JÖNKÖPING UNIVERSITY
School of Engineering

Beating the odds

Machine Learning for football match prediction

PAPER WITHIN *Computer Science*
AUTHORS: *Emil Christoffersson*
TUTOR: *Ulf Johansson*
JÖNKÖPING *June 2023*

This exam work has been carried out at the School of Engineering in Jönköping in the subject area Computer Science. The work is a part of the Two-year Master of Science in Engineering programme. The authors take full responsibility for opinions, conclusions and findings presented.
Examiner: Vladimir Tarasov
Supervisor: Ulf Johansson
Scope: 30 credits
Date: 2023-06-14

Mailing address:
Box 1026
551 11 Jönköping

Visiting address:
Gjuterigatan 5

Phone:
036-10 10 00 (vx)

Abstract

This study aimed to compare the accuracy of machine learning models with the probabilities generated by sports betting companies in predicting the outcome of football matches. The study also investigated the impact of different feature combinations on the performance of machine learning models for predicting football match outcomes. The study used data from various sources of the Swedish football league between the seasons 2018-2022. The comparison between the model's predictions and the probabilities generated by sports betting companies showed that the model's predictions were more accurate. Support Vector Machines(SVM) performed the best with an accuracy of 52.4 percent compared to the betting companies at 40.4 percent. The results also showed that different feature combinations can have a significant impact on the performance of machine learning models for predicting football match outcomes but the importance of features varied depending on the selection method used. The study used four different feature selection approaches: filter methods, Lasso, Ridge, and PCA, to identify the most important features for prediction.

Overall, the results of this study suggest that machine learning models can outperform sports betting companies in predicting football match outcomes and that the choice of feature combination can have a significant impact on model performance. Further research is needed to explore these findings in more detail and to investigate the usefulness of different feature selection techniques at different points in the season.

Keywords

Machine Learning, feature selection, prediction, sports betting, football

Acknowledgements

I would like to express my gratitude to my supervisor Ulf Johansson, for his guidance, support, and insights throughout this research.

Contents

1 Introduction

The sports industry has been a major source of entertainment and data generation, and the increased use of data analysis in sports has created a need for advanced methods of analysis. As noted by (Schumaker et al., 2010), the volume of data created in the world of sports, both from player performance statistics and fan interaction data, has raised the need for techniques of analysis that may assist companies in making better decisions. In this industry, machine learning has emerged as a critical tool, allowing sports companies to evaluate massive volumes of data, find trends, and make data-driven choices (Jordan & Mitchell, 2015).

Machine learning is utilized in the sports sector to evaluate team performances and game plans. Organizations may optimize their operations and even gain an advantage over competitors by processing massive volumes of data and spotting trends (Tichy, 2016). For instance, the Italian football team AC Milan reported a 90 percent reduction in injuries during the first year after collecting 60,000 data points on each player (Davenport, 2014). In order to save costs and improve team or player performance, businesses might utilize machine learning to predict potential adverse circumstances (Sarlis & Tjortjis, 2020). By providing useful insights, machine learning helps organizations to make data-driven decisions that have a significant impact on their performance. This has led to increased spending on machine learning technology and research by the sports sector.

Football has grown to be a massive industry, and as a result of its enormous popularity, more data is now being produced and utilized to guide business decisions. Utilizing this information, betting sites may determine the likelihood of several outcomes in a sporting event, including match and goal results as well as additional betting possibilities. The businesses' confidentially of this information offer them an advantage over competitors and enables them to provide more accurate odds while also generating more revenues.

Data sources can range from prior performance information, which is important for betting sites since it helps them to make well-informed judgments as well as understand the possible results of matches. Furthermore, this data can be used to create predictive models that will assist these companies in fine-tuning their strategies for better results. Betting sites can get a considerable competitive advantage in the market by utilizing such data. Online gambling in the European Union is a thriving industry with annual revenue of 38.2 billion euros in gaming gross wins. The market is expected to grow to 54.3 billion euros by 2027, with online sports betting being a major contributor accounting for 40 percent of the market size (European Gaming and Betting Association, 2022).

Additionally, the information produced by the sector can be used to better understand the sport and how it functions. This may help in the development of insights into certain strategies. Such data-driven insights may help sporting organizations, fans, teams, and players make better decisions and achieve a competitive advantage in this field.

Football game predictions are challenging to make since they must take into consideration a number of variables that might have an impact. For instance, injuries to key players can significantly affect the outcome, while extreme weather conditions can also have a crucial role to play by influencing the performance of teams. In addition to this having the support of the home fans and being more familiar with the surroundings can generate advantages as well. Another thing that is worth considering is the form. A team in good form is more likely to perform well in a match.

1.1 Problem Formulation

Football match prediction is a critical subject for a number of reasons. First of all, it may significantly affect the sports betting market as well as bets made by fans. Making educated judgments in the sports market may be improved by accurate projections, which can offer insightful information about the area of sports analysis, especially when taking the viewpoint of bookmakers or oddsmakers into account. The effectiveness of these models and their potential use in the sports market may be evaluated by comparing the accuracy of the model's predictions to the probability produced by sports betting organizations. Moreover, the concept of using statistical analysis for building cost-effective sports teams was popularized by the book "Moneyball," published in 2004 (Lewis, 2004). Since then, several methods have been created in this area, including machine learning models for forecasting football game results. Identifying the key features in the data that are most important for accurate predictions can result in the development of improved models for predicting match outcomes. This has significant effects on machine learning and data science researchers since it can assist them in better grasp the capabilities of these models. Furthermore, taking into account the randomness that exists in football matches, studying this phenomenon can provide insights into the reasons why the underdog often wins in football games, despite statistical odds being against them.

This study also offers insights into the performance of various machine-learning models and algorithms in addition to its impacts for the sports business. This has implications for machine learning and data science researchers since it can assist them in understanding the capabilities of these models.

1.2 Purpose and Research questions

This study aims to investigate the potential of using machine learning methods for predicting football matches. The research is lacking in its expertise and understanding of the potential and effectiveness of applying machine learning models to predict football game results and compare their performance to the probability produced by sports betting organizations. By examining the usage of machine learning techniques for football match predictions and identifying the most crucial features and the most effective algorithms for precise forecasts, the study seeks to close this gap. The following research questions are addressed by the study:

> *RQ1: How accurate are the model's predictions compared to the probabilities generated by sports betting companies?*
>
> *RQ2: What is the impact of different feature combinations on the performance of machine learning models for predicting football match outcomes?*

This study has the ability to advance our understanding of the possibilities of utilizing machine learning techniques to foresee football game results and compare those predictions to the probabilities produced by sports betting organizations. By examining the use of machine learning techniques together with the key features for accurate predictions, this study's ultimate purpose is to make a contribution to the field of machine learning. There will also be an evaluation of the model's prediction performance and effectiveness, including any potential flaws and areas for development. The insights obtained from this analysis will be crucial in ensuring that the model is up to par and delivering the desired results.

1.3 Outline

Section 2 provides a review of related studies that address the challenge of making sports game predictions, offering valuable context and insights into existing approaches.

Section 3 goes into the methodology of the project, explaining the approaches used and the steps taken to achieve the desired results.

Section 4 presents the results of the project, showcasing its findings and accomplishments. This section offers an in-depth examination of the project's outcomes and what they mean for the field of sports game prediction.

Section 5 presents the conclusion of the work, summarizing the key findings and highlighting its significance and impact. This section provides insights into future directions for research in the field of sports game prediction.

Lastly, Section 6 offers a comprehensive discussion and analysis of the work performed. This section provides a critical evaluation of the results and methods used and explores the implications and potential applications of the work.

2 Theoretical Background

This section explores the theoretical background of sports betting and odds, as well as the field of machine learning. In addition, numerous sports prediction techniques that have emerged and been put to use recently are examined. By examining these topics, a more thorough understanding of the complexities involved in sports betting and the techniques that are being used to predict sporting outcomes can be gained.

2.1 Sports betting and odds

Sports betting has a rich and fascinating history, with roots dating back to ancient civilizations. It has developed and gained popularity throughout the years, becoming an essential component of modern sports culture as well as changed over time as cultures and technologies have advanced. In the 18th and 19th centuries, many of today's popular sports were created and refined, and betting on sporting events soon became common (Forrest, 2012). Sports betting has progressed along with technology. The growth of online betting has been affected by both the popularity as well as the availability (Lopez-Gonzalez & Griffiths, 2016). The arrival of online bookmakers and betting apps for mobile devices has simplified the betting process, making engagement easier and more convenient for users. These developments have elevated sports betting to an important part of the market, enabling simple and smooth interaction among bettors all over the world. The increased accessibility of online betting has greatly contributed to its growing popularity and the convenience offered through websites and mobile apps has made the world of sports betting more accessible and appealing to a wider audience. Along with this, gambling ads have become more frequent, especially on television and through sponsorship. This goes to show that 95 percent of football shirts with English football clubs are endorsed by gambling businesses (Bunn et al., 2019). The rise of online betting has also led to increased competition, resulting in more betting options for customers. One of the key concepts in sports betting is odds, which in other words is a different expression of a probability (Grimes & Schulz, 2008). According to (Cortis, 2015) there are various methods for showing odds such as European, English, or Hong Kong. European odds, sometimes known as decimal odds, are the inverse of probability and show the likelihood of an event occurring.

$$EuropeanOdds(Decimal) : \frac{1}{probability} \tag{1}$$

When both numbers are integers, English or fractional odds show the ratio of the wager to the amount that would have been won if the wager had not been placed. The ratio of the chance that an outcome won't occur to the likelihood that it will be used to compute them.

$$EnglishOdds(Fractional) : \frac{probability}{1 - probability} \tag{2}$$

Hong Kong odds display the net winnings from a one-unit wager, as opposed to the total payout including the original wager, as with European odds. As a result, Hong Kong odds are always one unit less than their European counterparts.

$$HongKongOdds : \frac{1 - probability}{probability} = \frac{1}{probability} - 1 \tag{3}$$

In the field of sports betting, the importance of odds cannot be overstated since they act as a vital check and balance between the interests of the gambler and the bookmaker. Offering accurate odds and guaranteeing a steady profit margin are the bookmaker's two main goals (Štrumbelj, 2014). Keeping the integrity of the betting process needs a complex balancing act since the chances must be equal for both sides. In order to do this, bookmakers carefully consider all of the potential influences on a sporting event's result and modify their odds as necessary. Along with ensuring fairness and balance in sports betting, odds are

also employed to determine the amount of money that will be returned on a wager. The amount a bettor can anticipate winning on a winning bet is calculated based on the odds that the bookmaker has set. A bettor who places a winning bet will receive the amount of the payback. Bettors now have access to a greater variety of alternatives as online sports betting gains popularity. There are now several sorts of sports betting, giving people additional opportunities for participation in the sports betting industry.

It is also important to remember that the odds might shift. There are both opening lines i.e. when the sport betting companies release the odds of a sporting event as well as closing lines which are the odds that is available right before an event starts. In addition to this, there are live odds as well, this is when the event is already underway and you can in real time bet on different possibilities.

2.2 Predictive Modelling

An important part of predictive analytics which is a field in data analytics that uses both historical and present data to predict future behaviors, trends, and activities, is predictive modeling (Kuhn, Johnson, et al., 2013). Although producing accurate predictions is the main objective of predictive modeling, understanding the model and how to interpret its results are equally important. Predictive modeling is typically applied to data that is in tabular format, where each row represents an instance and each column represents a feature (Guyon et al., 2008). In predictive modeling, creating a mathematical function that connects feature values to the target variable is the main goal. The target variable can either be continuous, which is known as regression, or categorical, which is referred to as classification. Predicting the category or class of an input based on a collection of characteristics is the objective of a categorical target (D. Powers & Xie, 2008). As stated by (Montgomery et al., 2021), regression models are commonly used to predict a numerical value (i.e., a continuous target variable) based on a set of predictor variables. For example in a football game, the features may include the names of the home team, the away team, the date of the game, the number of goals scored by each team, etc. The objective might be continuous, such as the number of goals scored (regression), or categorical, such as whether a team won, lost, or drew the match (classification). In situations when the target variable is known and there is a large amount of labeled data available, supervised learning is a popular method for predictive modeling (Goodfellow et al., 2016). This describes how the algorithm in predictive modeling learns given a collection of characteristics and known target values for each instance in the data. The model then employs this knowledge to anticipate the goal value for new unforeseen scenarios. The goal is to create a model that predicts the target value for every instance in the dataset as well as for any future cases the model might run into.

For predictive modeling to produce the best outcomes, there are factors to consider when selecting a modeling technique. Considerations such as data complexity and volume are important. While some algorithms work better on smaller datasets there are also algorithms that work better on larger datasets. This also applies to models where some are better suited for simple issues while others are more suited for more complicated problems. To see how well a model will work on new data, cross-validation can be used. During cross-validation, data is separated into numerous groups and make use of a training as well as a validation set (Hastie et al., 2009). For example, if we use k-fold cross-validation, the data is split into k groups. Then the algorithm is trained on k-1 groups and the remaining group is used for validation. This is done k times, each time using a different group for validation. Cross-validation can help to obtain a more reliable estimate of the algorithm's performance, particularly when the available data is limited, and can also help to avoid overfitting the training data.

2.3 Metrics

To evaluate the performance of a classification model, various measures are used, such as Accuracy, Precision, Recall, and F1-Score (D. M. Powers, 2020). The percentage of the model's correct predictions out of all of its predictions is what is known as accuracy. Precision is the percentage of true positives

out of all positive predictions, while Recall counts the percentage of actual positives among all positives. F1-Score on the other hand combines both Precision and Recall to provide a balanced evaluation. It helps to balance the tradeoff between these two metrics and is calculated as the weighted average of Precision and Recall, with equal weights assigned to each measure.

Assessing the performance of a classification model with these metrics is important for determining the model's ability to correctly predict the class labels of unseen data across various domains. For instance, in sports betting, properly forecasting the result of a football game is essential for betting organizations to make precise payouts. Since it assesses the model's performance as a proportion of all predictions, it would be a key metric to consider in this case. On the other side, while evaluating player performance, it's crucial to correctly identify the best players during a match for team selection and training. As it reflects the model's capacity to accurately forecast the top performers out of all positive predictions, Precision would be an important parameter to take into account in this situation. The metrics chosen to evaluate the performance of a classification model are both dependent on the problem and the type of model. By using appropriate metrics to evaluate the performance of a classification model, decisions can be made about the model's effectiveness and areas for improvement can be identified.

2.4 Feature Selection

In order to create machine learning models, it is very important that features which are most useful for predicting the target variable is chosen first. Both the accuracy as well as the effectiveness of the models can be increased by applying feature selection techniques. It is also shown that feature selection can increase the predictive model accuracy while at the same time lowering the computing cost of the learning process (Yu & Liu, 2003). Another research by (Dash & Liu, 1997) also demonstrated how feature selection may greatly decrease the time needed for model training and prediction. Furthermore, (Guyon & Elisseeff, 2003) points out that feature selection may find and eliminate unimportant or distracting characteristics, resulting in a more precise and understandable model. Moreover, lowering the number of features through feature selection might assist in overcoming the dimensionality curse and enhance a model's generalization performance. By doing this the model may be made simpler and easier to understand by feature selection by reducing unnecessary features.

Filter approaches analyze the significance of features independently of any specific learning algorithm and pick the most relevant characteristics based on statistical metrics. According to (Chandrashekar & Sahin, 2014), filter approaches are quick and computationally efficient, but they may not always choose the optimum collection of features for a specific learning algorithm. On the other hand, Wrapper approaches treat feature selection as a search issue and use a specialized learning algorithm to evaluate the performance of different subsets of features. Wrapper approaches, according to (Guyon & Elisseeff, 2003), can be more accurate than filter or embedding methods, although they can be computationally costly and prone to overfitting. Lastly, embedded methods, which include feature selection directly into the learning process, generally use regularization techniques. These approaches represent a sort of embedded feature selection since feature selection and parameter estimate are conducted together throughout the optimization process. Although embedded approaches are computationally costly, they frequently outperform filter or wrapper methods by successfully decreasing overfitting and boosting model accuracy through feature selection.

Method	***Type***	***Description***
Chi-squared test	Filter	Measures the dependence between the feature and target
F-test (f_classif)	Filter	Computes the ANOVA F-value between the feature and target
Mutual information	Filter	Measures the mutual information between the feature and target
Recursive Feature Elimination (RFE)	Wrapper	Selects features by recursively considering smaller and smaller sets of features and ranking them by importance using a given machine learning algorithm
Genetic Algorithm (GA)	Wrapper	Iteratively select features by evolving a population of solutions and ranking them by their fitness using a genetic algorithm
Principal Component Analysis (PCA)	Embedded	Linear dimensionality reduction technique that transforms the data into a lower-dimensional space while retaining most of the original variance
Lasso	Embedded	Linear regression model that penalizes the absolute size of the regression coefficients
Ridge	Embedded	Linear regression model that penalizes the squared size of the regression coefficients

Table 1. Feature selection methods in machine learning.

As seen in table 1, various feature selection methods in machine learning are categorized based on their types, Filter, Wrapper, and Embedded. The table provides information about the name of the method, its type, and a brief description of what it does. The methods listed include classic techniques such as the Chi-squared test and F-test, as well as more advanced methods such as Recursive Feature Elimination and Genetic Algorithm.

2.5 Modeling Techniques

For machine learning projects to be successful, modeling approaches are essential. They are used to create predictions, sort data into various categories, and find patterns and correlations in data. The type of data, the issue being handled, and the individual project needs all influence the modeling approach that is used. It is crucial to thoroughly weigh the advantages and disadvantages of several modeling approaches before choosing the one that is best suitable for the task at hand.

One popular modeling technique is the decision tree which is a simple yet powerful supervised learning algorithm used for classification and regression tasks. The decision trees work by recursively splitting the input data subsets based on the values of features, creating a tree-like structure that eventually leads to a prediction (Kingsford & Salzberg, 2008). The resulting tree can be interpreted and visualized, making it useful for understanding the relationships between the input and the output. While using it parameters such as the maximum depth of the tree, which determines the number of splits and the complexity of the model are important to look into. Another important parameter is the splitting criterion used to determine how to divide the data at each node, such as Gini impurity or information gain. Additionally, decision trees can be prone to overfitting, so techniques such as pruning or using ensemble methods like random forests can improve the model's performance. In a variation of decision trees known as a "random forest," which is built during training, a large number of decision trees are built. The output of the random forest is a class that represents the mean of the classes (classification) or mean prediction (regression) of all the individual trees (Breiman, 2001). The number of trees in the forest, the depth of each tree, and the

number of features at each split are some critical factors to consider while building a random forest. Because it might impact the model's accuracy and stability, the number of trees is a particularly crucial element. While increasing the number of trees can improve accuracy but at the same time, it can increase the complexity and training time. The depth of each tree determines the level of detail in the decision rules learned by the model. A deeper tree can capture more complex relationships in the data but may also lead to overfitting. While random forests are an effective ensemble method that can improve the performance of decision trees, another popular ensemble method is gradient boosting. Unlike random forests, which combine multiple decision trees in parallel, gradient boosting involves combining multiple weak learners such as decision trees sequentially to correct the errors of the previous trees. This iterative approach results in a strong predictive model that can handle complex relationships and produce accurate predictions (Natekin & Knoll, 2013). It works by iteratively fitting new models to the residual errors of the previous model so that each subsequent model focuses on the hardest examples that the previous model got wrong. The final prediction is the weighted sum of the predictions of all the models. However, when using gradient boosting, it is important to consider several important parameters to optimize the model's performance for a given dataset and target variable. These include the number of iterations or trees, the learning rate or shrinkage, the depth of each tree, and the subsampling rate of the training data. The number of trees controls the complexity of the model, while the learning rate controls the contribution of each new tree to the final prediction. The depth of each tree determines the balance between bias and variance, and the subsampling rate controls the amount of randomness and diversity among the training instances. Another similar approach is LogiBoost which combines AdaBoost, a well-known boosting method, with logistic regression, a famous classification algorithm for binary outcomes. This method trains the logistic regression model on the original dataset and uses it as a base learner. The weak logistic regression models are then combined using the AdaBoost method to create a stronger ensemble model, which can provide predictions for fresh data that are more accurate (Collins et al., 2002). Following retraining of the logistic regression model using the weighted data, the AdaBoost method weights the data points according to their classification mistakes. Each iteration of this procedure adds a new logistic regression model to the ensemble, and the process is repeated a predetermined number of times. The number of base learners (i.e., logistic regression models) that are used in the ensemble, the learning rate, the maximum depth of the AdaBoost algorithm's decision trees, and the regularization parameter utilized in the logistic regression models are some crucial factors when utilizing LogiBoost.

Another popular modeling technique is Support Vector Machines (SVMs) which are supervised learning algorithms that may be applied to classification or regression problems (Noble, 2006). The fundamental principle of SVMs is to identify a hyperplane in a high-dimensional space that divides the various classes in the training data. The hyperplane is selected to maximize the margin, which measures the separation between the hyperplane and the nearest data points for each class. Support vectors, which are the data points closest to the hyperplane, are used to establish the hyperplane's location. The kernel function that is used to convert the input data determines whether an SVM is linear or nonlinear. The linear kernel, which maps the data into a linear hyperplane, and the radial basis function (RBF) kernel, which is nonlinear and maps the data into a higher-dimensional space, are the two most often used kernel functions. When using SVMs, some parameters to consider are the kernels to be used, the regularization parameter C, and the kernel parameter y(RBF). Even though SVMs can be useful they can at the same time be computationally expensive and can require careful tuning of the hyperparameters to achieve good performance. Naive Bayes, on the other hand, is an effective probabilistic technique for classification problems. It is founded on the Bayes theorem, which estimates the likelihood that an event will occur given previous knowledge of circumstances that may be relevant to the occurrence (McCallum, Nigam, et al., 1998). The algorithm assumes independence between the input data's features, simplifying the calculations and making it very efficient for large datasets. There are various crucial parameters to take into account while utilizing Naive Bayes. Which type of Naive Bayes algorithm to use is one of these parameters. This is because there are different varieties of it such as Gaussian Naive Bayes and Multinomial Naive Bayes, that may be better suitable for certain types of data. The smoothing parameter which helps in dealing with the problem of zero probabilities is also an important parameter to consider. This can occur when a feature in the test data has not been observed in the training data. K-nearest neighbors (k-NN) which is another popular non-parametric algorithm used for classification tasks. The k-NN can be used both for classification as

well as regression. It works by finding the k closest points in the training data to the new input and using the label or value of those points to make a prediction. When using the k-nearest neighbors algorithm, there are several important parameters to consider (Peterson, 2009). The algorithm's performance can be significantly impacted by the choice of distance metric used to compare the similarity of data points. Furthermore, it is important to carefully choose the value of k because a greater number may result in overfitting.

Another method for tackling classification and regression issues in machine learning is artificial neural networks (ANNs). Neurons, which are typically connected components of an ANN, translate a series of inputs into the appropriate output. What gives ANN its power is the non-linearity of the hidden neurons in changing the weights that go into the final decision. An ANN's output typically depends on the input features as well as other network components like weights. Weights linked with connected components are continually changing to attain high levels of prediction accuracy (Witten et al., 2005). To make use of an ANN, there are different parameters that are important to understand that can be adjusted in order to optimize the performance. Some parameters to take into account can be the number of hidden layers and neurons which determines the capacity and complexity of the model, the activation function which controls the non-linearity of the hidden neurons and can have a significant impact on the model's performance, the learning rate which determines the step size of weight updates during training and the regularization parameters, which can help prevent overfitting. Furthermore, the loss function that is used during training can also have an impact on how well the model performs. Multilayer Perceptrons (MLP) are one form of ANN frequently employed in machine learning. They consist of numerous layers of linked "neurons," or nodes, where each neuron in one layer is connected to every neuron in the following layer. Weights are assigned to the connections, which indicate how important the inputs are to the neuron's output. In order to reduce the loss function that measures the difference between the predicted output and the actual output, the MLP is trained using a supervised learning algorithm that can modify the weights of the connections between the neurons (Haykin, 2009). When using an MLP, there are several important parameters to consider that can greatly affect the performance of the model. The number of hidden layers and neurons, activation function, learning rate, and regularization strength are all important factors to consider. Tuning these parameters can optimize the performance of MLP for a given task and dataset.

2.6 Related Work

Through the years there have been many attempts to predict sports results with the help of machine learning. One of the earliest models to predict sports outcomes was conducted in 1996 with the help of an artificial neural network (ANN) (Purucker, 1996). Purucker, the article's author, offered advice on how to distinguish between superior and inferior teams using unsupervised techniques built on clustering when predicting the result of National Football League (NFL). The model was trained using the backward propagation method, which makes it easier to determine a loss function's gradient regarding all of the network weights (Rumelhart et al., 1986). The accuracy score in this article was 61 percent, which was lower than the expert's accuracy score of 72 percent. All of the information utilized was gathered from the tournament's first eight rounds. No team that was superior to its opponent in every category has ever lost, hence those categories in particular were picked. A team generally has a higher chance of winning a game the more categories it controls. Thus, the relative strength of teams may be determined by contrasting these statistical characteristics. Even though this study was conducted in 1996, it is important because it was one of the earliest attempts to predict sports outcomes using machine learning techniques. This study demonstrates the potential of using artificial neural networks to predict the outcome of sports events, and it laid the groundwork for future studies that built upon this approach. Additionally, the focus of the work on clustering and unsupervised learning techniques for finding patterns in sports data is still relevant today. These strategies may be used to discover essential features of successful teams and players in a variety of sports and events. The study also emphasizes the significance of feature selection in forecasting sports results. Despite utilizing just data from the first eight rounds of the tournament, the model was able to achieve good accuracy by concentrating on key factors that are most predictive of a team's relative

strength. This technique can assist guide the feature selection in current machine-learning models for sports prediction.

According to a follow-up study in 2010, a similar approach to Purucker's work was used to forecast the football matches of the 2006 World Cup. The study employed a Multilayer Perceptron (MLP) and backward propagation to anticipate the winning percentage of two teams based on official statistical data from earlier stages (Huang & Chang, 2010). In contrast to traditional categorical classification methods used in football, the authors used binary classification, where one class was for the home team and the other for a draw result with the away team. This shows that machine learning approaches may be tailored to individual demands and obtain higher results by using binary classification rather than categorical classification. Their approach achieved an overall accuracy average of 76.9 percent by utilizing only eight of the original 17 categories in the data. This is important to note because it shows that machine-learning techniques may be used to predict sports results with high accuracy.

The fuzzy categorization technique offered an alternative method for predicting the results of sporting events. To predict National Basketball Association (NBA) games, ten fuzzy rule learning algorithms were chosen, assessed, and contrasted with traditional linear regression (Trawiński, 2010). The text employs a majority vote method together with feature selection techniques to pinpoint the most illustrative characteristics. They make use of a Knowledge Extraction based on Evolutionary Learning (KEEL) (Alcalá-Fdez et al., 2008) system to contrast linear regression with the fuzzy approaches. They utilize the KEEL system because it has a variety of algorithms for data preparation, experiment design and execution, postprocessing, evaluation of outcomes, and presentation of results. The characteristics of the model were selected by the author using the open-source, nonprofit Waikato Environment for Knowledge Analysis (WEKA) data mining tool (Hall et al., 2009). Two rounds of testing were carried out for the paper's experiment section. Predicting the outcome of the first test's game using data from the previous three contests was the objective. The team's most recent performance and season-long statistics were used to estimate the results for the second test. For predicting the results of basketball games, the algorithm (Clas-Fuzzy-Chi-RW) with the highest accuracy on both the test and training sets were selected.

These implementations are all reliant on particular sport-specific data categories. After an evaluation of components associated with various sports, a combination of comparable data categories was chosen where 11 categories were included in all sports (McCabe & Trevathan, 2008). The research underlines the significance of choosing the relevant data categories for a certain sport. For the actual sports it was decided to choose American football, football, and rugby as the sports in question. An MLP model trained using backward propagation and conjugative-gradient techniques were employed by McCabe and Trevathan. Rugby was a sport where the model thrived, with an average accuracy of 67.5 percent. In addition, this evaluation included football stats from the English Premier League. When the outcomes of this study were compared to those of a competition called TopTipper, where top human experts also made weekly predictions on the same games, the model was able to get the highest percentile among the other human experts. The significance of this research lies in its capability to conduct a comprehensive comparison of various sports using a consistent set of categories, and demonstrate superior performance than even the most skilled human experts.

The issue of forecasting the results of NBA games has also been resolved by the use of data mining tools. This was looked at by Miljkovic et al. (Miljković et al., 2010) in a classification issue. To improve the system's classification performance, feature selection and normalization were used. The classification algorithms K-nearest neighbors, Naive Bayes, Decision trees, and Support vector machines (SVM) were all tried. Every time a game was played, the system's data was updated, and the algorithm then provided a probabilistic prediction of subsequent games based on the updated dataset. The Naive Bayes classifier with feature selection and normalization delivered the best outcomes. Over 778 games, the system's accuracy was 67 percent. In a comparative analysis of several models for forecasting college basketball games (Shi et al., 2013), the authors highlight how characteristics seem to be more important than models. The attributes used and how they were computed were the main factors influencing performance differences. Although different averaging techniques employed during the season have an influence on the quality of

the findings, both feature selection techniques and the performance of classifiers trained on those representations are utilized to validate factors. When other or additional characteristics were included, the results were weaker. Decision trees, rule learners, ANNs, particularly an MLP, Naive Bayes, and random forest were the models that were contrasted in this work. Their study shows that using MLP produces a high level of accuracy. Both these studies hold significant importance as it showcases the potential of data mining tools in resolving complex issues. By utilizing advanced classification algorithms and data preprocessing techniques such as feature selection and normalization, the study demonstrates the effectiveness of machine learning in improving the accuracy of predictions. This can have the potential to help answer one of the research questions of this report with the help of the data preprocessing techniques that were conducted.

The issue has also been addressed using Bayesian networks. In their study, (Constantinou & Neil, 2012) use Bayesian modeling to analyze football game outcomes over the course of two Premier League seasons. Their approach considers a number of factors for both home and away teams, including team strength, team form, team psychology, and tiredness to produce the outcome prediction. There are 21 different factors that make up these parts, and each one has an impact on the prediction's result. The study demonstrates how including subjective input considerably increased the model's capacity for forecasting. Their research also highlights the value of Bayesian networks, which enable the representation and display of subjective data.

There have also been projects where locating the most important elements has been a priority. In their work, (Baboota & Kaur, 2019) create a feature set for determining the most important elements for predicting the results of a football game in their study. A very precise prediction system is then created using feature engineering and exploratory data analysis. They divide all of the characteristics into Class A and Class B categories in order to find the most useful ones. Class B has differential traits, whereas Class A contains all of the individual characteristics for both the home and away teams. They performed SVM, random forest, Naive Bayes, and gradient boosting models on the feature set, fine-tuning each model for maximum effectiveness. The gradient boosting model turned out to be the best of all the models when the same models were evaluated for the actual prediction.

By classifying the players at different stages of play for both teams. a supervised learning technique that makes use of an SVM model with various kernels, such as linear, nonlinear poly, and RBF, to predict the result of a cricket match against a certain side (Jayanth et al., 2018). The best outcomes were achieved by the SVM utilizing the RBF kernel, with accuracy and precision of 75 percent and 83.5 percent, respectively. The importance of this approach lies in its ability to fine-tune different models, such as SVM, random forest, Naive Bayes, and gradient boosting, to improve the prediction accuracy of forecasting sports prediction. By fine-tuning these models, this research can improve its understanding of how the model works best in forecasting sports outcomes.

(Hucaljuk & Rakipović, 2011) created a system where multiple tests were run to determine the best method for predicting the results of football matches in order to settle on the appropriate combination of characteristics and classifiers. The results of prior games, the number of goals scored and allowed, and the current form were all taken into account. They evaluated several classifiers using these characteristics, including ANN, random forest, Naive Bayes, Bayesian networks, and LogitBoost. With a prediction accuracy of 68 percent, ANN produced the best results.

Since the initial studies in the 1990s, the application of machine learning techniques in sports prediction has come a long way. The experiments discussed in this section show that machine learning approaches have the capacity to properly forecast the results of sporting events. These studies also highlight the importance of feature selection, data preparation, and model evaluation in achieving high accuracy in sports prediction which will be taken into account in this report.

2.7 Research Gap

The related work area contains a variety of alternative approaches for forecasting sports match results that have been presented. Even though there has not been much study done to compare the odds provided by sports betting organizations with the forecasts made by machine learning algorithms, it has been demonstrated that these models are successful at producing reliable predictions. How well these predictions compare to the odds offered by bookmakers is unclear. This comparison is crucial since it may reveal how effective these models could be as well as suggest areas where they could be more accurate. As a result, a research that investigates the relationship between predictions based on machine learning and sports betting odds is clearly needed in order to shed light on the effectiveness of these prediction models.

3 Method and implementation

This chapter examines several techniques for gathering data as well as techniques for analyzing it. A thorough knowledge of the research approach employed in this study will be presented through a detailed analysis of these techniques.

3.1 Data Understanding

A detailed overview of the distribution of real football game outcomes and the data source utilized in this study are provided in this part to help the readers. Understanding the data source is essential for evaluating the validity and dependability of the findings and recommendations made in this study. The distribution of results from actual football games will help readers better understand and evaluate the statistics offered later in this book.

3.1.1 Data Collection

The study's data were created by combining data from three different sources. Using the Python module Beautiful Soup, which extracts data from files including HTML and XML (Richardson, 2007), two of these datasets were collected through web-scraping from two separate websites, Dataset 1[1], Dataset 2[2]. The third dataset was collected from a third website, Dataset 3 [3], which was not web scraped as the other datasets.

The first dataset, Dataset 1, provides a thorough examination of four seasons of Allsvenskan, Sweden's highest division in football. The dataset makes use of several aspects including the number of shots, shots on target, goals scored and allowed by each side, as well as other elements like time, day, location, possession, yellow and red cards, yellow and red cards, fouls, crosses, and formation. Every match appears twice in the statistics, totaling 960 matches and 1920 rows, to guarantee that the metrics for each side are appropriately recorded. This method enables a more extensive and in-depth examination of each team's performance throughout the course of the seasons.

The second dataset, Dataset 2, includes data from the same four football seasons as the previous dataset. However, instead of focusing on match statistics, this dataset provides ratings for each team. These ratings were obtained from FIFA, a popular computer game created by Electronic Arts (EA). The ratings included in this dataset cover different aspects of a team's performance, such as attack, midfield, defense, and overall rating. These ratings are essential in the game, reflecting a team's real-life performance on the field. By utilizing these ratings, the dataset provides an additional layer of analysis of the teams' performance, allowing for a more in-depth examination of their strengths and weaknesses.

The third dataset, Dataset 3, includes information from 968 football matches, including the season, date, time, teams playing at Home and Away, and the corresponding Home and Away goals, as well as the match result. Additionally, it includes the probabilities for Home, Away, and Draw outcomes in the form of odds, sourced from the betting website Pinnacle, with consideration given to the maximum odds available across various betting sites, and the average odds calculated from different betting sites. All these odds are the closing lines of that specific game i.e. the odds that is available just before the start of the game. It is worth noting that this dataset also incorporates qualifiers from the lower league into Allsvenskan, and therefore does not match the number of matches in Dataset 1.

[1] https://fbref.com/
[2] https://www.fifaindex.com/
[3] https://www.football-data.co.uk/

Besides these features that have been fetched from different sources, other features have been manually put together with the help of these datasets. Features such as the average points for each team when playing at home and away, and the form of the team have also been added which can be defined as how a team has performed over the last couple of games. These were also furthered extended to get the difference between the two teams. For example instead of average points for each team when playing at home and away another feature was added that was the difference between the two. This will extend the number of features in total but important to add these to see if these features can provide additional information. In total, there are 109 different features to choose from. This total also ramped up because the average of x amount of games in certain features was calculated. For instance, shots taken in the last 5 and 10 games were calculated for both the home and away teams. All features can be found in the appendix.

3.1.2 Distribution

The frequency of different match outcomes (Home Win, Draw, Away Win) can be examined across all the matches in the dataset to analyze the actual distribution of the games available from the data. This can be done by calculating the percentage of matches that fall into each category.

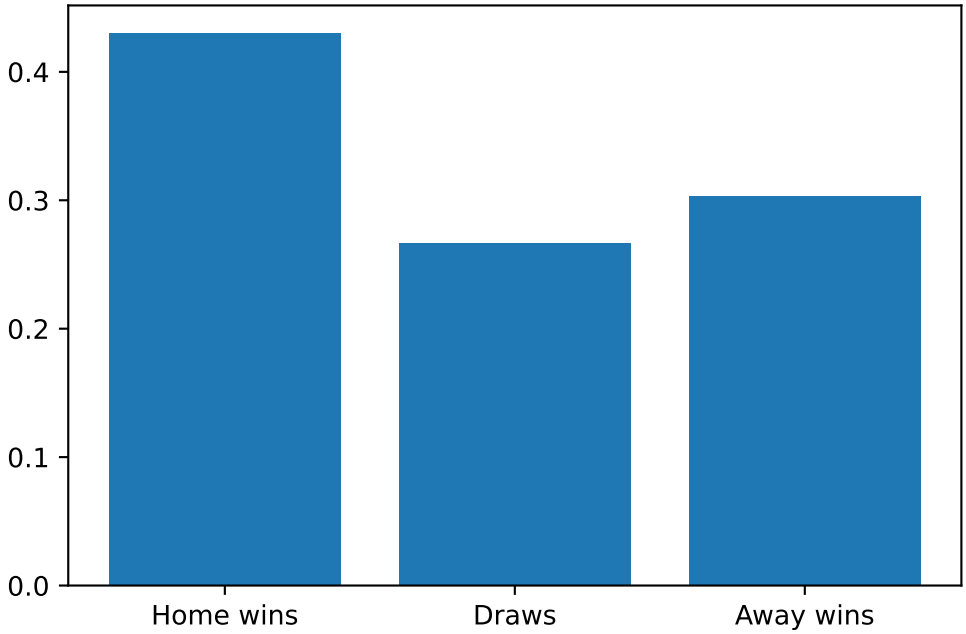

Figure 1. Figure over the outcome distribution for all 960 games in Allsvenskan.

Based on the available data, Home wins occurred in 43 percent of matches, Away wins occurred in 30 percent of matches, and Draws occurred in 27 percent of matches, see figure 1. These percentages suggest that Home teams have a higher likelihood of winning compared to Away teams, and Draws occur less frequently than either Home or Away Wins. It is worth noting that the actual distribution of match outcomes may be influenced by a range of factors. Understanding the actual distribution of match outcomes can provide valuable insights into the overall performance of teams in the league and help identify any underlying trends or patterns that may be useful for predicting match outcomes.

3.2 Data Cleaning

To integrate the three datasets, it was necessary to clean the data and confirm that the columns shared by the datasets had consistent values. Two crucial columns were selected for this purpose: the date column and the team column. Nevertheless, following closer analysis, it was discovered that the team column had a number of distinct versions for team names, which would need to be standardized before the datasets could be effectively combined.

Dataset 1	***Dataset 2***	***Dataset 3***
AIK	AIK Stockholm	AIK Solna
Falkenbergs	Falkenberg	Falkenbergs FF
Malmo FF	Malmo	Malmö FF
Ostersunds	Ostersund	Östersunds FK

Table 2. Different variations of team names.

Data inconsistencies can cause errors and complications in data analysis, making it difficult to get reliable findings. This can be demonstrated in table 2 where different names for the same football team occur. For the datasets to be properly merged and evaluated, it is essential to standardize team names. The data can be made more consistent and dependable by locating and resolving these differences, enabling accurate analysis.

Team	***Opponent***	***Shots***	***ShotsTarget***	***GoalsForward***	***GoalsAgainst***
HIF	MFF	10.0	3.0	1.0	2.0
MFF	HIF	18.0	7.0	2.0	1.0
DIF	AIK	17.0	5.0	2.0	2.0
AIK	DIF	9.0	3.0	2.0	2.0

Table 3. Features of the same game for both teams.

Home	***Away***	***Shots Home***	***Shots Target Home***	***Shots Away***	***Shots Target Away***	***Goals Home***	***Goals Away***
HIF	MFF	10.0	3.0	18.0	7.0	1.0	2.0
DIF	AIK	17.0	5.0	9.0	3.0	2.0	2.0

Table 4. Merged features for both teams.

Since every match appeared twice in Dataset 1 as seen in table 3 where the data columns represent the statistic for a specific team this had to be aggregated so one match only appeared once. This is important for several reasons. For starters, it simplifies data analysis and interpretation. Having two rows for each match, one for each team might make comparing the performance of the two teams in a given match difficult. Aggregating the data into a single row for each match enables a more clear study of one team's performance with respect to the other. Second, combining the data makes it possible to calculate statistics like the average number of shots or goals in each game more precisely. The same statistics are counted twice when a match appears twice in the dataset, which might bias the final findings. By combining the data, each game is only reported once, resulting in statistics that are accurate and true to the team's real performances which are shown in table 4.

As noted earlier Dataset 3 had 968 unique matches to the 960 matches in Dataset 1. This disparity in the number of matches may be traced to the qualifying matches that were present in Dataset 2 but not in Dataset 1. These qualification matches had to be eliminated to make the datasets comparable and consistent. Both datasets had the same amount of matches once the qualification matches were eliminated, enabling efficient comparison and analysis.

3.3 Feature Selection

Before model training and after data cleaning, the feature selection procedure was carried out. The process aimed to minimize the dimensionality of the dataset and find the most useful features for the predictive model in order to enhance model performance. Because the total amount of features staggered up to 109 different features some sort of feature selection had to be made. This is because having too many features might cause the model to suffer from the curse of dimensionality, which can lead to overfitting and poor generalization. By picking only the most informative features, the dimensionality may be minimized of the dataset and increase the performance of the model (Li et al., 2017). As mentioned in 2.4 there are different ways to go in order to choose the right feature. By integrating various feature selection approaches, a more robust and trustworthy assessment of the value of each characteristic can be obtained. The first approach made was to combine the filter methods, such as the chi-squared test, ANOVA F-test, and mutual information. The scores acquired from each filter technique are then normalized to guarantee comparability and equal weightage in the final results. Normalization reduces the effect of changing scales between features on feature ratings. The mean score was then calculated by averaging the results of all filter techniques. As it captures the strengths of each filter technique, this approach gives a more thorough evaluation of the value of each characteristic. They also allow us to rapidly and effectively discover the most relevant and least relevant features respectively. To keep the most relevant features a cutoff value, 0,127, which was the average from the filter methods importance scores was used. From the filter methods, the number of features was reduced from 109 to 34. To further build on this the same technique of combining different feature selection methods was used again but this time with Lasso, Ridge, and PCA from the remaining 34 features to see which features would contribute the most to the end goal.

To decide how many characteristics are best to employ with each classifier. The models were iterated through the reduced features, and for each iteration, the top n features were selected based on their importance scores as discussed in (Guyon & Elisseeff, 2003). The authors point out the variable ranking method, which involves ordering variables according to a scoring algorithm that determines their worth. This method is used to produce subsets of variables with decreasing importance that may be used as predictors. The models were then evaluated using a k-fold cross-validation method for each classifier and feature count. For each of the 10 folds, the models were tested on the validation set after being trained on the training set. To measure the best performance of each model, the highest accuracy along with the matching number of features, was used after iterating over all the models and features.

Although Recursive Feature Elimination (RFE) and Genetic Algorithms (GA) were discussed in 2.4, they were not used in this study. This choice was made after considering the tradeoff between computational efficiency and model performance. While RFE and GA may be capable of identifying a more optimal subset of features, they require significantly more computational resources and may not result in a significant improvement in model performance. Therefore, the combination of filter methods and regularized regression was deemed to be a suitable approach for this study, as it provided a balance between feature selection and model complexity.

3.4 Model Selection

In section 2.6 (Shi et al., 2013) mentioned that features were said to be more significant than the actual model but model selection is still a crucial step in machine learning, especially in choosing the best model for predicting outcomes. Different models also react differently to what features they are getting so a comparison between different models will take place. The models that will be tested are:

- Random Forest
- Logistic Regression

- Support Vector Machines (SVM)
- Gaussian Naive Bayes
- Gradient Boosting
- XG Boost

The chosen models cover a range of popular machine learning algorithms and have been successful in similar task as mentioned in 2.6. Each of these models has advantages and disadvantages and can perform differently depending on the data and situation at hand. As a result, it is important to evaluate the performance of many models in order to pick the best one for the specific purpose of forecasting football match results. For all the models in this research, they were run using their default settings.

In total Three seasons of Allsvenskan (2019,2020,2021) with 720 matches will be utilized for training, and one season (2022) with 240 matches will be used for testing. Neural networks may be an efficient model for this particular problem, but they frequently need a lot of data and computing capacity to train well. In this instance, there are just three seasons of Allsvenskan available for training, which might not be enough to train a complicated neural network which is why it is not included as a model.

3.5 Determination of better odds

Predicting the outcomes of sporting events is a difficult task, and many bettors rely on the odds supplied by sports betting organizations to influence their wagers. However, the odds supplied by sports betting organizations are not always correct and may be biased toward specific outcomes. Predictive models, on the other hand, can integrate several data sources to provide forecasts that may be more accurate and unbiased. As a result, it is critical to assess the predictive model's accuracy and compare it to the odds given by sports betting organizations. One method to assess these sources' accuracy is to compare their predictions for all three outcomes in a multi-class classification. In this experiment, a comparison between the predictions given by the model to the lowest odds set by sports betting companies will be made. This experiment intends to highlight the potential benefits of utilizing predictive models in sports betting and to provide insights into their accuracy and reliability. The experiment specifically seeks to establish whether the model can generate predictions that are more precise than the odds supplied by sports betting organizations.

In addition to analyzing the accuracy of predictive models and sports betting organizations in multi-class classification, it is also important to analyze their performance in binary classification. Although draws occurred in a significant number of matches which was seen in figure 1, sports betting businesses frequently underestimate the likelihood of a draw and favor home or away wins in their predictions as demonstrated in figure 2. To address this issue, the second experiment will assess a predictive model's ability to determine when sports betting organizations are correct or incorrect in their forecasts for binary outcomes. The experiment will specifically forecast whether or not betting companies successfully predict the outcome of a match. This experiment can provide insights into the potential profitability of utilizing the model to wager on binary outcomes by comparing the predictive model's performance with that of sports betting organizations.

Prediction	***HomeTeam***	***AwayTeam***	***AvgOdds***	***ModelOdds***
H	IFE	GIF	1.34	1.11

Table 5. Model odds edge over the average odds.

In addition to these methods, odds can be compared to one another. This means that probabilities generated by the model are converted to odds which are then compared to the odds generated by the sports betting companies. To compare the probabilities generated by the model with the odds offered by sports

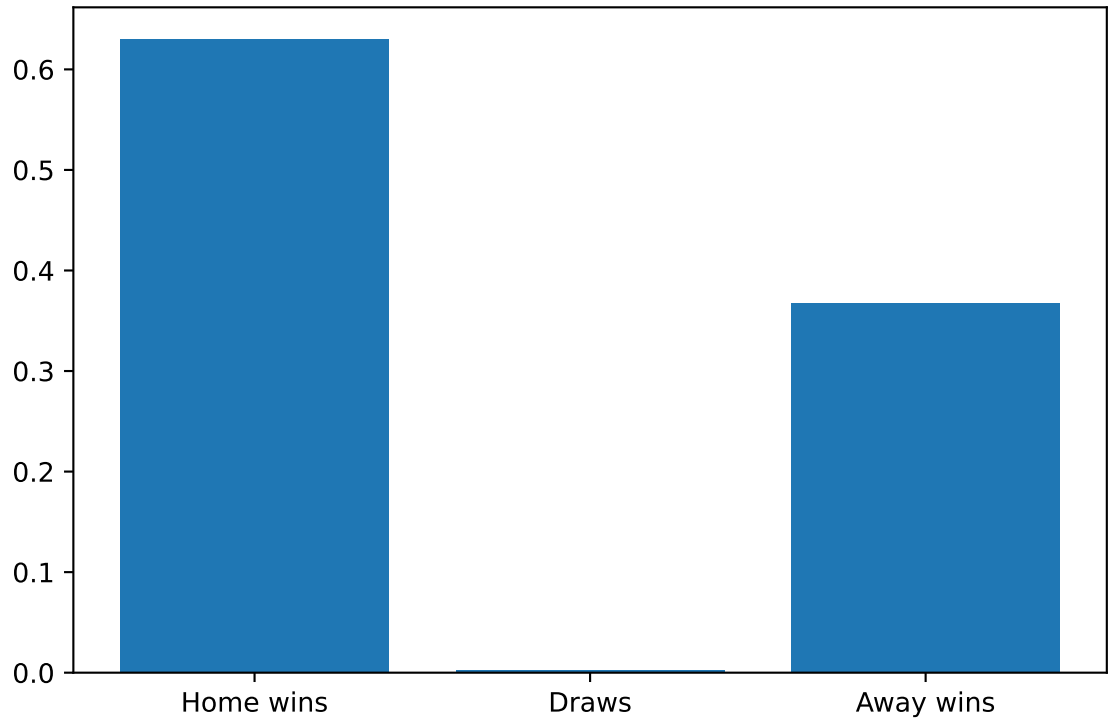

Figure 2. Figure over the bookie prediction distribution for all 960 games in Allsvenskan.

betting companies, it is important to ensure that they are in the same format. Since sports betting companies display their probabilities as odds, it is necessary to convert the model's probabilities to odds as well. To accomplish this, Equation 1 was used, as described in section 2.1. The conversion to odds enables a more direct comparison with the odds provided by sports betting businesses and assists in identifying the result with the highest probability of occurring. The comparison procedure is made easier and more understandable by transforming the model's probabilities into odds. It makes sure that both sets of probabilities are on the same scale and makes it possible to see more clearly how likely certain scenarios are. When the model's odds are lower than the actual odds, it has an advantage over the average odds. As a result, when the model projected a Home victory, just the average odds for the home team and the model's odds for the home team were used, as shown in table 5. As seen in the table, the model's odds are lower than the betting sites' odds, giving it an advantage. This is crucial since it allows a direct comparison between the model's predictions and the odds generated by sports betting businesses. This comparison may help measure the accuracy of the model's predictions and highlight scenarios in which the model outperforms the odds. Furthermore, the comparison can also assist in identifying successful betting chances. In order to test this for multi-class classification a bet of 1000 SEK will be placed every time an advantage occurs. That is, if the actual result differed from the prediction, you would lose 1000 SEK. If the actual result was right, these 1000 SEK were multiplied by the odds set by the bookies. When calculating the profit of binary classification a bet is placed whenever the model predicts if the lowest bookie odds are correct or not correct in their prediction of the lowest odds. If the model agrees with the lowest odds from bookies then a bet of 1000 SEK was placed on that outcome. If the model predicts the lowest bookie odds to be wrong then two bets are made, 1000 SEK each on the other two outcomes resulting in at least one bet for every game.

Finally, the expected value (EV) formula is a strategy used to calculate whether a bet has a positive or negative expected value depending on the probability of the event and the betting company's odds. The EV formula can be represented as:

$$EV = (P * O) - 1 \tag{4}$$

where:

- EV is the expected value of the bet
- P is the probability of the outcome
- O is the decimal odds offered by the betting company

For example, if a betting company offers odds of 2.5 for a home team win and the model predicts a probability of 0.6 for a home team win, the bet value calculation for a home team win would be:

$$EV = (0.6 * 2.5) - 1 = 0.5 \tag{5}$$

A positive EV value shows that the bet has a positive expected value, implying that a bet holds value. A negative EV number shows that the bet has a negative anticipated value, implying that a bet does not have value. With this method, a more proper measure of the profitability of each model and a comparison of their performance can be made. This experiment is important because the bettor can evaluate if a bet has a positive or negative anticipated value by calculating the expected value of the bet based on the likelihood of the result and the betting company's odds. This knowledge is critical for making informed betting selections and optimizing earnings. Furthermore, by utilizing the EV formula to evaluate the model's performance, its effectiveness can be directly examined and compared to that of sports betting businesses. To get the results from the EV formula a bet of 1000 SEK was made for an outcome every time it was the highest of all three possible outcomes and a positive number. If the EV number is a positive one this means that placing a bet on this outcome holds value.

4 Results

This section will detail the outcome of the strategy provided in the method section and responses to the research questions.

Research Questions:

> *RQ1: How accurate are the model's predictions compared to the probabilities generated by sports betting companies?*
>
> *RQ2: What is the impact of different feature combinations on the performance of machine learning models for predicting football match outcomes?*

As mentioned in 3.5 there were four possible ways to determine if the model's odds are better than the sports betting odds:

- Compare Accuracy in multi-class classification.
- Binary classification.
- Compare the model odds to sport betting odds.
- The expected value (EV) strategy.

4.1 Comparision of accuracy

The accuracy of various machine learning models for multi-class and binary classification problems will be compared in this section. The accuracy of the lowest odds of bookies will be compared to models such as random forest, logistic regression, SVM, Naive Bayes, gradient boosting, and XGBoost.

4.1.1 Multi-class Classification

	Lowest odds from Bookies	***Random Forest***	***Logistic Regression***	***SVM***	***Naive Bayes***	***Gradient Boosting***	***XG Boost***
Accuracy	0.404	0.511	0.520	0.524	0.518	0.493	0.478
Num of Features	-	24	7	32	12	10	31

Table 6. Performance on Multi-class classification.

Looking at table 6 one can see that the lowest odds provided by the sports betting companies are not very accurate. Only around 40 percent of these are correct while all the machine learning models hover between 50 to 52 percent thus making these models more accurate than the bookies which could lead to more accurate probabilities and odds in Allsvenskan. The highest accuracy comes from SVM with an accuracy of 52.4 percent. The results from the multi-class classification show that the machine learning models outperform the bookies in terms of accuracy. This indicates that machine learning models can be useful in predicting football match outcomes and provide more accurate probabilities and odds than the lowest odds from sport betting companies. This answers RQ1 that it is possible to obtain a higher accuracy on football matches compared to sport betting companies for the Swedish football league Allsvenskan.

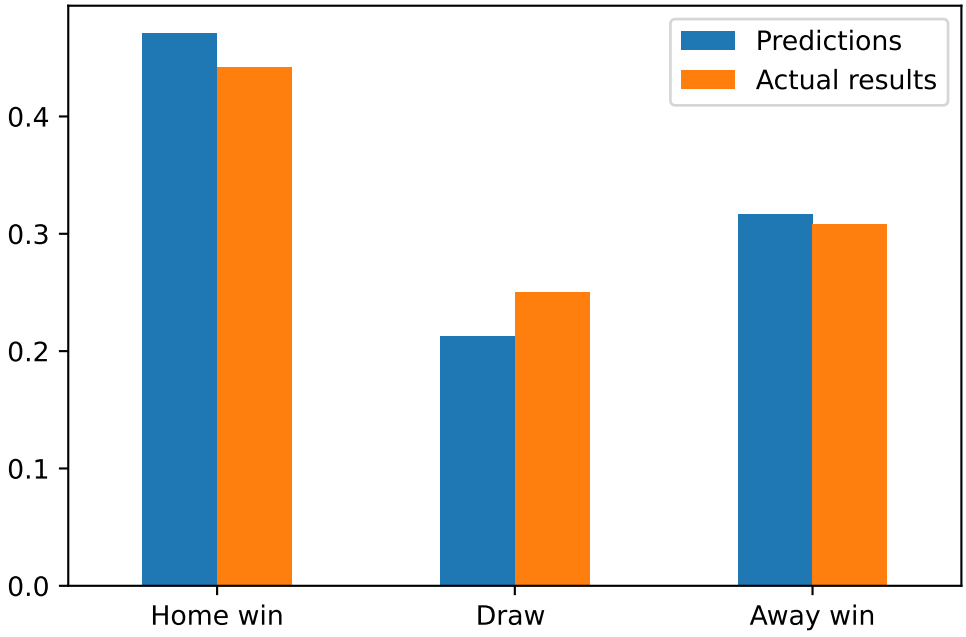

Figure 3. Figure over the outcome distribution of season 2022 for the actual result and the predicted result for Naive Bayes.

The most evident difference in the distribution of the outcomes was that home wins were projected somewhat more and draws were predicted significantly less, but away results remained roughly the same see in figure 3.

4.1.2 Binary Classification

For the binary classification, it is predicted when the bookie prediction is wrong or correct i.e. if the odds for the home team are lowest and the result is either a draw or away win this would be regarded as a wrong prediction. If the result is a home team then the prediction is regarded as correct. As mentioned earlier we know that the bookie accuracy is 40 percent thus making them their prediciton wrong 60 percent of the time.

	Random Forest	***Logistic Regression***	***SVM***	***Naive Bayes***	***Gradient Boosting***	***XG Boost***
Accuracy	0.622	0.637	0.641	0.614	0.618	0.604
Num of Features	18	7	4	5	4	24

Table 7. Performance on binary classification.

Table 7 provides information on the performance of different machine learning algorithms. SVM has the greatest accuracy score of 64.1, followed by Logistic Regression, which has a score of 63.7. The amount of features utilized in the classification job for each method is also included in the table. This information reveals the model's complexity and the number of characteristics required for accurate categorization. SVM, for example, employs only four characteristics, but XG Boost needs 24, showing that SVM may be a simpler model with fewer input data requirements. According to the findings of this experiment,

using binary classification could also provide more accurate forecasts and probabilities than sports betting companies.

4.2 Profitable models

The comparison of how profitable all the models are for both multi-class and binary classification will be evaluated in this section. Since the model's accuracy and precision were so close to each other as previously seen in table 6 and table 7 all the models will be evaluated. In addition to this, the EV strategy will also be presented.

4.2.1 Multi-class Classification

	Lowest odds from Bookies	***Random Forest***	***Logistic Regression***	***SVM***	***Naive Bayes***	***Gradient Boosting***	***XG Boost***
Profit (SEK)	-3,290	+11,830,	+17,850	-1,250	+6,060	+2,330	+1,610
Matches (Edge Occured)	240	65	66	64	101	72	108
Accuracy	-	0.58	0.58	0.44	0.56	0.51	0.53
Total Home pred (Accuracy)	-	44 (0.61)	38 (0.58)	43 (0.44)	63 (0.57)	49 (0.52)	71 (0.56)
Total Draw pred (Accuracy)	-	3 (0.33)	6 (0.67)	4 (0.25)	5 (0.60)	3 (0.33)	5 (0.20)
Total Away pred (Accuracy)	-	18 (0.55)	22 (0.55)	17 (0.47)	33 (0.52)	20 (0.56)	32 (0.50)

Table 8. Profit from the models when model odds are lower than actual odds.

As seen in table 8 the models were able to accurately anticipate the outcomes of the matches better than the lowest odds from the bookies. The Logistic Regression model produced the largest profit, generating a profit of 17,850 SEK over 66 matches. The 66 matches are in fact when the model odds are lower than the sport betting companies' odds as mentioned in the method. The Random Forest model also did well, gaining 11,830 SEK. When betting on the lowest odds from the sports betting companies it generated a loss of 3,290 SEK. This is also quite expected since sport betting companies wants to make money. This is important to include in this table since it serves as a benchmark to compare with the models. When the models were tested on individual outcomes, the Random Forest model predicted home victories the best, with an accuracy of 61 and a total of 44 total matches. With an accuracy of 67 and a total of 6 matches, the Logistic Regression model was the most accurate at predicting draws. The Gradient boosting model, on the other hand, predicted away victories the best, with an accuracy of 56 and a total of 20 matches. In terms of profitability, it's worth noting that the Logistic Regression as well as the Random Forest models,

which had the best accuracy in forecasting home victories, was also the most lucrative models. This shows that correctly anticipating home victories might be especially useful for betting. Nonetheless, the findings indicate that some machine learning models have the potential to beat traditional techniques, such as bookmakers' odds, and might be useful tools for forecasting football match outcomes.

4.2.2 Binary Classification

	Random Forest	***Logistic Regression***	***SVM***	***Naive Bayes***	***Gradient Boosting***	***XG Boost***
Profit (SEK)	+16,100	+26,550	+20,150	+3,740	-10,560	+4,420
Profit (Home)	+5,460	+5,320	-3,880	+5,610	-8,710	-1,280
Profit (Draw)	-6,020	+2,260	-12,140	-16,200	-19,000	-16,740
Profit (Away)	+16,660	+18,970	+36,090	+14,330	+17,150	+22,440

Table 9. Profit from the models in binary classification.

The total profit of the six algorithms is included in table 9. Since a bet is made for all matches as described in the method, every model in this experiment was tested on the whole test data. The findings reveal that Logistic Regression made the most money (+26,550 SEK), followed by SVM (+20,150 SEK). Gradient Boosting, on the other hand, made a negative profit margin of -10,560 SEK. The table gives useful information on the economic advantages of adopting machine learning algorithms for business applications. What is really interesting in the table is that all models made the biggest profit when a bet was made on the away team. This can be because the odds tend to be higher on away teams thus grating greater profit. This is also true for predicting a draw result but draws tend to be much harder to predict than a winning team thus showing a negative profit on most of the models. By providing insights into the profitability of different algorithms, this analysis can help businesses and individuals make more informed decisions when it comes to betting on soccer matches on Allsvenskan.

4.2.3 Expected Value (EV)

	Random Forest	***Logistic Regression***	***SVM***	***Naive Bayes***	***Gradient Boosting***	***XG Boost***
Profit (SEK)	+10,540	+11,040	-30,290	+5,710	+10,060	-33,690
Matches (Value in bet)	240	240	240	240	240	240
Total Home pred (Accuracy)	87 (0.42)	104 (0.36)	104 (0.30)	92 (0.47)	97 (0.39)	108 (0.43)
Total Draw pred (Accuracy)	72 (0.33)	39 (0.44)	45 (0.24)	77 (0.31)	72 (0.34)	62 (0.19)
Total Away pred (Accuracy)	81 (0.22)	97 (0.23)	91 (0.19)	71 (0.31)	71 (0.27)	70 (0.27)

Table 10. Profit from the models with the use of EV.

For the EV method table 10 shows that Random Forest, Naive Bayes, Gradient Boosting, and Logistic Regression were profitable, while XG Boost and SVM resulted in losses. The most profitable model was Logistic Regression with a profit of +11,040 SEK, followed by Random Forest and Gradient Boosting with a profit of +10,540 SEK and +10,060 SEK respectively. In terms of prediction accuracy, Naive Bayes had the most accuracy in forecasting home outcomes, with 47 percent right predictions out of 92 matches. Logistic Regression had the highest accuracy for draws, with 44 percent accurate predictions out of 39 matches. Finally, with 31 percent right predictions out of 71 matches, Naive Bayes had the greatest accuracy in forecasting away outcomes. What is really interesting in this part of the result is that the amount of draws is so high with the Naive Bayes model predicting the highest amount of matches (77) compared to table 8 where bets on draw games rarely happen. One possible explanation for the large proportion of draws projected by the models is that the odds for a draw are often higher than the odds for a home or away win. This makes betting on draws a more appealing choice in terms of possible payoff. As a result, it's plausible that the models are picking up on this tendency and forecasting more draws as a result. However, it is crucial to highlight that the models' profitability might be dependent on their ability to anticipate draws. Looking at the models that did not perform well there seems to be a correlation where they did not predict the draw games at the level the other models that were profitable did. This is interesting because it suggests that predicting draw games accurately is a factor in determining the overall profitability of a model. It can also be because the odds of a draw generally are higher than home and away wins, indicating that predicting draw outcomes somewhat accurately may impact profitability. Another thing that is interesting in this table is that all matches have a positive expected value for every model, meaning a bet will be made for every match which is a promising finding for those looking to use machine learning models to inform their betting strategies. Overall, the findings of this table indicate that machine learning models may be effective at forecasting the outcomes of betting events and that the model selected can considerably impact the profitability of the betting strategy.

4.3 Feature selection

The study's goal was to create a model that properly predicts football match outcomes and which features mattered the most in that setting. As explained in section 2.4 the filter methods were utilized to find the features that were most connected with the target variable, whereas Lasso, Ridge, and PCA employed regularization to find the most relevant features.

Feature Name	***Mean Score***
Average Odds Away	0.725
Average Odds Home	0.696
Diff In FIFA Overall	0.522
Difference In Goals By Average	0.424
Difference In Points By Average	0.408

Table 11. Mean importance score of filter feature selection.

Feature Name	***Mean Score***
Diff In FIFA Overall	0.691
Difference In Points By Average	0.635
Difference Shots	0.548
Difference In Standing Previous Season	0.528
Difference In Goals By Average	0.502

Table 12. Mean importance score of Lasso, Ridge, PCA feature selection for multi-class classification.

From the filter methods, the top features and the mean score of these features can be seen in table 11 where its seen that the average home and away odds by bookies have the highest importance mean score alongside the difference in the FIFA overall, goals by average and points by average.

Looking at table 12 there is a shift in which features matter the most. This implies that various feature selection approaches might provide different findings and that it is necessary to combine numerous strategies to gain full knowledge of the essential features. The inclusion of FIFA data in the research generated surprising results, demonstrating that overall team stats had the predictive potential for match outcomes. This shows that even simple indicators of team quality, such as total team ratings, may be valuable predictors of match results. This conclusion might be explained by the fact that overall team ratings represent a wide variety of player traits that can influence team performance. A team with better overall ratings, for example, may have more competent and experienced players, giving them an edge in a contest. While the study did not look into the individual player characteristics that influence total team evaluations, this might be an intriguing area for future research. A more detailed knowledge of the factors that lead to team success in football might be achieved by evaluating which individual player attributes are most closely connected with match outcomes.

Feature Name	***Mean Score***
Average Odds Away	0.731
Average Odds Home	0.683
Diff In FIFA Overall	0.530
Difference Shots	0.517
Difference In Goals By Average	0.492

Table 13. Mean importance score of Lasso, Ridge, PCA feature selection for binary classification.

The mean significance score of Lasso, Ridge, and PCA feature selection for binary classification in football matches is presented in table 13. The top five attributes are shown, along with the mean ratings for each. In the binary classification scenario, the findings demonstrate that the average odds away and average odds home are the two most relevant factors for this scenario. This is interesting because in 12 these features do not even make the top 5. This shows that bookmaker odds may be more useful predictors of match outcomes in a binary classification context rather than in multi-class classification.

5 Conclusions

In this thesis, two main issues were examined. Firstly the study aimed to evaluate the accuracy of machine learning models in predicting football match outcomes and secondly to explore the impact of different feature combinations that affect match outcomes. An analysis of 960 matches in Allsvenskan spread across four seasons was conducted to achieve this. The results of predicting home, away and draw showed that the models were able to predict the match outcomes at a higher accuracy rate than the sport betting companies with Support Vector Machines(SVM) having the highest accuracy of 52.4 percent compared to the betting companies at 40.4 percent. The most important features affecting match outcomes were FIFA ratings, the difference in points by average, and the difference in shots by average. To further evaluate this, binary classification was also used to predict when the betting companies were correct or not. The results of this analysis showed that SVM performed at the highest level with an accuracy of 64.1 percent and the most important features being the average odds for away and home teams as well as FIFA ratings. Regarding the feature combinations, the results showed that the importance varied depending on the selection method used but simple indicators of team quality can play a significant role in predicting football match results.

These findings may have important practical implications for the sports industry, particularly for football coaches and analysts. The study revealed the important elements impacting match outcomes, which may be utilized to improve team selection and tactical plans. Coaches and analysts may use this data to obtain a better knowledge of the game and make better judgments. Profit estimates were included in the study to give insight into the practical uses of predictive algorithms for sports betting. However, the potential ethical implications of using these models should be carefully considered.

Overall, this work illustrates the potential of machine learning models for forecasting football match results and gives useful insights into the elements that impact these outcomes. By doing this our understanding of football and sports in general can be increased by refining and improving these models.

6 Discussion

Throughout the papers provided in 2.6, the type of classifier and the characteristics placed into the classifier are held to a high level. This demonstrates that various classifiers produce the best results in different scenarios. It may also suggest that the characteristics incorporated into the classifiers are more essential than the classifier itself. Looking at the outcomes of these papers, which features that were included in their specific challenge determined which classifier was best suited for that task. It is also important to note that since different leagues play different styles of football other features may be extra vital in those leagues. Another critical observation is that better football leagues for example Premier League or La Liga might probably be tougher to analyze. This is because these leagues generally have more data to go off and generate more profit from sponsorships, which could lead to betting companies having more accurate odds. While the current study offers useful insights into the use of predictive models for predicting football match results, one potential limitation is that it only covers a single season of Allsvenskan. This study's conclusions may not apply to other leagues or seasons within the same league. It will be interesting to see if the study's findings are consistent over seasons or if it is limited to a single year. Comparing the performance of the models across leagues could help in determining whether there are any league-specific factors that affect the accuracy of the models. Furthermore, it is also important that kind of study would require more data and resources. But if it is done correctly such analysis would provide valuable insights into predictive models for football match predictions. Finally, more research is required to fully understand the strengths and weaknesses of predictive models for football match forecasting across various leagues and seasons.

One issue that should be studied more in future studies is the usefulness of the feature selection technique early as opposed to late in the football season. Certain features might be more important earlier in the season compared to later in the season. This uncertainty, however, also applies to the bookies' odds. This might be considered and corrected for future research by integrating more current data as the season continues. Another intriguing concept for future study is to make use of the fact that today we have many prediction models. Combining models has been proven to be effective in a variety of scenarios, and it might be utilized in future studies to reduce biases and estimate variances. The average forecast of all the models, for example, might be used to eliminate bias, whilst the variation among the models may be used to obtain insight into the amount of uncertainty in the predictions. In future investigations, this technique might lead to more robust and accurate forecasts. Another thing is that tuning the hyperparameters of the models employed in this study is one potential future project to enhance model performance. The models were used with their default settings, however, a variety of hyperparameters can be changed to enhance their performance.

A note that is seen in the results is that the average odds for the away team play a bigger role for the models than the average odds for the home team in the binary setting. This observation is an interesting one and can be further explored. It is possible that the models give more weight to the odds of the away team because there may be a higher likelihood of upsets or unexpected wins by the away team since the away odds tends to be higher than the home odds. Another possibility is that the bookmakers may be more accurate in predicting the outcomes of matches where the home team is favored, leading to less profit potential for the models. However, when the away team is favored, the bookmakers' predictions may be less accurate, leading to greater profit potential for the models. This might be true since the models in the binary classification gained the most profit from the away odds.

What is interesting that is touched upon in the results is the FIFA data importance rating in the feature selection. This discovery is important because it shows that using additional sources of information, like FIFA data, can improve the accuracy of predicting sports match outcomes. While usual factors like previous match results and home/away advantage are important, adding more data can help make better predictions. It's important to remember that FIFA is a widely recognized and accepted measure of team strength but they are not without limitations. Nevertheless, this finding suggests that external data sources like FIFA ratings may be an important consideration for researchers moving forward.

Another thing that might be interesting in future work is features that can capture the performance of different positions and interactions between different positions. Since football is a team sport and every team plays differently depending on the players they have this might be an area of research.

Regarding the accuracy of machine learning models for sports event prediction, it is worth noting that such models are primarily used by large corporations and kept confidential. This suggests a lack of publicly available research in this area. In their work (Shi et al., 2013), suggests that the ceiling for forecast accuracy in predicting sports outcomes is around 75 percent. None of the papers included in 2.6 have achieved a substantially higher accuracy rate, making their argument quite compelling. However, due to numerous uncertainties that can occur during a game, such as injuries, weather conditions, last-minute changes to the team, and the quality of the pitch, obtaining even higher accuracy rates would be remarkably challenging.

The inclusion of profit calculation in the study's results offers insight into the potential practical application of predictive models for sports betting. However, the research question did not directly address the profitability of the models. This is because the primary objective was to evaluate the accuracy of the models in predicting match outcomes. The profit calculation was meant to provide the readers with an idea of how the models would perform against bookmakers. While these models can provide a competitive edge over other bettors, some might argue that it may encourage gambling problems or that it is unfair to use algorithms to gain an advantage. It is crucial to weigh the potential benefits of using predictive models for sports betting against these concerns. It's important to remember that the tables showing profits in this study were based on a specific way of betting. While this strategy was successful in this study, it may not work well in all situations. Different betting strategies might work better for different datasets. Before using any specific betting strategy, it's important to consider its advantages and limitations. With that said, the profit from the binary classification stands out compared to the multi-class classification which is intriguing that predicting when the betting companies are wrong can have such an impact. Further work needs to be done since these findings suggest that there is still much to be learned about the relationship between sports betting odds, machine learning models, and football match outcomes. This can also be extended to future work about binary settings in football matches that can be betting on over/under odds for goals, cards, corners, etc.

A limitation that is worth mentioning is that the odds used in this project were the average odds across different sports betting companies. But what is important to mention is that these sport betting companies that are combined to create the average are not specifically mentioned. However, the odds supplied are very similar to the pinnacle odds that also were available making it likely that the average odds used in this project included odds from prominent sports betting companies.

The models established in this work might have real-world effects on sports teams and betting organizations in particular. These models might help teams acquire insight into their opponents' playing styles and make smart judgments. Betting businesses might potentially utilize these algorithms to better accurately modify their odds and boost their long-term profitability. Incorporating these models into current systems may require more resources and expertise. It's important to consider the benefits and drawbacks before using them. In addition to this, these models may need to be updated regularly with new data to ensure accurate predictions.

References

Alcalá-Fdez, J., Garcıa, S., Berlanga, F. J., Fernández, A., Sánchez, L., del Jesus, M. J., & Herrera, F. (2008). Keel: A data mining software tool integrating genetic fuzzy systems. *2008 3rd International Workshop on Genetic and Evolving Systems*, 83–88.

Baboota, R., & Kaur, H. (2019). Predictive analysis and modelling football results using machine learning approach for english premier league. *International Journal of Forecasting*, *35*(2), 741–755.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Bunn, C., Ireland, R., Minton, J., Holman, D., Philpott, M., & Chambers, S. (2019). Shirt sponsorship by gambling companies in the english and scottish premier leagues: Global reach and public health concerns. *Soccer & Society*, *20*(6), 824–835.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, *40*(1), 16–28.

Collins, M., Schapire, R. E., & Singer, Y. (2002). Logistic regression, adaboost and bregman distances. *Machine Learning*, *48*(1-3), 253–285.

Constantinou, N. F. A., & Neil, M. (2012). A bayesian network model for forecasting association football match outcomes. *Knowledge-Based Systems*, *36*, 322.

Cortis, D. (2015). Expected values and variances in bookmaker payouts: A theoretical approach towards setting limits on odds. *The Journal of Prediction Markets*, *9*(1), 1–14.

Dash, M., & Liu, H. (1997). Feature selection for classification. *Intelligent data analysis*, *1*(1-4), 131–156.

Davenport, T. H. (2014). Analytics in sports: The new science of winning. *International Institute for Analytics*, *2*, 1–28.

European Gaming and Betting Association. (2022). Online gambling market report [Retrieved from EGBA website]. https://www.egba.eu/

Forrest, D. (2012). Betting and the integrity of sport. In P. M. Anderson, I. S. Blackshaw, R. C. Siekmann, & J. Soek (Eds.), *Sports betting: Law and policy* (pp. 14–26). T.M.C. Asser Press. https://doi.org/10.1007/978-90-6704-799-9_3

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Grimes, D. A., & Schulz, K. F. (2008). Making sense of odds and odds ratios. *Obstetrics & Gynecology*, *111*(2 Part 1), 423–426.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, *3*(Mar), 1157–1182.

Guyon, I., Gunn, S., Nikravesh, M., & Zadeh, L. A. (2008). *Feature extraction: Foundations and applications* (Vol. 207). Springer.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The weka data mining software: An update. *ACM SIGKDD explorations newsletter*, *11*(1), 10–18.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Vol. 2). Springer.

Haykin, S. (2009). *Neural networks and learning machines, 3/e*. Pearson Education India.

Huang, K.-Y., & Chang, W.-L. (2010).
A neural network method for prediction of 2006 world cup football game.
The 2010 international joint conference on neural networks (IJCNN), 1–8.

Hucaljuk, J., & Rakipović, A. (2011). Predicting football scores using machine learning techniques.
2011 Proceedings of the 34th International Convention MIPRO, 1623–1627.

Jayanth, S. B., Anthony, A., Abhilasha, G., Shaik, N., & Srinivasa, G. (2018).
A team recommendation system and outcome prediction for the game of cricket.
Journal of Sports Analytics, *4*(4), 263–273.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects.
Science, *349*(6245), 255–260.

Kingsford, C., & Salzberg, S. L. (2008). What are decision trees? *Nature biotechnology*, *26*(9), 1011–1013.

Kuhn, M., Johnson, K., et al. (2013). *Applied predictive modeling* (Vol. 26). Springer.

Lewis, M. (2004). *Moneyball: The art of winning an unfair game*. WW Norton & Company.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017).
Feature selection: A data perspective. *ACM computing surveys (CSUR)*, *50*(6), 1–45.

Lopez-Gonzalez, H., & Griffiths, M. D. (2016).
Is european online gambling regulation adequately addressing in-play betting advertising?
Gaming Law Review and Economics, *20*(6), 495–503.

McCabe, A., & Trevathan, J. (2008). Artificial intelligence in sports prediction.
Fifth International Conference on Information Technology: New Generations (itng 2008), 1194–1197.

McCallum, A., Nigam, K., et al. (1998).
A comparison of event models for naive bayes text classification.
AAAI-98 workshop on learning for text categorization, *752*(1), 41–48.

Miljković, D., Gajić, L., Kovačević, A., & Konjović, Z. (2010).
The use of data mining for basketball matches outcomes prediction.
IEEE 8th international symposium on intelligent systems and informatics, 309–312.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis*.
John Wiley & Sons.

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, *7*, 21.

Noble, W. S. (2006). What is a support vector machine? *Nature biotechnology*, *24*(12), 1565–1567.

Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia*, *4*(2), 1883.

Powers, D., & Xie, Y. (2008). *Statistical methods for categorical data analysis*.
Emerald Group Publishing.

Powers, D. M. (2020). Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.

Purucker, M. C. (1996). Neural network quarterbacking. *IEEE Potentials*, *15*(3), 9–15.

Richardson, L. (2007). Beautiful soup documentation. *Dosegljivo: https://www. crummy. com/software/BeautifulSoup/bs4/doc/.[Dostopano: 7. 7. 2018]*.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986).
Learning representations by back-propagating errors. *nature*, *323*(6088), 533–536.

Sarlis, V., & Tjortjis, C. (2020).
Sports analytics—evaluation of basketball players and team performance.
Information Systems, *93*, 101562.

Schumaker, R. P., Solieman, O. K., & Chen, H. (2010). *Sports data mining* (Vol. 26). Springer.

Shi, Z., Moorthy, S., & Zimmermann, A. (2013).
Predicting ncaab match outcomes using ml techniques–some results and lessons learned.
ECML/PKDD 2013 Workshop on Machine Learning and Data Mining for Sports Analytics.

Štrumbelj, E. (2014). On determining probability forecasts from betting odds.
International Journal of Forecasting, *30*(4), 934–943.
https://doi.org/https://doi.org/10.1016/j.ijforecast.2014.02.008

Tichy, W. (2016). Changing the game: Dr. dave schrader on sports analytics. *Ubiquity*, *2016*(May), 1–10.

Trawiński, K. (2010). A fuzzy classification system for prediction of the results of the basketball games.
International conference on fuzzy systems, 1–7.

Witten, I. H., Frank, E., Hall, M. A., Pal, C. J., & DATA, M. (2005).
Practical machine learning tools and techniques. *Data Mining*, *2*(4).

Yu, L., & Liu, H. (2003).
Feature selection for high-dimensional data: A fast correlation-based filter solution.
Proceedings of the 20th international conference on machine learning (ICML-03), 856–863.

Appendices

List of Features

Feature Description
Average Odds Home from different sport betting companies
Average Odds Draw from different sport betting companies
Average Odds Away from different sport betting companies
Home Team Points
Away Team Points
Home Team's previous match result was a Draw
Home Team's previous match result was a Loss
Home Team's previous match result was a Win
Home Team's last 2 match results were Draws
Home Team's last 2 match results were Losses
Home Team's last 2 match results were Wins
Home Team's last 3 match results were Draws
Home Team's last 3 match results were Losses
Home Team's last 3 match results were Wins
Away Team's previous match result was a Draw
Away Team's previous match result was a Loss
Away Team's previous match result was a Win
Away Team's last 2 match results were Draws
Away Team's last 2 match results were Losses
Away Team's last 2 match results were Wins
Away Team's last 3 match results were Draws
Away Team's last 3 match results were Losses
Away Team's last 3 match results were Wins
Home Team Goal Difference
Away Team Goal Difference

Feature Description
The difference in points from the previous 5 games
The difference in League Positions
Total shots by the Home team in the last 5 games
Total shots on target by the Home team in the last 5 games
Total goals scored by the Home team in the last 5 games
Total goals conceded by the Home team in the last 5 games
Total yellow cards received by the Home team in the last 5 games
Total red cards received by the Home team in the last 5 games
Total fouls committed by the Home team in the last 5 games
Number of fouls committed by the Home team in their last 5 matches
Number of fouls drawn by the Home team in their last 5 matches
Number of offside calls against the Home team in their last 5 matches
Number of crosses attempted by the Home team in their last 5 matches
Number of interceptions made by the Home team in their last 5 matches
Number of successful tackles made by the Home team in their last 5
Number of interceptions made by the Away team in their last 5 matches
Number of successful tackles made by the Away team in their last 5
Total shots by the Away team in the last 5 games
Total shots on target by the Away team in the last 5 games
Total goals scored by the Away team in the last 5 games
Total goals conceded by the Away team in the last 5 games
Total yellow cards received by the Away team in the last 5 games
Total red cards received by the Away team in the last 5 games
Total fouls committed by the Away team in the last 5 games
Number of fouls committed by the Away team in their last 5 matches
Number of fouls drawn by the Away team in their last 5 matches
Number of offside calls against the Away team in their last 5 matches
Number of crosses attempted by the Away team in their last 5 matches
Number of interceptions made by the Away team in their last 5 matches
Number of successful tackles made by the Away team in their last 5 matches

Feature Description
The average number of goals scored by the Home team at Home
The average number of goals conceded by the Home team at Home
The average number of goals scored by the away team away from home
The average number of goals conceded by the away team away from home
The average number of shots taken by the home team at home
The average number of shots against by the home team at home
The average number of shots taken by the away team away from home
The average number of shots against by the away team away from home
The average number of shots on target taken by the home team at home
The average number of shots on target against by the home team at home
The average number of shots on target taken by the away team away from home
The average number of shots on target against by the away team away from home
The average number of fouls committed by the home team at home
The average number of fouls against by the home team at home
The average number of fouls committed by the away team away from home
The average number of fouls against by the away team away from home
The average number of fouls drawn made by the home team at home
The average number of fouls drawn against the home team at home
The average number of fouls drawn made by the away team away from home
The average number of fouls drawn against the away team away from home
The average number of crosses attempted by the home team at home
The average number of crosses against by the home team at home
The average number of crosses attempted by the away team away from home
The average number of crosses against by the away team away from home
The average number of interceptions made by the home team at home
The average number of interceptions against the home team at home
The average number of interceptions made by the away team away from home
The average number of interceptions against the away team away from home
The average number of points by the home team at home across the seasons
The average number of points by the away team away from home across the seasons
The average number of points by the home team at home the last 6 games
The average number of points by the away team away from home in the last 6 games
The average number of points by the home team at home the last 3 games
The average number of points by the away team away from home in the last 3 games
Home team's encoder value
Away team's encoder value
Home team's FIFA overall Rating
Home team's FIFA Attack Rating
Home team's FIFA Midfield Rating
Home team's FIFA Defense Rating
Away team's FIFA overall Rating
Away team's FIFA Attack Rating
Away team's FIFA Midfield Rating
Away team's FIFA Defense Rating
The overall difference between the two teams in FIFA ratings
The difference in goal difference between the two teams
The difference in shots taken between the two teams
The difference in shots on target between the two teams
The difference in fouls committed between the two teams
The difference in interceptions between the two teams
The difference in crosses between the two teams
The difference in fouls drawn between the two teams
The difference in points average in last 6 games between the two teams
The difference in points average in last 3 games between the two teams