YILDIZ TECHNICAL UNIVERSITY

FACULTY OF ARTS AND SCIENCES

MACHINE LEARNING AND DATA ANALYTICS IN SPORTS: THE USE OF ARTIFICIAL INTELLIGENCE-BASED METRICS IN PERFORMANCE ANALYSIS AND SCENARIO PREDICTION WITH MACHINE LEARNING

Can İPEK

UNDERGRADUATE THESIS

Department of Statistics

Advisor

Res. Asst. Dr. Coşkun PARİM

May, 2024

In the study titled " MACHINE LEARNING AND DATA ANALYTICS IN SPORTS: THE USE OF ARTIFICIAL INTELLIGENCE-BASED METRICS IN PERFORMANCE ANALYSIS AND SCENARIO PREDICTION WITH MACHINE LEARNING" prepared by me under the responsibility of my advisor Res. Asst. Dr. Coşkun PARİM, I obtained the necessary legal permissions for data collection and data use, I showed the information I received from other sources in the main text and references, and I did not distort and/or falsify the research data and results, I declare that I have acted in accordance with the principles of scientific research and ethics during my study. If my statement proves otherwise, I accept any legal consequences.

Can İPEK

Signature

ACKNOWLEDGEMENTS

I am particularly grateful to my advisor, Asst. Dr. Coşkun PARİM, for his help and invaluable expertise that guided me through the thesis journey. Also, with deepest gratitude, I am thankful for the family whose staunch support was a source of strength; particularly, my father, who instilled in me the true purpose behind all my endeavors. Finally, I would like to thank all my friends within the Department of Statistics at Yıldız Technical University, whose comradeship and encouragement were truly inspiring.

Can İPEK

CONTENT

ABBREVIATION LIST

ANN	Artificial Neural Networks
KKT	Karush–Kuhn–Tucker
KNN	K-Nearest Neighbours
MAE	Mean Absolute Error
MLB	Major League Baseball
MSE	Mean Squared Error
NBA	National Basketball Association
NP	Non-Penalty
OLS	Ordinary Least Squares
PPDA	Passes Allowed Per Defensive Action
RMSE	Root Mean Squared Error
SVM	Suport Vector Regressor
SVR	Support Vector Regressor
Xa	Expected Assists
xG	Expected Goals

ABSTRACT

MACHINE LEARNING & SPORTS ANALYTICS: USAGE OF AI-BASED METRICS WITH MACHINE LEARNING ALGORITHMS IN PERFORMANCE ANALYSIS & SCENARIO PREDICTION

Can İPEK

Department of Statistics

Undergraduate Thesis

Supervisor: Advisor: Res. Asst. Dr. Coşkun PARİM

The first concept that comes to mind in terms of modern analytics methods is machine learning. Today, with the dynamics changing every moment in the game, the sports sector is incomplete without machine learning—the success factor for making the earliest decision through fast predictions. Bluntly testing the predictive power of sports analytics, it has been diagnosed with the high development of varying momentum: from the time it initiated the journey to new techniques emerging with the development of artificial intelligence in recent times. Using latest-generation metrics combined with machine learning for forward-looking analyzes, thus further increases the value of the obtained outputs.

This study includes a literature analysis of some important studies in the field of sports data analytics and major machine learning methods. It also encompasses machine learning applications aimed at performance analysis and match events prediction in football. The study discusses the purposes for which sports data analytics is used in different branches, the types of studies conducted in the past, and the impacts of these studies not only on the sports world but also on other disciplines. In the application part of the study, data obtained from the understat.com site were used. The dataset covers 380 matches played in the English Premier League 2023-2024 season. The dataset maintains data in 5 different main data sets with granularity ranging from overall season player and team statistics to minute-by-minute data. Among the variables it includes are categorical statistics such as shot type and attack form, as well as numerical statistics like the number of goals and passes near the goal. It also contains new generation metrics based on artificial intelligence, such as expected goals (xG) and expected assists (xA), which have recently gained popularity.

Python was used in all processes in this study. Initially designed to obtain descriptive outputs, the raw datasets were used to create new datasets in an appropriate form for predictive analysis with necessary data preprocessing steps. Subsequently, various outputs were obtained regarding the characteristics of some variables in the dataset through visualizations, and comments were made on these outputs. The representation power of the expected goal statistic for actual goal probability was tested, and differences in consistency were discussed on an individual basis among players. Later, machine learning models such as linear regression, random forest regression, support vector regression, and artificial neural networks were used to predict some events within the match, and the models were compared in terms of accuracy based on the obtained data. The contributions that all these analyzes can provide in technical decision-making were discussed.

Keywords: Machine Learning, Sports Analytics, Performance Analysis, Event Prediction, Python, Expected Goals, Linear Regression, Regression Tree, Support Vector Regressor, Artificial Neural Networks

1
INTRODUCTION

1.1 Literature Review

Machine learning, as a discipline at the intersection of computer science and statistics, is the process of creating software that learns through experience [31]. Big data analytics is the process of examining, cleaning, transforming, and modeling big data to discover and communicate useful information, suggest conclusions, and support decision-making [32]. These two disciplines offer innovative solutions in sports analytics, especially in the age of big data. Data analytics in sports provides benefits such as improving athlete performance, making administrative decisions, and understanding these decisions [45].

Performance analysis involves evaluating teams and players using a variety of metrics to understand the game, optimize performance indicators, and make better choices for team layout [50]. Using machine learning models such as linear or logistic regression, enables the determination of the impact of factors on performance and the analysis of descriptive statistics [7]. It aims to provide a better understanding of gaming behaviors to improve future outcomes [16].

Injury prevention is vital for athletes to sustain their careers. Data analytics in injury prevention helps identify common injuries and their impact on player and team performance [54]. It is possible to predict future disabilities using historical disability data and biomechanical analyzes.

In team sports, game strategies are one of the important factors affecting match results. Data analytics and machine learning are used to optimize the team's strategies by analysing the game strategies of opponents [35]. These insights allow coaches and analysts to make data-driven decisions.

1.2 Sports Analytics and Developments in Some Different Sports

Data analytics in sport includes field-level analyzes of athletes, coaches and referees, decisions by administrations, and analysis of a number of literatures in the fields of economics and psychology [45].

Sports analytics has become a $780 million industry as of 2019 [51]. According to the data obtained, the use of data analytics in the sports sector is increasing and it is predicted that it will continue to increase [55].

1.2.1 Baseball & Data Analytics

Baseball is one of the first sports that comes to mind when it comes to sports analytics. In the 1960s, the publication of mathematician Earnshaw Cook's "*Percentage Baseball*" was indicative of the early development of data analytics in baseball [2]. The 2003 book "*Moneyball: The Art of Winning an Unfair Game*" by Michael Lewis and its adaptation to the cinema in 2011 made Billy Beane's sabermetric approaches to the Oakland Athletics widely known [47].

Billy Beane, in his club with a low-budget team, preferred players with certain characteristics that he identified as critical and managed to achieve success by signing low-cost starting pitchers [62]. Beane's strategies include teams' decisions such as optimising runs and signing starting pitchers at low cost [10].

The Moneyball approach has shown that analysing baseball data provides significant advantages in player selection and team management, while baseball data analytics provides a competitive advantage based on organizational knowledge and can maintain its strategic advantage even after this information spread throughout the community. [52].

1.2.2 Basketball & Data Analytics

Data analytics has found its place in basketball history since the early periods, but until the 1990s, the data used provided basic information such as box scores, points, rebounds and assists. While these statistics provided only a superficial analysis of the game, they were not sufficient for more in-depth analyzes [50].

Since the late 1990s, with technological developments and innovations in data science, more detailed and advanced analyzes have begun to be made in basketball. Today, thanks to advanced data analytics techniques, teams are able to make more informed and data-driven decisions in many areas, from player selection processes to game strategies [56]. As a result, while data analytics in basketball was mostly limited to descriptive analysis in the past, today it is based on more advanced analysis techniques.

A large part of the NBA audience believes that there have been serious changes in the understanding of the game in the 2010s. Since then, the proportion of three-point shots has skyrocketed, and this change has become central to game plans. Players have started to avoid mid-range shots and tend to be more inclined towards layups and three-point shots.

Figures 1.1 and 1.2 show on-court visualizations of randomly selected samples of 10000 shots before and after 2017 from a dataset containing approximately 5 million shots taken in the NBA between 1997 and 2020. The green dots indicate accurate shots, and the red dots indicate inaccurate shots.

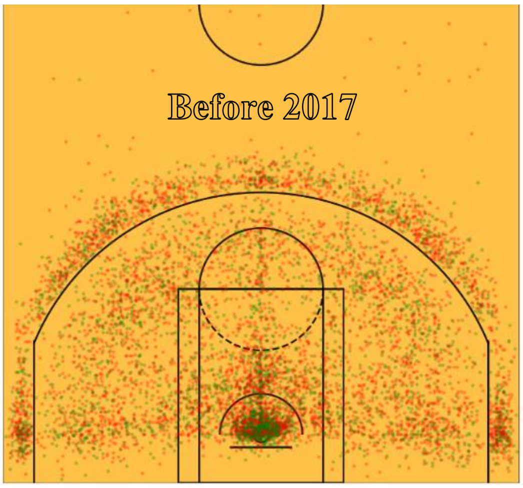


Figure 1.1

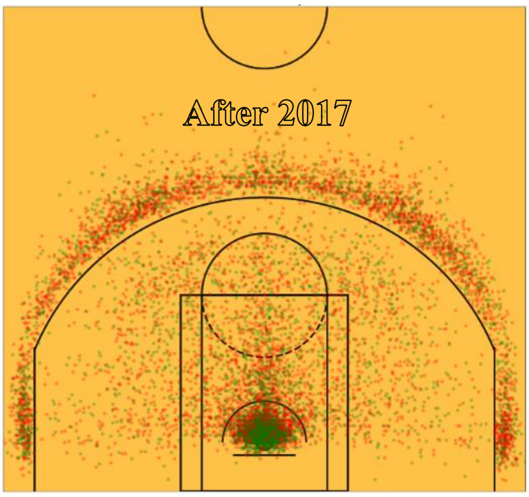


Figure 1.2

When the intense areas on the maps are compared, the trend towards under-the-basket and three-point shooting rather than mid-range shots can be observed in the post-2017 period.

Daryl Morey is a statistician who served as the general manager of the Houston Rockets from 2007 to 2020. Morey's background in statistics and data analytics has allowed him to develop innovative strategies in the NBA. During this period, the Houston Rockets focused prominently on 3-point shooting, basing their style of play on statistical analysis [40].

While Morey's data-driven approach has been criticized by some commentators and fans, the Rockets have had significant success. For example, in the 2017-2018 season, the Rockets advanced to the Western Conference finals despite injuries and finished first in the Western Conference by breaking the all-time record for the number of 3-pointers made in a season in the same season.

Figures 1.3 and 1.4 show visualizations on the basketball court of a randomly selected sample of 8300 shots for the Houston Rockets and other NBA teams during the 2017/2018 season. The green dots indicate accurate shots, and the red dots indicate inaccurate shots.

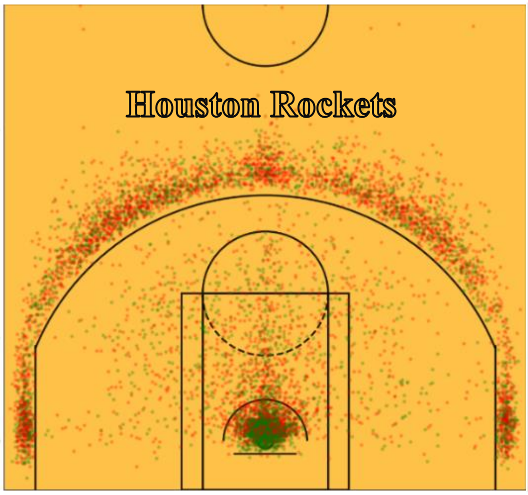


Figure 1.3

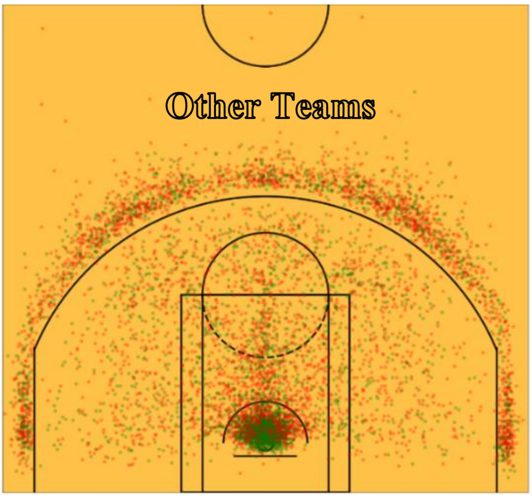


Figure 1.4

When the maps are compared, one can observe the difference between the Rockets' tendency towards three-point and under-the-basket shooting and the tendency of their opponents.

Academic research on Morey's strategic decisions and the Rockets' performance shows that the technical experience and training of general managers have a significant impact on team success [40].

Morey's approach can also be explained by the interpretation of descriptive statistics. Figure 1.5 shows some of the odds for nearly 5 million shots scored in the NBA between 1997 and 2020. The odds shown in blue on the map show the success rate of shots fired from the relevant area, while the number shown in red on the right shows the average number of shots per shot. For example, if one out of every three shots taken from a 3-point shooting zone is successful, the average number per shot is 1.

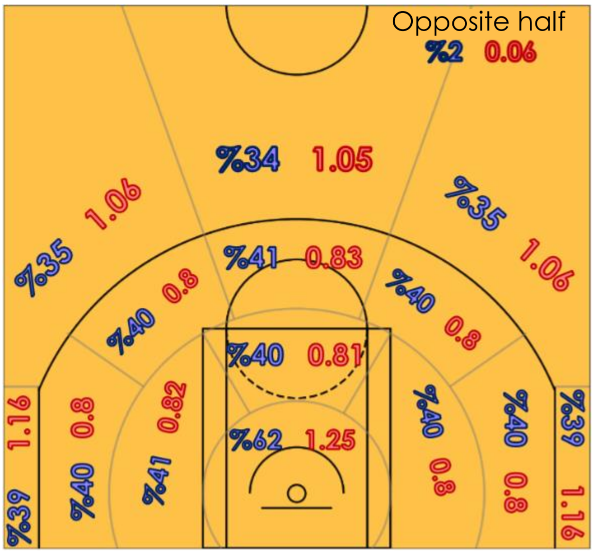


Figure 1.5

Although the map generally shows that the probability of a shot being successful increases as the distance between the basket and the shooting point decreases, when the numbers are involved, the loss in mid-range shots is noticeable.

1.2.3 "Hot Hand" Discussions

A 1985 study by Thomas Gilovich, Robert Vallone, and Amos Tversky suggested that the fact that basketball players make consecutive successful shots does not increase the likelihood of success of subsequent shots, meaning that the concept of "hot hand" is actually a misconception. Controlled experiments conducted on different teams, especially the Philadelphia 76ers, found no such relationship between consecutive shots [3]. This study was also the beginning of a debate between researchers and basketball fans that continues to this day. Those who argue that a hot hand is an illusion have presented evidence about human behaviour as well as statistical evidence. Bruce D. Burns, in his 2004 study, associated the hot hand with the "gambler's fallacy", associating it with people's search for order in random series [11]. If a fair coin toss is repeated five times and all of them are written, it may surprise some people that the sixth throw will also come up, but in fact, it is the result of a 50% probability that the sixth throw will come to the text. According to Burns, in more complex problems, such as the hot hand, people tend to ignore such statistical facts. More recent research has examined the perception created by the hot hand phenomenon and how it affects players' behaviour. For example, some NBA players have been found to exhibit changes in their behaviour depending on the outcome of their previous throwing. In one study, while there was no evidence of a hot hand effect in in-game shots, a weak hot hand effect was observed in free throw attempts [53].

Perceptions of the hot hand and the gambler's fallacy have been the subject of economics research in the past, and in behavioural economics, they have been put forward as evidence that investors are often driven into bad decisions by seeing patterns that are not real [15]. The article "*The Gambler's and Hot-Hand Fallacies: Theory and Applications*" lays out the theory and potential financial applications of the hot-hand fallacy [19].

Despite all these studies, research conducted in the last decade with advanced analysis methods has presented some findings about the existence of a hot hand.

A paper published by Yaari and Eisenmann in October 2011 found that a large dataset of more than 300,000 NBA free throws showed "strong evidence" for the "hot hand" phenomenon at the individual level. The result of analyzing all free throws used during five regular NBA seasons from 2005 to 2010 found that in a two-shot series, players were likely to hit the second shot compared to the first. It has also been found that in a set of two consecutive shots, the probability of hitting the second shot is higher if the previous shot is accurate [20].

In November 2013, researchers at Stanford University used data from MLB and found that there was "strong evidence" that the hot hand in baseball existed in ten different statistical categories [36].

In 2015, a review of the 1985 study by Joshua Miller and Adam Sanjurjo found flaws in the methodology of the 1985 study and showed that hot hands may in fact exist. The researchers said this could be attributed to the improper application of statistical techniques. The authors argued that people were right to believe that hot hands existed in basketball [37].

A 2021 study using data from NBA Three-Point Contests from the period 1986–2020 found "substantial evidence of the existence of the hot hand phenomenon among players" [57].

In 2014, a paper presented by three Harvard alumni at the Sloan Sports Analytics Conference, which for the first time used advanced statistics in basketball games that could control variables such as the player's shot location and the defender's position, showed a "small but significant" hot hand effect [26].

However, no evidence of the presence of hot hands has been observed in other recent studies [53]. Moreover, the evidence obtained in general suggests that only a small subset of players can show a "hot hand", which means that the effect size tends to be small [63].

In conclusion, research on the hot hand phenomenon yields mixed results. Some studies support this phenomenon, while others suggest that the hot hand is actually an illusion. The general judgment in basketball is that the hot hand effect at the level of perceived effect by the audience is an illusion, but there is a small amount of hot hand effect in series shots, especially in fixed conditions such as free throws or three-point contests. Hot hand discussions reveal the complex face of data analytics in sports and its relationship with different disciplines such as psychology.

1.2.4 Football & Data Analytics

The origin of data analysis in football dates back to the 1950s. In 1956, during a match between Swindon Town and Bristol Rovers, Charles Reep began to take notes of a number of numerical data in order to produce effective solutions for Swindon. In this match, he observed that Swindon had a total of 147 attacking attempts and scored only one goal. Reep noticed the negative correlation between the number of passes and the chance to score in the development of offenses and determined that the effect of long passes on the score was more positive. Reep became professional football's first performance analyst, constantly keeping statistical notes on matches. Brentford's manager, Jackie Gibbons, was able to keep the team in League Two by hiring Reep as a consultant. Thanks to Reep's analysis, Brentford have doubled their goal rate per game [9]. Reep's long-ball strategy was adopted by Charles Hughes, who was appointed director of training and coaching for the English Football Association in 1983. *In his book "The Winning Formula*", Hughes argued that 85% of goals come with five passes or less, and that it is necessary to move the ball to what he described as the "highest chance zone" as soon as possible [46].

The methods of Reep and Hughes were also applied by Graham Taylor. Taylor led Watford to second place in Division One and became head coach of the English national team in 1990. However, after England's failure to qualify for the 1994 World Cup, Taylor resigned [9].

Although Reed's understanding began to lose its validity with these events, the founding of Opta in 1996 marked a major step forward in the field of football data analytics. Opta developed a comprehensive system that collects and analyzes match statistics and makes this data available to clubs, media and fans [27].

It is known that technical managers in the UK began to realize the value of data analytics during this period. In 2001, Alex Ferguson abruptly sold defender Jaap Stam to Lazio, making him the first major signing in the press based largely on statistics. In his autobiography, Dutch footballer Dennis Bergkamp recounted an argument between him and French coach Arsene Wenger over Wenger's "obsession" with numbers [25].

Thanks to today's advanced facilities, it has become easier to produce actionable data in sports. Therefore, using the data correctly makes it a critical factor to use the right analysis methods in order to produce useful results in a wide range of analyzes from injury estimation to scouting and performance analysis. In addition, traditional quantitative data from matches offers more useful outputs than before with advanced visualization methods such as heat maps. These methods allow athletes and coaches to better understand and improve their game strategies. In addition to all these developments, the "expected goals" (xG) metric, which has emerged as a fruit of technological developments in recent years, has revolutionized the field of data analytics in football and has become one of the most valuable measures in performance analysis from the first moment it was used. The xG metric will be mentioned in more detail in the next section.

1.2.5 AI-Based Metrics

Artificial intelligence is a branch of computer science that allows machines to work efficiently and analyze complex data [64]. Sub-branches such as machine learning and deep learning are technologies that support the advanced capabilities of Artificial Intelligence.

Artificial intelligence-based metrics in sports have also emerged thanks to these methodologies. Using a large amount of data, the probability of any outcome of the position being predicted can be predicted based on the results of hundreds of thousands of previous position records and the factors that influence these outcomes.

Expected goals (xG) is a metric that gives the probability that a shot with certain characteristics scored in a football match is a goal. The distance at which the shot was made is calculated taking into account the angle, shot speed and other factors. This metric offers a different perspective to objectively assess the performance of teams and players [61].

Different metrics such as expected goals metric, xA (expected assists) and npxG (expected goals without penalties) can also be derived. In the application section, these metrics will be used to make predictions on performance analyzes and match scenarios.

2
MACHINE LEARNING

2.1 What is Machine Learning?

Machine learning is one of the fastest-growing technical study fields today, at the intersection of computer science and statistics, at the heart of artificial intelligence and data science [31]. As a general definition, machine learning is the field of study that enables the computer to gain the ability to perform this operation without being specially coded for it [1]. Although the concept of machine learning emerged for the first time with this definition made in 1959, this concept has gained significant meaning in recent years. This has been supported by the development of new learning algorithms and theories, as well as the rise of online data and low-cost computing power [31]. In particular, advances in deep learning and big data analytics have increased the popularity of machine learning [41].

Machine learning automates the process of building an analytical model and produces the desired outcome based on problem-specific "train" data [59].

Machine learning uses algorithms to learn from machine-readable data. These are examined under 4 main headings: supervised, unsupervised, semi-supervised and reinforcement learning [48].

2.2 Supervised Machine Learning

It has been mentioned in the previous section that machine learning is the field of study that enables the computer to reach the result by extracting meanings from the data in its hands. In supervised learning, the data on which the algorithm is trained also includes the expected outcomes, which are referred to as 'labels' [48]. By using the relationships between these labels and other variables, predictions are made on unknown labels.

There are two typical supervised learning functions, classification and regression. Different algorithms are available for both functions. In some cases, regression algorithms can be used for classification and vice versa [48].

Determining whether the e-mails in your mailbox are spam is an example of a popular classification problem. Your email provider does this thanks to a classification algorithm that learns from a massive dataset of spam and non-spam emails. Here, the data is your mails, and the label is whether they are spam or not.

An algorithm that can predict the market value of a player of interest by training on a data set that includes the market value, age, position, and match statistics of hundreds of athletes is an example of regression algorithms. The tag here is the market value of each player.

Supervised machine learning can be likened to a math equation. There are inputs and a result obtained. Unlike supervised machine learning, this result is not calculated and instead predicted, but instead based on results previously obtained with similar inputs. In classification, there are varieties of results in the training data that are divided into several specific categories. The goal is to predict which of these different categories are the unknown and the outcomes that are being tried to be predicted. In regression, on the other hand, the number of possible numerical results are infinite, and the aim is to predict the results that are tried to be predicted in the closest way to the truth.

In supervised learning, the outcome variable that is tried to be predicted is called the 'dependent variable' and the variables used in prediction are called the 'independent variable'.

In the application part of this study, supervised machine learning algorithms were used. Definitions of some of the wide used supervised machine learning algorithms are given in the subheadings.

2.2.1 Linear Regression

A linear regression model describes the linear relationship between the dependent variable Y and one or more independent variables (X). The model is usually expressed as $Y=\beta0+ \beta1X+ \varepsilon$ [24].

β0: Slope, β1: Intercept, ε: Error

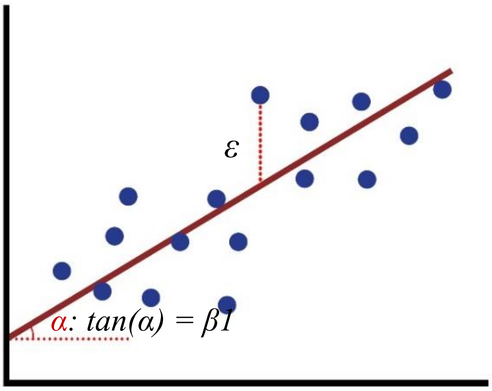


Figure 2.1

Estimation of parameters is usually done by the Ordinary Least Squares (OLS) method. This model determines the linear relationship by minimizing the sum of the squares of the errors.

The terms mentioned in Figure 2.1 are represented on a simple linear regression model. The blue dots represent observations. If red is true, it represents the regression model. This line is the linear model that best describes the relationship between the independent variable and the dependent variable.

Linear regression can outperform more complex models, especially in situations with small data sets or a low signal-to-noise ratio. It can also be extended with data transformations, making it more flexible [17].

2.2.2 Logistic Regression

Logistic regression is an effective way to analyze the impact of a group of independent variables on a binary outcome by measuring the unique contribution of each independent variable [21].

This method is used to obtain the odds ratio, i.e. the ratio of the probability of an event to the probability that it will not occur, and it can determine these ratios even in the presence of more than one explanatory variable [28].

The coefficients are estimated using the maximum likelihood method and the well-being of the model is evaluated with various goodness-fit measures [21].

Figure 2.2 shows the distribution of observations with only two values and the model of logistic regression in the form of a sigmoid function. The value that the sigmoid function takes on the y-axis between 0 and 1 indicates the probability that the value of interest belongs to the supergroup.

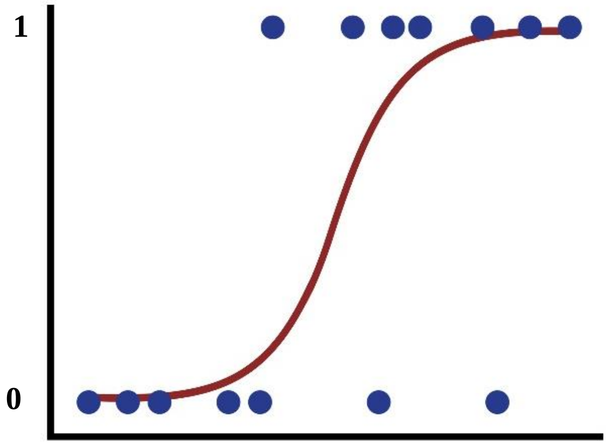


Figure 2.2

Logistic regression is used in sports analytics for problems such as predicting the outcome of the match or obtaining the expected goals (xG) ratio by using the goal/not goal variables as the dependent variable [65].

2.2.3 Support Vector Machine

A Support Vector Machine (SVM) is an algorithm used in classification problems between two groups. SVM maps the input vectors in a nonlinear manner into a very high-dimensional feature space and creates a linear decision surface in that feature space. The special properties of this surface make the algorithm have a high generalization ability [4].

The purpose of SVM is to find the hyperplane that separates the two classes and provides the maximum margin (the largest distance between the two classes). This hyperplane relies on support vectors determined by the training data. Support vectors determine the decision boundary as the closest points [12].

SVM transforms nonlinear data into a linear-separating hyperplane using kernel functions. Kernel functions project data into a higher-dimensional space, making nonlinear boundaries linear. Common kernel functions include polynomial, radial-based function (RBF), and sigmoid [13].

Karush-Kuhn-Tucker (KKT) Conditions: KKT conditions play an important role in the implementation of SVM. These conditions play a critical role in finding the solution in the optimization process and, unlike other pattern recognition algorithms, make SVM algorithms more efficient [38].

The ability of SVMs to create linear decision surfaces in feature space increases their generalization capabilities. This is especially true for data that is not linearly separable [4]. SVM regression can also be used in regression with good generalization performance using analytical parameter selection from training data [14].

2.2.4 Decision Trees and Random Forests

Decision trees are predictive models that divide the space of the covariate into subspaces, using each subspace for a different prediction function, and can be used for both classification and regression tasks. [39]. Classification trees are designed to predict a certain number of unordered values, while regression trees are for continuous or ordinal discrete values; Forecast error is measured by the square difference between the observed and predicted values [22].

Random forests are a combination of tree predictions, where each tree is connected to a random vector that is sampled independently and has the same distribution for all trees in the forest [8]. It is a general-purpose classification and regression method that averages and aggregates estimates from different trees [33].

Figures 2.3 and 2.4 visualize random forest models that predict goals/not goals by classification and expected goals by regression, respectively, using the data of the shots taken by Erling Haland, the top scorer in the English Premier League 2023-2024 season.

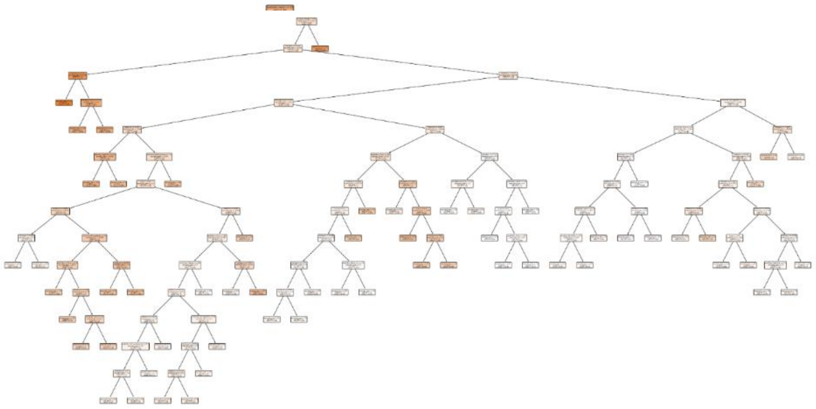


Figure 2.4

2.2.5 K-Nearest Neighbor

The K Nearest Neighbor (kNN) method is a simple application of machine learning that is used for classification, regression, and incomplete data assignment in data mining and machine learning applications [42]. Here, each dimension refers to an argument that is used to predict the dependent variable. The dependent variable 'k' is classified according to the density and proximity of the classes between its number of neighbors. In the graph below, values belonging to two different classes and three unspecified values are visualized on a two-dimensional plane. The classification of unknown values according to different 'K' values is visualized below as representative.

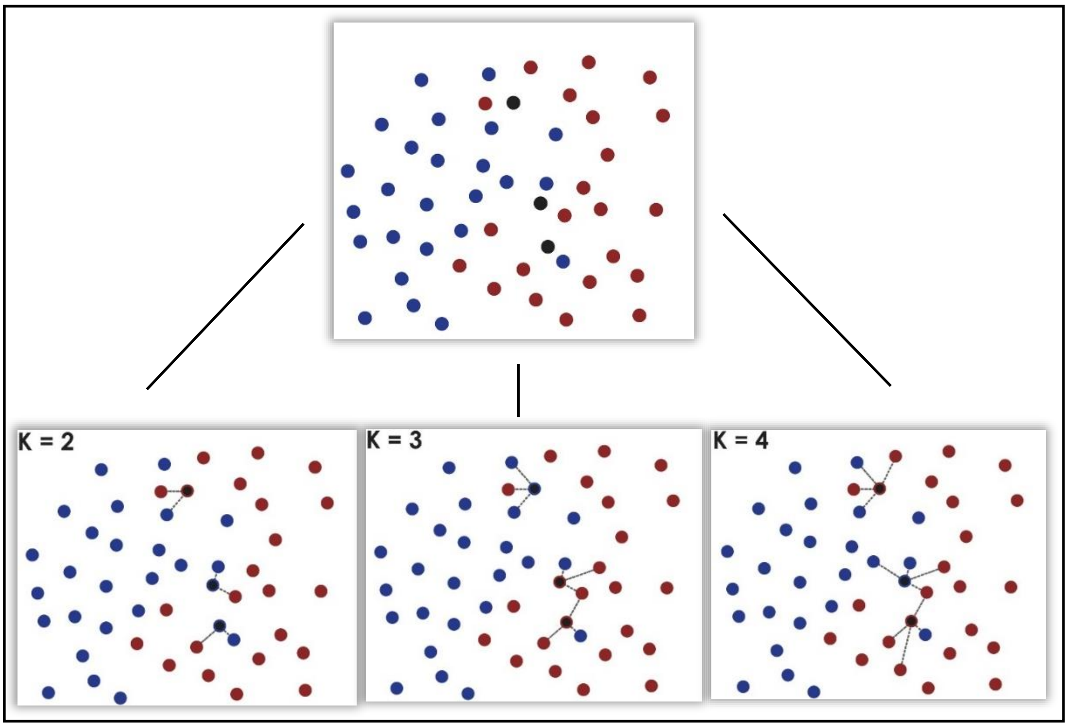


Figure 2.5

2.3 Unsupervised Machine Learning

In unsupervised machine learning, training data is unlabeled. The system tries to learn without a specific teacher [48]. Unsupervised learning has various functions such as clustering and anomaly detection. Unsupervised machine learning involves clustering and association rules that are used to identify hidden patterns and groups in datasets [43]. Although unsupervised learning algorithms were not used in this study, unsupervised machine learning algorithms are used in various subjects [60]. In the next section, the K-Means algorithm will be added as an example of unsupervised machine learning algorithms.

2.3.1 K-Averages

The K-means algorithm is a simple clustering method used to classify similar data in the same cluster [44]. The K-means algorithm is an approach to clustering that dynamically adds one cluster center at a time through a deterministic global search procedure [23]. A center is determined for the specified number of clusters, then each observation is assigned to a cluster according to the center it is closest to. The types of intimacy can vary. After clustering, a new center point is determined for the newly formed clusters. Afterwards, the clustering process is repeated according to the new center points. This process continues repeatedly until the same result is obtained, and new clusters are formed that do not have any information about them in the training set [58].

In Figure 2.6, the clustering of a data group in two different ways with 2 and 3 k values using the Euclidean distance with the K-means algorithm is visualized as a representation.

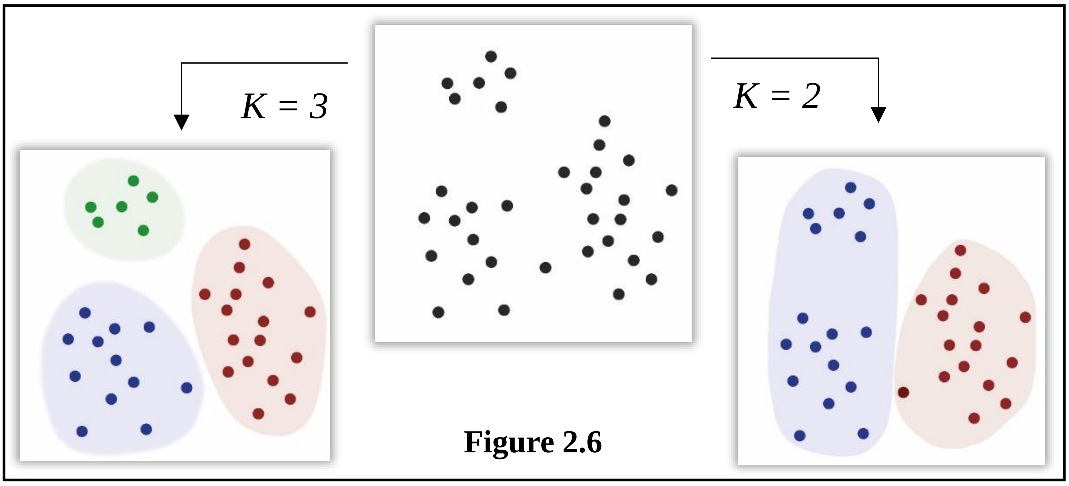


Figure 2.6

2.4 Semi-Supervised Machine Learning

Semi-supervised learning is a type of machine learning in which labeled and unlabeled data are used together to improve supervised learning tasks when labeled data is scarce or expensive [18].

2.5 Reinforcement Learning

Reinforcement learning, one of the most active research areas of artificial intelligence, is a computational approach to learning in which the machine, called an agent, tries to maximize the total amount of reward it receives when interacting with a complex, uncertain environment [6].

In reinforcement learning, a software agent interacts with its environment, takes action, and receives negative or positive results as a result of each action. Positive results are called 'rewards' and negative results are called 'punishments'. The agent records the data of his actions in his experiences and the rewards he obtains as a result, and in a way, he creates his own training data. Then, using this data, it creates an action plan (policy) with algorithms similar to machine learning to optimize the amount of reward. This process is visually represented in figure 2.7.

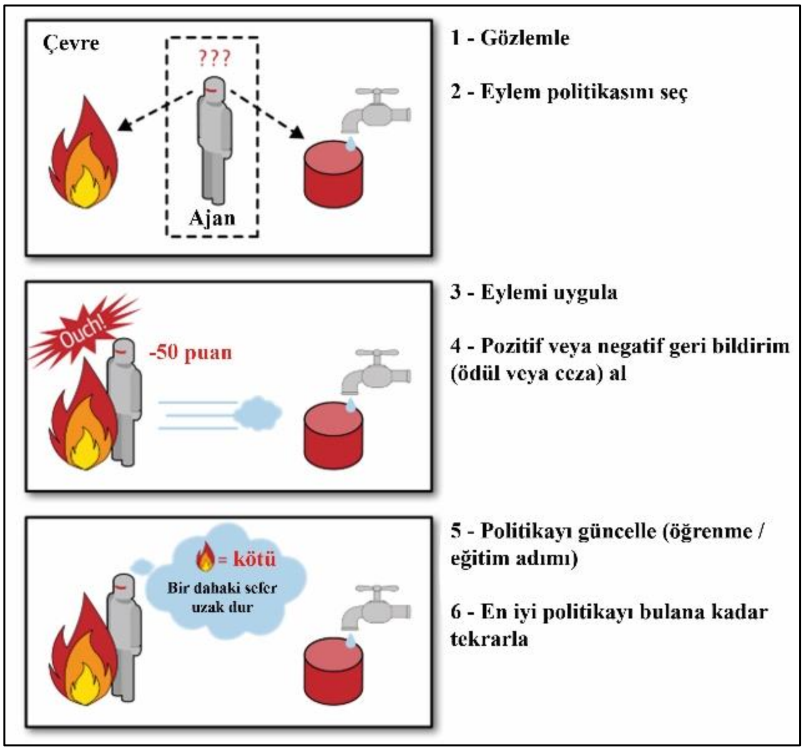

Çevre
???
Ajan
1 - Gözlemle
2 - Eylem politikasını seç
Ouch!
-50 puan
3 - Eylemi uygula
4 - Pozitif veya negatif geri bildirim (ödül veya ceza) al
= kötü
Bir dahaki sefer uzak dur
5 - Politikayı güncelle (öğrenme / eğitim adımı)
6 - En iyi politikayı bulana kadar tekrarla

Figure 2.7

In 2017, Stockfish, the most successful chess AI known up to that time, faced off against AlphaZero, which was released for the first time. While Stockfish was trained with the data of thousands of chess matches played by humans in the past, AlphaZero was a reinforcement learning agent that set itself move policies by creating algorithms that would optimize the rewards and punishments it obtained by playing repeated matches against itself in just a few hours without seeing any chess matches with reinforcement learning. More than 80% of the matches played between the two AIs ended in a draw, while almost all of the rest ended in an AlphaZero victory.

2.6 Deep Learning and Artificial Neural Networks

Deep learning is a class of machine learning techniques that use multiple levels of representation to automatically discover complex functions such as image recognition and speech recognition in high-dimensional data [34].

Artificial neural networks are a distributed network of computing elements that can identify relationships in input data and predict relationships in newly presented datasets, modeled after a biological nervous system [5].

Artificial neural networks typically consist of an input layer, one or more hidden layers, and an output layer. They use backpropagation for training, which involves adjusting weights based on error rates to improve accuracy over time. The input layer is the layer that transmits raw data (e.g., pixel values, text data) to the rest of the network. This layer simply takes the data and transfers it to the next layer. Hidden layers process data to extract higher-level features. These layers use non-linear functions to process input and extract important information. There can be more than one hidden layer in an artificial neural network, and each layer performs a certain level of feature inference [32]. The output layer is the layer that produces the final output of the network. This layer takes the information from the hidden layers and makes a specific decision or classification [48]. The diagram representing the layer structure in artificial neural networks is available in figure 2.8.

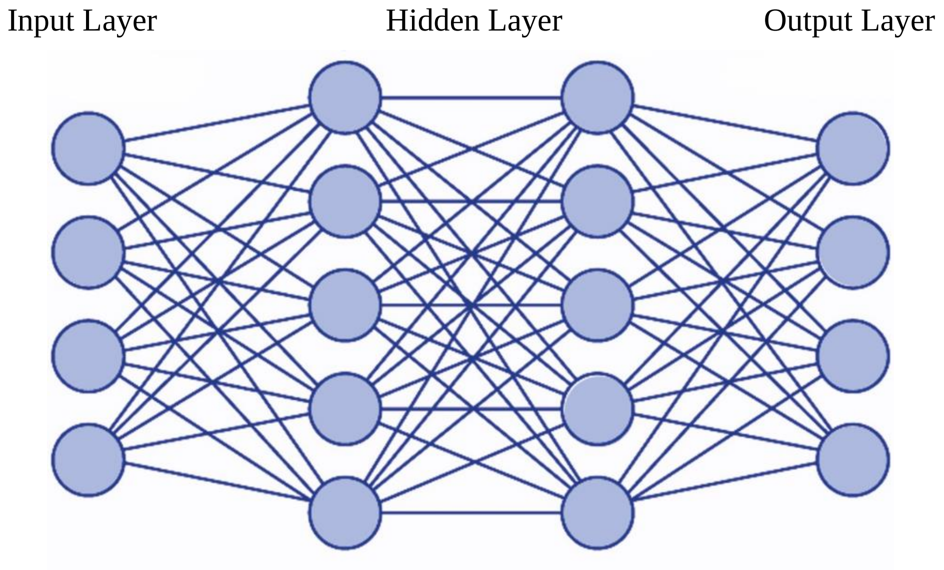


Figure 2.8

Artificial neural networks often need extensive data and computational resources for training. They include complex architectures and training processes, such as deep learning, which can contain millions of parameters. This complexity can lead to challenges such as overfitting, but advances in computing power and techniques such as drop-off have alleviated these problems [29].

2.7 Overfitting Case

Generalization is a basic requirement in machine learning to develop models that can learn from one or several different but related domains. Overfitting is a fundamental problem in supervised machine learning that prevents models from being generalized to fit observed data and unseen data due to noise, limited training set size, and classifier complexity [49].

Figure 2.9 shows representative graphs of three different states for a regression model.

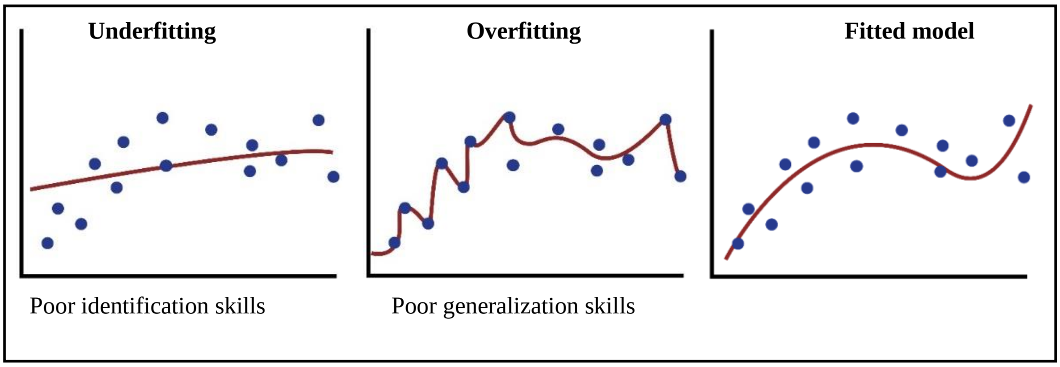


Figure 2.9

Figure 2.10 shows representative graphs for a classification model of these three states.

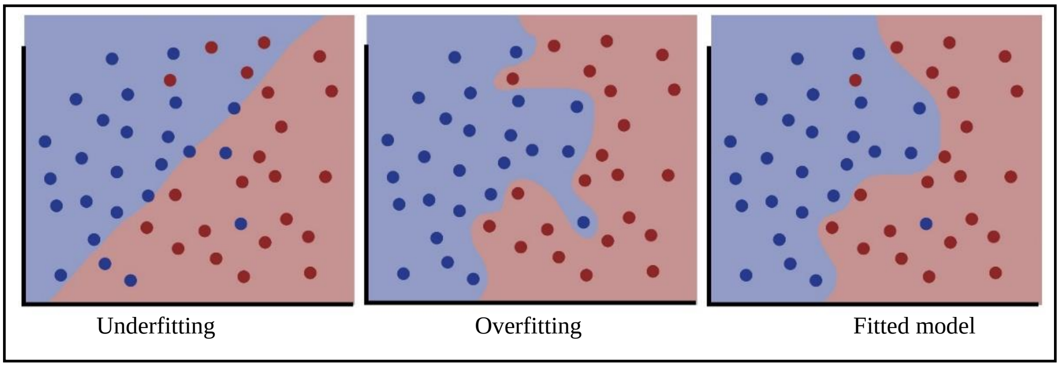


Figure 2.10

The hold-out method is used to detect overfitting and to test the generalization ability of the model. In this method, the data set to be used for the model to be trained is divided into two as training and testing. A large part is reserved for training. The model is trained on the training data, and then the accuracy of its predictions on the test data is measured. If the accuracy obtained by testing is lower than the accuracy obtained by training, then this is a sign of overfitting [30].

2.8 How is Expected Goals (xG) calculated?

After mentioning all these machine learning concepts, we can talk about how to calculate expected goals from artificial intelligence-based metrics. Expected goals predict the probability of a football team or player finding a goal in a given set of positions. This model uses factors such as the distance of the shot, the angle, the position in which the shot was made, etc., to assess the probability that a shot is a goal. The calculation is usually done with logistic regression or other machine learning models [61], [65], [66]. Variables such as the distance of the shot from the goal, the position of the goal, and the placement of the defenders are used to determine the probability of a goal. Machine learning algorithms calculate these probabilities by learning from past match data [61]. xG models can be enhanced with previously untested features such as player and team abilities [66]. In other words, setting a value of 0.25 for expected goals means that a quarter of the shots that the AI detects to have the same characteristics as this shot, which were simply randomly selected to train the AI model that determines the value, resulted in a goal.

3
APPLICATION

3.1 Description

In this section, using the data of the English Premier League 2023-2024 season taken from the understat.com website, what inferences can be made for performance analysis from these data, which are used for descriptive statistics and data visualization on the relevant website, will be examined, xG metric will be discussed, and machine learning models will be used to make predictions as high as possible.

3.1.1 Tools Used

Python language was used in the whole study. The definitions of some of the libraries and modules used in Python are as follows:

1- **Numpy**: Used for linear algebra calculations.
2- **Pandas**: Used for data analysis and manipulation.
3- **Matplotlib**: Used to create graphs and maps.
4- **Seaborn**: Used to create graphics.
5- **Scipy.stats**: Used for statistical analyzes.
6- **Scikit-learn**: Used for machine learning models.
7- **Math**: Used for basic mathematical operations
8- **Json**: Used to process data in Json format.
9- **Requests**: Used to create http requests in the Web Scrapping process.
10- **BeautifulSoup**: Used to parse HTML and XML files in the Web Scrapping process.
11- **MplSoccer:** Used to visualize football data.
12- **GridSearchCV:** Used for hyperparameter optimization of Machine Learning models.

3.1.2 Obtaining the Raw Data

The data was obtained by Web Scrapping method using the BeautifulSoup library. This was done by creating requests to different script indexes on the understat.com site, bringing the data obtained in 'json' format into the appropriate form and converting it into a 'Pandas DataFrame' object.

3.1.3 Tables and Their Contents

As a result of the Web Scrapping process, a data set consisting of 5 tables was obtained from the undertat.com site. The tables are named as team_matches, matches, player_perf, players and shots, respectively. Descriptive statistics of the numerical data in the tables and definitions of all data are indicated in the subheadings for each table.

3.1.3.1 "Players"

- This dataset keeps a record for each player.
- In total, there are 570 records and 18 variables.
- The descriptive statistics of numerical variables are as shown in table 3.1:

	GAMES	TIME	GOALS	XG	ASSISTS	XA	SHOTS	KEY PASSES	YELLOW CARDS	RED CARDS	NPG	NPXG	XG CHAIN	XG BUILDUP
COUNT	570	570	570	570	570	570	570	570	570	570	570	570	570	570
MEAN	19,97	1318,6	2,1	2,29	1,5	1,66	18,38	13,75	2,79	0,1	1,93	2,15	6,64	4,08
STD	11,74	1036,22	3,64	3,72	2,37	2,35	23,08	17,96	2,78	0,32	3,21	3,32	6,83	4,37
MIN	1	1	0	0	0	0	0	0	0	0	0	0	0	0
25%	10	360	0	0,09	0	0,06	2	1	1	0	0	0,09	1,22	0,59
50%	21	1205,5	1	0,88	0	0,79	9	7	2	0	1	0,87	4,65	2,81
75%	31	2154	3	2,68	2	2,16	26	19,75	4	0	2	2,62	10,14	5,92
MAX	38	3420	27	31,65	13	13,34	122	113	13	2	20	25,56	31,88	26,29

Table 3.1

- The variables are as follows:

1. id: The code of the respective player

player_name 2: Name of the respective player

3. games: The total number of matches that the respective player has played

4th time: Total time taken by the respective player

5. goals: The total number of goals scored by the respective player

6. xG: Total expected goals of the respective player

7. Assists: Total number of assists by the respective player

8. xA: Expected total assists of the respective player

9. shots: Total number of shots by the respective player

10. key_passes: The total number of key passes made by the respective player

11th yellow_cards: Total number of yellow cards for the player concerned

12th red_cards: Total number of red cards by the player concerned

13. position: The position of the player concerned

14. team_title: The team of the respective player

15. npg: Total number of goals by the respective player, excluding penalties

16. npxG: Total expected goals of the respective player, excluding penalties

17. xGChain: Total expected goals in all positions in which the respective player is involved

18. xGBuildup: Total expected goals in all positions in which the player is involved, except for the corresponding key passes and shots

3.1.3.2 "player_perf"

- This dataset keeps a record for each match in which each player takes the field.
- In total, there are 11384 records and 21 variables.
- The descriptive statistics of numerical variables are as in 3.2:

	GOALS	OWN	SHOTS	XG	TIME	YELLOW	RED	KEY	ASSISTS	XA	XG	XG
		GOALS				CARD	CARD	PASSES			CHAIN	BUILDUP
COUNT	11384	11384	11384	11384	11384	11384	11384	11384	11384	11384	11384	11384
MEAN	0,11	0	0,92	0,11	66,02	0,14	0,01	0,69	0,08	0,08	0,33	0,2
STD	0,35	0,07	1,31	0,25	33,1	0,35	0,07	1,1	0,29	0,19	0,42	0,31
MIN	0	0	0	0	1	0	0	0	0	0	0	0
25%	0	0	0	0	38	0	0	0	0	0	0,03	0
50%	0	0	0	0	90	0	0	0	0	0	0,16	0,07
75%	0	0	1	0,1	90	0	0	1	0	0,07	0,5	0,29
MAX	4	1	12	2,93	90	1	1	9	4	1,78	3,64	2,89

Table 3.2

- The variables are as follows:

1. id: The corresponding match-specific code of the respective player

2. goals: The number of goals scored by the respective player in the respective match

3. own_goals: Number of own goals scored by the respective player in the respective match

4. shots: The number of shots taken by the respective player in the respective match

5. xG: Expected goals of the respective player in the respective match

6. time: The time (minutes) taken by the relevant player in the relevant match

player_id 7: Code of the respective player

team_id 8: The code of the respective player's team

9. position: The position of the player in question in the relevant match

10. player: The name of the respective player

11. h_a: The side of the player concerned in the respective match (home/away)

12. yellow_card: Number of yellow cards received by the player in question in the relevant match

13. red_card: Number of red cards received by the player in question in the relevant match

14th roster_in: The code of the login record of the player who entered instead of the player who left

15. roster_out: The code of the exit record of the player who left instead of the player who entered

16. key_passes: The number of key passes made by the relevant player in the relevant match

17. assists: The number of assists of the respective player in the respective match

18. xA: Expected assists of the relevant player in the relevant match

19. xGChain: Total expected goals in the respective match in the positions in which the respective player is involved

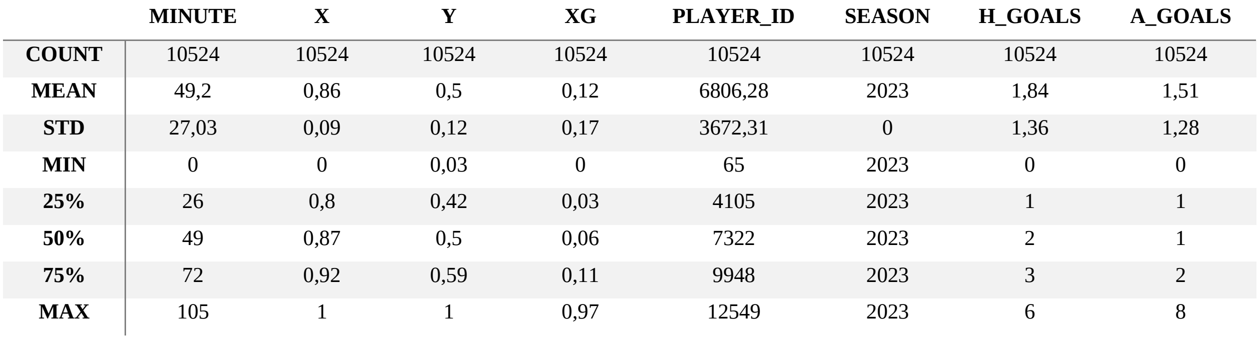
20. xGBuildup: Total expected goals in all positions in which the player is involved in the respective match, excluding the relevant key passes and shots

21. positionOrder: The rank of the position played by the respective player in the respective match

3.1.3.3 "shots"

- This data keeps a record for every single shot taken during the English Premier League 2023-2024 season.
- In total, there are 10524 records and 17 variables.

- The descriptive statistics of numerical variables are as shown in table 3.3:



	MINUTE	X	Y	XG	PLAYER_ID	SEASON	H_GOALS	A_GOALS
COUNT	10524	10524	10524	10524	10524	10524	10524	10524
MEAN	49,2	0,86	0,5	0,12	6806,28	2023	1,84	1,51
STD	27,03	0,09	0,12	0,17	3672,31	0	1,36	1,28
MIN	0	0	0,03	0	65	2023	0	0
25%	26	0,8	0,42	0,03	4105	2023	1	1
50%	49	0,87	0,5	0,06	7322	2023	2	1
75%	72	0,92	0,59	0,11	9948	2023	3	2
MAX	105	1	1	0,97	12549	2023	6	8

Table 3.3

- The variables are as follows:

1. id: The code of the shot

2nd minute: The minute the shot is taken

3. result: The result of the shot (save, escape, goal, etc...)

4. xG: Expected goal of shot

5. player: The name of the player who took the shot

h_a 6: Shooting side (home/away)

player_id 7: The code of the player who took the shot

8. situation: The situation in which the shot was taken (open play, free kick, etc.)

9. shotType: Type of shot (head, right foot...)

10. match_id: The code of the match from which the shot was taken

11th h_team: Home team

12th a_team: Away team

13. date: Date

14th player_assisted: Shooting passer

15. lastAction: Last activity (head pass, tackle...)

16. X: The X coordinate of the area where the shot was taken on the field

17. Y: The Y coordinate of the area where the shot was taken on the field

3.1.3.4 "matches"

- This dataset maintains a separate record for each match played.
- In total, there are 380 records and 12 variables.

- The descriptive statistics of numerical variables are as follows:

	GOALS.H	GOALS.A	XG.H	XG.A
COUNT	380	380	380	380
MEAN	1,8	1,48	1,92	1,46
STD	1,37	1,28	1,03	0,9
MIN	0	0	0,09	0,03
25%	1	1	1,11	0,8
50%	2	1	1,76	1,29
75%	3	2	2,63	1,93
MAX	6	8	6,67	5,11

Table 3.4

- The variables are as follows:

1. id: The code of the match

2. datetime: Date

h.id 3: Home team code

4. h.title: Home team name

5. h.short_title: Home team abbreviation

a.id 6: Away team code

7. a.title: Away team name

8. a.short_title: Away team abbreviation

9. goals.h: Home team goals

10. goals.a: Away team goals

11. xG.h: Home expected goals

12. xG.a: Away expected goals

3.1.3.5 "team_matches "

- This dataset maintains a separate record for the 38 matches played by each team.
- In total, there are 760 records and 21 variables.

- The descriptive statistics of numerical variables are as follows:

	XG	XGA	NP	NP	DEEP	DEEP	SCORED	MISSED	XPTS	WINS	DRAWS	LOSES	MON	NP	OPPDA	PPDA
			XG	XGA		ALLOWED								XGD		
COUNT	760	760	760	760	760	760	760	760	760	760	760	760	760	760	760	760
MEAN	1,69	1,69	1,58	1,58	8,3	8,3	1,64	1,64	1,4	0,39	0,22	0,39	1,39	0	13,64	13,64
STD	1	1	0,91	0,91	5,52	5,52	1,33	1,33	0,88	0,49	0,41	0,49	1,35	1,47	11,07	11,07
MIN	0,03	0,03	0,03	0,03	0	0	0	0	0	0	0	0	0	-4,93	2,57	2,57
25%	0,93	0,93	0,9	0,9	4	4	1	1	0,65	0	0	0	0	-0,94	8,19	8,19
50%	1,51	1,51	1,42	1,42	7	7	1	1	1,35	0	0	0	1	0	11,57	11,57
75%	2,28	2,28	2,16	2,16	11	11	2	2	2,12	1	0	1	3	0,94	15,62	15,62
MAX	6,67	6,67	5,66	5,66	37	37	8	8	3	1	1	1	3	4,93	193	193

Table 3.5

- The variables are as follows:

1st h_a: The side of the respective team in the relevant match (home/away)

2. xG: Total expected goals of the respective team in the respective match

3. xGA: Expected goals of the opposing team in the respective match

4. npxG: Expected goals of the respective team in the relevant match, excluding penalties

5. npxGA: Expected goals of the opposing team in the respective match, excluding penalties

6. deep: The number of passes made by the relevant team less than 20 meters from the goal in the relevant match

7. deep_allowed: Number of passes made by the opponent less than 20 meters from the goal in the relevant match

8. scored: A goal scored by the relevant team in the relevant match

9. missed: A goal missed by the relevant team in the relevant match

10. xpts: Expected points of the respective team in the respective match

11. result: The result for the relevant team in the relevant match (win / draw / loss)

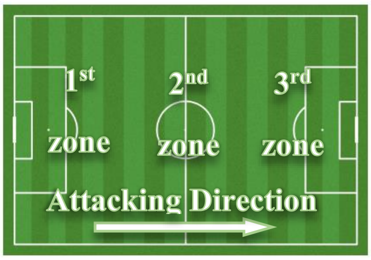


Figure 3.1

12th date: The date on which the match was played

13. wins: Related team win (0/1)

14. draws: Draw for the respective team (0/1)

15. Loses: Loss to the team involved (0/1)

16. pts: Points earned by the respective team in the respective match

17. npxGD: Expected goal difference without penalty in the respective match

18: Code of the respective team

19. team_title: Name of the relevant team

20. ppda: The ratio of the number of passes made by the opposing team in the first and second zones to the defensive actions taken in the same zone. This metric is used to measure the pressing power of the teams at the front, so the smaller the number resulting from this specified ratio, the better the pressing power of the team at the front

21. oppda: The ratio of the number of passes made by the relevant team in the first and second zones to the defensive actions taken by the opponent in the same zone. (Competitor's ppda value)

3.1.4 Data Processing Derived Data

A variable 'isGoal' has been created so that the variable 'result' in the 'shots' table takes the value of 'Goal' in the rows where it takes the value 'Goal' and the value 0 in the other rows.

The indoor/away variables in all tables containing side information have been changed to take the value of 1 in the case of home and 0 in the case of away.

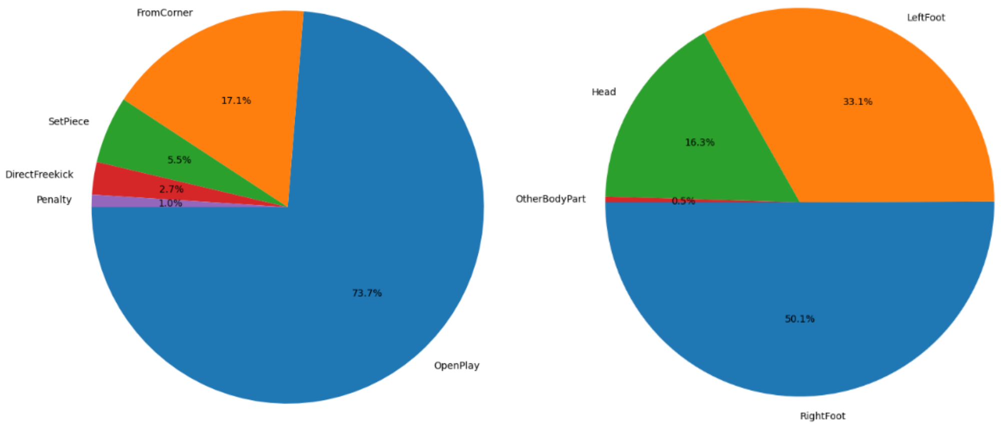
An 'entered' variable has been created that shows the minute the players start the game. The 90-X values in the 'player_perf' table, where the variable 'roster_out' is 0 in the rows where it is filled and the 'time' value is X in the rows where it is empty, are overwritten the newly created variable with the lambda function.

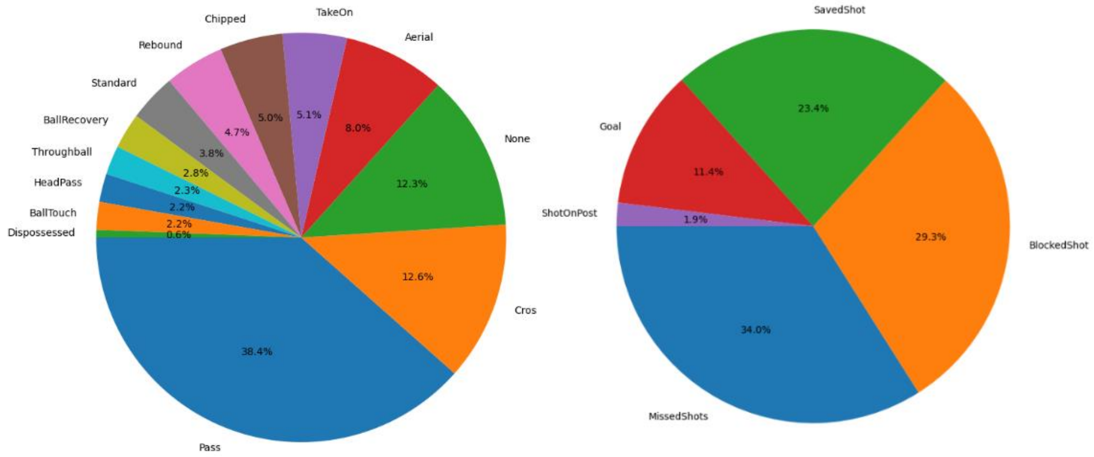
The variables in the team_matches table were grouped according to the teams and the numerical variables were averaged. In this way, a 20-row table was created for the numerical statistics on a match basis, containing the seasonal average of each team on a single match scale, i.e. a record for each team. This table was then recombined with the one-to-many relationship team_matches table. In this way, a new table called 'games_reg_match' has been created, which includes the numerical statistics obtained by the teams for that match, as well as the seasonal averages of both sides, in the new table, where each row contains a match. This process allowed the use of seasonal averages to predict machine learning match-specific statistics.

In order to strengthen the performance of the models created with the 'games_reg_match' table if the players who will play in the team are known, a new table called df_xg has been created with numerical statistics specific to the squad. For this, seasonal average variables obtained from the player_perf table according to the top 11 players in the rows where the roster_out variable is blank were created. With this process, it is aimed to create better performing models by using the season-wide averages of the squad that the team fielded for that match, instead of the season-wide averages of the team in question.

3.2 Data Visualization and Performance Analysis

In order to examine the proportional distributions of four different categorical variables expressing the characteristics of all shots taken in the league in the 'Shots' table, a frequency table was first created using the 'value_counts()' function to create a pie chart. Afterwards, the graphs obtained with the matplotlib library look as follows.





According to the expected goal statistic, the frequency distribution of shots is visualized in the histogram graph in figure 3.6.

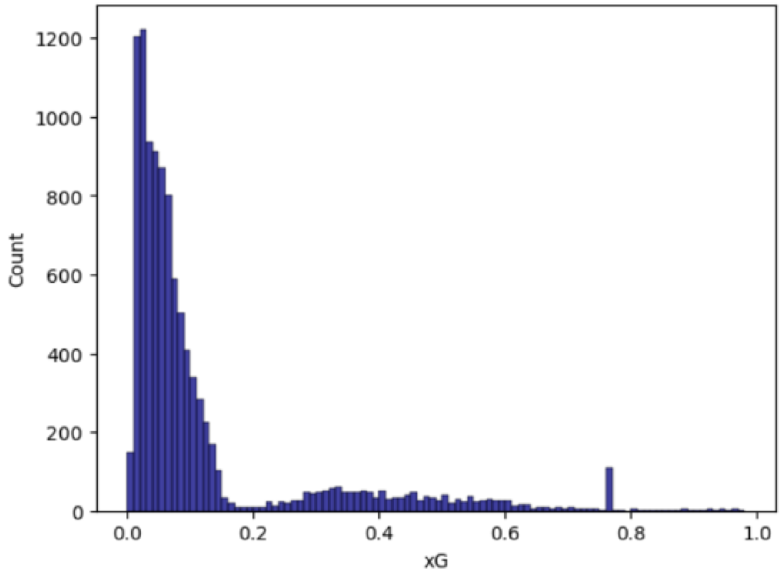


Figure 3.6

In the graph made with the histplot function in Seaborn, the parameters are set in such a way that each column expresses an expected goal of 0.01. While it is noteworthy that shots with an expected goal value of less than 0.2 are concentrated, the increase in the column at this point is striking due to the fact that the expected goals in penalty shootouts is always 0.76.

A logistic regression model was created with the Scikit-learn library to measure the ability of the expected goals in the dataset to represent the distribution of real goals. Then, using the matplotlib library, the distribution of the data was visualized together with the sigmoid function of the model (Figure 3.8). The confusion matrix, which is estimated on the 'X' axis and shows the frequency of the actual values on the 'Y' axis, was visualized with the Seaborn library (Figure 3.7).

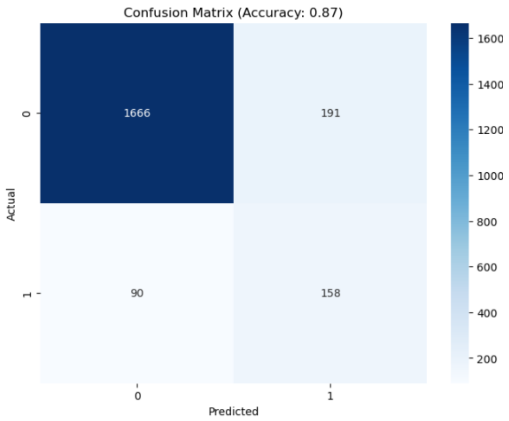


Figure 3.7

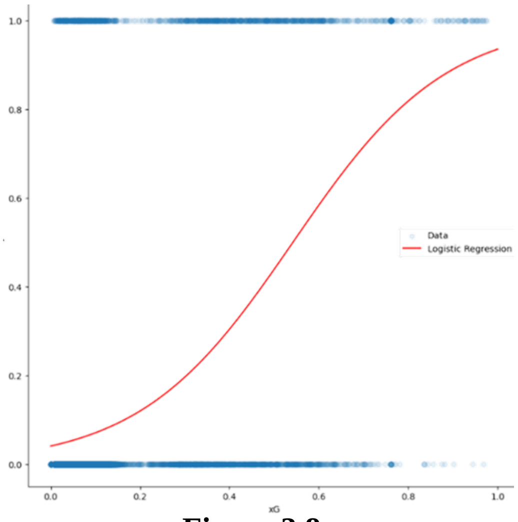


Figure 3.8

Since the distribution of shots with and without goals in the dataset was uneven (shots without goals were about 9 times more), the SMOTE technique was used to multiply the samples in the minority 'goal' class when running the logistic regression model.

With the xG obtained, the goal prediction model works with 87% accuracy, while the majority of the correct predictions belong to 'non-goal' shots, which make up the majority. While the accuracy of non-goal shots is 90%, the accuracy of shots that result in a goal remains at 63%. If we look at the distribution and sigmoid function graph in Figure 3.8, shots that are not on goal tend to be concentrated at low expected goal levels, while shots that end with goals show a more homogeneous distribution.

Figure 3.9 visualizes the total number of goals per minute according to the time the players stay in the game and the total number of goals by match minutes. For this process, the total number of goals was taken separately for each point value of the 'entered' and 'time' variables. On the X axis, there are the values of these two variables, and on the Y axis, there are the total number of goals. The dashed line shows the number of goals according to the time the players stay in the game, and the straight line shows the number of goals according to the minutes of the match.

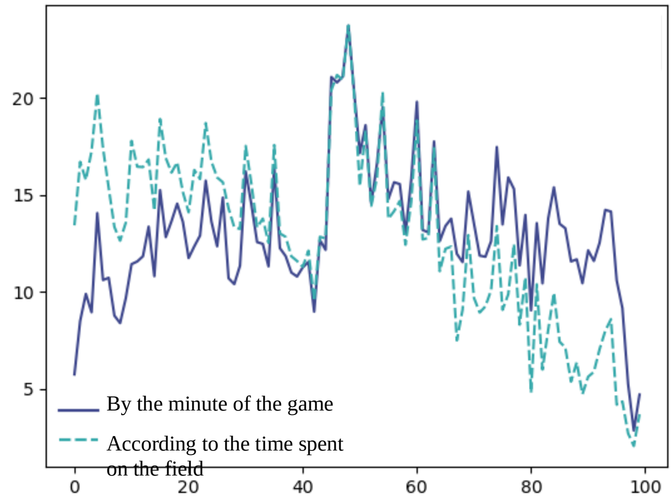
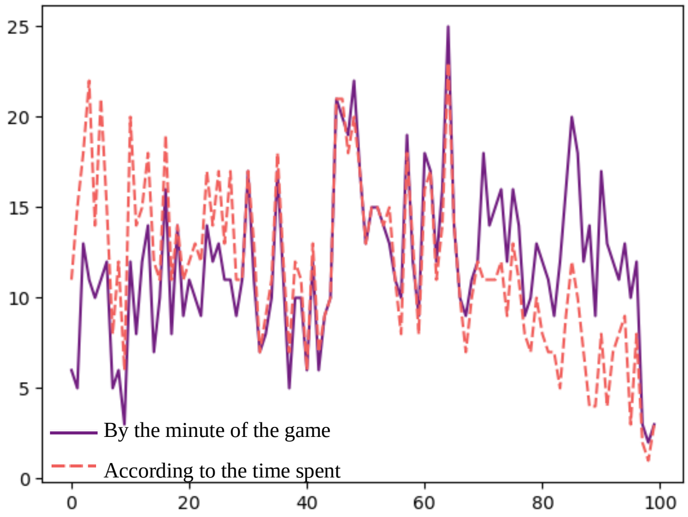


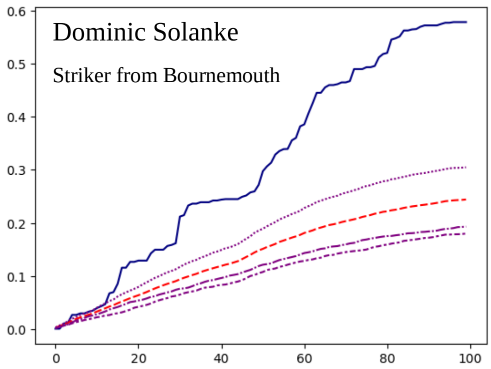
Figure 3.9

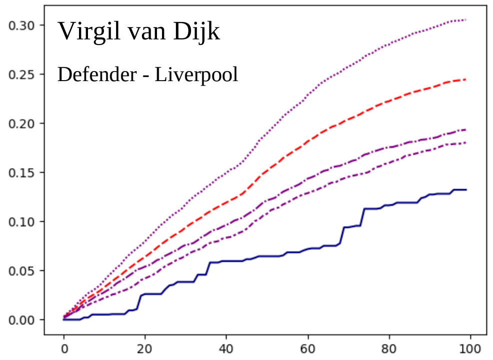
It is observed that the number of goals decreases as the players get time and the number of goals falls below the match minutes.

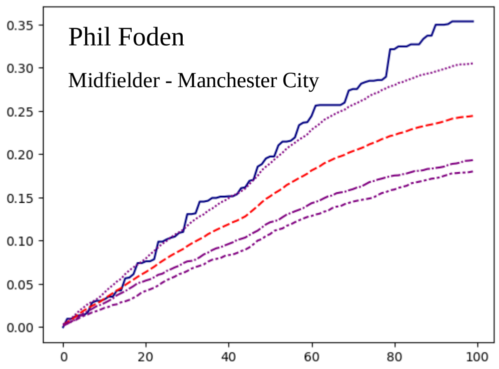
In Figure 3.10, the same parameters are compared, this time in terms of expected goals instead of number of goals. A similar relationship applies when the expected goal variable is used instead of the number of goals.

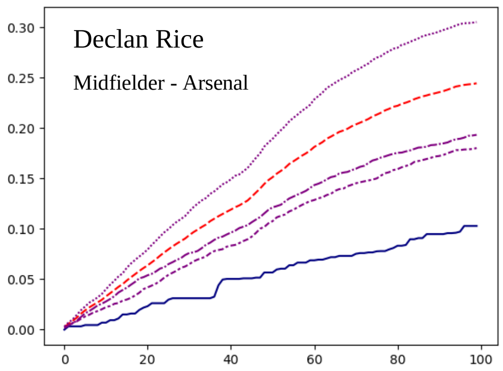


The graphs below visualize the cumulative increase in expected goals of some players over the course of an average match, comparing them to the league's cumulative expected goals increases on an overall and positional basis. For this process, the method used for the expected goals graph on a minute basis was used in the same way after filtering by position and player. A cumulative distribution is obtained by adding the expected goals in each minute with the previous minutes. Each group was then scaled according to its own total minutes to obtain match-by-match averages. The dark blue colored lines show the cumulative expected goal increase of the respective player, the purple colored lines show the cumulative increase in the expected goals of the players in three different positions (forward, midfielder, defender), and the red line shows the overall cumulative expected goal increase.









The 5 players who completed the 2023-2024 season with the most goals in the English Premier League and the differences between their goals and expected goals are listed as they appear in table 3.6 by applying appropriate filters to the players table

PLAYER NAME	GOAL	XG	TEAM	GOAL XG DIFFERENCE
ERLING HAALAND	27	31.653997	Manchester City	-4.653997
COLE PALMER	22	17.832245	Chelsea, Manchester City	4.167755
ALEXANDER ISAK	21	22.074266	Newcastle United	-1.074266
DOMINIC SOLANKE	19	21.406831	Bournemouth	-2.406831
PHIL FODEN	19	11.307983	Manchester City	7.692017

Table 3.6

According to the table, Erling Haaland, who finishes at the top of the top scorer race with 27 goals, has 4.65 less goals in the shots he has scored throughout the league. This indicates that the player has made the most of his shooting opportunities with subpar success. This is a striking fact for a player who has finished at the top of the league.

Let's go back to the random forest examples in the Machine Learning section, where we symbolically created an algorithm that predicts Erling Haaland's goal and expected goals values. To compare the independent variables used in these algorithms in order of importance, figure 3.15 shows a histogram chart showing the importance of the independent variables in the random forest regression model used for goal expectation estimation.

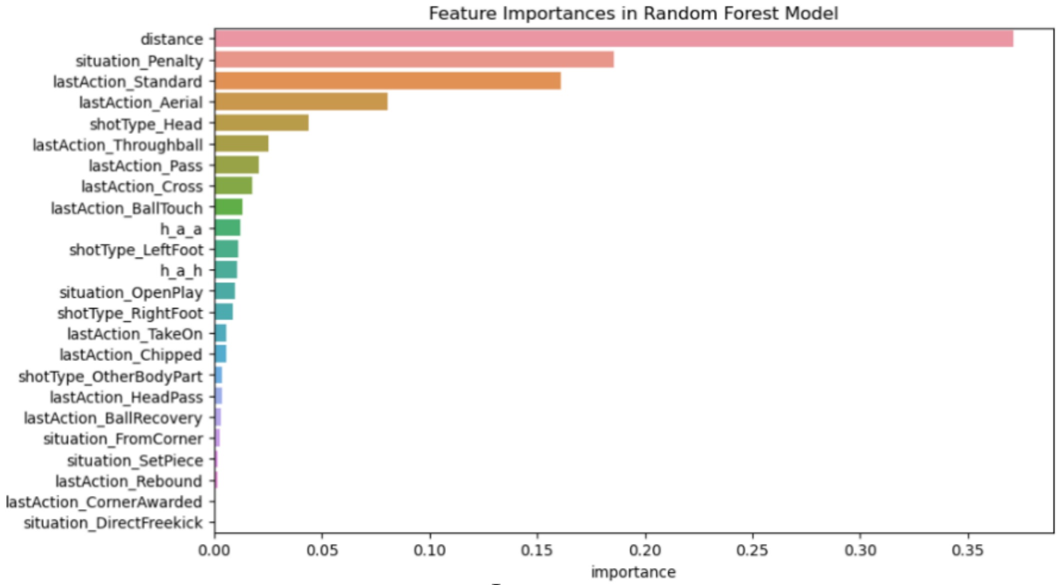


Figure 3.16 shows a histogram chart showing the importance of the independent variables used for the random forest classification model used for lake estimation.

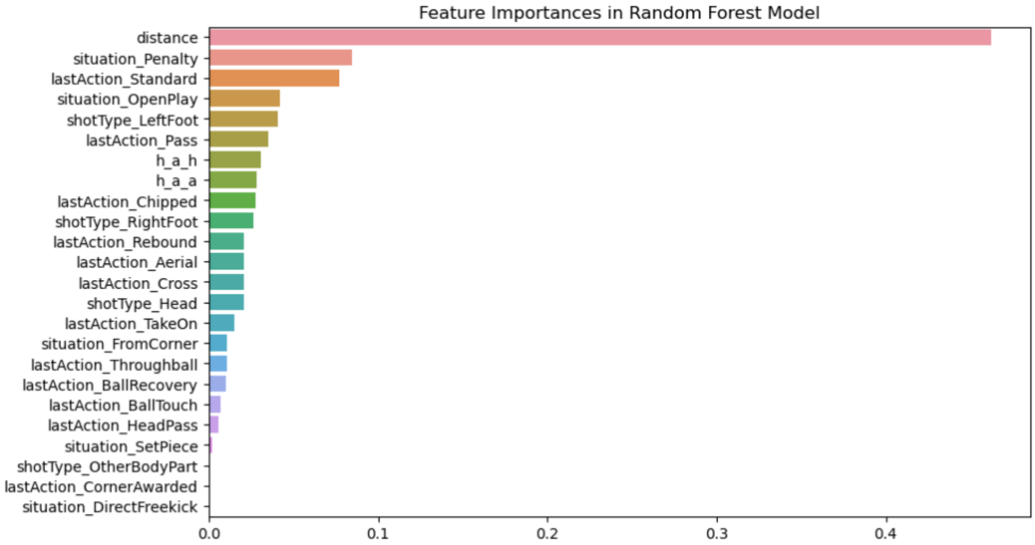


Figure 3.16

Depending on the impact levels of the variables in both graphs, more ideas can be obtained for the player's playing style. By determining the direction of the effect of highly influencing factors, it can be tried to be optimized specifically. The impact of variables can also be supported and elaborated by alternative descriptive methods such as heat maps. For example, the heat map in figure 3.17 was created with the Mplsoccer library and shows the heat map of all shots taken in the English Premier League during the Season.

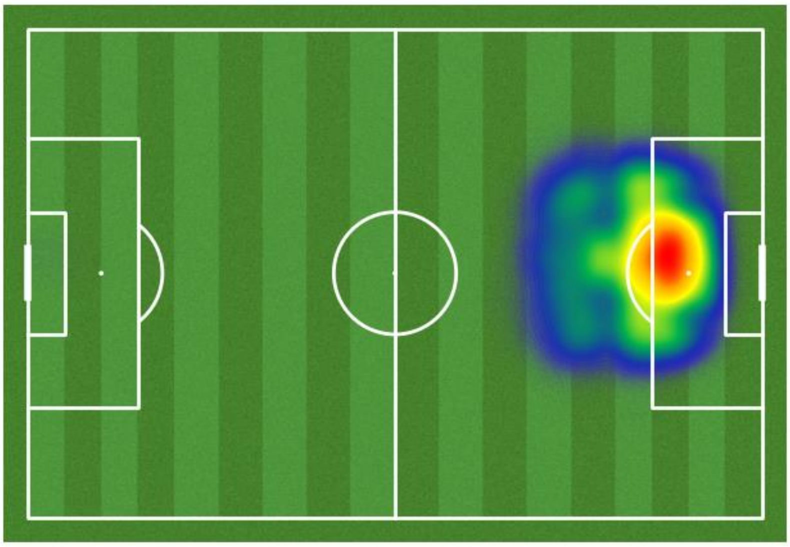


Figure 3.17

As expected, the more than 10000 shots taken have an almost symmetrical and homogeneous distribution according to the target. However, when the data is filtered according to the players or the main characteristics of the shot taken using pandas, the distinctive features of the heat map emerge. These maps can be used to obtain location-related information about the outputs from the random forest model we use to determine the effectiveness of factors that affect Haaland's ability to take advantage of scoring opportunities. The figures below are some examples of heatmaps of Erling Haaland's shots, which can be filtered out with an unlimited number of combinations with different characteristics.

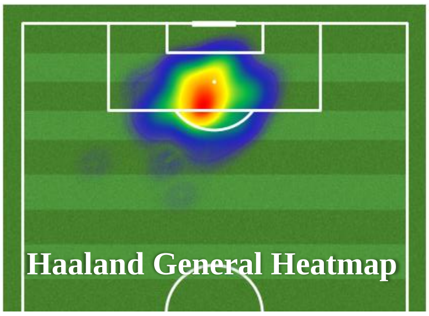


Figure 3.18

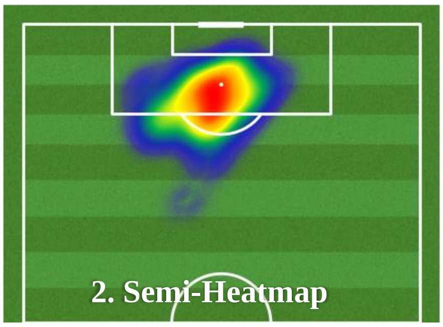


Figure 3.19

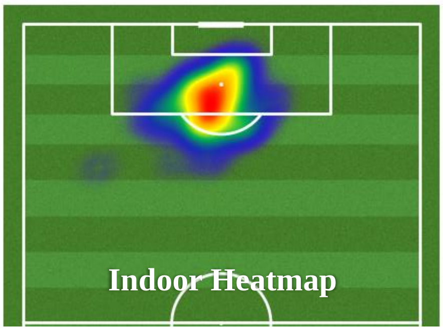


Figure 3.20

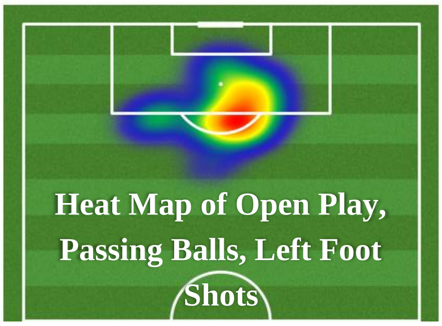


Figure 3.21

When the two tables showing the variable effects of the random forest models created for Gol and xG are compared, the differences between the importance and degrees of the variables are striking. For example, in the expected goal algorithm, the distance factor seems to have less effect, while Haaland's ability to convert positions into goals seems to have more of an effect.

From this, it can be concluded that the expected goal statistic and representation power may vary from player to player. Although the dataset used in this study includes a small set of shots taken for each player in a season, when training data or more retrospective detailed data are collected, player-specific expected goal algorithms can be developed, and efficient results can be obtained in the field of performance analysis by working on the optimization of the factors that affect these algorithms in the most positive way.

3.3 Match Scenario Prediction with Machine Learning

In this section, four different regression models are created with linear regression, decision trees, support vector machines and artificial neural networks for scenario predictions that will provide technically useful information before a match. While the definitions of the variables and the correlation coefficients were used in the selection of independent variables, the results obtained in the experiments with different combinations were compared. In order to define the performance of each model, MSE (Mean of squares of error), MAE (Absolute mean of error), RMSE (Root mean square deviation) R² (Descriptive coefficient) criteria were examined.

3.3.1 Prediction of Passing Actions Close to the Goal

The correlation coefficients, which indicate the linear relationship between the variable of the number of passes made by the opponent within 20 meters of the goal and the data of the team of interest and the opposing team, are as shown in

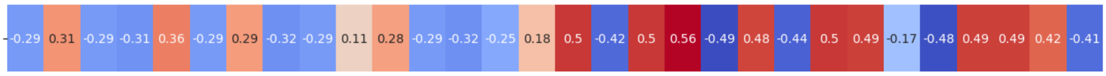


Figure 3.22

The variables used are as follows:

1. The average number of passes made by past opponents of the team of interest within 20 meters of the goal
2. Average expected goals tolerance of the team of interest
3. Average score expectation of the team of interest
4. Win rate of the team of interest

5. The number of passes made by the opposing team within 20 meters of the goal in the past matches
6. Average expected goals of the opposing team without penalties
7. Average score expectation of the opposing team
8. Side of the team of interest (home/away)

Statistics on the consistency of the regression models created are as shown in table 3.7.

	LINEAR REGRESSION		REGRESSION TREE		SVR		ANN	
	Test	train	test	train	test	train	test	train
MSE	12.66	17.78	13.59	16.09	11.91	17.86	12.64	17.8
MAE	2.78	3.18	3.0	2.94	2.7	3.05	2.84	3.18
RMSE	3.56	4.22	3.69	4.01	3.45	4.23	3.56	4.22
R²	0.46	0.45	0.42	0.5	0.49	0.44	0.46	0.45

Table 3.7

3.3.2 Expected Goals Tolerance Estimation

The correlation coefficients, which indicate the linear relationship between the tolerance variable shown to the opponent's expected goals and the data of the interested team and the opposing team, are as shown in figure 3.26

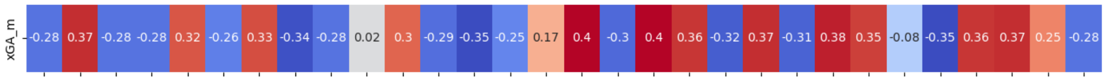


Figure 3.23

The variables used are as follows:

1. Average expected goals tolerance of the team of interest
2. Average expected goals of the opposing team
3. Average score expectation of the opposing team
4. Side of the team of interest (home/away)

The statistics on the consistency of the regression models created are as shown in table 3.8.

	LINEAR REGRESSION		REGRESSION TREE		SVR		ANN	
	test	train	test	train	test	train	test	train
MSE	0.78	0.62	0.88	0.58	0.86	0.61	0.79	0.61
MAE	0.69	0.62	0.71	0.59	0.71	0.59	0.69	0.62
RMSE	0.88	0.79	0.94	0.76	0.93	0.78	0.89	0.78
R²	0.34	0.34	0.25	0.38	0.27	0.35	0.33	0.34

Table 3.8

3.3.3 Score Expectation Estimation

The correlation coefficients, which indicate the linear relationship between the match-based score expectation variable and the season-based average data of the interested team and the opposing team, are as shown in figure 3.24.

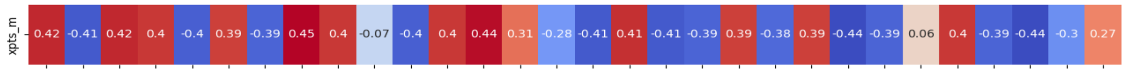


Figure 3.24

The variables used are as follows:

1. Average score expectation of the team of interest
2. Average expected goals of the opposing team
3. Average score expectation of the opposing team
4. Side of the team of interest (home/away)

The statistics on the consistency of the regression models created are as shown in table 3.9.

	LINEAR REGRESSION		REGRESSION TREE		SVR		ANN	
	test	train	test	train	test	train	test	train
MSE	0.38	0.42	0.47	0.4	0.4	0.42	0.37	0.42
MAE	0.5	0.54	0.55	0.51	0.49	0.5	0.49	0.54
RMSE	0.62	0.65	0.69	0.63	0.63	0.65	0.61	0.65
R²	0.49	0.46	0.38	0.49	0.48	0.46	0.51	0.46

Table 3.9

3.3.4 Expected goals Prediction

The correlation coefficients, which indicate the linear relationship between the expected goals variable on the basis of the match and the average data of the team of interest and the opposing team on a season basis , are as shown in figure 3.25.

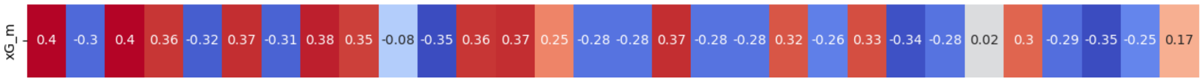


Figure 3.25

The variables used are as follows:

1. Average expected goals of the interested team
2. Average score expectation of the team of interest
3. Opponent's average expected goals tolerance
4. Average score expectation of the opposing team
5. Side of the team of interest (home/away)

Statistics on the consistency of the regression models created are as shown in table 3.10.

	LINEAR REGRESSION		**REGRESSION TREE**		**SVR**		**ANN**	
	test	train	test	train	test	train	test	train
MSE	0.57	0.67	0.73	0.68	0.58	0.69	0.58	0.69
MAE	0.61	0.65	0.67	0.64	0.61	0.64	0.61	0.66
RMSE	0.75	0.82	0.86	0.82	0.76	0.83	0.76	0.83
R²	0.35	0.34	0.16	0.33	0.34	0.33	0.33	0.32

Table 3.10

3.3.4.1 Data Processing to Improve the Expected Goals Prediction Model

With data pre-processing, it is possible to create better-performing models with more efficient metrics by taking advantage of the information at hand. In the model created in the previous section to predict the expected goals on a match basis, the prediction process was made with the seasonal averages of both sides. However, in some cases, the players on the field may be different from the season as a whole. So knowing the players who will take the field can help improve the model. In this section, in addition to the seasonal averages, the seasonal expected goal variable, which is specially created for the players in the squad, is used. With the new variables, the correlation coefficients indicating the linear relationship between the expected goal variable on a match basis and all other variables are as shown in figure 3.26.

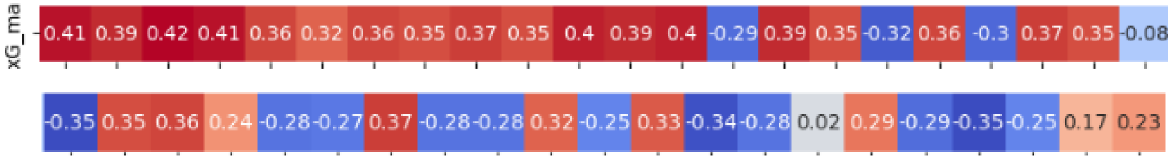


Figure 3.26

The variables used are as follows:

1. Average expected goals of the squad in which the team of interest takes the field
2. Average rating of the team of interest
3. Opponent's average expected goals tolerance
4. Average score of the opposing team
5. Side of the team of interest (home/away)

Statistics on the consistency of the regression models created are as shown in table 3.11.

	LINEAR REGRESSION		**REGRESSION TREE**		**SVR**		**ANN**	
	test	train	test	train	test	train	test	train
MSE	0.58	0.72	0.65	0.74	0.59	0.68	0.59	0.73
MAE	0.62	0.67	0.62	0.68	0.6	0.63	0.61	0.67
RMSE	0.76	0.85	0.81	0.86	0.77	0.83	0.77	0.86
R²	0.36	0.34	0.29	0.32	0.35	0.37	0.36	0.33

Table 3.11

3.3.5 Result Evaluation

Thanks to machine learning, it is possible to make predictions about the events that will occur in unplayed matches. Machine learning models with descriptive levels between 30% and 50% have been created with the data kept on the Understat.com website to provide descriptive statistics and to be used in data visualization. It has been observed that these models can be improved with the right data engineering in line with the available information. With deep industry knowledge, it is possible to obtain more consistent models with larger purpose-built data sets and more comprehensive pre-processing methods. These models can provide valuable insights to on-site managers.

4 RESULT

Although it is known that data analytics in sports has a history of more than 50 years, it has become a field of study that is rapidly changing with the developments in technology and progressing with rising momentum. The data density arising from the fact that it is a field of study that concerns large masses has brought the sports industry to a valuable position for data science. So much so that some studies in sports analytics have been the subject of studies in different fields such as psychology and behavioral economics. In sports, the exposure of data science, which has a wide range of uses from injury predictions to off-field management, to the technical understanding of the game, has made its impact strongly felt in sports branches followed by large masses, especially after the 2000s.

Recently, the use of artificial intelligence-based metrics in various sports, especially football, has opened a new page in sports analytics. As a result-oriented approach, it has been determined in the findings of the study that expected goals is a metric with a high consistent predictive power on the way to the goal. It is possible that the use of such metrics in other sports branches will become widespread. In the 2010s, the new approach to game understanding in basketball by names such as Daryl Morey is the result of findings based on the correct analysis of region-based scoring expectation, similar to the expected goal metric in football. In this context, just as expected goals vary depending on the various characteristics of the shot, the recent analysis of basketball players from mid-range shots to three-point shots and layups is no different from a reinforcement machine learning model determining new behavioral policies to optimize its results.

Although the expected goal metric has now become an actively used metric in football, findings have been obtained in the study that it can be used in more efficient forms for different purposes. The number of goals of players who manage to reach a high level of goals compared to the average may be below the expected goals. While this may be related to a player's ability to take advantage of the opportunities they get, it could also mean that the consistency of the model, which reveals the expected goal metric trained on shots from players at many different levels, can vary from player to player. According to the findings obtained in the study, differences were found between the factors affecting the ability of players with a high number of goals to convert positions into goals and the factors that increase the expectation of goals. This situation revealed that the predictive power of the model that reveals the expected goal metric will increase if it is trained with more original data specific to the player, the player's playing style or the league, provided that it is sufficient in number. Future studies in this area with larger data sets will be useful in developing tactical plans suitable for the players, identifying missing points and optimizing the results for the score.

In addition to expected goals, numerous numerical variables that can be extracted from a football match can be used to predict match scenarios with machine learning. In this study, it was seen that concrete statistics such as "number of passes to be made near the goal" or derived statistics such as expected goals were estimated with consistency coefficients of up to 50%, and the results could be improved with appropriate data engineering techniques that were open to elaboration with deeper technical knowledge. Future studies in this area will provide more comprehensive technical insights to in-field managers.

BIBLIOGRAPHY

[1] Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. IBM Journal of Research and Development, 3(3), 210-229. https://doi.org/10.1147/rd.33.0210

[2] Cook, E. (1966). *Percentage Baseball*. MIT Press.

[3] Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology, 17*(3), 295–314. https://doi.org/10.1016/0010-0285(85)90010-6

[4] Cortes, C., Vapnik, V. Support-vector networks. Mach Learn 20, 273–297 (1995). https://doi.org/10.1007/BF00994018

[5] Itchhaporia, D., Snow, P., Almassy, R., & Oetgen, W. (1996). Artificial neural networks: current status in cardiovascular medicine.. Journal of the American College of Cardiology, 28 2, 515-21 . https://doi.org/10.1016/0735-1097(96)00174-X.

[6] Sutton, R.S., & Barto, A.G. (1998). Introduction to Reinforcement Learning.

[7] Atkinson, G., & Nevill, A. (2001). Selected issues in the design and analysis of sport performance research. Journal of Sports Sciences, 19, 811 - 827. https://doi.org/10.1080/026404101317015447.

[8] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

[9] Pollard, R. (2002). Charles Reep (1904-2002): pioneer of notational and performance analysis in football. Journal of Sports Sciences, 20(10), 853–855. https://doi.org/10.1080/026404102320675684

[10] Lewis, M. (2003). *Moneyball: The Art of Winning an Unfair Game*. W.W. Norton & Company.

[11] Burns, B. D. (2004). Heuristics as beliefs and as behaviors: The adaptiveness of the "hot hand". *Cognitive Psychology, 48*(3), 295–331. https://doi.org/10.1016/j.cogpsych.2003.07.003

[12] Cortés, C., & Vapnik, V.N. (2004). Support-vector networks. Machine Learning, 20, 273-297.

[13] Smola, A.J., Schölkopf, B. A tutorial on support vector regression. Statistics and Computing 14, 199–222 (2004). https://doi.org/10.1023/B:STCO.0000035301.49549.88

[14] Cherkassky, V., & Ma, Y. (2004). Practical selection of SVM parameters and noise estimation for SVM regression. Neural networks : the official journal of the International Neural Network Society, 17 1, 113-26.

[15] Sundali, J., & Croson, R. (2006). Biases in casino betting: The hot hand and the gambler's fallacy. Judgment and Decision Making, 1(1), 1–12. https://doi.org/10.1017/S1930297500000309

[16] McGarry, T. (2009). Applied and theoretical perspectives of performance analysis in sport: Scientific issues and challenges. International Journal of Performance Analysis in Sport, 9, 128 - 140. https://doi.org/10.1080/24748668.2009.11868469.

[17] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.

[18] Zhu, X., & Goldberg, A.B. (2009). Introduction to Semi-Supervised Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning.

[19] RABIN, M., & VAYANOS, D. (2010). The Gambler's and Hot-Hand Fallacies: Theory and Applications. The Review of Economic Studies, 77(2), 730–778. http://www.jstor.org/stable/40587644

[20] Yaari, G., & Eisenmann, S. (2011). The hot (invisible?) hand: Can time sequence patterns of success/failure in sports be modeled as repeated random independent trials? PLOS ONE, 6(10), e24532. https://doi.org/10.1371/journal.pone.0024532

[21] Stoltzfus, J. (2011). Logistic regression: a brief primer.. Academic emergency medicine : official journal of the Society for Academic Emergency Medicine, 18 10, 1099-104 . https://doi.org/10.1111/j.1553-2712.2011.01185.x.

[22] Loh, W. (2011). Classification and regression trees. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1.

[23] Xie, J., Jiang, S., Xie, W., & Gao, X. (2011). An Efficient Global K-means Clustering Algorithm. J. Comput., 6, 271-279.

[24] Su, X., Yan, X. and Tsai, C.-L. (2012), Linear regression. WIREs Comp Stat, 4: 275-294. https://doi.org/10.1002/wics.1198

[25] Bergkamp, D., & Winner, D. (2013). Stillness and Speed: My Story. Simon & Schuster UK.

[26] Cohen, B. (2014, February 27). Does the 'hot hand' exist in basketball? Wall Street Journal. Retrieved May 6, 2016, from https://www.wsj.com/

[27] Coutts, A. J. (2014). Evolution of football match analysis research. Journal of Sports Sciences, 32(20), 1829–1830. https://doi.org/10.1080/02640414.2014.985450

[28] Sperandei S. (2014). Understanding logistic regression analysis. Biochemia medica, 24(1), 12–18. https://doi.org/10.11613/BM.2014.003

[29] Schmidhuber, J. (2014). Deep learning in neural networks: An overview. Neural networks : the official journal of the International Neural Network Society, 61, 85-117 .

[30] Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2014). Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal, 13, 8–17. https://doi.org/10.1016/j.csbj.2014.11.005

[31] Jordan, M., & Mitchell, T. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349, 255 - 260. https://doi.org/10.1126/science.aaa8415.

[32] Min Cao, Roman Chychyla, Trevor Stewart; Big Data Analytics in Financial Statement Audits. *Accounting Horizons* 1 June 2015; 29 (2): 423–429. https://doi.org/10.2308/acch-51068.

[33] Erwan Scornet. Gérard Biau. Jean-Philippe Vert. "Consistency of random forests." Ann. Statist. 43 (4) 1716 - 1741, August 2015. https://doi.org/10.1214/15-AOS1321

[34] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. Nature, 521, 436-444. https://doi.org/10.1038/nature14539.

[35] R. Stanojevic and L. Gyarmati, "Towards Data-Driven Football Player Assessment," *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW),* Barcelona, Spain, 2016, pp. 167-172, doi: 10.1109/ICDMW.2016.0031.

[36] Green, B., & Zwiebel, J. (n.d.). The hot hand fallacy: Cognitive mistakes or equilibrium adjustments? Evidence from baseball. Stanford Graduate School of Business. Retrieved May 6, 2016, from https://www.gsb.stanford.edu/

[37] Miller, J. B., & Sanjurjo, A. (2016). Surprised by the gambler's and hot hand fallacies? A truth in the law of small numbers. IGIER Working Paper, 552. https://doi.org/10.2139/ssrn.2627354

[38] Mammone A., Turchi M., Cristianini N. Support vector machines. Wiley Interdisciplinary Reviews: Computational Statistics 2009; 1(3): 283-289.

[39] Rokach, L. (2016). Decision forest: Twenty years of research. Information Fusion, 27, 111-125. https://doi.org/10.1016/j.inffus.2015.06.005

[40] Juravich, M., Salaga, S., & Babiak, K. (2017). Upper Echelons in Professional Sport: The Impact of NBA General Managers on Team Performance. Journal of Sport Management, 31, 466-479.

[41] Lee, J.H., Shin, J., & Realff, M.J. (2017). Machine learning: Overview of the recent progresses and implications for the process systems engineering field. Comput. Chem. Eng., 114, 111-121.

[42] Zhang, S., Li, X., Zong, M., Zhu, X., & Cheng, D. (2017). Learning k for kNN Classification. ACM Transactions on Intelligent Systems and Technology (TIST), 8, 1 - 19. https://doi.org/10.1145/2990508.

[43] Hodeghatta, U. R., & Nayak, U. (2017). Business Analytics Using R - A Practical Approach. Apress.

[44] Yu, S., Chu, S., Wang, C., Chan, Y., & Chang, T. (2017). Two improved k-means algorithms. Appl. Soft Comput., 68, 747-755. https://doi.org/10.1016/j.asoc.2017.08.032.

[45] Morgulev, E., Azar, O., & Lidor, R. (2018). Sports analytics and the big-data era. International Journal of Data Science and Analytics, 5, 213-222. https://doi.org/10.1007/s41060-017-0093-7.

[46] Pollard, R. (2019). Invalid interpretation of passing sequence data to assess team performance in football: Repairing the tarnished legacy of Charles Reep. The Open Sports Sciences Journal, 12(1), 17-21. https://doi.org/10.2174/1875399X01912010017

[47] Duquette, C. M., Cebula, R. J., & Mixon, F. G. (2019). Major league baseball's *Moneyball* at age 15: a re-appraisal. *Applied Economics*, *51*(52), 5694–5700. https://doi.org/10.1080/00036846.2019.1617399

[48] Geron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (2nd ed.). O'Reilly Media.

[49] Ying, X. (2019). An Overview of Overfitting and its Solutions. Journal of Physics: Conference Series, 1168.

[50] Sarlis, V., & Tjortjis, C. (2020). Sports analytics - Evaluation of basketball players and team performance. Inf. Syst., 93, 101562. https://doi.org/10.1016/j.is.2020.101562.

[51] Hoege, J., Lansing, M., Nelson, S., Ungerleider, D., Iyer, R., Rhodes, C., Metzger, B., Worcester, P., Chandra, A., Leonard, J., Kreitzer, R., & Scherer, W. (2020). An Interdisciplinary Approach to Sports Analytics in a University Setting. 2020 Systems and Information Engineering Design Symposium (SIEDS), 1-6. https://doi.org/10.1109/SIEDS49339.2020.9106647.

[52] Elitzur, R. (2020). Data analytics effects in Major League Baseball. *Omega*, 90, Article 102021. https://doi.org/10.1016/j.omega.2018.11.010

[53] McNair, B., Margolin, E., Law, M., & Ritov, Y. (2020). The hot hand and its effect on the NBA. arXiv:2010.15943 [stat. AP].

[54] Sarlis, V., Chatziilias, V., Tjortjis, C., & Mandalidis, D. (2021). A Data Science approach analysing the Impact of Injuries on Basketball Player and Team Performance. Inf. Syst., 99, 101750. https://doi.org/10.1016/J.IS.2021.101750.

[55] Watanabe, N. M., Shapiro, S., & Drayer, J. (2021). Big Data and Analytics in Sport Management. Journal of Sport Management, 35(3), 197-202. Retrieved May 27, 2024, from https://doi.org/10.1123/jsm.2021-0067

[56] Branga, V. (2021). Big Data Analytics in Basketball Versus Business. Studies in Business and Economics,16(3) 24-31 https://doi.org/10.2478/sbe-2021-0042.

[57] Miller, J. B., & Sanjurjo, A. (2021). Is it a fallacy to believe in the hot hand in the NBA three-point contest? European Economic Review, 138, 103771. https://doi.org/10.1016/j.euroecorev.2021.103771

[58] Cebeci Z., Yildiz, F., & Kayaalp G., (2015). Selecting the Optimum K Value in K-Means Clustering. 2nd National Management Information Systems Congress (pp.231-242). Erzurum, Turkey

[59] Janiesch, C., Zschech, P. & Heinrich, K. Machine learning and deep learning. Electron Markets 31, 685–695 (2021). https://doi.org/10.1007/s12525-021-00475-2

[60] Koshkina, M., Pidaparthy, H., & Elder, J.H. (2021). Contrastive Learning for Sports Video: Unsupervised Player Classification. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 4523-4531.

[61] Umami, I., Gautama, D., & Hatta, H. (2021). implementing the Expected Goal (xG) model to predict scores in soccer matches. International Journal of Informatics and Information Systems, 4(1), 38-54. https://doi.org/10.47738/ijiis.v4i1.76

[62] Pinheiro, R., & Szymanski, S. (2022). All Runs Are Created Equal: Labor Market Efficiency in Major League Baseball. Journal of Sports Economics, 23(8), 1046-1075. https://doi.org/10.1177/15270025221085712

[63] Pelechrinis, K., & Winston, W. (2022). The hot hand in the wild. PLOS ONE, 17(1), e0261890. https://doi.org/10.1371/journal.pone.0261890

[64] Bhattamisra, S., Banerjee, P., Gupta, P., Mayuren, J., Patra, S., & Candasamy, M. (2023). Artificial Intelligence in Pharmaceutical and Healthcare Research. Big Data Cogn. Comput., 7, 10. https://doi.org/10.3390/bdcc7010010.

[65] Hewitt, J. H., & Blackbird, O. (2023). A machine learning approach for player and position adjusted expected goals in football (soccer). Frontiers in Artificial Intelligence and Operations Research, 2, 100034. https://doi.org/10.1016/j.fraope.2023.100034

[66] Mead J, O'Hare A & McMenemy P (2023) Expected goals in football: Improving model performance and demonstrating value. Muazu Musa R (Editor) <i>PLOS ONE</i>, 18 (4), Art. No.: e0282295. https://doi.org/10.1371/journal.pone.0282295

Additional Resources

- Zhang, T. (2001). An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. A.I. Mag., 22, 103-104.
- Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.
- Baca, A., & Kornfeind, P. (2012). Stability analysis of motion patterns in biathlon shooting. *Human movement science*, *31*(2), 295–302. https://doi.org/10.1016/j.humov.2010.05.008
- Provost, F., & Fawcett, T. (2013). Data Science and Its Relationship to Big Data and Data-Driven Decision Making. *Data Science Journal*, 1(1), 51-59.
- Sampaio, J. E., Lago, C., Gonçalves, B., Macãs, V. M., & Leite, N. (2014). Effects of pacing, status and unbalance in time motion variables, heart rate and tactical behaviour when playing 5-a-side football small-sided games. *Journal of science and medicine in sport*, *17*(2), 229–233. https://doi.org/10.1016/j.jsams.2013.04.005
- Batra, R., & Verma, S. (2020). A study of the adoption of artificial intelligence by the marketing professionals of India. *Omega*, *90*, 102021. https://doi.org/10.1016/j.omega.2018.09.005
- https://understat.com/