PREDICTING FOOTBALL MATCH OUTCOMES: A COMPARATIVE ANALYSIS BETWEEN LA LIGA AND SERIE A

PASCHALIS KALLIMANIS

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

2098581

COMMITTEE

dr.Paris Mavromoustakos Blom
dr.Seyed Mostafa Kia

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

January 22th, 2023

WORD COUNT

8558

PREDICTING FOOTBALL MATCH OUTCOMES: A COMPARATIVE ANALYSIS BETWEEN LA LIGA AND SERIE A

PASCHALIS KALLIMANIS

Abstract

This research discusses the prediction of football outcomes based on match statistics. The data used spans seven seasons from the Italian and Spanish leagues, *Serie A* and *La Liga*, respectively. An extended data pre-processing is being conducted to enhance the preexisting literature by creating several features that are capable of improving the results. Additionally, a combination of feature selection methods is used to find the best feature set for the models. Regarding the modeling part, there will be a classification procedure in which three classifiers are utilized. The models being deployed are Random Forest, Support Vector Machine, and Decision Tree, which were selected after a literature review. By virtually betting a standard amount of 10 euros, the profitability of each of the models is measured, and through a comparison procedure, the best model for each league arises. The most efficient predictors of the best models in each league are also being discovered, and the models are compared in terms of accuracy and profitability. Moreover, a comparison between the performances of the models for each league was performed. Random Forest was the best model for Serie A, achieving an accuracy score of 0.575. In La Liga, SVM was the most efficient model, with an accuracy score of 0.539 and a negative profit margin of -1.43%. Random Forest achieved the best accuracy score in Serie A using 5 features, while SVM used 9 features to predict that for La Liga.

1 INTRODUCTION

The world of football is not only a source of immense passion for fans but also a fertile ground for applying data-driven techniques and machine-learning algorithms. The motivation behind this research is twofold, en-

compassing both the scientific and societal aspects of the football world.
From a scientific perspective, accurately predicting football match outcomes is a challenging problem with substantial implications. Traditionally, football betting has thrived on speculation and intuition. However, in recent years, there has been a growing interest in using the power of data and machine learning to make informed predictions.
Generally, it can be argued that for the prediction of football-match outcomes, some researchers use team statistics as Alfredo and Isa (2019), Rodrigues and Pinto (2022), others depend more on player statistics as Holmes and McHale (2023), and some others use both, such as Steijns (2020) and Remmen (2022). The current thesis will exclusively rely on team statistics. It has been noticed that there is a gap in the existing literature according to the features used to make predictions based on team statistics. This gap is expected to be covered with the extended data processing that is being conducted. Moreover, a pioneering combination of feature selection methods is used, which tries to find the best feature set for the models.
From a societal perspective, accurate match outcome predictions have the potential to reshape the football betting landscape and provide valuable insights for sports enthusiasts, punters, and even football teams and managers. In addition, the comparative analysis applied between the two leagues, La Liga and Serie A, presents an opportunity to assess the applicability of predictive models across different leagues, shedding light on the potential transferability of insights. Factors that can play an instrumental role in the predictions of each league will be shown, giving people the opportunity to better understand the aspects that affect the result of the game.
This thesis aims to apply machine learning methods to football match data to predict football match outcomes. The most important factors will be found, which may estimate the football results, and there will be a comparison between the two leagues. More specifically, data from all the games played in La Liga (First Division of Spain) and Serie A (First Division of Italy) from the 2014-2015 season up to and including the 2020-2021 season will be obtained. Our dataset encompasses seven football seasons in total. For each league, 80% of the data will be used to train the models, and the remaining 20% will be allocated for testing. There will be a classification procedure; hence, we will use three classifiers showing promising performances on related works. The models that will be used are Random Forest, Support Vector Machine, and Decision Tree.
For each model, the profit will be measured by betting a standard price (10 euros) on each football match that will be predicted. After that, the model with the highest accuracy will be discovered for each league and consequently, the most profitable model. The key factors of our predictions

will be found, and afterward, there will be a comparison of the prediction results between the two leagues. To assist us in achieving our aim, the following research questions and sub-questions have been formulated:

Research Question 1: To what extent can machine learning techniques be effectively applied to football statistics and match results data to propose profitable bets?

Research Question 2: Which are the most important predictors among all the features?

Research Question 3: How does each model perform when applied to a different league?

Sub-Question: Is the same model the best for both leagues?

Sub-Question: Are the same features the most important for both leagues?

2 RELATED WORK

Through searching for similar purposeful works, the research of Rodrigues and Pinto (2022) came to the surface, since it aimed to predict football results using diverse machine learning algorithms. The authors analyze data from five seasons of the Premier League. They use match statistics and odds, proposing feature engineering methods that are adapted and integrated into the current thesis. More specifically, they compare the models to find the best one based on accuracy (65.26%). There is a constant virtual bet on each game's prediction, which measures each model's total profit. These results are promising, considering the profit margin (26.74%). This current thesis applies some of the models used in Rodrigues and Pinto (2022) to serve the purpose of football outcome prediction. Rodrigues and Pinto (2022) have selected Naive Bayes, K-nearest neighbors, Random Forest, Support Vector Machine, Decision Tree, and Xgboost as their algorithms. In this current thesis, the dataset is different due to including data from La Liga and Serie A and not Premier League data. Moreover, in Rodrigues and Pinto (2022) odds from BET365 are used as predictors, whereas in the current thesis, they are implemented solely for the observation of each model's profit.
In the research of Patil et al. (2023), the authors discuss the application of various machine learning algorithms (KNN, Random Forest, Linear SVC) to predict the outcomes of Premier League football matches for the

2021-2022 season. In Patil et al. (2023) the final opinion is that recent performance and home advantage are the most essential factors. Based on that, some new variables will be created for the teams' recent performance. Generally in sports, recent performance can play an instrumental role, and that can be proven by Patil et al. (2023), Rodrigues and Pinto (2022), Steijns (2020), since for the better prediction of football matches, variables for recent performance need to be measured.

In Steijns (2020) work, the focus is on the effect of several team-specific and player-specific features on the outcome of a football match. In the current thesis, there are only team-specific features. However, the data pre-processing shares commonalities with Steijns (2020) work. They use team-specific data that resemble those that are demonstrated in the current thesis. Afterward, by creating some new variables and using Pearson correlation, they find the most important features. In the correlation test, it is noticeable that Steijns (2020) uses some variables that have significant values (difference in points in the current season, difference in points in the previous season, home team points, away team points). Moreover, he measures the corresponding conceded statistics for all the features. For instance, in the feature 'homeGoals' he measures the home goals conceded, and thereupon he discovers the difference between goals scored by the home team and the goals conceded. These variables appear in this current thesis as well. It is worth noting that in Steijns (2020) thesis, profits are not being measured. Furthermore, in the current thesis, more extended data processing takes place, which can lead to more efficient results.

Holmes and McHale (2023) present a new model to predict football match outcomes, focusing on player-specific data instead of team-based methods. A rating system is deployed to compare players across different leagues. The model faces the limitations of team-based models by directly modeling the dynamics of matches through player interactions, eliminating the need to account for inconsistent team strengths. In other words, by directly considering changes in team line-ups and short-term player performance, it provides a more reliable and stable approach to football match prediction. While the model required a sizeable amount of data, the availability of player ratings made it practical for accurate predictions. The model showed promising forecasting performance in tests, outperforming bookmakers in predicting match outcomes. It is tested with betting outcomes, producing positive results in the odds market.

In the field of football match outcome prediction, Samba (2019) explores the performance of deep learning algorithms. The research is based on a dataset that includes 20,000 matches from prominent football leagues like Premier League, Championship, League 1, and League 2. Samba (2019) combines team-specific and team-independent features in its features. The

team-specific features include important factors such as team strength, cumulative season performance metrics, form, and time gaps between matches. On the other hand, team-independent features encompass external influences such as the referee, bookmakers' odds, season, and division. The target value is divided into the classes home wins, away wins, or draws. Samba discusses various multilayer perceptrons to assess how effectively deep learning algorithms can predict outcomes, with differences in depth, number of neurons per layer, and output configuration. A key finding in his study showed that the neural network was the best model with three output neurons, achieving an accuracy of 48%. This outperforms a neural network with a single output neuron, which has a comparatively lower accuracy of 43%.

In the related study of Stübinger et al. (2020), the top five European football leagues are discussed for the seasons from 2006 to 2018. Stübinger et al. (2020) use player statistics and betting odds to construct machine learning models for football match betting. In their research, there are 4 machine learning models, including Random Forest, Support Vector Machine, Boosting, and Linear Regression, that achieve profitable outcomes. The Random Forest model achieved the best accuracy and the highest profit among the other models.

In the study of Cintia et al. (2015), there are football match data from four seasons of the 4 major football leagues in Europe (England, Italy, Spain, and Germany). Several player statistics serve as valuable predictors of match outcomes. The selection of the most effective algorithm for predictions depends on the specific football league under consideration. In the investigated leagues, the random forest algorithm demonstrates favorable accuracy, achieving 0.55 in Italy, 0.53 in Spain, and 0.58 overall, while in the German Bundesliga, the KNN algorithm outperformed others with an accuracy of 0.6.

Gomes et al. (2021) retrieves football data from 13 seasons of the Premier League having statistical information from 4940 football matches, aiming to create a decision support system to predict football games. The match statistics used in this study exhibit similarities with those presented in the current thesis. From Gomes et al. (2021) derived the idea of creating two features related to the points that the home and away teams accumulate when they play at home or away correspondingly. Generally, the data processing of Gomes et al. (2021) has mutual properties with the equivalent processing in the current thesis. Gomes et al. (2021) use Decision Tree, Naive Bayes, and Support Vector Machine to achieve the best result with an accuracy score of 0.51. Furthermore, they virtually bet 100 euros in each game of a total of 70 games, gaining a profit of 1,409 euros with a profit margin of 0.20.

Remmen (2022) uses Decision Tree, Random Forest, and gradient boost models to predict football match outcomes. To achieve that, they used two databases, one with match statistics from the top 5 leagues in Europe (England, Spain, Italy, France, and Germany) and one with player statistics. There is a feature called the rating of each team that plays a significant role in the accuracy score. When this rating value is removed, the score of gradient boosting has an accuracy score of 0.647 when it is used only with match statistics and data from midfield players. When this rating value exists as a predictor, an accuracy score of 0.864 is obtained using match statistics and player data from midfield players. After Remmen (2022) removed the match statistics to make a feature analysis only in the player statistics, it was noticed that all the algorithms achieved a higher score using data from the attacking players.

3 DATA EXPERIMENTAL SETUPS

3.1 *Data Source*

The dataset [1](Football Database) utilized in this thesis is publicly available on Kaggle.com. This dataset is separated into 7 different csv files containing football-related information from the Top-5 leagues in Europe for the years 2014-2021 (7 seasons). For this project, only data from La Liga (First Division of Spain) and from Serie A (First Division of Italy) is incorporated. Moreover, some data is extracted from Transfermarkt.com [2] is used.

3.2 *Software*

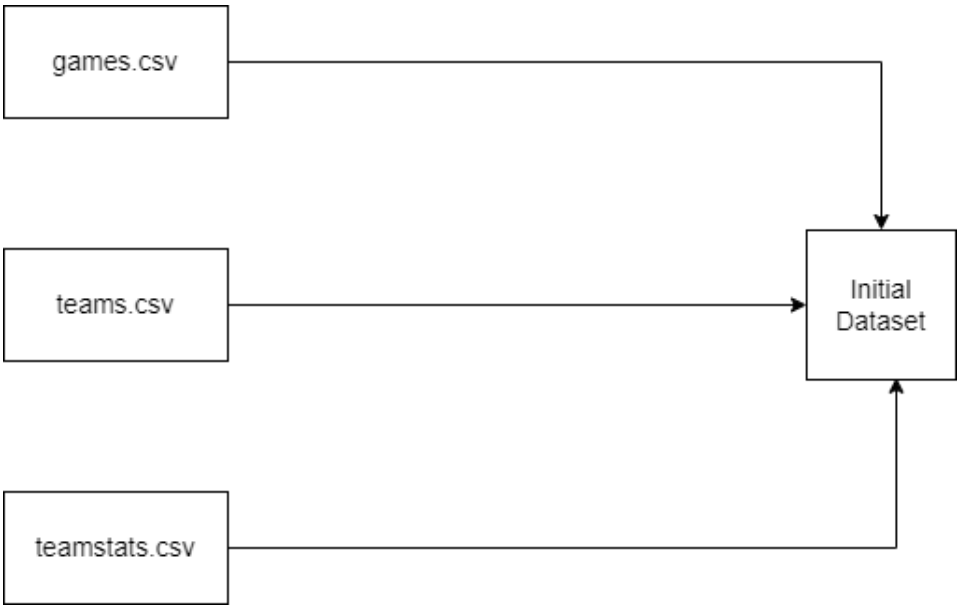
Python Programming language is used for all the procedures. The program selected for coding in Python is Anaconda Navigator "Anaconda Software Distribution" (2020), an open-source distribution of different programming languages. The module Pandas McKinney (2010) is exercised extensively in this thesis and is mainly implemented in data pre-processing. Scikit-learn Pedregosa et al. (2011) library was utilized to launch the models.

[1] https://www.kaggle.com/datasets/technika148/football-database
[2] https://www.transfermarkt.com/

3.3 *Data Pre-Processing*

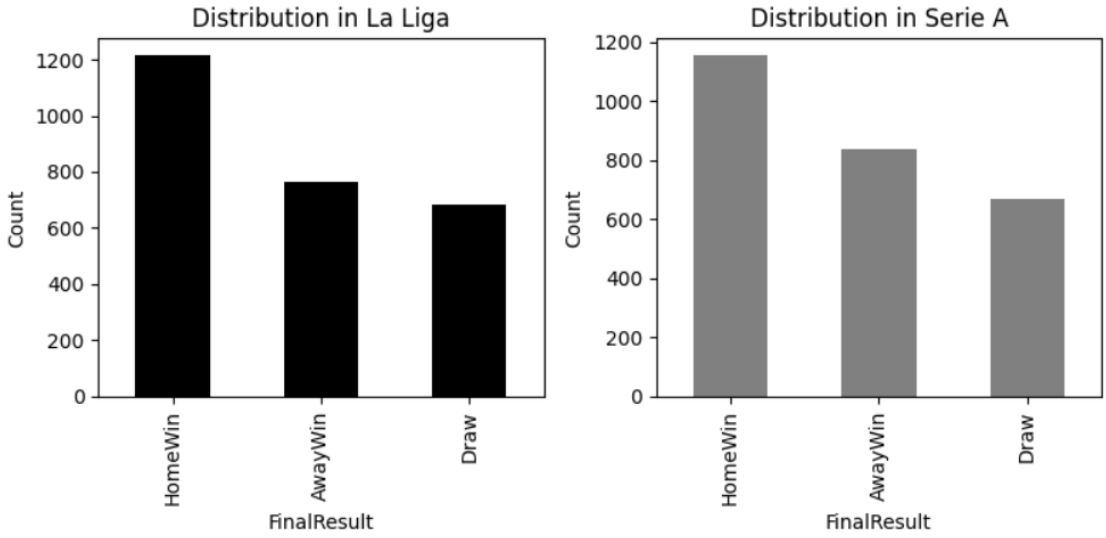
For this thesis two separate datasets have been used, the one containing the data of La Liga (Spanish league) and the other the data of Serie A (Italian league), but since the data pre-processing is identical for both leagues, the procedure will be addressed once responding to both cases. As mentioned above, 3 csv files are used from a total of 7.
The first csv file named games.csv contains 34 features but only 11 of them will be utilised. This csv file contains the odds of each game from the bet365 website and some other features, namely leagueID, teamIDs, gamesID, away goals, home goals, season, and date. The second csv file named teamstats.csv contains 16 features in total, with most of them being based on the match statistics of each game. The third csv file is named teams.csv has 2 features with the team name and the team ID, which could make the dataset more comprehensible to handle. Throughout the csv files, there are mutual features that make it feasible to merge them.

Figure 1: Merging the csv files



Handling away goals and home goals from the already merged dataset, the target value was created, which was named FinalResult. The distribution of the target value can be seen.

Figure 2: Distribution of target value for both leagues



The objective is to conclude a final dataset in which there will be statistics for the home team and the away team to pursue the prediction of football match outcomes. Hence, a data pre-processing procedure was essential to creating features about the home team and the away team. Moreover, it should be stressed that the aim is to create features that concern statistics before the game focuses on prediction, an operation identical to the equivalents in the research of Rodrigues and Pinto (2022), Gomes et al. (2021) and Steijns (2020). After the above procedure, we conclude the initial dataset with the following features:

Table 1: Description of features

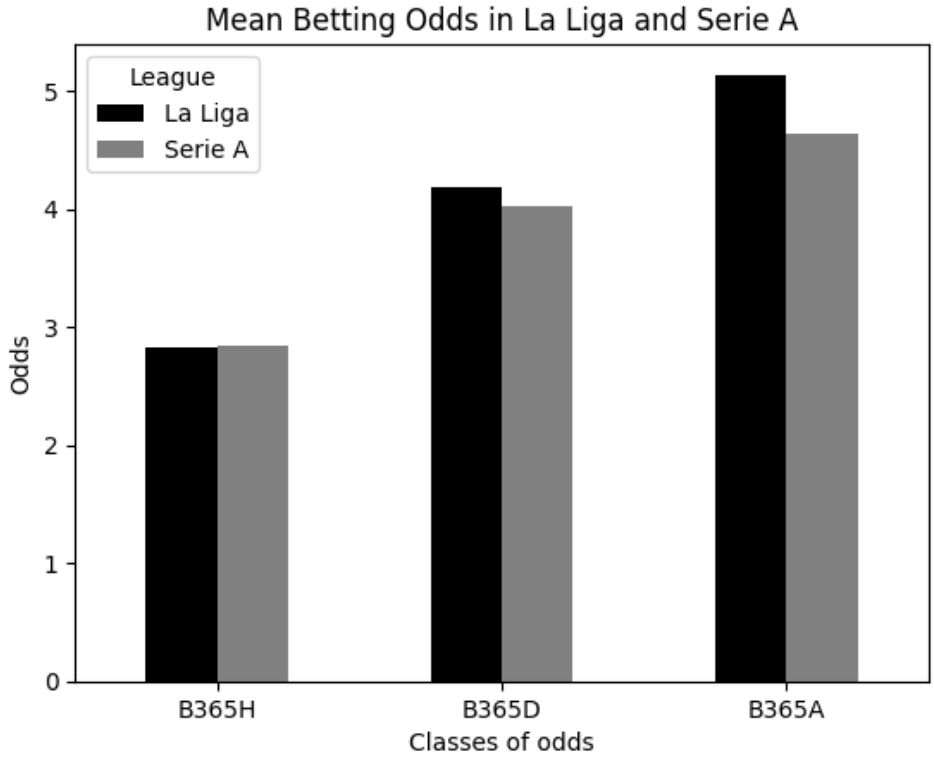
HomeGoals (Number of goals scored by the home team)
AwayGoals (Number of goals scored by the away team)
B365H (initial odd of home team win from BET365)
B365A (initial odd of away team win from BET365)
B365D (initial odd of draw from BET365)
HomeTeamShotsOnTarget (Number of shots on target by the home team)
AwayTeamShotsOnTarget (Number of shots on target by the away team)
HomeDeep (Passes completed within an estimated distance of 20 yards of goal for home team)
AwayDeep (Passes completed within an estimated distance of 20 yards of goal for away team)
HomePPDA (Passes allowed per defensive action in opposition half for home team)
AwayPPDA (Passes allowed per defensive action in opposition half for away team)
HomeShots (Shots for the home team)
AwayShots (Shots for the away team)
HomeXGoals (Expected goals for the home team based on Understat)
AwayXGoals (Expected goals for the away team based on Understat)
HomeYellowCards (Yellow cards received by the home team)
AwayYellowCards (Yellow cards received by the away team)
HomeRedCards (Number of red cards received by the home team)
AwayRedCards (Number of red cards received by the away team)
HomeFouls (Number of fouls commited to the home team)
AwayFouls (Number of fouls commited to the away team)
HomeCorners (Corners for home team)
AwayCorners (Corners for away team)
date (Date of the match)
season (the season that the football match is played)
HomeTeamID (Id of home team)
AwayTeamID (Id of away team)
gameID (Id of the game)
leagueID (Id of the league)
FinalResult (HomeWin, Draw or AwayWin)

Through the data processing procedure, the initial features are modified to create new, appropriate features for the current work. After that, the data is split (80/20) into a training set and a test set. A feature selection procedure will occur to find the most efficient combinations to produce the classifiers' highest accuracy scores.
Data Pre-processing was the most time-consuming procedure in this thesis because not any of the initial features have been used as a match-outcome predictor; however, they were the means to create other more efficient features to forecast the final results of football matches. A framework follows that briefly demonstrates all the procedures

As we have already referred, odds from the website BET365 are used for this thesis, but only in the procedure in which we measure the profits of each model and not in the prediction as predictors, considering that what

we desire to use as predictors only match statistics. It follows a figure with the distribution of each class of odds for each league.

Figure 3: Distribution of odds



We notice that the odds are higher in La Liga for the draw and away class, while in Serie A, home class is slightly higher. Those differences can affect the profits produced in La Liga and Serie A, considering that the same correct predictions for both leagues in the Away class will possibly produce more profits for the model that made predictions for La Liga.
In the research of Steijns (2020), two variables played a significant role in prediction based on the ranking of the teams. Steijns (2020) created a feature that demonstrated the difference in ranking of the teams from the previous season, giving the teams that were promoted from a lower division and did not have a league ranking in the previous year a ranking comparable to that of the lower-ranked team. Using the same logic, in the current thesis, the features that contain the difference in points from the rankings of the previous season are created. In the current thesis, this procedure is executed with points and not with rankings to achieve higher precision. For the teams that did not have rankings from the previous season, due to their promotion from the lower division, it is assumed that the ranking

points from the previous season were 38 points. Generally, observing final league rankings in Transfermarkt.com [3] a which accumulated 38 points is on the verge of relegation in terms of league division. A similar approach was followed by Steijns (2020) but they assigned to those teams rankings of the lowest positions. For the creation of this feature, a manual collection of data is needed. In particular, the dataset did not have data from the season 2013–2014, the previous year of the season 2014-2015 (this season refers to 2014 in the dataset). Through transfermarkt.com [4], the points for the teams for season 2013-2014, of the two leagues were collected, and the DPPS (Difference in Points from the Previous Season) originated without missing values.

Moreover, Steijns (2020) created a feature that quantifies the difference in points between the home team and the away team during a season. Likewise, the current thesis contains a similar feature; in particular, this feature is named DiffPts, representing the subtraction of the away team's points from the home team's points until the next upcoming match. For instance, taking the football season 2016–2017 as our reference point and supposing there is a scheduled football match between the home team and the away team on a specific date, the gaining points of the home team must be measured for the season 2016–2017 up to the point of that particular match. The same operation has to be executed for the away team, and then predictably, the subtraction of home team points - away team points follows to display the deduction of points between the two.

HomePoints and AwayPoints are two additional features of this thesis. HomePoints measure the points that the home team has gained in its home games for the specific season, and awayPoints demonstrate the points that the away team has gained in its away games for the specific season. This creation is based on the logic that some teams perform differently in home games and away games. This can be perceptible from the target value distribution that is shown above. A feature of similar logic emerged in Gomes et al. (2021) where they measured the number of wins in the last five games of the home team in home games and wins in the last five games of the away team in away games

Similarly with Gomes et al. (2021), features of recent performances of teams are measured. HomeRecentFormPoints and AwayRecentHomePoints were created. These features measure the points that home and away teams gained in the last 8 weeks. It follows a table with those 6 newly generated features that have to be mentioned

[3] https://www.transfermarkt.com/
[4] https://www.transfermarkt.com/

Table 2: New features

HomePoints
AwayPoints
HomeRecentFormPoints
AwayRecentFormPoints
DPPS (Difference in points from the previous season)
DiffPts (Difference in points this season

New features were created to improve the prediction of football results. Similarly with Steijns (2020), for each of the match statistics, the corresponding conceded evidence is measured. For instance, we measured the conceded goals for the home team, the conceded shots, the conceded cards, etc. Afterward, we measured the new features taking into account the difference; for instance, we measured 'DifferenceHomeTeamShotsOnTarget' = 'HomeTeamShotsOnTarget' – 'ConcededHomeTeamShotsOnTarget'. This is done for all the features that are considered match statistics (goals, shots, corners, etc.). The initial match statistics dropped because the objective was to predict the statistics before the game. After the above procedure, those were the features of the dataset.

Table 3: Description of features

gameID
leagueID
season
date
B365H
B365D
B365A
HomeTeamID
AwayTeamID
FinalResult
HomeTeamName
AwayTeamName
DPPS
DiffPts
HomeRecentFormPoints
AwayRecentFormPoints
HomePoints
AwayPoints
HomeGoalDifference
AwayGoalDifference
HomeShotsDifference
AwayShotsDifference
HomeShotsOnTargetDifference
AwayShotsOnTargetDifference
HomeXGoalsDifference
AwayXGoalsDifference
HomeDeepDifference
AwayDeepDifference
HomePPDA_difference
AwayPPDA_difference
HomeFoulsDifference
AwayFoulsDifference
HomeCornerDifference
AwayCornerDifference
HomeYellowCardsDifference
AwayYellowCardsDifference
HomeRedCardsDifference
AwayRedCardsDifference

The potential problem with the new features (meaning the features from 'HomeGoalDifference' to 'AwayRedCardsDifference' in the above table) is that they are measured totally from the first match of a team until the last one through the dataset. For instance, the 'HomeCornersDifference' measured the corners won by a home team, subtracted the conceded corners of this home team, and if that is a variable of the 2020-2021 season (the last season of the dataset), then, it is not certain whether it can result in an accurate picture of this home team. A team can play well from 2014 until 2019

but then display a decline in performance. To resolve such problems, some new features have been created. Those previously modified features can be modified again to create even more efficient features. More specifically, for each feature such as 'HomeXGoalsDifference', 'AwayGoalDifference' etc, new features have been developed for the last 20 home and away games respectively; for instance, when we are dealing with a home-related feature (ex. 'HomeCornerDifferenceLast20', 'HomeYellowCardsDifference'), the home-related statistics will be estimated for the last 20 home games of the home team. The same is applied to away-related features. The number 20 is determined after experimentation with the logic that the last 20 home or away games correspond to the final year (During a season in Italy and Spain, they play 19 home and away games).
After a Pearson correlation test, it was found that the newly created features are more correlated with the final result compared to the previous. The most recently created features that depicted the last 20 home or away games are derivative of the others, creating high values of multicollinearity in the data. For example, HomeDeepDifferenceLast20 is highly correlated with HomeDeepDifference. To resolve that problem, we dropped the following features, considering that Pearson correlations were less important in comparison to the new features, which correspond to the last 20 home and away games. In addition, assisting features such as leagueID, date, gameID, HomeTeamId, and awayteamID were dropped. In addition, the odds of BET365 will be held in a separate dataset, considering that we want them to measure profit, but they will be dropped from the current dataset. After the above procedure, we conclude with the following feature set:

Table 4: New features

FinalResult
DPPS
DiffPts
HomeRecentFormPoints
AwayRecentFormPoints
HomePoints
AwayPoints
HomeGoalDifferenceLast20
AwayGoalDifferenceLast20
HomeShotsDifferenceLast20
AwayShotsDifferenceLast20
HomeShotsOnTargetDifferenceLast20
AwayShotsOnTargetDifferenceLast20
HomeXGoalsDifferenceLast20
AwayXGoalsDifferenceLast20
HomeDeepDifferenceLast20
AwayDeepDifferenceLast20
HomePPDADifferenceLast20
AwayPPDADifferenceLast20
HomeFoulsDifferenceLast20
AwayFoulsDifferenceLast20
HomeCornerDifferenceLast20
AwayCornerDifferenceLast20
HomeYellowCardsDifferenceLast20
AwayYellowCardsDifferenceLast20
HomeRedCardsDifferenceLast20
AwayRedCardsDifferenceLast20

3.4 *Evaluation and Metrics*

Three different classifiers have been chosen to predict football match outcomes, and their effectiveness has been assessed in terms of both accuracy and profitability. The Random Forest and the Support Vector Machine were selected as predictive models, with the Decision Tree algorithm being the baseline model for comprehensive performance comparison. The scikit-learn library has been employed for the implementation of each model.
The decision to employ SVM (Support Vector Machine) in this study originated from its promising performance in Rodrigues and Pinto (2022). In their work, SVM demonstrated promising results with an accuracy of 0.639, making it an appropriate choice for the current thesis. It is important to point out that only Random Forest outperformed SVM in both accuracy and profitability scores among the seven models compared to the study of Rodrigues and Pinto (2022). Given the similarities between their dataset and that in the current thesis, SVM can be considered a reasonable choice.

Random Forest as a predictive model is preferred from several related works with noticeable results. Stübinger et al. (2020), Cintia et al. (2015), and Alfredo and Isa (2019) have all reported encouraging findings with Random Forest as a predictive model. Therefore, by measuring the success of Random Forest in these works, we anticipate a positive impact of this model on the predictive accuracy of football match outcomes in our study. The Decision Tree model has been selected as the baseline for its simplicity and interpretability (Remmen, 2022), providing a benchmark against which the performance of more complex models can be evaluated. By opting for Decision Tree, we aim to assess whether the increased complexity of Random Forest and SVM drives significant improvements in predictive accuracy.
All the models have been executed for both La Liga and Serie A. In total, 3 random forest models were launched for each league, 3 Support Vector Machine models were executed for each league, and 1 Decision Tree was used for each league, which is the baseline model. In total, 14 models have been executed. The execution of each model is based on feature selection methods, which are demonstrated in the next few paragraphs.
After applying standardization and normalization to the data, the accuracy scores decreased compared to the unprocessed data, suggesting that these pre-processing techniques may not be beneficial for improving model performance in this specific case. However, the computational time for the Support Vector Machine was long, and for this model, scaled data were used.

3.5 *Methodology*

The dataset was divided into two subsets: 80% for the training set and 20% for the testing set. While a stratified split could ensure a similar class distribution in both sets, it does not align with the unpredictability and variability of football matches. Hence, it is considered that a random split is the most appropriate for the specific work. Moreover, as we can detect from the final dataset, there are no team identifiers as team names, which can lead to team-biased results; thus, we have no such concerns about the split of the dataset (specifically, the selection process for the training and test sets is unbiased, as no feature in the dataset reveals the identity of the teams behind the features).
The core methodology followed for the feature selection methods used was inspired by the work of Alfredo and Isa (2019) in which they involved a process of iterative feature elimination from a set of 14 features, where the least important feature was removed in each of the 14 iterations. In

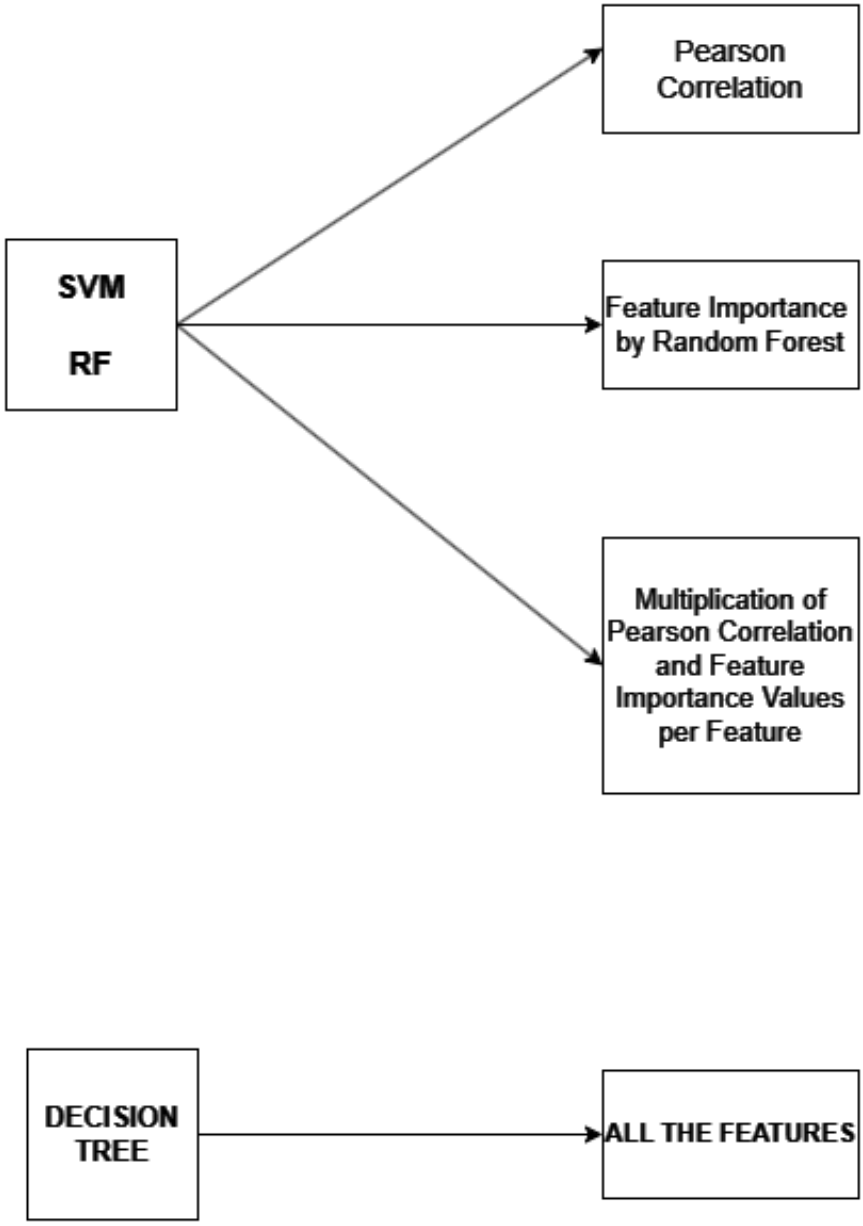
this study, a reversed procedure is used. More specifically, in building ranked lists based on feature selection methods, a stepwise feature selection method is implemented. This process begins with the incorporation of the top-ranked feature and incrementally adds the next most significant feature. The addition of features is performed iteratively, cross-validation is applied for each iteration, and hyperparameter tuning is conducted for each iteration. It has to be mentioned that this procedure is applied to the training set, and the result of that procedure is to find the feature set and the hyperparameters that achieved the best cross-validation score (the best mean score of the folds).

The feature selection strategy involves initially sorting the features according to their respective values derived from some feature selection methods. For example, using the Pearson correlation method, features are organized from the most significant (exhibiting the highest correlation with the target variable) to the least significant. In this thesis, three feature selection methods are used: Pearson correlation, feature importance by Random Forest, and the multiplication of both aforementioned. Those three methods are applied by SVM and Random Forest. More analytically, Random Forest was employed to analyze and learn patterns from the provided dataset. By training on the training set, the Random Forest model gained insights into feature importance, demonstrating the significance of each predictor in influencing predictions. The resulting importance values were then organized in a descending-sorted format, providing an understanding of the most influential features of Random Forest. This method, like the other two, is applied both by SVM and Random Forest, and this is not a usual procedure considering that the feature-importance method from Random Forest is expected to be used only via Random Forest. Similar to Dewi (2019) we experimented with using this feature selection method performed by Random Forest for SVM. In the same manner as Steijns (2020), Pearson correlation is used to apply one more feature selection method. The features are sorted from the feature with the highest value to the lowest. Moreover, an experimental feature selection method is applied, multiplicating the values between the Pearson correlation value and the feature importance value for each feature independently. The results of this multiplication are sorted in descending order, as in the previous two methods. This last procedure has not been detected in any previous work but could potentially lead to the discovery of a feature set that depends both on correlation and important scores that have not been investigated in the aforementioned methods.

As mentioned, those three feature selection methods are applied by both SVM and Random Forest but not by Decision Tree. The decision tree,

the baseline model, uses all the features. The following table shows a framework for the feature selection strategy.

Figure 4: Feature selection methods



Having sorted lists, a stepwise procedure is operated both for the Random Forest and for the Support Vector Machine. Each model is operated 3 times to correspond with the 3 lists. Hence, starting from the most important feature for each list, for each iteration, the next more important feature is

added to the feature set. In total, there are 26 features; hence, 26 iterations are conducted.
For each iteration, cross-validation is utilized to validate the reliability and generalizability of the models. It helps in assessing the performance of the models across different data subsets of the training set, demonstrating that our findings are not biased based on the data of a single dataset partition. More specifically, the training set is split into 10 folds. One fold is used as the validation set, and the remaining 9 folds are used for training. This process is systematically rotated such that each fold serves as the validation set at some point, guaranteeing that every data point contributes to both training and validation. The performance of each hyperparameter set is assessed based on its average accuracy across all 10 folds of cross-validation. This approach ensures that the chosen parameters show consistent performance across different subsets of the data.
For each newly formed subset of features, hyperparameter tuning is conducted using GridSearchCV to find the best parameters specific to each feature combination. For each iteration, where an additional feature is added, a comparison is performed with the 10-fold-validation score of the model obtained with the current set of features and its best hyperparameters against the best cross-validation score achieved in previous iterations. When a specific combination of features and hyperparameters achieves higher performance than the previous best performance, as indicated by a better 10-fold validation score, the model is updated to integrate this more effective combination.
The next procedure is to test the model on unseen data. The test set was not involved in any part of the training process, considering that it served as an independent benchmark for model evaluation. The model uses initial splitting (80% for the training set and 20% for the test set) with the best combination of features and hyper-parameters from cross-validation, testing it on unseen data (the test set). This evaluation of unseen data is crucial for noticing the generalisability of the model beyond the data it was trained on.

3.6 *Hyper parameter tuning*

The hyperparameter tuning was applied only for the Support Vector Machine and Random Forest. For the SVM, the hyperparameter tuning is focused on searching a range of values for 'C', the regularization parameter, which includes the following values: 0.1, 1, 10, and 100. The next parameter is 'gamma', which is important for determining the effect of a single training example, with values set at 1, 0.1, 0.01, and 0.001. In addition, the

kernel parameter has two options, 'rbf' and 'linear', allowing the SVM to assess the effect of different kernel operations on the performance of the model.
For the Random Forest, the tuning procedure contains 'n_estimators', where we experimented with the values 100, 200, and 300. 'max_depth' is the other parameter, representing the maximum depth of the trees, including the values None, 10, and 20. The third parameter for Random Forest is the 'min_samples_split', which represents the minimum number of samples required to split an internal node, with options of 2, 5, and 10.

Table 5: Hyperparameters for Random Forest models

Hyper parameters, La liga			
	max_depth	min_sampled_split	n_estimators
RF1	10	10	300
RF2	10	10	100
RF3	10	5	200
Hyper parameters, Serie A			
	max_depth	min_sampled_split	n_estimators
RF1	20	10	300
RF2	None	2	200
RF3	10	5	100

Table 6: Hyperparameters for Support Vector Machine models

Hyper parameters, La liga			
	C	gamma	kernel
SVM1	1	0.1	rbf
SVM2	100	0.001	rbf
SVM3	0.1	0.1	rbf
Hyper parameters, Serie A			
	C	gamma	kernel
SVM1	0.1	0.01	rbf
SVM2	100	0.001	rbf
SVM3	0.1	0.01	rbf

In the next two tables, we represent the values of features for each of the three feature selection methods being used, and they are ranked based on the values obtained by multiplication between Pearson correlation and feature importance using Random Forest.

Table 7: Feature selection, La Liga

Feature	Feature Importance	Pearson Correlation	Feature Importance * Pearson Correlation
DPPS	0.064	0.389	0.025
DiffPts	0.049	0.342	0.016
AwayXGoalsDifferenceLast20	0.047	0.282	0.013
HomeXGoalsDifferenceLast20	0.046	0.246	0.011
HomeDeepDifferenceLast20	0.044	0.246	0.011
AwayGoalDifferenceLast20	0.040	0.267	0.010
AwayShotsOnTargetDifferenceLast20	0.041	0.263	0.010
AwayDeepDifferenceLast20	0.038	0.261	0.010
HomeGoalDifferenceLast20	0.038	0.246	0.009
AwayPPDADifferenceLast20	0.042	0.219	0.009
AwayShotsDifferenceLast20	0.039	0.218	0.008
HomeShotsOnTargetDifferenceLast20	0.036	0.238	0.008
HomeCornerDifferenceLast20	0.039	0.196	0.007
HomePPDADifferenceLast20	0.040	0.180	0.007
HomeShotsDifferenceLast20	0.036	0.192	0.007
AwayCornerDifferenceLast20	0.039	0.173	0.006
HomeRecentFormPoints	0.030	0.174	0.005
AwayFoulsDifferenceLast20	0.040	0.122	0.004
AwayYellowCardsDifferenceLast20	0.033	0.134	0.004
AwayRecentFormPoints	0.028	0.148	0.004
AwayPoints	0.030	0.131	0.004
HomeYellowCardsDifferenceLast20	0.035	0.111	0.003
HomeFoulsDifferenceLast20	0.039	0.093	0.003
HomePoints	0.030	0.121	0.003
AwayRedCardsDifferenceLast20	0.022	0.069	0.001
HomeRedCardsDifferenceLast20	0.021	0.032	0.000

Table 8: Feature selection, Serie A

Feature	Feature Importance	Pearson Correlation	Feature Importance * Pearson Correlation
DPPS	0.061	0.417	0.025
DiffPts	0.047	0.349	0.016
HomeDeepDifferenceLast20	0.050	0.298	0.015
HomePPDADifferenceLast20	0.054	0.264	0.014
AwayDeepDifferenceLast20	0.046	0.273	0.012
AwayXGoalsDifferenceLast20	0.043	0.283	0.012
HomeShotsOnTargetDifferenceLast20	0.040	0.282	0.011
HomeShotsDifferenceLast20	0.039	0.284	0.011
AwayShotsOnTargetDifferenceLast20	0.039	0.271	0.010
HomeCornerDifferenceLast20	0.041	0.263	0.010
AwayShotsDifferenceLast20	0.041	0.261	0.010
HomeXGoalsDifferenceLast20	0.038	0.281	0.010
AwayPPDADifferenceLast20	0.041	0.253	0.010
AwayCornerDifferenceLast20	0.041	0.249	0.010
AwayGoalDifferenceLast20	0.034	0.283	0.009
HomeGoalDifferenceLast20	0.032	0.268	0.008
HomeRecentFormPoints	0.030	0.200	0.006
HomeYellowCardsDifferenceLast20	0.034	0.166	0.005
AwayRecentFormPoints	0.030	0.175	0.005
AwayYellowCardsDifferenceLast20	0.031	0.161	0.005
AwayPoints	0.029	0.142	0.004
HomePoints	0.028	0.132	0.003
AwayFoulsDifferenceLast20	0.037	0.053	0.001
HomeRedCardsDifferenceLast20	0.022	0.056	0.001
HomeFoulsDifferenceLast20	0.039	0.028	0.001
AwayRedCardsDifferenceLast20	0.021	0.040	0.000

4 RESULTS

Research Question 1: To what extent machine learning techniques can be effectively applied to football statistics and match results data to propose profitable bets?

To answer this research question, the total profit of each model has to be measured. To achieve that, for each prediction, a virtual bet of 10 euros is placed. In particular, for each prediction, there is a specific odd, which is also used as a predictor. If the bet is in favor of the home team winning, the calculation for the specific case will be 10*B365H, where "B365H" shows the odds for the home team's victory. For example, if for this specific match, the eventual outcome is a home win, then the profit for this game will be (10*B365) - 10 because the monetary amount of 10 euros was the initial bet and a clear profit is needed. The summary of the bet returns is the total

profit or loss of each model.
The following table presents the accuracy and profit outcomes for three classifiers (Random Forest, Support Vector Machine, and Decision Tree) when applied to unseen data (test set) in La Liga.

Table 9: La Liga Models (Test set), accuracy score and profits

Models	Accuracy (La Liga)	Profit (La Liga)
RF1	0.522	-138.29
RF2	0.526	-96.70
RF3	0.520	-179.99
SVM1	0.530	-115.89
SVM2	0.539	-31.69
SVM3	0.531	-92.69
DT	0.417	-315.4

The following table includes the mean accuracy from 10-fold cross-validation, along with the maximum and minimum fold scores, for each model based on the training set (In the first column there is also the score from the test set).

Table 10: Results from models operated in La Liga from the 10-fold validation (training set)

Models	Accuracy (Test Set)	Mean CV score	Min Fold	Max Fold
RF1	0.522	0.521	0.460	0.571
RF2	0.526	0.517	0.462	0.570
RF3	0.520	0.519	0.460	0.566
SVM1	0.530	0.526	0.478	0.580
SVM2	0.539	0.524	0.480	0.581
SVM3	0.531	0.524	0.478	0.594
DT	0.417	0.424	0.394	0.450

Examining the La Liga models from the above table, the Random Forest models (RF1, RF2, and RF3) do not achieve profitability, with losses ranging from -96.70 euros to -179.99 euros and accuracy scores from 0.522 to 0.526. All the Support Vector Machine models are also not profitable, with losses ranging from -31.69 euros to -115.89 euros. The baseline model, the Decision Tree, has a loss of -315.4 euros with an accuracy of 0.417.
They follow 2 equivalent tables as overhead for the league of Serie A.

Table 11: Serie A Results

Models	Accuracy (Test set)	Profit (Serie A)
RF1	0.575	395.60
RF2	0.550	62.70
RF3	0.565	199.40
SVM1	0.574	146.09
SVM2	0.562	-71.09
SVM3	0.546	-112.19
DT	0.462	335.59

Table 12: Results from models operated in Serie A from the 10-fold validation (training set)

Models	Accuracy (Test set)	Mean CV score	Min Fold	Max Fold
RF1	0.575	0.539	0.491	0.589
RF2	0.550	0.538	0.507	0.589
RF3	0.565	0.540	0.483	0.584
SVM1	0.574	0.541	0.491	0.580
SVM2	0.562	0.541	0.490	0.566
SVM3	0.546	0.540	0.488	0.551
DT	0.462	0.433	0.401	0.475

Peering into the results of models executed for Serie A, the Random Forest achieves profits in all cases, with profits ranging from 62.70 euros to 395.60 euros. Their accuracy score ranges from 0.550 to 0.575 euros. The Support Vector Machine models contributed to the profits in one case (SVM1) with 146.09 euros and an accuracy of 0.575. The other two models (SVM2 and SVM3) had losses of -71.09 and -112.19 euros, with accuracy scores of 0.562 and 0.546, respectively. The Decision Tree model had an accuracy of 0.462 and a profit of 335.59 euros, surpassing all the SVM models in terms of positive returns.

We can conclude that, in La Liga, all the machine learning models that are used cannot produce profits. In Serie A, the results are slightly better, with some profitable models. Choosing the best model in each league, SVM2 with an accuracy score of 0.539 had a loss of 31.69 euros in La Liga, and RF1 with an accuracy score of 0.575 achieved a profit of 395.60 euros in Serie A.

It can be noticed that in La Liga, while Support Vector Machine models have accuracy scores higher than those of Random Forest, profits vary at almost the same levels. For instance, the less efficient SVM in La Liga (SVM1) has bigger losses than the most efficient Random Forest, with SVM1 having higher accuracy than RF3. Moreover, noticing the profit of the decision tree in Serie A, we can notice that this model achieves higher profits than all SVM models and two Random Forest models, which

had higher accuracy scores. Odds can explain that, since some models can correctly predict odds of higher values, resulting in higher profits than those from models capable of more correct predictions with odds of lower values. As such, what we discuss in this paragraph can become more comprehensible in the next section (Error Analysis), we will present Confusion Matrices and Receiver Operating Characteristic (ROC) Area Under the Curve (AUC) charts. These will demonstrate the models with the highest accuracy scores from our Random Forest, SVM, and Decision Tree analyses for both Serie A and La Liga.

Research Question 2: Which are the most important predictors among all the features?

To answer this question, the combinations that produced the best accuracy scores for each league will be demonstrated; two tables with the best combinations for each league are provided below. About the best combinations, we mean those that achieved the best accuracy scores for each league.

Table 13: Features produced the best accuracy, La Liga

DPPS	0.38
DiffPts	0.34
HomeGoalDifferenceLast20	0.24
HomeDeepDifferenceLast20	0.24
AwayShotsDifferenceLast20	0.21
HomeCornerDifferenceLast20	0.19
HomeShotsDifferenceLast20	0.19
HomePPDADifferenceLast20	0.18
HomeRecentFormPoints	0.17

It can be observed that the Support Vector Machine used those 9 features from the total 26 predictors to achieve the best possible score for La Liga. The values next to each feature are the Pearson Correlation value of each feature. These 9 features can be characterized as the most important predictors for La Liga.

Table 14: Features produced the best accuracy, Serie A

DPPS	0.41
DiffPts	0.34
HomeDeepDifferenceLast20	0.29
HomePPDADifferenceLast20	0.26
AwayDeepDifferenceLast20	0.27

In Serie A, the Support Vector Machine used 5 features from a total of 26 features to achieve the best accuracy score. These 5 features can be characterized as the most important predictors for Serie A. As mentioned in the methodology section, those sets of features was identified during the training phase through a 10-fold cross-validation procedure.

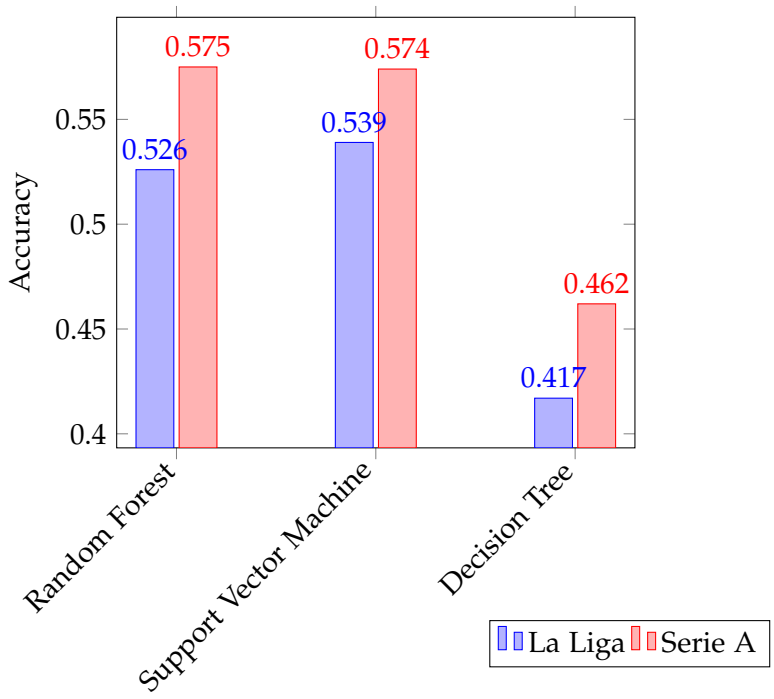
Research Question 3: How does each model perform when applied to a different league?

In La Liga, the Random Forest models showed moderate accuracy levels ranging from 0.522 to 0.526 but were not profitable, with losses between -96.70 to -179.99 euros. On the other hand, in Serie A, these models performed significantly better, both in terms of accuracy (0.550 to 0.575) and profitability (profits ranging from 62.70 euros to 395.60 euros). SVM models in La Liga showed slightly better accuracy (up to 0.539) but were unprofitable, having losses from -31.69 to -115.89 euros. In Serie A, only one SVM model (SVM1) showed profitability, while others reported losses. The Decision Tree model showed a better performance in Serie A compared to La Liga. In La Liga, it had the lowest accuracy with 0.417 and the largest loss with -315.4 euros. On the other hand, in Serie A, a similar model achieved a profit of 335.59 euros despite a low accuracy of 0.462.

Table 15: Accuracy score of models for La Liga and Serie A

	La Liga	Serie A
Random Forest	0.526	0.575
Support Vector Machine	0.539	0.574
Decision Tree	0.417	0.462

Figure 5: Accuracy score of models for La Liga and Serie A

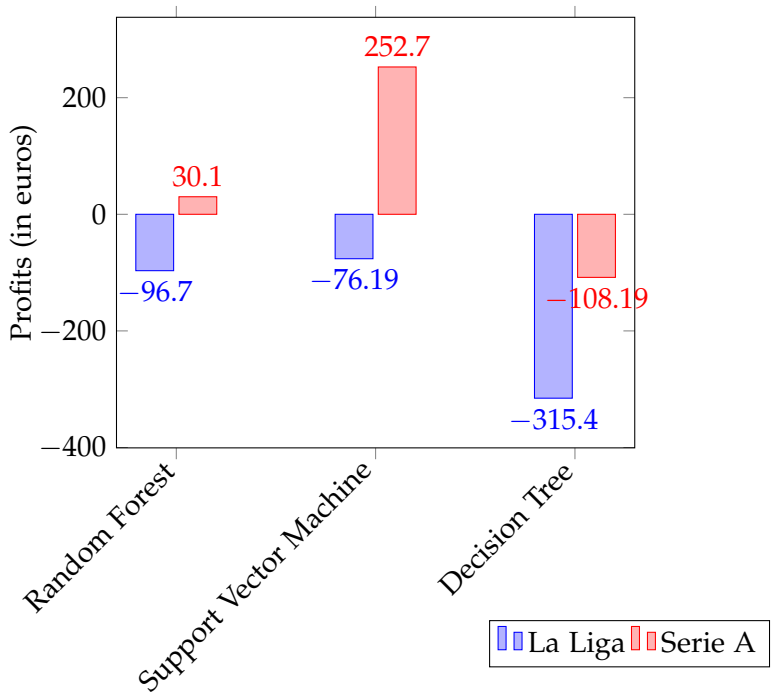


We can notice that for all the models, there is an increased accuracy for Serie A compared to La Liga. Below, there is a table and a bar plot that illustrate how each algorithm performed in each league in terms of profits.

Table 16: Profits of models for La Liga and Serie A

	La Liga	Serie A
Random Forest	-96.70	395.60
Support Vector Machine	-31.69	146.09
Decision Tree	-315.4	335.5

Figure 6: Profits of models for La Liga and Serie A



From the above table, an obvious improvement in the profits of all models can be detected in Serie A compared to La Liga.

Sub-Question: Is the same model the best for both leagues?

No, the Support vector machine has the best accuracy score in La Liga, while Random Forest has the highest accuracy score in Serie A. For the two best models in each league, the profit margin is measured to be able to compare the results of those models with the results of other works. The profit margin for SVM in La Liga is -1.43%, while the profit of Random Forest in Italy is 7.43%. Through profit margin, we can make precise comparisons with other works, avoiding the problem of different amounts of bets.

Sub-Question: Are the same features the most important for both leagues?

The features considered most important for achieving the best accuracy scores vary between La Liga and Serie A. The SVM model, utilized for feature selection, identified different sets of features for each league. For

La Liga, the SVM model identified 9 features from the total 26 predictors as important for achieving the best accuracy score. On the other hand, for Serie A, the Random Forest model utilized only 5 features from the total 26 to achieve the best accuracy score. In summary, while there is some overlap in important features, the variation in the selected predictors demonstrates distinctions in the two leagues, which leads to different feature selections.

5 ERROR ANALYSIS

5.1 *Confusion Matrices*

Table 17: Confusion Matrix for the models, La Liga

Decision Tree Confusion Matrix

	Predicted Away	Predicted Draw	Predicted Home
Actual Away	60	47	63
Actual Draw	28	34	52
Actual Home	46	74	128
Accuracy	0.44	0.21	0.52

Random Forest Confusion Matrix (RF2)

	Predicted Away	Predicted Draw	Predicted Home
Actual Away	71	18	81
Actual Draw	22	14	78
Actual Home	39	14	195
Accuracy	0.53	0.30	0.55

SVM Confusion Matrix (SVM2)

	Predicted Away	Predicted Draw	Predicted Home
Actual Away	65	0	105
Actual Draw	23	0	91
Actual Home	26	0	222
Accuracy	0.57	0	0.53

In the confusion matrix of the Decision Tree model for La Liga, the accuracy scores for predicting away and draw outcomes are low (0.44 and 0.21, respectively). This indicates that the model struggles to distinguish between these classes, leading to poor performance. The model performs better in predicting home wins, achieving an accuracy of 0.52. In addition, it has been noticed that this model has made more predictions in the draw class than the other two models

The Random Forest demonstrates a moderate performance with an accuracy of 0.55 in predicting home wins, 0.53 in away wins, and 0.30 in draws. Random Forest surpasses the Decision Tree scores in each class.

The confusion matrix of the SVM model can be characterized as unusual, considering that it has not made any predictions about the draw class. The zero predictions in the draw class suggest that the model might struggle to learn patterns associated with the draw class, possibly due to its lower representation in the dataset. However, in this way, a moderate accuracy score is achieved based on previous works. The SVM achieved better overall accuracy compared to Random Forest and the baseline model, having a 0.57 accuracy score for away wins and a 0.53 accuracy score for home wins.

Table 18: Confusion Matrices for the models, Serie A

Decision Tree Confusion Matrix

	Predicted Away	**Predicted Draw**	**Predicted Home**
Actual Away	83	43	53
Actual Draw	33	40	47
Actual Home	44	62	123
Accuracy	0.51	0.27	0.55

Random Forest Confusion Matrix (RF1

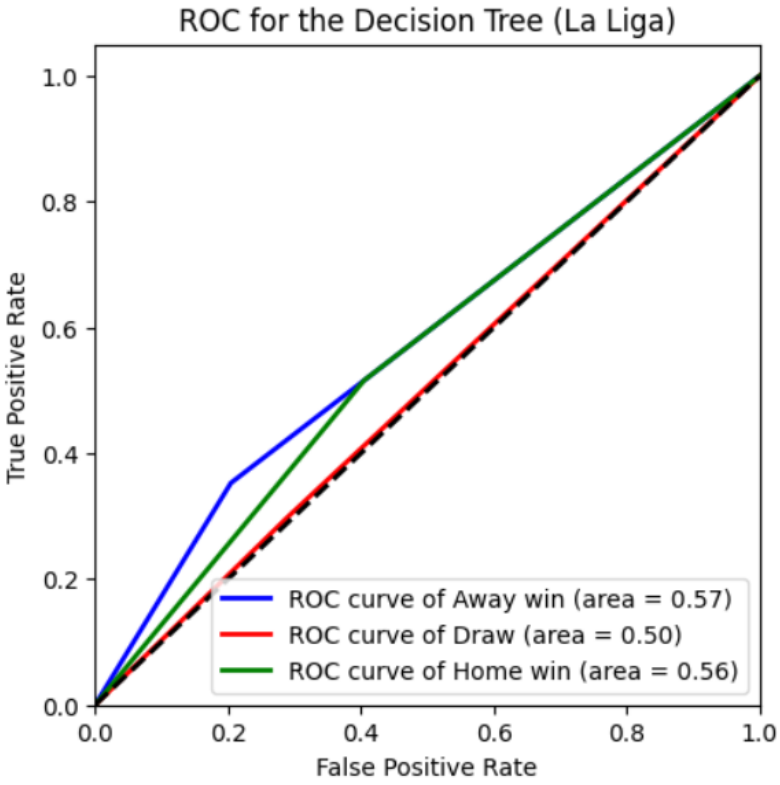
	Predicted Away	**Predicted Draw**	**Predicted Home**
Actual Away	105	12	66
Actual Draw	42	13	65
Actual Home	33	8	188
Accuracy	0.58	0.39	0.58

SVM Confusion Matrix (SVM1

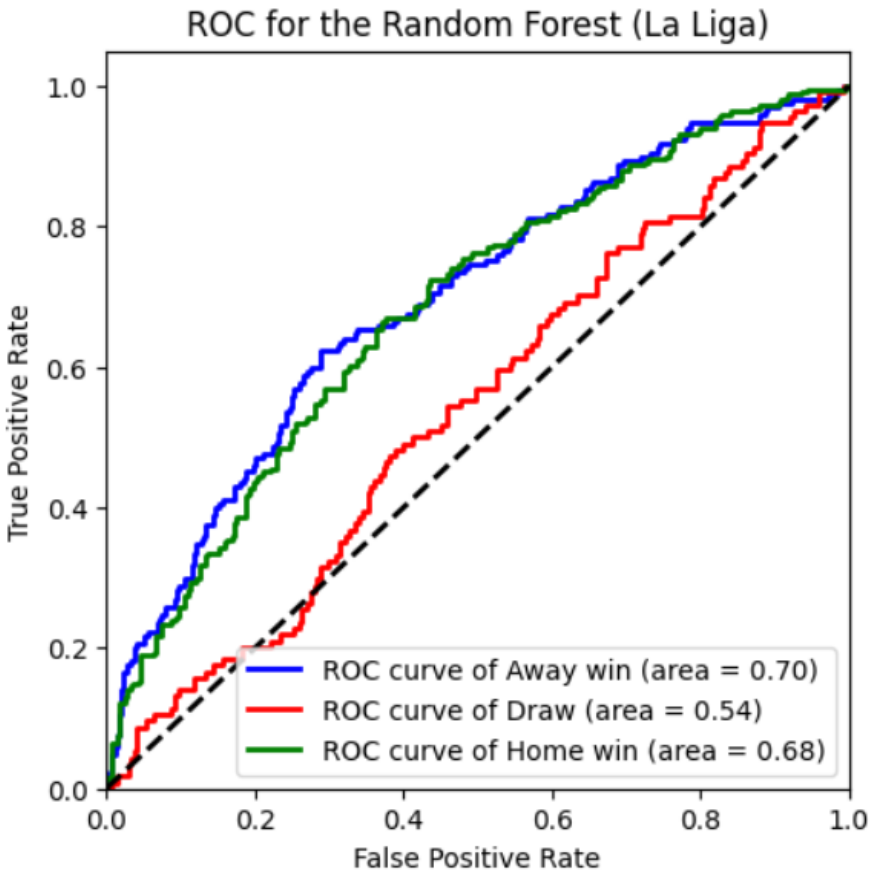
	Predicted Away	**Predicted Draw**	**Predicted Home**
Actual Away	94	0	89
Actual Draw	32	0	88
Actual Home	17	0	212
Accuracy	0.65	0	0.54

The Decision Tree model for Serie A struggled to accurately predict draws, achieving only 0.27 accuracy in this class and predicting home and away wins with accuracies of 0.55 and 0.55, respectively. The Random Forest model performed better across all classes than the Decision Tree. However, because we also measure the profits, we can notice that the Decision Tree predicted correctly 40 draws something that can explain why the Decision Tree has more profits than SVM and is also very close to reaching the profits of Random Forest. Random Forest achieved 0.39 accuracy in predicting draws, 0.58 for away wins, and 0.58 for home wins. The SVM model in Serie A demonstrated strength in predicting away wins, achieving an accuracy of 0.65 and 0.54 in home wins. However, the model did not make any predictions for draw outcomes, as in La Liga.

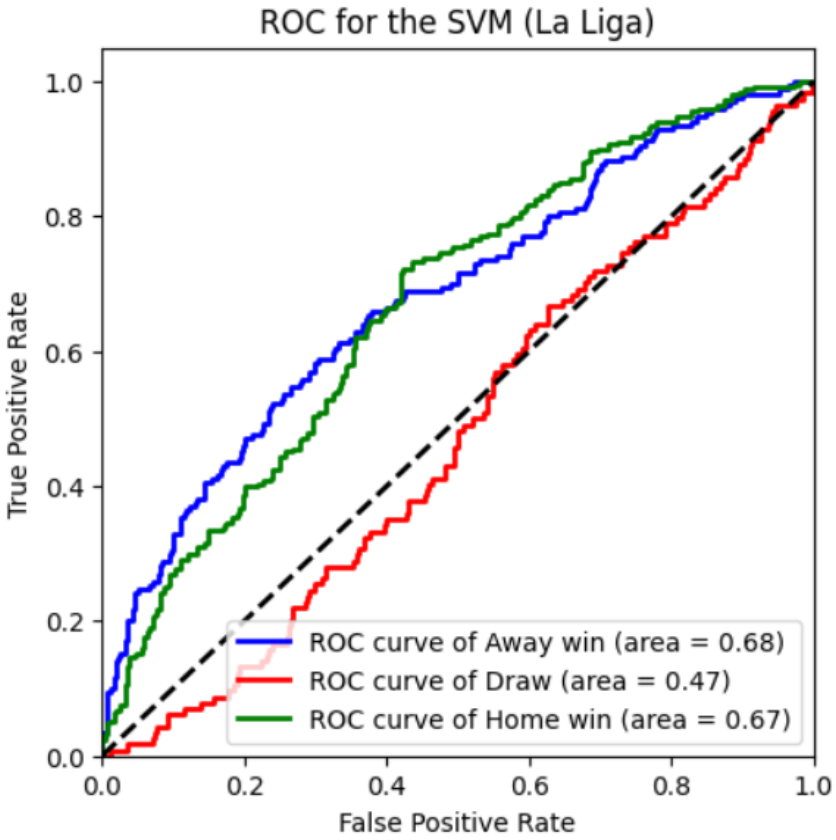
5.2 *Analysis of ROC AUC*



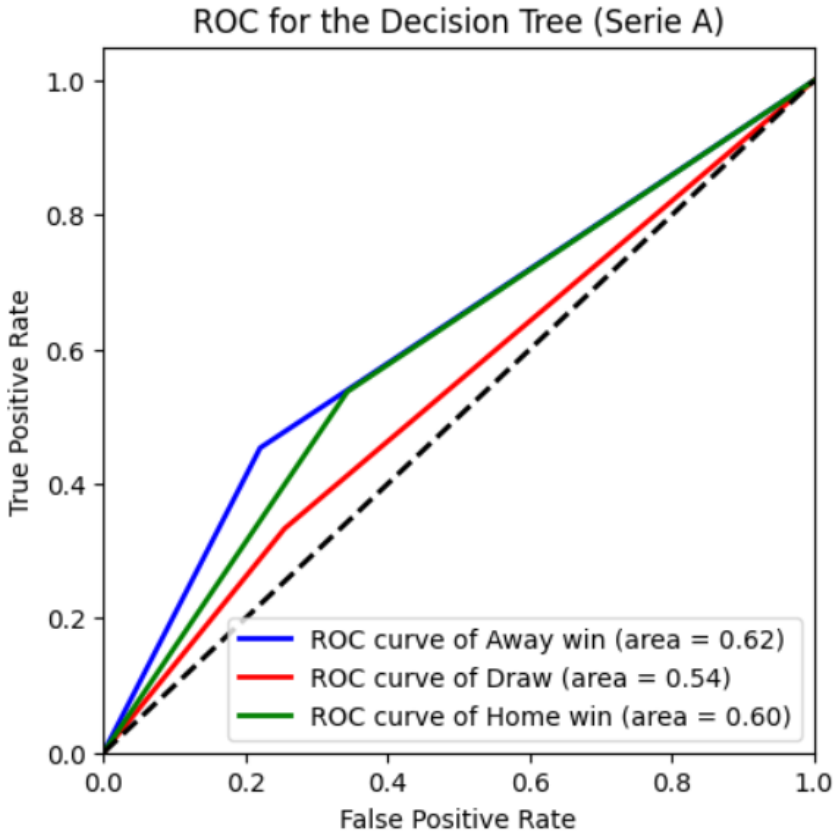
The ROC AUC values for the Decision Tree model in the La Liga dataset show poor performance in class discrimination. The ROC AUC value for the away win is 0.57, slightly above the threshold of random chance (0.5). The ROC AUC value for the draw class is exactly 0.50, showing no discriminative power above the random prediction. Home win class is similarly low at 0.56.



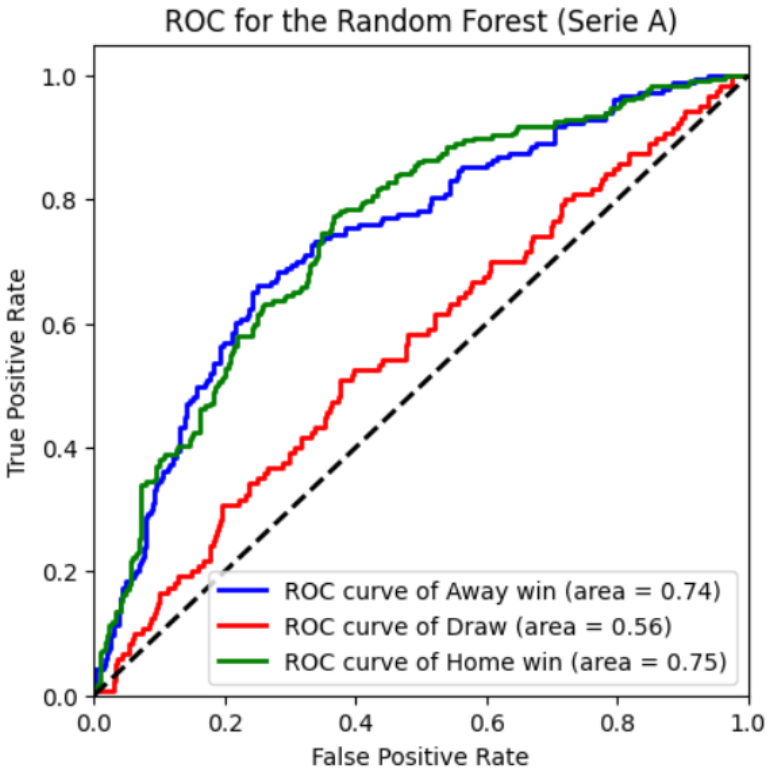
The ROC AUC values for the Random Forest model for La Liga show an improvement over the Decision Tree with ROC AUC values of 0.70 for away win and 0.68 for home win, suggesting a modest ability to distinguish these classes. However, the draw class has a ROC AUC value of 0.54, which is only marginally better than a random prediction.



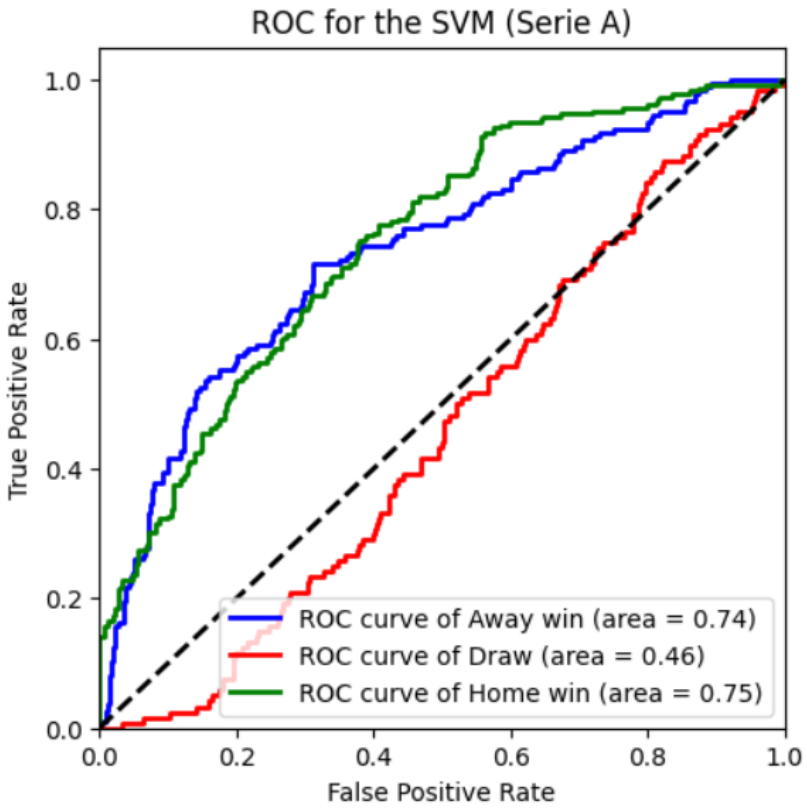
The SVM for La Liga produces ROC AUC values of 0.68 for the away win and 0.67 for the home win class suggesting the results as moderate. However, for the draw class, ROC AUC has a value of 0.47, indicating a lack of ability to recognize draws.



The Decision Tree in Serie A shows a ROC AUC value of 0.62 for the away win class. The draw class, with a ROC AUC value of 0.54, slightly surpasses random predictions for this class. For the home win class, the predictions are slightly better, with a ROC AUC value of 0.60.



The Random Forest ROC AUC values for Serie A demonstrate a moderate performance in identifying away win and home win classes, with ROC AUC values of 0.74 and 0.75, respectively. However, its ability to predict draws is weakened, with a ROC AUC value of 0.56, only marginally surpassing random prediction (0.5).



The ROC AUC values for the SVM in Serie A perform modestly in distinguishing away and home wins, with ROC AUC values of 0.74 and 0.75, respectively. However, it notably underperforms in the draw class, with a ROC AUC value of 0.46.

6 DISCUSSION

This research aimed to create features that can result in accuracy scores that outperform previous works. In La Liga, all the models used did not demonstrate profitability. However, SVM and Random Forest achieved moderate accuracy scores based on previous works. In Serie A, one of the SVM models that operated achieved profitability while the other two had losses. All the Random Forest models achieved profitability, with the best one achieving an accuracy score of 0.575. These results support the hypothesis that, in some cases, machine learning can be effectively applied to football statistics for profitable betting, as we have seen in previous research. The analysis of feature importance revealed variations in the important predictors for achieving the best accuracy scores in La Liga and

Serie A. While some overlap exists, there are differences in the predictors of the two leagues. The feature 'DPPS (difference in points in the previous year) had the most important role compared with any other feature, being the best predictor in La Liga and Serie A. The second most important feature, both in La Liga and in Serie A is the 'DiffPts' which shows the difference between the two opponent teams in terms of points.
SVM increased its score in Serie A by 4%, compared to the accuracy score in La Liga, while Random Forest increased its score in Serie A by almost 5%. The same happened with SVM. The baseline model (Decision Tree) increased its accuracy score in Serie A in comparison with the accuracy score in La Liga by almost 5%. As the accuracy scores increased in Serie A, the same occurred with profits. In general, there is an increased performance in Serie A compared to La Liga in terms of both accuracy and profitability. In Steijns (2020) there is a reversed tendency compared with this thesis, considering that in La Liga they achieved better results compared to Serie A. The only exception was Random Forest, which achieved the same score in both leagues.
The experimental feature selection method, which multiplied Pearson correlation values with feature importance scores generated by Random Forest, was effective. This method resulted in a feature set that achieved the highest accuracy score both for Random Forest and SVM models used for Serie A analysis. This is a promising result, considering that it shows that this innovative method can have significant outcomes.
Comparing the results of this thesis with prior studies, such as Rodrigues and Pinto (2022) and Steijns (2020), the models of the current thesis demonstrated a moderate performance. A Random Forest model is seen to have a profit margin of 26.78% euros with an accuracy score of 0.652 in the research of Rodrigues and Pinto (2022), outperforming the best Random Forest model in the current thesis (for Serie A), which had an accuracy score of 0.575 with a profit margin of 7.43%. However, we have to mention that Rodrigues and Pinto (2022) used a dataset from football matches in the Premier League (English first division), which may have remarkable differences from that in La Liga and Serie A. In Steijns (2020), Xgboost achieved a 0.56 accuracy score in La Liga, outperforming the best SVM of the current thesis for La Liga (accuracy score 0.539). In addition, in Steijns (2020) the XGboost algorithm achieved an accuracy score of 0.54, which was outperformed by the Random Forest of the current thesis (accuracy score of 0.575).
In Remmen (2022), they used only match statistics to make predictions, reaching an accuracy score of 0.632, but it is not clear if they used match data before the game to make predictions. Hence, we cannot compare this accuracy score with the best scores produced in the current thesis.

Moreover, the current thesis outperformed Cintia et al. (2015) in terms of accuracy in Serie A, considering that they had an accuracy score of 0.55 and almost the same accuracy in La Liga at 0.53. In the study of Gomes et al. (2021), a decision tree model predicted outcomes of Premier League football matches, achieving an accuracy of 0.549. This performance was outperformed by the Random Forest model of the current thesis (accuracy of 0.575), which was applied to Serie A matches. However, Gomes et al. (2021) achieved a larger profit margin (20%) than the best model in the current thesis (7.43%).

6.1 *Limitations and future works*

A limitation of this research is the functionality of the Support Vector Machine which made 0 predictions for the draw class both in La Liga and in Serie A. It is acknowledged that a more in-depth investigation of SVM may lead to an improvement in accuracy. A potential contributing factor to the observed limitation could be the class imbalance problem, demanding further exploration.
Generally, a limitation of the current thesis is the poor performance of the models to predict draw class. That aligns with other related works, which also mention that as a limitation. In future work, an investigation into improving the predictions of draw class can help the field make more accurate predictions. It is worth mentioning that in terms of profit, that can help in gaining important profits because traditionally, the odds of draw class are relatively high. In addition, this thesis focused on the accuracy score, trying to make the best possible prediction based on the accuracy metric. However, smaller accuracy can achieve higher profits, and that is noticed by Decision Tree. The baseline model has more profits than all SVM in Serie A, even with a smaller accuracy score. From that, we can infer that a good betting strategy could produce optimistic profits.
Moreover, there is a limitation to the data that is used. More features can be found that predict football match outcomes efficiently. Market values could be one of those, considering that they can be related to the results between two teams. The salaries of the coaches of the teams could be an interesting one, given the notion that teams that take the selection of coaches so seriously could have demanding objectives. In addition, further investigation of the teams can shed light on the staff that plays an important role in the performance of a team.
The data pre-processing in this thesis was extended. However, a more thorough analysis of the features could lead to better results. For instance, for the features that corresponded to the last 20 home and away games,

respectively, there may be another number (10, 30, 40, etc.), which may achieve a higher correlation with the target value.
Another limitation is the selection of hyperparameters for the Random forest. More precisely, for the random forest, an unlimited 'max_depth' was selected in the current study. However, it is essential to note that there is a potential risk of overfitting associated with an unlimited tree depth. Future research could explore alternative hyperparameter configurations, including constrained values for 'max_depth', to confirm a balance between accuracy and overfitting.

6.2 *Conclusion*

This thesis is based on match statistics to predict football match outcomes before the beginning of the football game. Several new features emerge to achieve the best feasible result, while 3 feature selection methods are employed to reveal the best feature set for the models. Random Forest, Support Vector Machine, and Decision Tree are employed to predict football outcomes, resulting in some profitable models for Serie A and without profitable models for La Liga. SVM achieved the best accuracy score and profit for La Liga, while Random Forest was the most profitable model in Serie A. Both the best models for each league used different feature sets to achieve their accuracy scores.

REFERENCES

Alfredo, Y., & Isa, S. (2019). Football match prediction with tree based model classification. *International Journal of Intelligent Systems and Applications*, *11*, 20–28. https://doi.org/10.5815/ijisa.2019.07.03

Anaconda software distribution. (2020). https://docs.anaconda.com/

Cintia, P., Pappalardo, L., Pedreschi, D., Giannotti, F., & Malvaldi, M. (2015). The harsh rule of the goals: Data-driven performance indicators for football teams. https://doi.org/10.1109/DSAA.2015.7344823

Dewi, C. (2019). Random forest and support vector machine on features selection for regression analysis. *International journal of innovative computing, information control: IJICIC*, *15*, 2027–2037.

Gomes, J., Portela, F., & Santos, M. F. (2021). Decision support system for predicting football game result. https://api.semanticscholar.org/CorpusID:52993204

Holmes, B., & McHale, I. G. (2023). Forecasting football match results using a player rating based model. *International Journal of Forecasting*. https://doi.org/https://doi.org/10.1016/j.ijforecast.2023.03.002

McKinney, W. (2010). Data Structures for Statistical Computing in Python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 56–61). https://doi.org/10.25080/Majora-92bf1922-00a

Patil, S., Kate, A., Wavare, K., Gujar, M., & Bachav, G. (2023). Predicting football match results using machine learning. *International Journal of Creative Research Thoughts*, *11*, g407–g414. http://www.ijcrt.org/papers/IJCRT2304812.pdf

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Remmen, J. (2022). *Predicting and analyzing football outcomes with match and player statistics: Using several machine learning algorithms* [Unpublished doctoral dissertation]. Tilburg University [Cognitive Science and Artificial Intelligence]. http://arno.uvt.nl/show.cgi?fid=161491

Rodrigues, F., & Pinto, Â. (2022). Prediction of football match results with machine learning [International Conference on Industry Sciences and Computer Science Innovation]. *Procedia Computer Science*, *204*, 463–470. https://doi.org/https://doi.org/10.1016/j.procs.2022.08.057

Samba, S. (2019). *Football result prediction by deep learning algorithms* [WorldCat: https://tilburguniversity.on.worldcat.org/oclc/1362447407]. http://arno.uvt.nl/show.cgi?fid=149223

Steijns, B. A. G. M. (2020). The characteristics of winning football matches: Determining a model that can predict the match outcome of football matches. http://arno.uvt.nl/show.cgi?fid=156292

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. https://doi.org/10.3390/app10010046