fouls conceded. Using an ANN, the best accuracy was 76.9%. The exclusion of draws, which is the most challenging class to predict, explains the higher performance of this model when compared to previous studies.

Knoll and Stübinger (2018) suggested a procedure for predicting the results of the topfive European football leagues and then back-tested the prediction outcomes on betting odds. The authors considered seasons 2013/2014 to 2017/2018 (8,082 total match instances) and a feature set including information about the general game, attack capacities, defense, pass behavior, and disciplinary measures. All the included features referred to football statistics. Given that the primary objective of this study was to identify an effective betting strategy, the initial approach involved the development of a regression model with the goal difference as the target variable. Subsequently, the problem was reformulated into a binary classification task, with only predictions deemed secure (i.e., those where the absolute value of the predicted goal difference exceeded two) being utilized. Therefore, the model classified games only as home or away win, while draws were not included since they were considered risky wagers. The random forest algorithm

achieved the best performance with an accuracy of 75.6%. In the back-testing study, this same model generated returns of 5.42% on average for each bet.

Betting odds have been proven valuable for match result prediction, even when employed as the exclusive predictor in a model (see Section 3.3.5).

Odachowski & Grekow (2012) predicted football results by employing information regarding fluctuations in betting odds. In particular, three-way betting odds (home win, away win, or draw) were drawn for ten hours before a game (with intervals of ten minutes). The information in the odds time series was converted by computing summary statistics into 32 features. The authors also balanced their dataset to have an equal quantity of observations for each possible outcome (372 instances for class, i.e., 1,116 in total), ultimately finding a best-performing model with an accuracy of 46%. Draws were found extremely difficult to predict correctly, and the accuracy increased to 70% once these observations were excluded.

Tax & Joustra (2015) explored data from seasons 2000/2001 to 2013/2014 of the Eredivisie (Dutch league) with 4,284 total game observations. The dataset included match-specific information referring to results and other details (not football statistics) from past games. In particular, the authors compared the performances of public match data, which had to be easily retrievable, and information contained in betting odds. Feature selection (Sequential Forward Selection and ReliefF) and feature extraction (Principal Component Analysis) methods were implemented. For the public match data, the best accuracy of 54.7% was achieved by the Naive Bayes and ANN in combination with PCA. The highest accuracy accomplished using bookmaker odds features exclusively was 55.3%. When both odds and match data were employed jointly, the accuracy was 56.1%, realized with LogitBoost and ReliefF.

Another relevant approach has been to develop a practice for assigning ratings to either teams or players in a match. While in the early research stages, these attributes were usually handcrafted by authors, in more recent studies, ratings have been procured mostly by considering knowledge from video games (in our analysis, analogous sources were employed in Section 3.3.6 and Section 3.3.7).

Prasetio (2016) employed logistic regression to predict the results of the Premier League using data from 2010/2011 to the 2015/2016 seasons (a total of 2,280 games). Four features were proposed in this study: both teams' offense and defense ratings. However, the author did not explicitly describe the methodology used to generate these attributes. Using these four variables in a logistic regression model, the author achieved an accuracy of 69.5%. Still, this study was conducted by omitting draws.

Danisik et al. (2018) analyzed data from four years of the Premier League, from 2011/2012 to the 2015/2016 season, incorporating 1,520 match instances. The prominent originality of this analysis was the use of player-level data obtained from the FIFA video game series. The dataset included both match-specific and player-level data, with a total of 139 predictive variables. The prediction accuracy of 52.5% resulted from applying a Long-Short-Term-Memory (LSTM) neural network model. The accuracy increased to 70.2% when not considering draws.

Stübinger et al. (2020) considered all matches of the five most significant European football leagues and the corresponding second leagues between 2006 and 2018 (47,856 total observations). This analysis employed match-specific and player attributes. Similarly to Danisik et al. (2018), player skills were acquired from FIFA video games. After solving the problem as a regression, with the response variable being the goal difference, the results were translated into a binary classification without draws, following the exact process from Knoll and Stübinger (2018). The highest accuracy of 81.26% was achieved using a random forest algorithm.

Rudrapal (2020) focused on 11,400 matches of the Premier League from season 2000/2001 to 2015/2016. The authors considered a total of 40 features that belong to three different main categories: team-related, player-related, and head-to-head match related. In general, the employed features referred either to FIFA video games information or to match specific knowledge (mainly table and form based). A Multi-Layer Perceptron (a common type of ANN) classifier with ten hidden layers had the best accuracy of 73.6%.

Another compelling and creative approach to this problem was taken by Godin et al. (2014) and later implemented by Schumaker et al. (2016). Both studies focused on sentiment analysis from Twitter micro posts to predict match results and create a

compelling betting strategy. In particular, they demonstrated the superiority of crowdsourcing compared to expert knowledge, notably in wagering decisions.

In their review paper on result prediction in sports, Haghighat et al. (2013) emphasized how the domain was characterized by ingrained complications in comparing the results from different analyses. There are numerous dimensions over which studies usually differ, such as the sport investigated, the input dataset (other matches, seasons, or competitions), the predictors, and the class label. Due to the scarcity of available benchmark datasets, researchers have considered diverse approaches to evaluate their empirical results, usually based on comparisons with baseline measures. The most common baseline models are built on either utilizing the results predicted by experts, assigning the class label to the outcome with the lowest betting odds, always selecting the majority class (mostly home win due to the home advantage phenomenon), or randomly selecting the result.

A recent attempt to address this problem was the comprehensive and public Open International Soccer Database (Dubitzky et al., 2019). The rationale behind this project was to provide a benchmark dataset for more robust comparisons between studies and to facilitate the creation of updated knowledge in the application domain. This dataset contains over 216,000 matches from 35 countries and 52 different leagues. The proposed challenge was to predict 206 future game results from 26 competitions. Here, we recap two of the most relevant research papers from this competition.

Hubáček et al. (2019) considered two different versions of rating measures: the pi-ratings (Constantinou & Fenton, 2013), which capture teams' historical capabilities by also accounting for the current form, and ratings based on PageRank (Page et al., 1999). To be more specific, pi-ratings compare the goal difference encountered in a fixture with the expected outcome based on the already existing ratings. This comparison produces the updated rating values, placing more attention (higher weights) on the result rather than on the goal margin. After computing the two rating measures for each team, Extreme Gradient Boosting (XGBoost) learning algorithms were considered in both a regression

and classification environment. The classification approach delivered the best performances over the validation and unseen test set, with a final test accuracy of 52.4%. Berrar et al. (2019) evaluated two disassociated feature sets distinguished by the nature of the information considered: recency features and rating features. Recency features were constructed by computing football statistics averages over the previous nine games, explicitly referring to attack and defensive force, home advantage, and opposition potency. Instead, rating features were assembled on the performance ratings of each team and updated after each observation by comparing the expected and observed results. The kNN algorithms were applied to both feature subsets separately and performed better on rating features than on recency variables. XGBoost algorithms were used only after the competition and corroborated the higher significance of rating features (also outperforming kNN models). The test accuracy achieved by kNN employing rating features was 51.9%.

Broadly, the research concerning the Open International Soccer Database delivered ingenious new approaches and techniques, especially combining rating-based methods and machine learning algorithms. Regardless of the massive dataset size, accuracies appeared to plateau. This remark is connected to the design choices made during the data collection. Notably, only easily retrievable information was gathered over the numerous leagues from various countries, including lower competitions. Consequently, the dataset does not include sport-specific features from in-game events or other highly refined and specialized fields. These developments should not be surprising since constructing more informative and discriminatory features through domain expertise is typically acknowledged as the soundest strategy for improving the performance of machine learning models. Indeed, when considering results prediction over a variety of sports and their related literature, it is clear how the sports that have experienced the most significant gains in accuracy over the years are the sports experiencing the most significant growth in feature sets. Conversely, reflections on the mere dataset size have been ambiguous across the literature since a vaster number of observations (like in the Open International Soccer Database) has yet to be linked with improved performances.

Almost 90% of the papers cited in Bunker (2019) considered engineering a richer set of features as the primary focus for future work. Additionally, when considering vaster

feature spaces, techniques for dimensionality reduction, like feature extraction and feature selection, acquire a fundamental role. Consequently, feature engineering combined with dimensionality reduction techniques is the primary driver for performance improvement in this (and many others) application domain.

Bunker (2019) also stresses how the outcome prediction for certain sports has been proven to be innately more complex than others. Even if football has been the most researched sport, its highest accuracy has been 78%, far from performances obtained in other sports domains. It has been proven that low-scoring sports tend to embody a more significant element of randomness in outcomes, lowering models' predictive performance. Low-scoring sports are also associated with a higher likelihood of drawn contests. Moreover, when predictions are formulated in a classification design with multiple labels (multiclass formulation) rather than a binary problem, the complexity of the learning task increases, and hence accuracies tend to decline. Differentials in performances among sports can also be significantly attributed to the characteristics (like the points-scoring system) and especially to the competitive context denoting a specific study.

# Chapter 3

# Methodology

## 3.1 Data Sources

Our study collected and analyzed football match data from the top five European leagues (Bundesliga, La Liga, Ligue 1, Premier League, and Serie A) from the 2016/2017 season to the 2021/2022 season. The scope of our research did not include games from domestic and European cups. However, observations from these competitions were nevertheless considered to create several features employed in the study.

The data was gathered from three online sources:

- Sportmonks football API (https://docs.sportmonks.com/football/)
- FIFA index website (https://www.fifaindex.com/)
- Visual Crossing Weather API (https://www.visualcrossing.com/weather-api)

From the Sportmonks football API, we acquired detailed information about football fixtures, including game results, statistics, betting odds, in-game events, and starting lineup players.

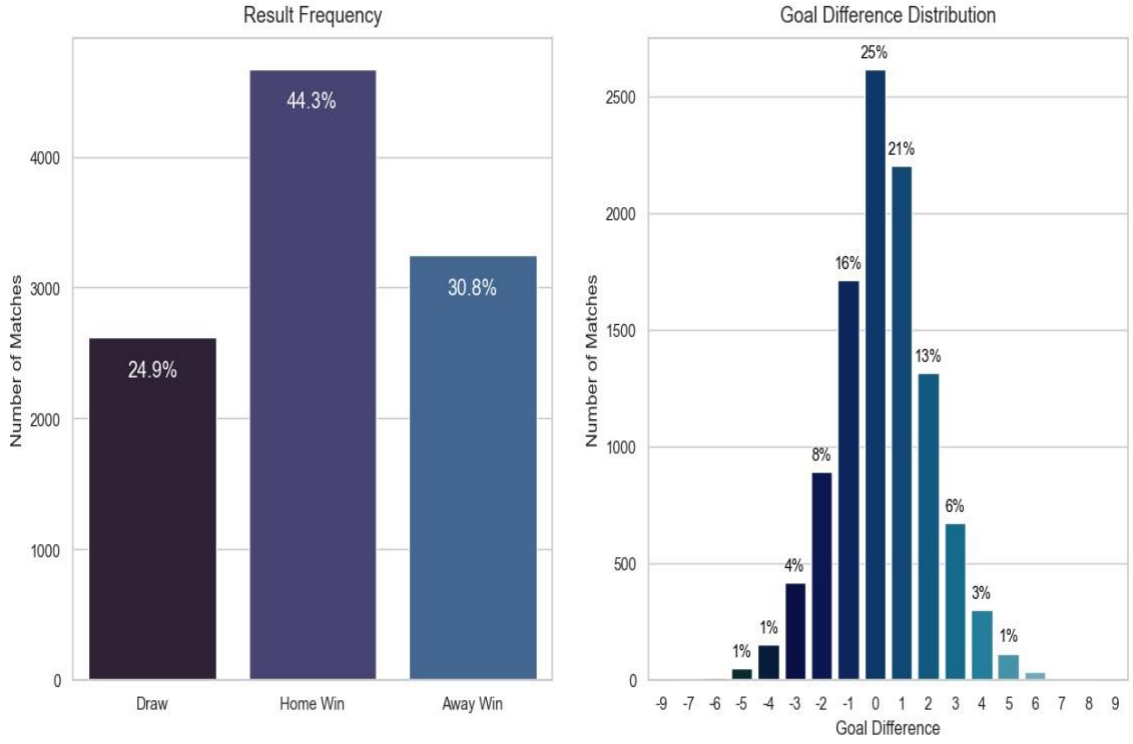
The FIFA index website contains details from the FIFA video games series referring to both teams' and players' attributes. The information from this source was extracted through web scraping, a technique to collect large amounts of internet data automatically. Since this process is highly time-consuming, we employed cloud computing through the Google Cloud Platform to accelerate the operation. The scraped data was organized into

two separate databases: one for teams with approximately 82,000 observations and one for players with roughly 1,250,000 instances (both available on Kaggle: https://www.kaggle.com/datasets/enricocattaneo/fifa-videogame-players-and-teamsdata). Both databases contain approximately weekly updated data (for each team or player), depending on the specific updates frequencies of the webpage.
From the Visual Crossing Weather API, we attained meteorological variables referring to the weather conditions for the matches considered in our analysis.

A total of 10,536 matches were considered in this analysis. However, since our study compared multiple datasets originated from and related to different aspects of the sport, some sub-datasets contained more petite samples due to missing values than others, depending on the completeness of the corresponding information.

The precise indications for each sub-dataset size and feature space dimension are provided in Section 3.3. Moreover, each feature subset data cleaning, feature engineering, and preprocessing processes were conducted and presented individually.

**Figure 1: Match Result Distributions**



Our analysis's primary focus is to predict football match results defined as a home win, draw, or away win. Figure 1 illustrates the distribution of our target variable and the goal difference. In total, 4,670 games resulted in a home team win (44.3%), 2,619 matches ended in a draw (24.9%), and 3,247 games resulted in an away team win (30.8%). This representation confirms the well-documented phenomenon of home advantage in sports. In addition to match outcomes, we also represented the distribution of the goal difference between the home and away teams. This more detailed version of the results variable revealed an asymmetric distribution, with more contests having a positive goal difference (the home team wins) than a negative goal difference (the away team wins). However, the most frequent outcome regarding goal difference was zero, corresponding to a draw.

## 3.2 Description of Common Methods Used across Sub-Models

The literature on sports prediction is divided on whether a classification approach based on results is more effective than a regression approach based on point difference. Studies such as Delen et al. (2012) and Valero (2016) have found that classification performed better, while Danisik et al. (2018) found that a numeric prediction approach was superior for their dataset. Given these conflicting findings, Bunker (2018) recommended that future researchers include a comparison between methods in their experimental design. Despite these debates, our study addressed the issue of predicting football outcomes exclusively using a classification framework since most research on football prediction has implemented classification practices. Besides, such a comparison was beyond the scope of our analysis, which focused more on evaluating features' predictive value.

Furthermore, we decided to utilize a multiclass classification approach rather than the binary approach implemented in a few previous studies. This decision was based on the intrinsic nature of football, which has a high probability of draws—considering such a category increased the learning task's complexity and caused a decline in accuracy, especially since draws are the most challenging class to predict. Also, we defined

accuracy as the performance metric to evaluate models. This decision was based on the fact that accuracy is employed in most research studies within the specific application domain. However, it should be noted that a small number of researchers have explored the use of alternative performance metrics.

As mentioned in Section 2, the literature on football match prediction has been afflicted by the inability to compare approaches from different papers due to the focus on different samples, caused primarily by the absence of a shared and comprehensive dataset. The division of our entire data into lower-dimensional feature spaces attempted to address this issue by comparing between sub-datasets. By reducing the initial dimensionality of the data, this approach facilitated the management of existing data and the development of new attributes while also providing a deeper understanding of different aspects of football match information. Notably, each sub-dataset was arbitrarily partitioned to contain knowledge referring to a specific field, differing in the nature and design of the included information or sharing a particular affinity. The proposed framework represents a significant contribution to research on football prediction since it enables the comparison of various forms of football-related knowledge. Another crucial novelty of our approach derives from utilizing a vast and comprehensive dataset. While previous studies have, in a few cases, used datasets of larger size in terms of the number of matches than our dataset, this study is novel in considering a more extensive feature space. This analysis significantly increased the data dimensionality by almost an order of magnitude compared to prior studies.

Our seven sub-datasets, described in more detail in Section 3.3, included information on the venue of the match (Section 3.3.1), league standings (Section 3.3.2), form and rest (Section 3.3.3), football statistics (Section 3.3.4), betting odds (Section 3.3.5), and team and player attributes (respectively Sections 3.3.6 and 3.3.7). Then these datasets were combined to create a comprehensive dataset (Section 3.4).

A shared set of techniques was employed throughout this study to ensure comparability across the results obtained from different datasets, including all the subsets and the comprehensive dataset. As described in the previous section, the majority class

in our dataset is a home win, accounting for 44.3% of observations. To serve as a benchmark, we supposed a model that solely predicts this class as a reference for submodel performances.

The first step of our standard approach was an 80/20 split of the available data into separate training and testing sets. Notably, the partition was accomplished through stratified sampling, guaranteeing that the proportion of different classes mirrors that of the overall dataset in both training and testing sets. As a result, these sets accurately represented the total population, thereby preventing class imbalance and enhancing model performance by minimizing the risk of overfitting and underfitting.

After the split and to conduct an exhaustive analysis, we employed a systematic procedure to identify the optimal model while improving its robustness and generalization. This approach involved a grid search, with five-fold cross-validation for evaluation, over various preprocessing techniques and classification algorithms. Additionally, hyperparameter tuning was applied for each model to enhance its performance.

Grid search, a widely adopted method in machine learning research and practice, involves an exhaustive search through a manually specified subset of the parameter space to determine the optimal combination of hyperparameters. The main limitations of this process include its sensitivity to the prescribed range of hyperparameters' values and its computationally intensive nature (also time-consuming). In our analysis, this search also considers diverse preprocessing techniques.

To evaluate the myriad combinations generated during the grid search, we opted to utilize the k-fold cross-validation (CV) method rather than additionally splitting the training set to obtain a separate evaluation set. This technique divides the data into k subsets (folds), using k-1 to train the model and the remaining for evaluation. This process is repeated k times, with each fold serving as the evaluation set once. The final performance metric is then computed by averaging the model's metric over the number of iterations, thus generating a more robust estimate of the model's performance.

Moreover, this method allowed us to use all the data points for training and evaluation, reducing the risk of overfitting and improving the model's generalization. In this study,

we chose a five-fold CV with stratified sampling, so each of the five folds was a balanced representation of the different result classes.
Data cleaning and feature engineering methods are described separately for each dataset in Section 3.3. Other preprocessing techniques necessary to prepare the data for training and evaluation were examined within the grid search to identify the most effective preprocessing procedure and a comprehensive optimal approach. Precisely, categorical encoding, feature scaling, and dimensionality reduction methods were included in the grid search process. Additionally, the grid search considered six different classifiers and their respective hyperparameter tuning. This strategy produced a more exhaustive evaluation of the model and aimed to enhance its performance.

The various specific preprocessing methods and the classification learning algorithms regarded within the grid search are presented in the next section.

### 3.2.1 Common Preprocessing Techniques for Predictive Modeling

The present and following sections provide a brief overview of the preprocessing techniques and classification learning algorithms included in the grid search context discussed previously. While this presentation is not intended to be exhaustive, it serves as an introduction for those unfamiliar with these methods. Instead, only a quick summary is necessary and provided here for those already knowledgeable about these techniques. As previously stated, the grid search considers various preprocessing techniques, such as categorical encoding (one-hot and ordinal) and feature scaling (normalization and standardization). Also, dimensionality reduction methods were assessed, including feature extraction (PCA) and feature selection processes (mutual information, f-test, decision tree-based, and random forest-based). Moreover, the grid search also evaluates several classifiers and respective hyperparameter tuning, including Naive Bayes, kNearest Neighbors (k-NN), Multinomial Logistic Regression, Decision Tree, Random Forest, and AdaBoost.

### *Encoding Categorical Variables and Scaling Features*

#### *Categorical encoding*

*Categorical encoding* is a meaningful preprocessing procedure to transform categorical data represented by a finite set of discrete values into a numerical expression suitable for machine learning algorithms utilization. Categorical features in our entire dataset are of three main kinds: binary, nominal, and ordinal. Since every binary variable considered is represented numerically, not requiring additional encoding, this paragraph focuses only on nominal and ordinal attributes. While nominal attributes have no quantitative representation being purely qualitative, ordinal features have an order of measurement with a scale indicating direction, but where distances cannot be quantified. For simplicity's sake, regardless of numerous and more sophisticated encoding methods, in our analysis, the choice of encoder felt between One-Hot encoding and Ordinal encoding:

- *One-Hot encoding* transforms each category into a vector of binary variables, each indicating the absence or presence of a specific category.
- *Ordinal encoding* assigns an integer to each category. Differently from OneHot encoding, with ordinal encoding, there is no increase in the dimensionality. However, it can be problematic when the encoded variable may not possess an inherent ordinal relationship. In such scenarios, ordinal encoding may impose an artificial order among categories, leading to a loss of information and potentially biased conclusions.

Although One-Hot encoding is one of the most common methods to encode categorical variables, it has various drawbacks. Mainly, the inappropriateness of One-Hot encoding has been empirically shown when dealing with high-cardinality categorical features (Cerda et al., 2018). This primary drawback could lead to poor performance in the predictive models. Our approach was to consider both encoding methods in the context of a grid search for those categorical features characterized by either high cardinality or unclear definition. However, for those categorical attributes with neither

high cardinality nor ambiguous structure, the most appropriate encoding method was chosen before the grid search based on feature analysis and domain knowledge.

*Feature scaling*

*Feature scaling* is a crucial preprocessing technique in machine learning, bringing all features in a dataset to the same scale. With rare exceptions (such as decision tree-based methods), many learning algorithms perform poorly when numerical input variables have vastly different scales, such as distance-based models. Additionally, feature scaling is essential when implementing Principal Component Analysis for dimensionality reduction.

Our analysis employed the most common methods to scale numerical attributes: normalization and standardization.

- *Normalization*, often called min-max scaling, is the simplest method that shifts and rescales feature values to constrain them between 0 and 1. This effect is accomplished by subtracting the feature's minimum value and dividing it by the feature's range, which is the difference between the maximum and minimum values.
- *Standardization*, also known as Z-score normalization, does not force values in a specific range. However, it rescales features to a standard normal distribution with a mean of zero and a standard deviation of one. This result is achieved by subtracting the feature's mean and then dividing it by its standard deviation so that the resulting distribution has a unit variance.

***Curse of Dimensionality***

The curse of dimensionality refers to the exponential increase in the number of samples needed to accurately estimate an arbitrary function as the number of input variables (i.e., dimensionality) increases. In machine learning, this expression depicts cases when learning a state of nature from a finite number of data samples in a highdimensional feature space (generally of a hundred or more) is necessary. Such

problems often bear complications in training models due to the sparsity in the data, which augments the risk of overfitting. The issue is deeply interconnected to the peaking phenomenon (Hughes, 1968), which states that for a fixed sample size, the predictive power of a model initially increases and then, beyond a certain dimensionality, begins to decline as the number of features grows (while the error follows a diametrically opposite trajectory). Theoretically, a potential solution is to increase the training set size to attain a sufficient density of training instances. However, in practice, the number of observations required to reach a given density and generalize accurately grows exponentially with the number of dimensions. A more functional approach to alleviate issues linked with high-dimensional data is through dimensionality reduction techniques. These procedures aim to transform the data into a lower-dimensional space while preserving most of the meaningful information in the original dataset, nevertheless inducing some information loss. By reducing redundancy and noise in the data, these approaches diminish the complexity of learning algorithms making the training phase considerably faster, and also improve the model's performance.

In our study, we focused on using feature selection and feature extraction as dimensionality reduction techniques, which proved to maintain most of the relevant information and generally improve the performances of our models.

*Feature selection*

*Feature selection* methods aim at reducing the number of input features when designing a predictive model by selecting a subset of the most valuable variables and removing redundant or non-informative attributes. As stated by Yu and Liu (2004), an optimal subset contains strongly relevant and weakly relevant but non-redundant features, removing completely irrelevant and noisy attributes and weakly relevant but redundant features. As previously stated, feature selection can enhance a model's performance and reduce complexity. Additionally, it improves understandability through knowledge discovery by selecting only the most relevant and informative variables. Feature selection can be classified into three categories based on the type of data used:

supervised, semisupervised, and unsupervised. Supervised feature selection, which utilizes labeled data, is further divided into three main subcategories based on the evaluation criterion employed: filter, wrapper, and embedded. In our research, we applied only methods belonging to filter and embedded feature selection techniques. The utilization of wrapper methods was not pursued due to their higher computational costs, susceptibility to overfitting, and lack of generalizability (valid only for the particular algorithm the procedure was "wrapped" around).

Filter methods assess the inherent characteristics of the features independently of the learning task. The nature of most filter procedures for feature evaluation is univariate. Among the supervised feature selection categories, these methods are known for their efficiency and scalability, making them suitable for large datasets. Additionally, the delivered solutions are independent of any specific learning algorithm, which results in better generalization properties since the bias in the feature selection is not correlated with the bias in the learning method. However, it is crucial to note that filter approaches neglect the relations between classifiers and features, which may lead to variations in performance when different learning algorithms use the selected subset of attributes.

In this study, we considered two filter methods in our grid search process:

- *Mutual Information*. In the study of information theory, mutual information is a symmetric measure applied in estimating the amount of information that two random variables contain about each other (Doquire et al., 2013). If the variables are independent, their mutual information is equal to zero. Feature selection based on mutual information was first proposed by Battiti (1994) and is become a prevalent method due to its interpretability and efficiency. This approach estimates the dependence between features and the target variable by computing the information gain between each attribute and the class labels (Tang et al., 2014), aiming to maximize the relevance of the selected features while minimizing their redundancy. An essential property of mutual information-based feature selection is its capacity to detect nonlinear relationships between variables.
- *F-test*. In practice, ANOVA, which stands for analysis of variance, is often used as a feature selection method in classification problems with numerical attributes. We employed an F-test in one-way ANOVA, which calculates the

ratio of the variance between groups and the variance within a group, where the groups are observations from the same target class. This feature selection creates a subset of features based on higher F-score values, where distances within the groups are lower compared to the distances between the groups. The result of this selection process is that the more independent variables from the target are removed from the dataset, while features presenting a significant difference between classes are selected. This feature selection method is widespread among filter approaches, even in the literature (Elssied et al., 2014; Dhanya et al., 2020).

Embedded methods are built-in feature selection instruments that embed the feature selection procedure inside the learning algorithm using its properties to conduct feature evaluation. These methods have several advantages over wrapper methods, including lower computational costs and reduced risk of overfitting while retaining comparable results. Additionally, embedded methods allow for interaction with the classifier by incorporating classifier bias into feature selection, as in wrapper methods, and permit feature interactions even of a high order, meaning interactions between more than two features. Among feature selection embedded methods, decision tree-based and random forest-based were considered in our analysis.

- *Decision tree-based* feature selection methods rank features on importance (like many filter methods do) based on metrics like the Mean Decrease Impurity or Mean Decrease Gini (Louppe et al., 2013).
- *Random forest-based* feature selection methods employ an ensemble of decision trees to determine the significant features by averaging the feature importance scores (MDI or MDG) across all trees in the forest.

Empirical studies have shown that both embedded feature selection techniques may encounter challenges in automatically removing redundant features, thus lowering performance (Kubus, 2019). Furthermore, as the number of attributes increases, the ability of these techniques to detect feature interactions is observed to decrease.

*Feature Extraction*

The objective of feature extraction techniques is to identify and extract salient information from data, reducing the dimensionality of the feature space while preserving relevant knowledge for machine learning algorithms implementation. In contrast to feature selection methods, feature extraction techniques do not isolate a subset of features deemed as most pertinent but instead transform the original data into a more appropriate representation for a specific model or task. Additionally, dimensionality reduction improves efficiency in data processing and storage. This study considered only Principal Component Analysis (PCA) among feature extraction methods.

*Principal Component Analysis*, an unsupervised, nonparametric statistical technique, is widely used for the dimensionality reduction of high-dimensional datasets (Jolliffe, 2002). The principal components capture the underlying structure of the data, and the dimensionality reduction is achieved by selecting a subset of these components. The objective of PCA is to retain as much information as possible by preserving the maximum amount of variance in the data. The main limitation of PCA is its linear nature, which may impede its ability to capture nonlinear relationships in the data. In such cases, alternative techniques such as Kernel PCA, which can perform complex nonlinear projections, should be considered. Despite this limitation, PCA remains a powerful and versatile technique for extracting meaningful knowledge from high-dimensional datasets.

### 3.2.2 Common Learning Algorithms for Classification Modeling

In our study, we evaluated the performance of six different classifiers. To ensure a proper comparison, we employed a shared evaluation protocol for each classifier, which applied hyperparameter tuning using a grid search method. The algorithms were chosen based on popularity and effectiveness in various applications.

*Naive Bayes* is a classification algorithm based on the maximum a posteriori (MAP) concept in Bayesian statistics. A naive Bayes classifier assigns the class with the maximum a posteriori probability to an unknown example given the observed data by applying strong (naive) independence assumptions. Despite being based on the unrealistic assumption of conditional independence, this classifier has been highly successful in practice, even compared to more sophisticated models. Furthermore, it has proven remarkably suited for high-dimensional data (Mehra & Gupta, 2013) and small datasets since it requires only a small number of instances to estimate the parameters (e.g., means and variances for quantitative features).

*k-Nearest Neighbors* (Fix & Hodge, 1951) is among the most straightforward and oldest nonparametric classification algorithms. To classify an unknown example, a distance metric measures the distance from the new instance to every training observation. Then, the k smallest distances are determined, and the most represented class is deemed the predicted class for the unknown example. The number of nearest neighbors (k) is a crucial hyperparameter that requires thorough tuning. Additionally, k-NN is referred to as a lazy (or instance-based) learner since it does not make any generalizations using training points. Despite its simplicity, k-NN has many advantages over more modern methods. The algorithm is highly interpretable, and its performance is often competitive with more sophisticated algorithms. Furthermore, k-NN can be applied to small datasets, and it does not require any assumptions about the data distribution. One of the main drawbacks of this model, which is designed for low-dimensional data, is its performance with sparse data, which complicates the computation of distances and identification of nearest neighbors.

*Multinomial Logistic regression* (also called Softmax regression) is a classification technique that generalizes the logistic regression model to support multiple classes (multiclass classification problems). The Softmax regression algorithm applies binary logistic regression to numerous classes at once, and it is obtained by replacing the sigmoid logistic function with the Softmax function. Like Logistic regression, the Softmax regression estimates the probability that an instance belongs to a particular class (the entire set sums to 1) and predicts the class with the highest estimated probability. A

fundamental assumption of this method is for categories to be mutually exclusive from each other.

A D*ecision Tree* is a potent classifier defined by a top-down and greedy approach called recursive binary splitting. It consists of nodes forming a rooted tree, implying a directed tree with a "root" node with no incoming edges and with all other nodes having only one incoming edge. Nodes with outgoing edges are defined as internal or test nodes, while all other terminal nodes are called leaves (or decision nodes). Each test node in a decision tree splits the instance space into two (or more) subspaces following a discrete function of the features' values. At the bottom of the tree, the majority class of the target value is designated for each terminal leaf. After training, unknown observations are classified by following a course from the root down to a leaf, with splitting nodes showing the way. Decision trees' main advantages are the easy interpretability of the decision process, the ability to handle both numerical and categorical features, and the unnecessity of feature scaling. Decision trees are also the fundamental components of many ensemble learning algorithms.

*Random Forest* (Breiman, 2001) is an ensemble bagging algorithm combining decision tree predictors (as base learners), which grow in randomly selected feature subspaces. Instead of searching for the best feature when splitting a node (like in decision trees), a random forest introduces extra randomness by using a random selection of attributes to split each specific node. The added randomness produces greater tree diversity, which trades a higher bias for a lower variance. After training, to classify an unknown instance, each tree provides classification, which is recorded as a vote. The votes from each tree are combined, and the predicted class for the new observation is chosen using majority voting. Random forests lose some of the simplicity in the interpretation and visualization that characterizes decision trees. However, Random Forests produce highly accurate predictors while remaining relatively easy and fast to implement (at least when the number of decision trees is not excessively elevated). They are good at handling categorical and numerical data without requiring any feature scaling. For our framework purpose, it is also essential to emphasize random forests' ability to

manage high-dimensional data without overfitting and their robustness in the presence of noise or missing data.

*AdaBoost* (Freund, 1996) is a boosting technique combining weak learners to construct a strong learner. This learning algorithm uses weighted versions of the training data, with all weights being initialized equally at the start. After each learning round, the relative weights of misclassified training instances are raised so that the next learner will focus on the complex observations in the training data. Given the adaptiveness of the learning cycle, it is crucial to choose weak learners as base classifiers. In the case of a strong base learner, if high accuracy is achieved, only outliers and noisy observations will have a significant enough weight to be learned in the succeeding rounds. Therefore, decision stumps (one-level decision trees) are usually picked as base learners. To make predictions, AdaBoost computes the weighted average of the weak learners. The AdaBoost's main advantage is its robustness to overfitting, which furthermore simplifies the hyperparameters tuning when compared to other learning algorithms. However, this learning algorithm requires particular quality datasets, given its susceptibility to noise and outliers in the data.

## 3.3 Preparation of Sub-Datasets for Football Knowledge Prediction

This section illustrates the different sub-datasets regarded in our analysis. As mentioned above, separating the entire feature space into smaller sets is a fundamental aspect of our study, permitting the comparison of the predictive value related to different types of football-related knowledge. Moreover, each dataset's data cleaning and feature extraction approaches are described exhaustively for reproducibility purposes.

### 3.3.1 Venue Dataset

The rationale behind the venue dataset design was to include exclusively locationbased features aiming to assess their aggregated predictive strength.

Among all the attributes in the dataset, only three did not closely relate to games' site details. Namely, these three categorical variables were the league name, the season name, and a code combining the information of the two mentioned above (distinct value for the same league on different seasons and vice-versa).

When describing the remaining 19 features, a division into groups is helpful for clarity. There were seven attributes associated with venue characteristics: a venue identifier, the name of the city and the country where the venue was, a match-specific attendance figure (which could be learned before the start of a game by examining ticket sales), the venue capacity (i.e., the maximum number of persons it was designed to hold), the attendance and capacity ratio (which provides an approximative measure of home fan support), and a binary variable indicating the field surface type (one when it was natural grass, and zero otherwise).

Then, eight meteorological features were attained from the Visual Crossing weather API. The included metrics were the following: temperature (in Celsius), cloud cover (a percentage), relative humidity, atmospheric pressure (in millibar), visibility (distance at which distant objects are visible), wind speed (in m/s), and wind direction. These variables delivered the climate conditions from the venue location, gauged at the beginning of a contest and not regarding subsequent variations in conditions. An additional feature was produced using astronomical details, precisely the sunset time. Specifically, a binary variable denoting night game took the value one if the sun fell in at most one hour from the event's start (meaning that at least one half was played at night) and zero otherwise.

Three other features were derived from travel considerations, two based on the away squad and one on the home team. First, a travel measure (in kilometers) was gathered by calculating the distance between the match venue and the away team's stadium. We utilized the respective coordinates of the two mentioned sites for this computation. Also, there were two additional dummy variables: one indicating if the away team had to travel

outside its country and the other denoting whether the home team played in its official stadium.
The dataset also contains two extra features, which were added even if not location-based. They were nominal variables, indicating each team kit's primary color. They were generated by converting the colors from the hexadecimal format into six categories based on resemblance.

After data preparation, our dataset had 23 features (12 categorical and 11 numerical) and 10,536 observations, i.e., the entire set of matches considered in our study.

### 3.3.2 Standings Dataset

The purpose of the standings dataset was to determine the predictive value of information found in league tables, which reflects the season-long performance of teams.

Data concerning standings information from the Sportmonks API had poor quality. As a remedy, we developed a program generating the league tables for each round of the five competitions in our analysis. The program uses match scores (both first-half and fulltime scores) and the round to which a specific game belongs as input information. To ensure the accuracy and completeness of the data, we captured a moment in time by generating snapshots of the league standings for each round and competition examined. These snapshots were subsequently compiled to create the standings dataset utilized in our study, enabling us to rigorously control the information and reduce the risk of any potential inconsistencies in the data.
For a thorough application and to account for home advantage (represented in section 3.1), we designed three diverse types of tables, each based on a different assessment. To be exact, we created a home table (where only games played at home are considered), an away table (contemplating only matches played on the road), and a comprehensive table (the classic league's table) where no venue-based distinctions were made. Furthermore, each table type was divided into three time-based sub-types: a sub-table that considers

games in their entirety and two sub-tables based respectively on first-half and second-half scores.

After creating the described tables, the standings dataset was acquired by merging match-specific and league standings data. The combining process linked both teams in a game with their related table information that emerged from the end of the previous round, avoiding look-ahead bias.

In particular, our dataset included standings information from the comprehensive table for both clubs. In contrast, the home and away tables were incorporated only for the respective team (each containing all three time period specifications). Since all the attributes in the dataset were absolute measures, we also computed relative features by dividing each variable by the number of games played to that point. Thus, these new features were expressed either as a percentage (when result-based) or as an average.

Some critical considerations were made while analyzing the tables' data. Mainly, league tables data, by nature, was plagued by a systematic un-informativeness of its observations. In fact, at the start of each season, the details from the previous year were entirely erased. For this reason, no valuable standings knowledge could be provided for the season's first round when interpreting classic tables. For perspective, a single round denoted approximately 2.6% of the matches in an entire season. Furthermore, the magnitude of this problem was even more prominent when considering the home and away tables instead of just the comprehensive table. Indeed, depending on the league schedule, the home and away tables could remain uninformative for multiple rounds (at least doubling or even tripling the quotas above), increasing the significance of the problem.

The data-cleaning processes were organized into two sequential phases to account for and try to correct this ingrained data issue: one principal and the other corrective. Foremost, missing values from early rounds were filled with data from the end of the prior season. Although this step rectified uninformative observations in the table's features, it neglected to consider the presence of newly promoted teams (11-15% of the total clubs in a league) that did not hold any standings details from the prior year. As a proposed adjustment to this first inaccurate procedure, missing data from initial rounds

belonging to recently promoted teams were filled by randomly selecting a value among the 20% worst observations in a variable-specific distribution. Depending on the intrinsic characteristics of each feature, the new values were randomly chosen from the fifth quintile (in the case of variables where higher values represent worse performances, like rank position) or from the first quintile (when lower values translate to poorer performances, like points). In practice, this preparation strategy supposed clubs coming from a minor competition (11-15% of the entire league) to be in the bottom 20% for every performance metric in the beginning stages of a season. Despite being based on a strong assumption, the procedure yielded more informative observations than assigning zero to every earlyround match.
An alternative solution would have been to drop each instance that did not retain information from that season's standings. Still, in our analysis, we opted for recovering seasons' beginning rounds since they were deemed particularly valuable.

To summarize, we considered four tables (the comprehensive table for both teams and the home or away table for each), each having three distinct time-based sub-types.
The total number of features extracted from each employed table was fifteen. The considered absolute features were won games, drawn games, lost games, goals scored, goals conceded, goal difference, and points made. From the features listed above, the following relative features were derived: win percentage, draw percentage, loss percentage, goals scored per game, goals conceded per game, goal difference per game, and points per game.

After completing the data preparation phase, our dataset contained 180 features (168 numerical and 12 categorical) and 10536 games, the entire set of matches considered in our study, since the cleaning procedure was designed to avoid inducing missing values. Within the grid search, no distinction was necessary for encoding categorical features since the only considered nominal variable is ordinal by nature (features referring to the league table rank position).

### 3.3.3 Form and Rest Dataset

With the form and rest dataset, we sought to investigate the relationship between information from previous events and subsequent games' results. Therefore, we assembled a dataset only utilizing elementary knowledge of past games to appraise its relevance and informativeness within our classification environment. Despite being constructed on similar footings, our features were conceptually divisible into two distinct families. Specifically, while variables referring to rest regarded the mere occurrence of a match as its only applicable information, form attributes also assessed game-related scores and outcomes.

The rationale behind the rest data was to approximate and account for squad fatigue. Before building these features, we contemplated two divergent approaches for estimating teams' rest in a sporting context. The first method adopted a short-sighted stance, defining rest by only counting the number of days since the previous game was played. Instead of just depending on the days since the last game, the second assessment strategy relied on counting the previous matches played during a fixed period (expressed in days).
Independently of the employed process, all the rest features were constructed by also accounting for matches played in cup competitions (both domestic and European). This choice differed from comparable decisions when gathering data in other sub-datasets, which excluded any cup data unequivocally. The main reason for such a different approach was that cup observations have been proven unreliable or just missing only when considering detailed information, which had no use in this case. However, these observations served the only purpose of feature creation and were not evaluated in the classification problem.
We arbitrarily chose numerous time intervals when contemplating the second definition of measuring rest. The picked time intervals were: one week, two weeks, one month, and two months. Each feature was computed separately for the home and away teams, accounting for ten total team rest features (five for each club). Five additional variables were formed, precisely the difference between each corresponding variable of the home and away teams.

Unlike the rest attributes, the form features were developed by examining the critical result-based information of each match to portray a team's past games performance. Practically, these variables were constructed by averaging the result-based key measures' values over a fixed number of previous observations. In our analysis, the four variables deemed as most valuable to summarize a team's form were: points, goals scored, goals conceded, and goal difference. Comparably to the standings dataset (see Section 3.3.2), each new feature type was computed twice for each team, first by considering all previously played games, then making a venue-based distinction, i.e., considering only matches played at home for the home club and away contests for the visiting squad. Furthermore, since, as recounted above in the rest data description, observations from cups were reliable and accurate for basic match information (such as scores and results), they were also employed in the computation of a features' batch. For each feature kind, the averages were calculated once only focusing on matches from our top-five leagues and once also regarding events from cups. The logic behind this approach was to account for the league-specific and cup-specific forms.

According to Buursma (2010), which used only form-based features for football games prediction, this group of attributes was more informative (delivering better performances) when regarding a more extended period (i.e., averaging over more matches) rather than a short-term perspective. Indeed, by computing the mean values over different periods for comparison, Buursma (2010) discovered that the best classification accuracy was achieved when considering the 20 previous games.

Following the literature, we performed the averages over multiple subsets of earlier contests, each concerning various instances. The adopted subgroups comprised the last one, three, five, ten, and twenty games.

For the initial observations in our dataset (beginning of the 2016/2017 season), the computations were accomplished using data from the 2015/2016 season, which is not directly part of our research coverage.

The described methods produced 160 form-based features, 80 for each team. Moreover, 80 additional variables were formed as the difference between home and away team corresponding attribute pairs.

In its totality, our form and rest dataset enclosed 255 numerical features (240 formbased and 15 rest-based). Moreover, no nominal variable was gathered, excluding the need for any categorical encoding technique within the grid search process. The dataset comprised 10,438 match instances out of the 10,536 considered globally in our study. A slight decrease in the dataset size occurred due to the nature of the feature extraction approach, which involved averaging values by examining past events. As a result, this method occasionally yielded missing values, primarily for newly promoted clubs where previous game information was unavailable.

### 3.3.4 Stats Dataset

The reasoning behind the stats dataset was parallel to the form and rest dataset, i.e., to inspect the impact of match-specific information from past games on current contests' results. However, while Section 3.3.3 was established on only elementary result-based knowledge (scores and outcomes), the stats dataset implemented detailed information on in-game events. This specific type of data is usually described as football statistics or simply stats, hence the name of our dataset.

Our analysis gathered football stats at a match level from the Sportmonks API. Turning them into valuable and unbiased information (i.e., avoiding look-ahead bias) was achieved with an analogous process to the one utilized for the form data (see Section 3.3.3). Therefore, the football statistics variables were constructed, in practice, by averaging their values over a fixed number of previous games. Another similarity with the feature creation process of the form data, borrowed from the tables' dataset approach, was that each football statistic feature was assembled twice for each team: once viewing all previously played matches without any distinction and once considering only home games for the home club and away games for the visiting club.
However, unlike the form and rest data preparation, no information from cup matches was used for the stats data. This impossibility in considering any cup-specific game was caused by its incompleteness and inaccuracy regarding detailed information such as

football statistics. Similarly, observations from the 2015/2016 season (also used for the form features preparation at the beginning of the 2016/2017 season) were incompatible with stats data due to many missing values. Indeed, the principal reason for choosing the seasons from 2016/2017 to 2021/2022 in the first place was the integrity and general quality of this sample's data from the Sportmonks API, especially compared to earlier football seasons. Another difference from the methods described for the form data (in Section 3.3.3) was using shorter ranges of past matches for calculating averages. Not having 2015/2016 season data to rely on to fill initial observations in our sample, the motivation behind this approach was to lose as few observations as possible, aiming to acquire relevant new information without excessively compromising our data's dimension or structure. In particular, our stats features were obtained from the last: one, three, and five games (form data also had considered the previous ten and twenty).

A total of 20 football stats were included in our analysis, including shots (on goal, off goal, blocked, inside the box, outside the box, and total), passes (total, accurate, and percentage), possession time, corners, offsides, tackle, fouls, cards (yellow, red, yellow card accumulation), goalkeeper saves, counts for total attacks and dangerous attacks only. Before computing the averages, to acquire as much knowledge as possible from football statistics, we created two additional sets of features relating to a team's performance relative to the opposing club. Specifically, pre-averaging differences were computed for each of the stats described above by subtracting away values from home values and viceversa.

When aggregated, the procedures described in this section generated 480 stats-based features. Additionally, 240 attributes were created as the post-averaging differences between home and away team corresponding variable pairs. Post-averaging differences were used to compare the past performances of the two clubs regardless of the opposition's behavior. In contrast, pre-averaging differences aimed to identify a team's past performance against previous opponents. It should be noted that these distinct methods of computation were not mutually exclusive and could be used in conjunction to provide a more comprehensive analysis of the teams' past performances.

In its totality, our stats dataset held 720 numerical features. Thus no categorical encoding was required during the subsequent grid search. The dataset contained 10,205 match instances out of the 10,536 considered globally in our study. Like the form and rest dataset, the feature extraction strategy caused this slight decline in the dataset size. Moreover, this dataset's reduction was more prominent due to the unattainability of statistics data from the 2015/2016 season.

### 3.3.5 Odds Dataset

The odds dataset was designed to evaluate the predictive value of pre-match betting odds for game results. For this purpose, no data from in-play betting was taken into account since we exclusively focused on knowledge generated before the start of a game to avoid look-ahead bias. In the literature, a few researches were already conducted employing similar information. Specifically, Odachowski & Grekow (2012) and Tax & Joustra (2015) only used details on match result betting (1X2 bets), which refers directly to the outcome of a match as defined in our multiclass problem (home win, away win, or draw). Moreover, previous comparable analyses were solely based on data retrieved from a single source, meaning that the betting odds for every instance came from a single bookmaker.

Our approach differed considerably from the priors due to the nature of the betting information recovered from the Sportmonks API. This analysis employed extensive knowledge of multiple betting odds types (outside of just 1X2) coming from various sources (i.e., bookmakers). Therefore, we aimed to exploit our data's more comprehensive details and assess its predictive power in our classification design.

Before any data preparation effort, the granularity of the raw betting odds information denoted its primary strength and, when added to its variability, the main obstacle to a straightforward evaluation. Our data required not only grouping odds of the same type together but also a method to control for the inconsistency shown in

observations' sources, meaning that different instances were likely to present information from other bookmakers.

As the proposed solution to regulate the intrinsic information variability, we regarded the odds coming from diverse bookmakers but concerning the same odd type as part of a betspecific distribution. For each match, this approach considered as many different distributions as the sum of the distinct possible outcomes in each considered odd type. For example, for the 1X2 type, three distributions were considered, one for each mutually exclusive result (i.e., home win, away win, or draw). The knowledge incorporated in the generated distributions was then translated into practical information by computing summary statistics, which constituted the final features of our odds dataset. Specifically, our bet-specific summary statistics were: the mean, the standard deviation, the variance, the quartiles (lower, median, upper, and interquartile range), and the maximum and minimum values (and their range).

The combination of betting odds allowed us to account for the different sentiments of various bookmakers and bets' variance, thus exploiting, at a small scale, the wisdom of the crowds.

After describing the methods employed to transform the initial knowledge into a usable and informative dataset, some reflections regarding the raw betting odds were required. First, the information acquired from the Sportmonks API was incomplete. Prematch odds were not available for the early stages of the 2016/2017 season, and also missing values were sporadically encountered for matches from later seasons. Second, distinct odd types or sources representation could be unequal and inconsistent across observations, even when provided with detailed data. Indeed, the more common betting markets and the more prominent bookmakers were less likely to be missing and were present more invariably.

Since, as said above, the incompleteness of the initial data already remarkably reduced the number of instances, we decided not to consider those produced features containing an excessive number of missing values to limit the underlying reduction in the dataset size. As a consequence, among the more than a hundred diverse betting odd types inspected during this analysis, the only ones ultimately considered for the final dataset were: 1X2, over/under, double-chance, first and last team to score, team to score, both

teams to score, highest-scoring half, win both halves, clean sheet, win to nil, the exact number of goals, correct result, three-way handicap (which involves assigning a handicap advantage or disadvantage to one of the competing teams), and Asian handicap (a specific type of handicap betting that seeks to remove the likelihood of a draw by awarding a head start to the underdog).

The final odds dataset had 776 numerical features. Due to the previously described presence of missing values, the set holds 8,415 observations out of the 10,536 instances considered globally in our analysis. The decrease in the data size was significant. However, it is crucial to notice how considerably the API betting odds data quality and completeness improved over the seasons. Mainly, this issue would have been reduced if we chose to adopt this same analysis only for later seasons or in the future.

### 3.3.6 Team Attributes Dataset

The motivation behind the team attributes dataset was to evaluate the predictive value of team-specific features in the context of our study. Specifically, we assessed the information regarding team ratings, conceptually similar to Prasetio (2016), and additional features referring to general team knowledge.

In line with Rudrapal's (2020) methodology, we considered information on team attributes from the FIFA video game series. As described in Section 3.1, this data was obtained via web scraping of the FIFA Index website and subsequently stored in a database. This database held team characteristics for the top-five European leagues evaluated in our analysis. However, to account for the promotion and relegation process ingrained in football leagues, the database also included the latest seasons' observations for each minor league team. In particular, these five minor leagues were Football League Championship (England), Bundesliga 2 (Germany), Serie B (Italy), Segunda Division (Spain), and Ligue 2 (France). When examining temporal characteristics, the database also included data from the latest available instances of the 2015/2016 season (previous

to our analysis time frame), allowing the evaluation of games belonging to the 2016/2017 season that occurred before any FIFA information for the stated season was recorded. The described incorporation of data related to minor leagues and the 2015/2016 season was conducted to avoid yielding any missing value.

The team attributes dataset was assembled by merging general match knowledge with the related data from the FIFA team database, joining clubs involved in a match with their team-specific attributes. This process was designed to avoid look-ahead bias by connecting a team with its nearest but prior observation (recorded before the match date) from the FIFA database. Moreover, a revision process was employed to establish consistency in team naming from the two sources and prevent the output of unwanted missing values during the merge. Since the number of different clubs was not elevated, these corrections were performed manually.

After the described merge and additional preparation steps, the team attributes dataset contained 28 features related to various team characteristics.

A total of 14 features were derived from FIFA video games, seven for the home team and seven for the away team. Each team's attributes included two categorical variables and five numerical variables.

One of the two binary variables indicated if the opposing squad coincided with the rival team, while the other dummy distinguished newly promoted clubs (from the same season a particular game was played in).

The five numerical features were the following: ratings for teams' defense, midfield, and attack (ranging from 0 to 100); a mock transfer budget; the width of teams' formation, which required converting to a new scale since the format of this attribute changed over the years.

Furthermore, six variables came from the original match data, three for each team (two numerical and one categorical). The two numerical attributes respectively referred to the year a club was founded and the age of the team's coach (assumed as an approximation of experience). The categorical variable denoted the squad formation, meaning players' relative position on the pitch.

For each of the seven pairs of related numerical features, a new attribute was created as the difference between the home and away values. Since other pairs' ratios were perfectly correlated with their differences, this additional variable was computed only for the transfer budget pair.

As mentioned above, the team attributes dataset had 28 features (six categorical and 22 numerical). Furthermore, it held 10,536 observations, covering the entirety of the matches considered in our analysis, thus implying the absence of any missing value.

### 3.3.7 Player Attributes Dataset

The creation of the player attributes dataset followed a similar rationale to the team attributes data (described in Section 3.3.6). While the latter attempted to evaluate the predictive value of team-specific information, the former focused on knowledge at the player level. A relative advantage of this last approach was the possibility of filtering just the players involved in a specific match. Precisely, our study assessed only players in the starting lineup since substitutes could not be known before the start of a game and therefore caused look-ahead bias.

Like Danisik (2018), our analysis employed player attributes from the FIFA video game series. The player data was retrieved from the same source and using similar methods of the team attributes. Like the team database and for the same underlying reasons described in the previous section, the players' database also contained data from the latest available instances of the 2015/2016 season. However, to maintain a manageable data size, observations from minor leagues were not retrieved for players due to the already massive size of the top-five leagues' data.

The methodology for creating the player attributes dataset was conceptually similar to the team attributes dataset. However, the process of merging general match knowledge with the FIFA player database had to be tailored to the specific characteristics of the players' information.

While inconsistencies in naming were managed manually for the teams' dataset, this approach was not possible for the players' dataset due to the large number of players. Hence, we automated the procedure to merge data from the two sources while managing inconsistencies in players' names between them. Whenever a connection based on the player's full name was unattainable, the program attempted to find an association by employing other looser characteristics. In such cases, the program searched sequentially inside a specific club for the presence of a player with the same last name or kit number. If an applicable connection was still not found, a specular search was carried out by skimming forward in time for the closest observation instead of looking only for prior observation as the previous method. If neither of the specular searches found a link between the two sources, the program would pick a random player belonging to the specific team and covering the same position on the field.

Although this strategy could produce look-ahead bias, it was the only solution to avoid missing values in the case of a newly transferred player from a league outside our analysis. While most of the connections were based on the player's full name, the automated procedure led to inconsistent associations, meaning merging the data on two distinct players, with a probability of around 1-2%. Furthermore, less than 1% of the links were found in the forward search producing any look ahead effect. As a result, this merging procedure caused some inaccuracies and biases, but due to their rarity, these minor issues were tolerated to ensure the overall utility of the resulting dataset.

After the merge, the dataset comprised player attributes contained in the FIFA database for each of the 22 players belonging to the starting lineups of a game. These attributes were then averaged across players of each team to deliver valuable features. In particular, most attributes were averaged over a team's entire lineup, while others were calculated by considering smaller groups of players within the lineup.

Twenty-two features were calculated by taking the average of each attribute across all players in the lineup. These attributes included both general details such as age, height, weight, value, wage, overall and potential player rating, as well as specific categories including psychological traits (aggression, reactions, attacking position, interceptions, vision), passing attributes (crossing, short pass, long pass), and physical characteristics (acceleration, stamina, strength, balance, sprint speed, agility, jumping).

Eight features resulted from averages only among midfielders and strikers: ball control, dribbling, heading, shot power, finishing, long shots, curve shots, and volley shots. Additionally, three features were obtained by averaging among only defenders and midfielders: marking, slide, and stand tackle.
Five features pertained exclusively to the goalkeeper's attributes and therefore did not require any computation: positioning, diving, handling, kicking, and reflexes. Alternatively, the features based on free-kick accuracy and penalties were formed by considering only the highest value among the players in the lineup since specialized players typically perform these actions.
Lastly, a feature was established by counting the number of players in a club who also played for their national team, hence identifying the proportion of players that represented their country internationally.
In total, 41 features were developed for each club. Additional features were computed as the differences of corresponding attribute pairs between the home and away teams.

The player attributes dataset comprises 123 numerical features and 10,536 observations. These observations encompass all matches included in our analysis, indicating the absence of missing values. Although our data collection process incorporated some forward-looking instances during the merge, it was still characterized by less look-ahead bias compared to previous studies that used player data referring to the end of each season or, as in the case of Rudrapal (2020), at the end of the specified analysis period.

## 3.4 Complete Dataset

The complete dataset was generated by combining all sub-datasets described in Section 3.3. This dataset had 2,105 features (including 30 categorical variables) and 8,241 observations out of the 10,536 instances considered globally in our analysis. The size reduction resulted from the prior considerations made separately for each sub-dataset.

The comprehensive feature space's dimensionality is a novel aspect compared to previous football prediction studies, which have typically considered more undersized feature sets. However, the high dimensionality of the dataset presented challenges. Therefore, dimensionality reduction techniques were deemed essential and implemented as a critical step in developing all models within our analysis.

# Chapter 4

# Results and Discussion

This section describes the highest performances obtained by evaluating each sub-dataset (presented in Section 3.3) and the comprehensive set (see Section 3.4), including the preprocessing techniques and classification algorithms employed. As the performance metric, we assessed test accuracy to evaluate the performance of models on different datasets since it was used in most studies within this specific application domain. This metric compares the predicted results to the actual results observed in the test set, computing the share of observations correctly classified by the model.

## 4.1 Results for Sub-Datasets

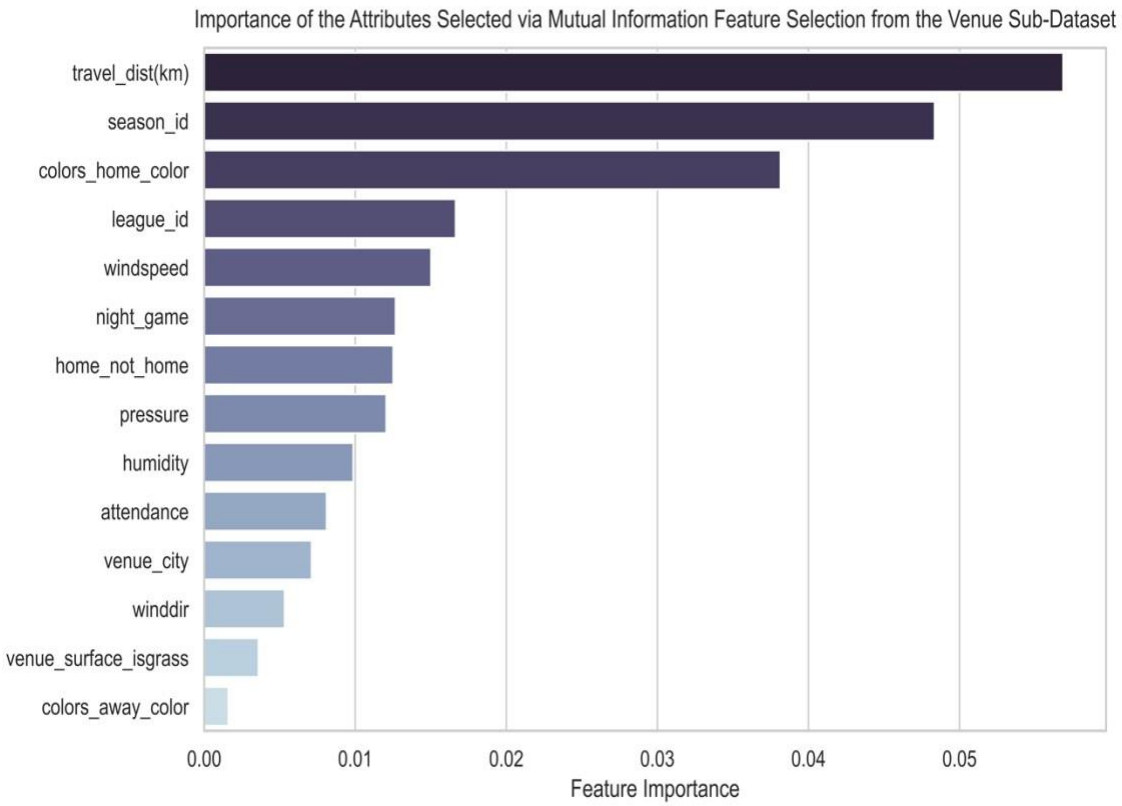
***Results for Venue Sub-Dataset***

Within the grid search, a Random Forest classifier achieved the best test accuracy of 48.7% for the venue sub-dataset when combined with ordinal encoding for categorical variables, normalization for feature scaling, and Mutual Information feature selection (see Section 3.2). Specifically, 14 features were chosen out of the initial 23 attributes. Figure 2 illustrates the importance score of the selected features, with features' names following their description given in Section 3.3.1. The three attributes identified as the most important within the venue sub-dataset were the distance traversed by the away

team (computed as the distance in kilometers between the venue of the away club and the actual match venue, typically the home team's venue), a season identifier (denoting the specific season during which the match was played), and the dominant color of the home team kit.

Despite having the lowest performance among all the other feature subsets, the model trained on the venue sub-dataset significantly improved the test accuracy over the benchmark model's performance. The benchmark model solely predicts the majority class in the dataset, which is a home win, achieving an accuracy of 44.3%. In contrast, the model trained on the venue sub-dataset yields a substantially higher accuracy rate, demonstrating the effectiveness of this feature subset in predicting the outcome of football matches.

**Figure 2: Importance of Venue Sub-Dataset Features**



***Results for Standings Sub-Dataset***

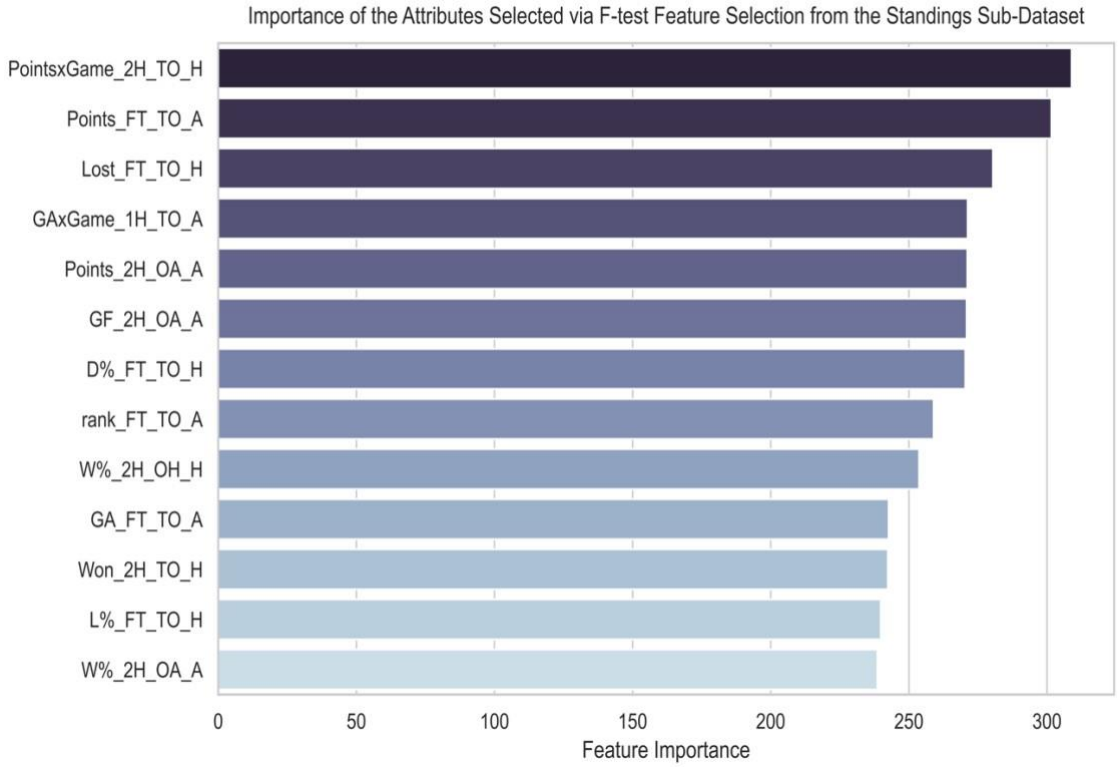
For the standings sub-dataset, a Multinomial Logistic regression classifier achieved the best test accuracy of 53.9% when combined with normalization for feature scaling and F-test feature selection (presented in Section 3.2). No categorical encoding method was evaluated in the grid search since all the nominal variables in the feature set were ordinal by nature (features referring to the league standings rank position).

Figure 3 depicts the 13 features selected from the 180 initial dataset's attributes, sorted by importance. The displayed features' names were composed to convey all the specific information by using underscores to separate the different characteristics of each specific feature. The first term of each name represents the feature type (see Section 3.3.2), e.g., GF stands for the goals scored expressed in absolute terms, while GAxgame denotes the goals conceded per game. The second term of each name defines the period over which each feature was computed; e.g., 1H and 2H stand for the two distinct halves of a game, while FT (full-time) considers a game in its entirety. The third term of each name expresses if the feature was calculated over only home games (OH), only away games (OA), or the totality of games without any venue-based distinctions (TO). Finally, the last term expresses whether the particular feature belongs to the home (H) or away (A) team. Of the 13 selected features, the distribution of attributes between the home and away team is relatively balanced, with six features related to the home team and seven related to the away team. Most selected features pertain to measures across all matches without distinction based on the venue (nine variables). Notably, only six variables provide information about entire games, while six attributes are specific to the second half of games, and one pertains solely to the first half. This finding emphasizes the importance of result-based information obtained exclusively from the second half of the games compared to that obtained from the first half.

Remarkably, the test accuracy of the standings-based model was similar to that of McCabe & Trevathan's (2008) model, which achieved an accuracy of 54.6% while considering only non-sport-specific features. These non-sport-specific features included

not only league standings-related features but also additional attributes, some of which were contained in the venue sub-dataset, such as travel distance. Such results suggest that the knowledge from the standings sub-dataset can offer valuable insights into predicting the outcome of football matches, despite consisting of superficial information.

**Figure 3: Importance of Standings Sub-Dataset Features**




***Results for Form and Rest Sub-Dataset***

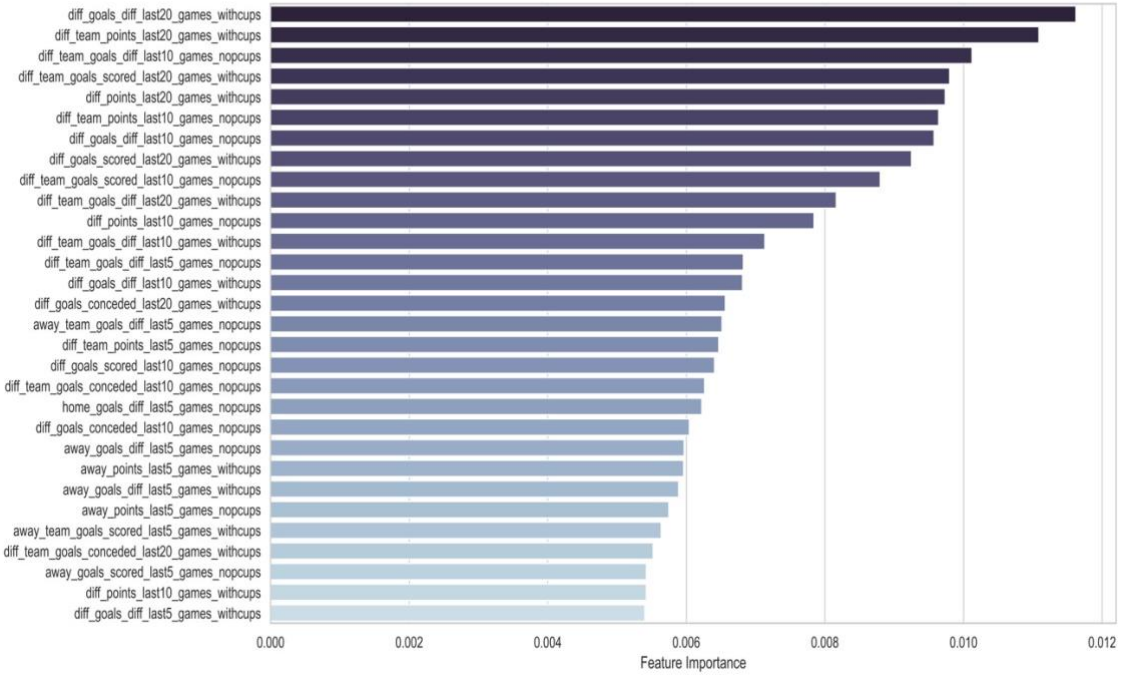
Within the grid search, a k-Nearest Neighbors classifier achieved the highest test accuracy of 54.5% for the form and rest sub-dataset, combined with normalization for feature scaling and random forest-based feature selection (see Section 3.2), which appointed 64 out of the 255 initial features. Since no nominal variables existed in the feature space, categorical encoding techniques were not considered during the grid search.

Figure 4 illustrates only the 30 most important features (for readability reasons) employed by the model. The features' names indicate each feature kind (described in detail in Section 3.3.3), including whether the feature was constructed as a difference

between the home and away value (by adding the prefix "diff"). Moreover, it is also specified if a feature considers only home or away games or all games without venue-based distinctions (in these cases, the feature names contain the word "team"). Additionally, the names exhibit the number of previous games considered when computing the average and whether cup matches were included in such measure's calculation.

**Figure 4: Importance of Form and Rest Sub-Dataset Features**

Importance of the Attributes Selected via Random Forest-Based Feature Selection from the Form and Rest Sub-Dataset



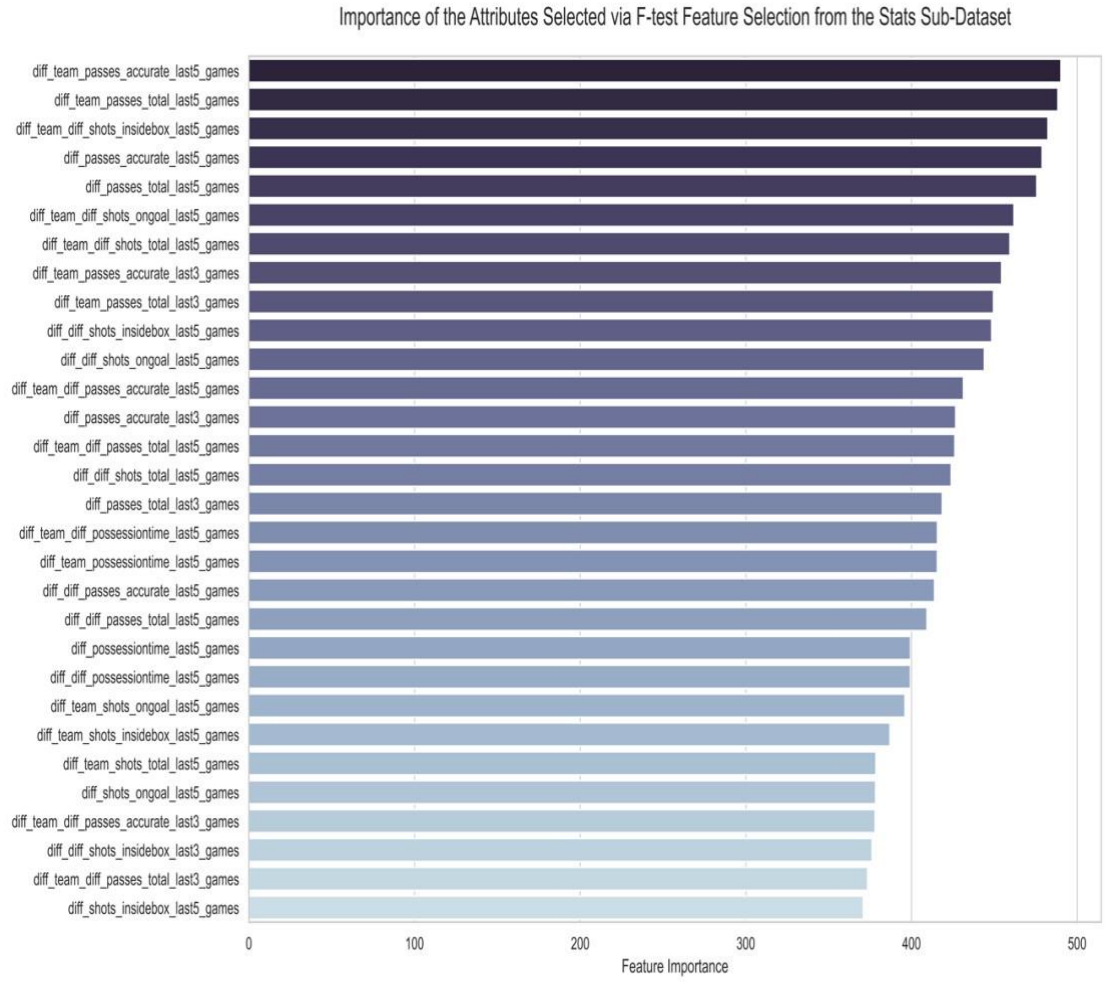

The model's test accuracy of 54.5% was comparable to the performance reported by Buursma (2010) in a similar analysis of form-related features on the Eredivisie Dutch league. Buursma (2010) discovered through experimentations that using form features calculated by averaging over the previous twenty games produced better prediction results than those considering a smaller number of prior instances. Our findings partially supported this conclusion, as the features with higher importance scores in Figure 4 were those calculated over twenty or ten previous observations. Remarkably, none of the restrelated features in the sub-dataset were selected and incorporated into the model, indicating their lack of significance in our prediction task. Moreover, all the 15 most important features and most other selected variables refer to differences between the home and away team values.

***Results for Stats Sub-Dataset***

For the stats sub-dataset, a Multinomial Logistic regression classifier achieved the best test accuracy of 52.7%, combined with standardization for feature scaling and F-test feature selection (described in Section 3.2), designating 88 out of the initial 720 features. As for the form and rest sub-dataset, no nominal attributes were present in the feature set, eliminating the need for any categorical encoding technique in the grid search.
For readability reasons, Figure 5 represents only the 30 most important features the model considered and their importance score. The attributes displayed in Figure 5 share parallel naming conventions with Figure 4. The only difference in features' names is that cup instances were omitted when computing features of the stats sub-dataset, hence bypassing the necessity to specify whether features contain these observations (see Section 3.3.4). All the 30 most important features were determined by computing differences between the home and away team values. These features pertained to either shots, passes, or possession time measures. Furthermore, 23 of the most noteworthy attributes were calculated as averages over the last five games, while seven were computed by averaging over the previous three games.

Notably, the model's test accuracy of 52.7% was significantly lower than the accuracy obtained by the model based on the form and rest set (54.5%). This observation raises questions about the differences in the two types of knowledge used to compute the features. While form features are established on only basic result-based knowledge, such as scores and outcomes, stats features contain detailed information on in-game events. One possible explanation for such a counterintuitive result is that, as described in Section 3.3.4, a smaller range of prior instances was considered in the computation of the stats features compared to the form attributes. Alternatively, this could result from the intrinsic characteristics of the two types of football-related knowledge. Further investigation is needed to understand the causes of the differences in performance, which could be the focus of future studies in this area.

**Figure 5: Importance of Stats Sub-Dataset Features**




***Results for Odds Sub-Dataset***

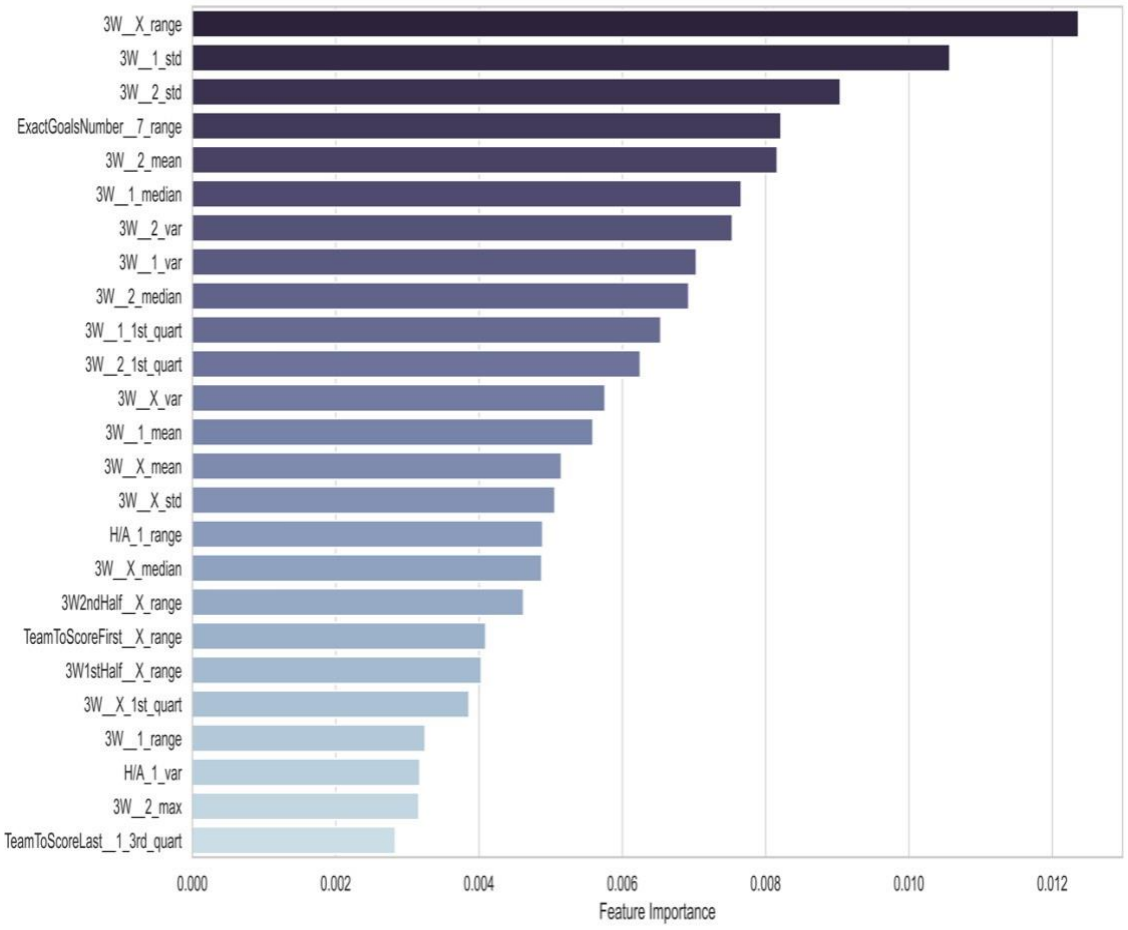
Within the grid search, a k-Nearest Neighbors classifier achieved the highest test accuracy of 60.0% for the odds sub-dataset (see Section 3.3.5), combined with standardization for feature scaling and random forest-based feature selection (described in Section 3.2), which selected 25 out of the 776 initial features. Categorical encoding techniques were not considered during the grid search since no nominal variables existed in the feature set.

Figure 6 displays the importance scores for the 25 selected features. Each feature's name contains a specification of the betting market as the first term (e.g., 3W stands for threeway betting), the specific outcome considered by the bet after a double underscore

(e.g., one for a three-way bet implicates a home team win), and, as the last term, the summary statistics calculated when constructing the specific feature (e.g., mean or standard deviation). Notably, the most critical features in Figure 6 represent bets in the three-way betting market. Specifically, 20 of the 25 selected features were found to be related to this market. This result was not unexpected since three-way betting shares the same possible outcomes as our classification problem, namely home win, away win, and draw.

**Figure 6: Importance of Odds Sub-Dataset Features**

Importance of the Attributes Selected via Random Forest-Based Feature Selection from the Odds Sub-Dataset




The odds model significantly outperformed all the models trained on individual subdatasets in our analysis. Moreover, the odds model substantially outperformed other models that exclusively considered betting odds information in previous studies, such as

Odachowski & Grekow's (2012) model (accuracy of 46%) and Tax & Joustra's (2015) model (accuracy of 56.1%), in terms of test accuracy. The high predictive value of the betting odds knowledge can be attributed to two primary aspects of our analysis. Firstly, the quality of the collected betting odds data from the Sportmonks API may have contributed to this result. Secondly, the strategy adopted in the feature extraction and preparation described in Section 3.3.5 may have also played a crucial role.

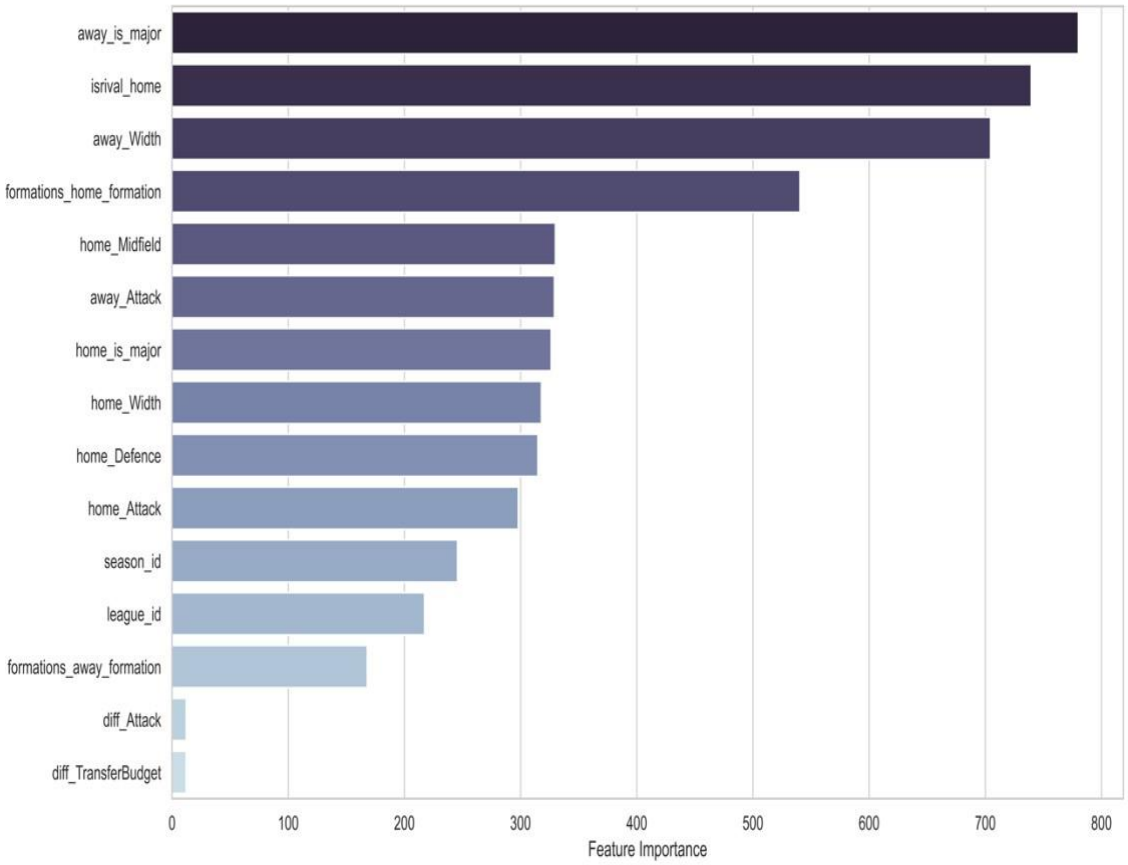
***Results for Team Attributes Sub-Dataset***

Within the grid search, a k-Nearest Neighbors classifier achieved the best test accuracy of 54.1% for the team attributes sub-dataset when combined with One-Hot encoding for categorical variables, standardization for feature scaling, and F-test feature selection (presented in Section 3.2), which designated 15 out of 28 features.
Figure 7 illustrates the importance scores of the selected features from the team attributes sub-dataset, with features' names closely following the description in Section 3.3.6. Within the team attribute sub-dataset, the two most crucial features referred to whether the away team had participated in the minor league during the previous season and whether the away team was the rival club of the home team.

Unfortunately, comparisons with similar models from the literature are unfeasible since the only paper considering a model based solely on team-related features, by Prasetio (2016), approached the classification problem as a binary classification (i.e., excluding draws) instead of a multiclass problem. Nevertheless, the team attributes model's performance is comparable to other models in this section.

**Figure 7: Importance of Team Attributes Sub-Dataset Features**

Importance of the Attributes Selected via F-test Feature Selection from the Team Attributes Sub-Dataset



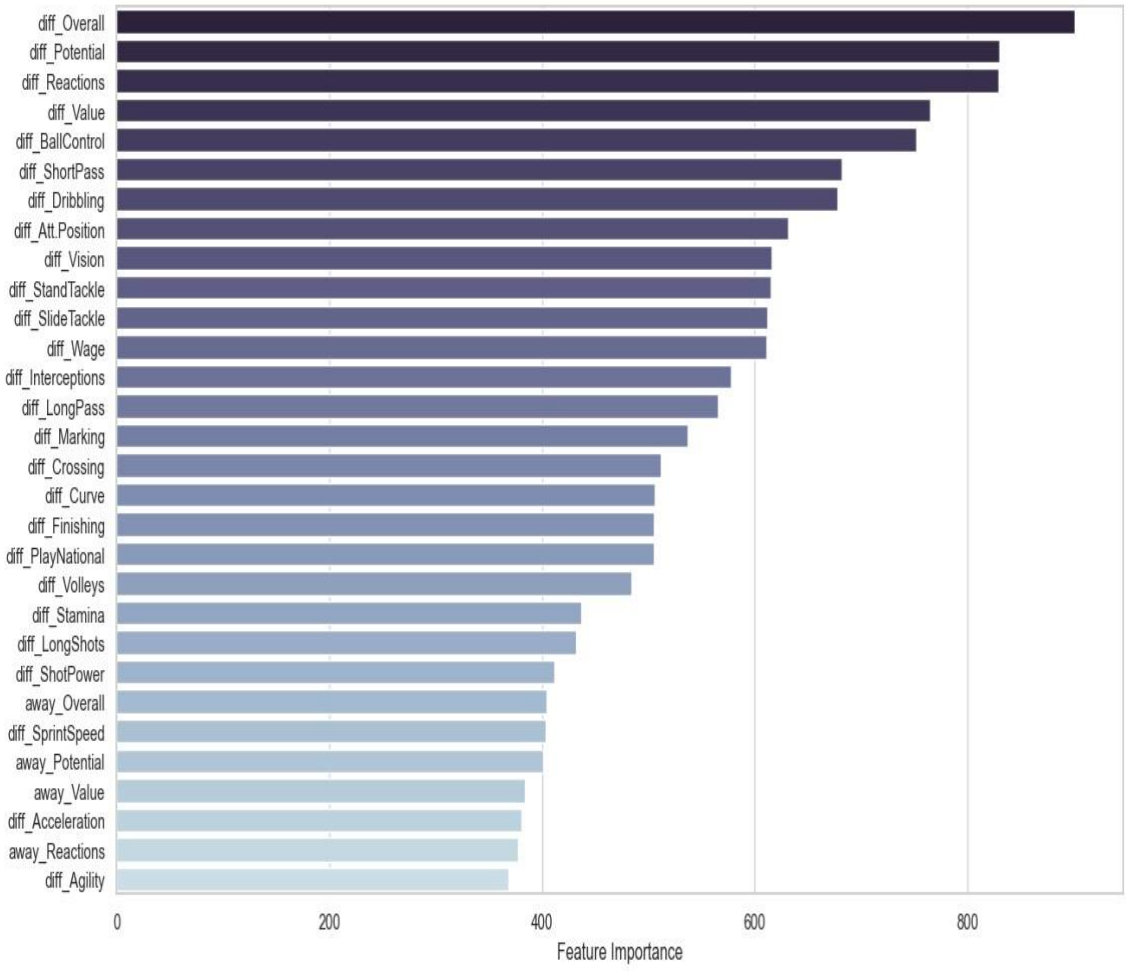

***Results for Player Attributes Sub-Dataset***

For the player attributes sub-dataset, a Multinomial Logistic regression classifier achieved the best test accuracy of 54.2%, combined with One-Hot encoding for categorical variables, standardization for feature scaling, and no feature selection technique since a better performance was accomplished using the entire feature space instead of a subset (see Section 3.2). Due to the characteristics of the One-Hot encoding method for nominal variables (described in Section 3.2.1), the feature set employed by the model was composed of 167 attributes generated from the initial 123 variables. Figure 8 represents the top 30 features of the player attributes sub-dataset by their importance. The features' names follow the description in Section 3.3.7, with the prefix "diff" indicating that a particular feature was generated as the difference between the home and away value.

Figure 8 represents the top 30 features of the player attributes sub-dataset by their importance. The features' names follow the description in Section 3.3.7, with the prefix "diff" indicating that a particular feature was generated as the difference between the home and away value. Four features stood out as the most important among the player attribute sub-dataset. These included the difference between the average values of the players' overall ratings for the home and away teams, their potential ratings, their reaction ratings, and their market value.

**Figure 8: Importance of Player Attributes Sub-Dataset Features**

Importance of the Attributes Selected via F-test Feature Selection from the Player Attributes Sub-Dataset




Notably, the model's test accuracy of 54.2% was comparable to the model's performance reported by Danisik et al. (2018) of 52.5%, employing both player-level and match-specific data. The similar performance of the player and team attributes models

conveys valuable insight. Notably, the player attributes sub-dataset contains knowledge of the specific lineup involved in a game. In contrast, by definition, the team attributes sub-dataset is constituted of measures referring to a team's entire roster. The similarity between the predictive value of these two football-related information types suggests that their contained knowledge is equivalent.

## 4.2 Results for Comprehensive Dataset

For the comprehensive dataset, a Multinomial Logistic regression classifier achieved the best test accuracy of 60.7%, combined with One-Hot encoding for categorical variables, normalization for feature scaling, and random forest-based feature selection (see Section 3.2), which selected 110 out of the 2,105 initial features.

**Figure 9: Confusion Matrix for the Comprehensive Model**

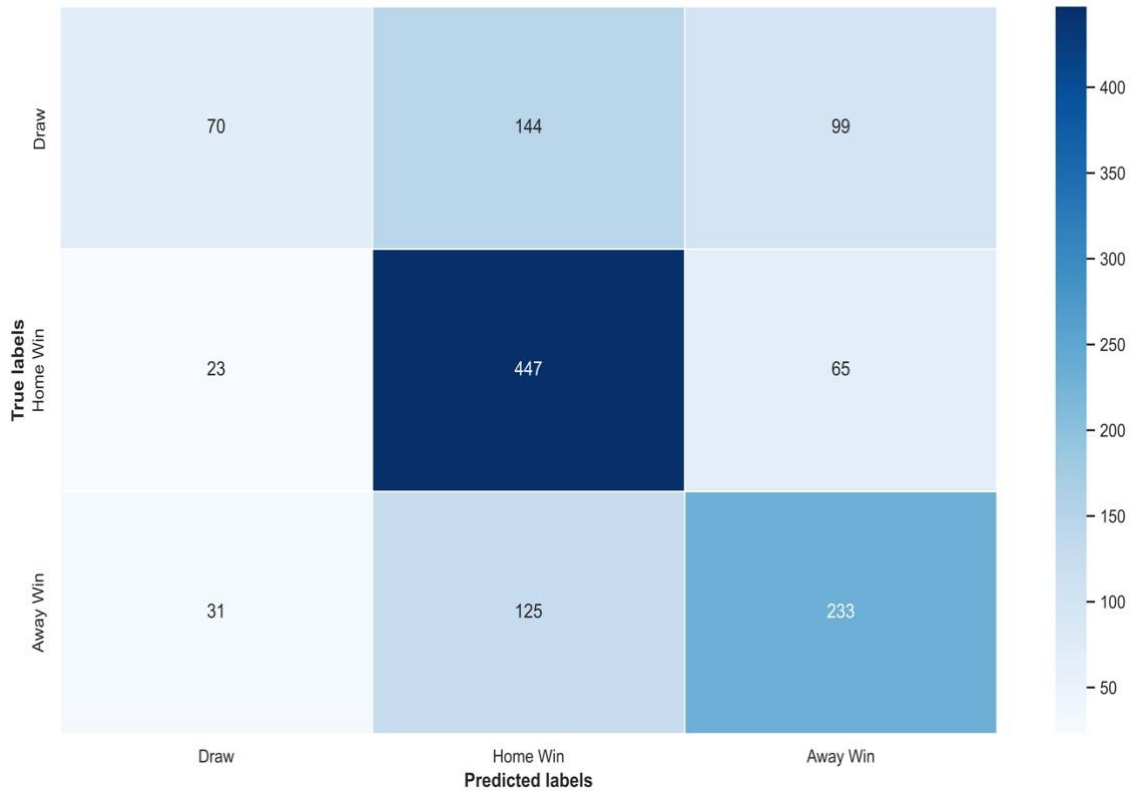


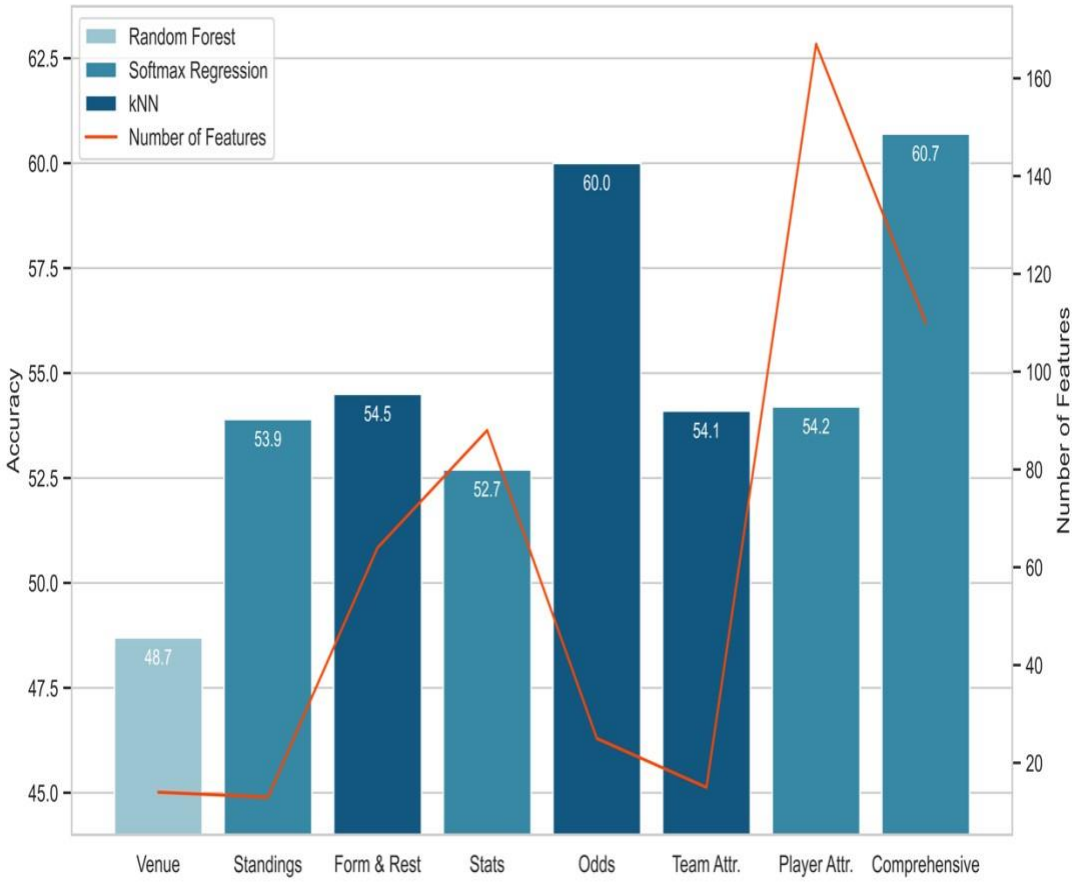
Figure 9 represents the confusion matrix for the best model trained on the comprehensive dataset, which is a tabular summary of its performance. The confusion matrix provides detailed information on the prediction performance for the various classes in the target variable (home win, away win, and draw) by comparing its actual labels with those predicted by the classification model, hence identifying areas where the model may underperform. As emphasized in the literature and defined in Section 2, football match prediction presents unique challenges due to its low-scoring nature, especially when considering draws in a multiclass design. The main issue in multiclass football match prediction is the complexity added when considering draws, which is the class where the comprehensive model severely underperforms.

Figure 10 offers valuable insight into the effectiveness of different types of footballrelated knowledge in predicting outcomes. Specifically, it enables us to compare and rank the predictive power associated with various feature sub-spaces considered in this study. The chart illustrates the best performances obtained using different feature sets, namely, the sub-datasets presented in Section 3.3 and the comprehensive dataset described in Section 3.4. The plot also provides information on the number of features selected via dimensionality reduction methods and classification learning algorithm, which yielded the best results. Interestingly, the best performances across different feature sets were obtained predominantly by implementing either Multinomial Logistic regression classifiers or k-Nearest Neighbors classifiers, while only once with a Random Forest classifier and never with the other classification algorithms mentioned in Section 3.2.2. Furthermore, it is worth noting that feature selection methods outperformed feature extraction techniques using Principal Component Analysis in all the analyzed feature sets. These findings suggest that Multinomial Logistic regression or k-Nearest Neighbors classifiers, combined with feature selection methods, are most effective in predicting football outcomes. In contrast, feature extraction techniques using PCA may be less effective in this context.

As shown in Figure 10, the model trained on the comprehensive dataset outperformed all the models trained on sub-datasets presented in Section 4.1. This model also performed better than most prior multiclass match prediction studies in the literature.

**Figure 10: Comparing Model Accuracies Across Datasets**

Comparison of the Model Performances Using Different Feature Sets, Indicating Classification Algorithm Employed and Number of Selected Features

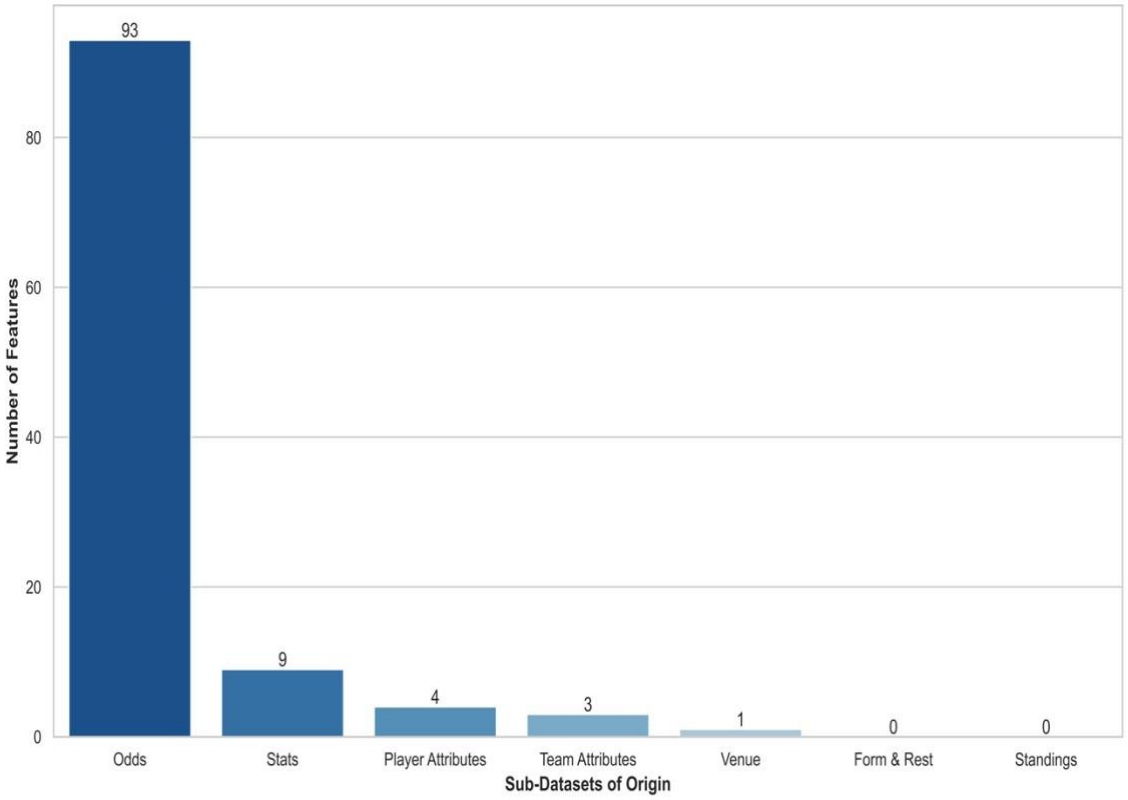



In the literature, betting odds have also been helpful for match result prediction, even when included as the sole model predictor by Tax & Joustra (2015). From our analysis, the model's performance using the odds sub-dataset confirms such a conclusion.
However, Hubáček et al. (2019) suggested that the betting odds should be omitted as a model predictor if the aim is to generate profit through betting strategies. Notably, our comprehensive model heavily relies on betting odds data, as shown in Figure 11, which illustrates the distribution of the 110 selected features for the comprehensive model based on their type of football-related knowledge. Therefore, testing the comprehensive model's effectiveness in yielding a profit via betting strategies represents a pragmatic continuation

of this study, aimed at questioning the deduction made by Hubáček et al. (2019). Moreover, this research could provide a helpful roadmap for navigating the vastness of football-related information for future studies, primarily through Figure 9's content. Another possible addition to the present study could be the consideration of a dataset generated by conducting sentiment analysis for football match prediction, as explored by Godin et al. (2014) and Schumaker et al. (2016).

**Figure 11: Number of Features by Sub-Dataset of Origin**

Count of the Features Included in the Comprehensive Model by Sub-Dataset of Origin (Football-Related Information Type)



# Chapter 5

# Conclusion

In this study, we aimed to expand upon existing research on football match prediction by using a more extensive feature set and designing a strategy to compare the predictive value of different forms of football-related knowledge. Our results demonstrated that the model using the comprehensive feature space achieved a test accuracy of 60.7%, outperforming many prior studies on multiclass match prediction. However, the comprehensive model's performance corroborated the earlier literature's reflections on the intrinsic complexity in predicting football match outcomes in a multiclass configuration due to complications added when considering draws which was the class where our model severely underperformed. We also partitioned the comprehensive dataset into seven smaller feature sets based on shared information-specific characteristics, using a standard set of preprocessing techniques and classification algorithms that allowed us to compare the predictive power of various types of knowledge. Notably, our findings showed that betting odds were the most valuable information type for the prediction task, supporting prior research by Tax & Joustra (2015).

Although we engineered a richer set of features compared to previous literature, future research should focus on further expanding the collection of relevant model predictors through domain knowledge. Additionally, alternative machine learning algorithms could also improve models' performances.

In conclusion, our study has demonstrated the value of using a more comprehensive feature set to predict football match outcomes and the importance of incorporating betting odds as a predictor. Our findings offer insights into the predictive factors influencing football match outcomes and could inform multiple future applications, including developing profitable betting strategies. Moreover, testing the comprehensive model's effectiveness in generating a profit through betting strategies represents a natural evolution of our analysis. By providing a foundation for future research in the football prediction domain, we hope these findings will be applied to develop practical solutions that improve the accuracy of football match prediction and contribute to a deeper understanding of this complex field.

# References


Aly, M. (2005). Survey on multiclass classification methods. Neural networks, 1-9.

Ang, J. C., Mirzal, A., Haron, H., & Hamed, H. N. (2016). Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 13(5), 971-989.

Bay, S. D. (1998). Combining nearest neighbor classifiers through multiple feature subsets. In International Conference on Machine Learning.

Berrar, D., Lopes, P., & Dubitzky, W. (2019). Incorporating domain knowledge in machine learning for soccer outcome prediction. Machine Learning, 108, 97-126.

Breiman, L. (2001). Random Forests. Machine Learning, 45, 5-32.

Bunker, R., & Susnjak, T. (2019). The application of machine learning techniques for predicting results in team sport: A review.

Buursma, D. (2011). Predicting sports events from past results: Towards effective betting on football matches.

Cerda, P., Varoquaux, G., & Kegl, B. (2018). Similarity encoding for learning with dirty categorical variables. Machine Learning, 107, 1477-1494.

Chen, L. (2009). Curse of dimensionality. In L. Liu & M. T. Özsu (Eds.), Encyclopedia of database systems (pp. 1-19). Springer.

Constantinou, A. C. (2019). Dolores: A model that predicts football match outcomes from all over the world. Machine Learning, 108, 49-75.

Constantinou, A., & Fenton, N. (2013). Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries. Journal of Quantitative Analysis in Sports, 9(1), 37-50.

Danisik, N., Lacko, P., & Farkas, M. (2018). Football match prediction using players attributes. 2018 World Symposium on Digital Intelligence for Systems and Machines (DISA), 201-206.

Delen, D., Cogdell, D., & Kasap, N. (2012). A comparative analysis of data mining methods in predicting NCAA bowl outcomes. International Journal of Forecasting, 28, 543-552.

Deloitte (2022, August). Annual Review of Football Finance 2022. Sports Business Group. https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/sports-businessgroup/deloitte-uk-annual-review-of-football-finance-2022.pdf

Dhanya, R., Paul, I. R., Akula, S. S., Sivakumar, M., & Nair, J. J. (2020). F-test feature selection in stacking ensemble model for breast cancer prediction. Procedia Computer Science, 171, 1561-1570.

Doquire, L. L., Gauthier, & Verleysen, M. (2013). Mutual information-based feature selection for multilabel classification. Neurocomputing, 122, 148-155.

Dubitzky, W., Lopes, P., Davis, J., & Berrar, D. (2019). The Open International Soccer Database for machine learning. Machine Learning, 108.

Elssied, N., Ibrahim, A. P. D. O., & Hamza Osman, A. (2014). A novel feature selection based on one-way ANOVA F-test for E-mail spam classification. Research Journal of Applied Sciences, Engineering and Technology, 7, 625-638.

FIFA (2021). The Football Landscape. FIFA Vision 2021.

https://publications.fifa.com/en/vision-report-2021/the-football-landscape/.

Fix, E., & Hodges, J. L. (1989). Discriminatory analysis. Nonparametric discrimination: Consistency properties. International Statistical Review / Revue Internationale de Statistique, 57(3), 238-247.

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. In International Conference on Machine Learning (pp. 148-156).

Grand View Research, Inc. (2021). Sports Betting Market Size, Share & Trends Analysis Report By Platform, By Betting Type (Fixed Odds Wagering, Exchange Betting, Live/InPlay Betting, eSports Betting), By Sports Type, By Region, And Segment Forecasts, 2023 - 2030. https://www.grandviewresearch.com/industry-analysis/sports-betting-marketreport.

Horvat, T., & Job, J. (2020). The use of machine learning in sport outcome prediction: A review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10.

Huang, K., & Chang, W. (2010). A neural network method for prediction of 2006 World Cup Football Game. The 2010 International Joint Conference on Neural Networks (IJCNN), 1-8.

Hubáček, O., Sourek, G., & Zelezný, F. (2018). Learning to predict soccer results from relational data with gradient boosted trees. Machine Learning, 108, 29-47.

Hucaljuk, J., & Rakipovic, A. (2011). Predicting football scores using machine learning techniques. 2011 Proceedings of the 34th International Convention MIPRO, 1623-1627.
Jolliffe, I. T. (2002). Principal component analysis (2nd ed.). Springer.

Knoll, J., & Stübinger, J. (2020). Machine-learning-based statistical arbitrage football betting. KI-Künstliche Intelligenz, 34(1), 69-80.

Louppe, G., Wehenkel, L., Sutera, A., & Geurts, P. (2013). Understanding variable importances in forests of randomized trees. Advances in Neural Information Processing Systems, 26.
McCabe, A., & Trevathan, J. (2008). Artificial intelligence in sports prediction. Fifth International Conference on Information Technology: New Generations (ITNG 2008), 1194-1197.

Mehra, N., & Gupta, S. (2013). Survey on multiclass classification methods.

Odachowski, K., & Grekow, J. (2012). Using bookmaker odds to predict the final result of football matches. International Conference on Knowledge-Based Intelligent Information & Engineering Systems.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. The Web Conference.

Prasetio, D., & Harlili, D. (2016). Predicting football match results with logistic regression. 2016 International Conference on Advanced Informatics: Concepts, Theory, and Application (ICAICTA), 1-5.

Reed, D., & O'Donoghue, P. (2005). Development and application of computer-based prediction methods. International Journal of Performance Analysis in Sport, 5(1), 12-28.

Rish, I. (2001). An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (pp. 41-46).

Rokach, L., & Maimon, O. (2005). Decision trees. In O. Maimon & L. Rokach (Eds.), Data Mining and Knowledge Discovery Handbook (pp. 1-22).

Rudrapal, D., Boro, S., Srivastava, J., & Singh, S. (2019). A deep learning approach to predict football match result. Advances in Intelligent Systems and Computing.

Stübinger, J., Mangold, B., & Knoll, J. (2020). Machine learning in football betting:

Prediction of match results based on player characteristics. Applied Sciences, 10(1), 46.

Tang, J., Alelyani, S., & Liu, H. (2014). Feature selection for classification: A review.

Data Classification: Algorithms and Applications, 37.

Tax, N., & Joustra, Y. (2015). Predicting the Dutch football competition using public data: A machine learning approach. Transactions on Knowledge and Data Engineering, 10, 1-13.

The Business Research Company (2023, January). Sports Betting Global Market Report 2023 – By Type (Line in play, Fixed Old Betting, Exchange Betting, Daily Fantasy,

Spread Betting, E Sports, Pari Mutuel, Other Types), By Sports Type (Football, Basketball, Baseball, Horse Racing, Cricket, Hockey, Other Sports Types), By Platform (Online, Offline) – Market Size, Trends, And Global Forecast 2023-2032. https://www.thebusinessresearchcompany.com/report/sports-betting-globalmarket-report

Valero, C. S. (2016). Predicting win-loss outcomes in MLB regular season games: A comparative study using data mining methods. International Journal of Computer Science in Sport, 15, 91-112.

Yu, L., & Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. Journal of Machine Learning Research, 5, 1205-1224.

Zaveri, N., Shah, U., Tiwari, S., Shinde, P., & Teli, L. K. (2018). Prediction of football match score and decision making process. International Journal on Recent and Innovation Trends in Computing and Communication, 6(2), 162-165.