INTERNATIONAL
HELLENIC
UNIVERSITY

Sports Analytics
Performance Prediction

Nikolaos Giannakoulas

SID: 3308210013

School of Science and Technology

A thesis submitted for the degree of

Master of Science (MSc) in Data Science

January 2023

Thessaloniki – Greece

INTERNATIONAL HELLENIC UNIVERSITY

Sports Analytics

Performance Prediction

Nikolaos Giannakoulas

SID: 3308210013

Supervisor:	Assoc. Prof. Christos Tjortjis
Supervising Committee Members:	Dr. Paraskevas Koukaras
	Dr. Christos Berberidis

School of Science and Technology

A thesis submitted for the degree of

Master of Science (MSc) in Data Science

January 2023

Thessaloniki – Greece

Abstract

This dissertation was written as part of the MSc in Data Science at the International Hellenic University.

Sports Analytics is a rapidly growing field. It is experiencing great development, and its applications are very useful to sports clubs. These clubs can collect information about the players and the game generally and then extract important insights that will help them improve.

There is a wide variety of daily football statistics. Thousands of football games happen every week, resulting in the production of numerous statistics. Sports analytics' responsibility is to collect those statistics, analyze them and then provide conclusions to football clubs.

The goal of this dissertation is to predict the performance of a player in the field of football. The purpose of the dissertations is to predict, as accurately as possible the number of goals a football player will achieve next season based on his previous years' performances.

Football players' data were collected from valid online sources and then analyzed [1]. After that, feature engineering was implemented to transform the collected data into the desired form. Then, three machine learning algorithms were used for predictions.

This dissertation is separated into two parts. The first part is theoretical, and it includes previous works on the field of performance prediction, not only in football but in other sports too, like basketball, volleyball and tennis.

For the second part, the practical one, we paid attention to the conducted experiments. The results were examined thoroughly and compared to understand them and determine which model had better performance. For this purpose, python was used for coding, specifically PyCharm.

Acknowledgments

At this point, I would like to thank my supervisor, Professor Christos Tjortjis, for his helpful advice throughout the dissertation. He gave me the appropriate feedback and he was very supportive whenever I had questions. In addition, I would like to thank PhD candidate Georgios Papageorgiou who was always helping me with efficient comments and suggestions to problems that emerged.

Moreover, I would like to thank my family for their support during those difficult and stressful months. They were really supporting me during this journey.

Nikolaos Giannakoulas

07 - 01 - 2023

Contents

List of Figures

List of Tables

1 Introduction

This dissertation comprises seven chapters. The first chapter is the Introduction. The second one, the Context, follows. In Context, general terms are discussed about Machine Learning (ML) and different models we used for the predictions. Next, in chapter three, Literature Review, we present previous works in the field of performance prediction and the historical background of sports analytics for various sports. Then, in chapter four, the problem and the approaches that have been used during the experiments are presented. After that, in chapter five, we present the models' results. Moreover, in chapter six, we discuss our findings and compare the results from the methods to find out which was the best. Lastly, conclusions and future work are discussed in the final chapter, chapter seven.

Over the last few years, sports analytics has become a vital part of sports. Subsequently, several works have been conducted in recent years regarding football, and these works have been proved vital for sports. Many historical and relevant data from players, games and scores are collected and analyzed through algorithms and ML models generally. Nowadays, many football clubs have created a group of data scientists in order to have access to meaningful information in order to improve themselves. The use of these technologies and statistics leads to better decision-making from clubs. The coaching staff can analyze the games and better understand their players' performance in order to make the appropriate decisions for the future. For this reason, football clubs know exactly their weaknesses because they are able to detect and fix them.

This dissertation focuses on football. Football is the most famous sport worldwide, with millions of fans and supporters. Billions of dollars are invested on football every year, specifically in players, stadiums, training grounds and TV broadcasts. The arising data on football generates interest from scientists who are keen on understanding this field.

In this dissertation, we aim to predict the performance of a player in terms of goals. For this purpose, we should be extremely careful in the collection of data. To be more specific, there is a huge amount of information and free data for this sport. So, it is not very easy to select the appropriate dataset for the research. For this reason, we collected data from a valid American organization, named Sports Reference. These data include

advanced statistics for more than 2.500 football players from season 2017-18 to season 2021-22 [1]. Besides that, we implemented feature engineering because we should make appropriate changes to the dataset to transform it into the desired form. The statistics about players in this dataset should appear in a specific form which is known as season-lag features and we will present it during the following chapters of this dissertation.

Finally, we had to choose and implement ML algorithms for the last stage of the research which was the predictions. The algorithms were evaluated with the appropriate metrics, and their results were examined and compared consequently. These metrics are Mean Absolute Error (MAE), Mean Square Error (MSE) and Root Mean Square Error (RMSE). Linear Regression was the algorithm with the best performance as it had the lowest MAE, MSE and RMSE comparatively with the other two algorithms. The second-best algorithm was Multilayer Perceptron which results were better than those of Random Forest. Generally, the results of all algorithms were good as their MAE, MSE and RMSE were really close to zero, which means that our predictions are close to the real values.

Additionally, it is important to mention that there is a difficulty that may affect the evaluation of the final results because there are many players that score a small number of goals in the season that we want to predict, while they used to score many goals during the previous seasons. A typical example is Lionel Messi, who used to score 31 goals on average during seasons 2017-18 to 2020-21 and he scored just 6 goals in season 2021-22.

So, it is obvious that there are many out-field factors that can affect the performance of a player. Examples of these out-field factors are the player's psychological condition, his injuries, weather conditions, team atmosphere, i.e. the relationship with his teammates and many others. Nevertheless, in this dissertation it is proved that it is possible to make a good prediction and achieve good accuracy for the goals of a football player. What makes this dissertation interesting is the use of many advanced statistics regarding those players and the process of converting them to historical data.

To conclude, sports analytics has achieved remarkable improvement throughout the years. However, there is still a place for enhancement. For instance, wearable devices should be used from as many football clubs as possible because with these, teams can learn more about their squad. So, they will improve the decision-making process, and achieve better performances.

2 Context

In this section general terms with regards to ML will be presented. In more detail, ML, Supervised Learning, Unsupervised Learning, Reinforcement Learning, Data Mining and Sports Analytics will be discussed. So, at the end of this section we will have a clear image about their meaning, importance and role to our research.

2.1 Machine Learning

Machine Learning is a subset of Artificial Intelligence and a subset of Computer Science consecutively that allows computers to become more autonomous in decision making through experience without being programmed to do so. Nowadays, ML is experiencing great development because it is one of the most growing technical fields worldwide [2]. Also, it has multiple applications in everyday life like fraud detection, speech recognition, manufacturing, health care, education and financial modeling [3].

There are three different ML approaches that a data scientist can follow depending on the type of data they want to predict. Supervised, Unsupervised and Reinforcement learning are these approaches.

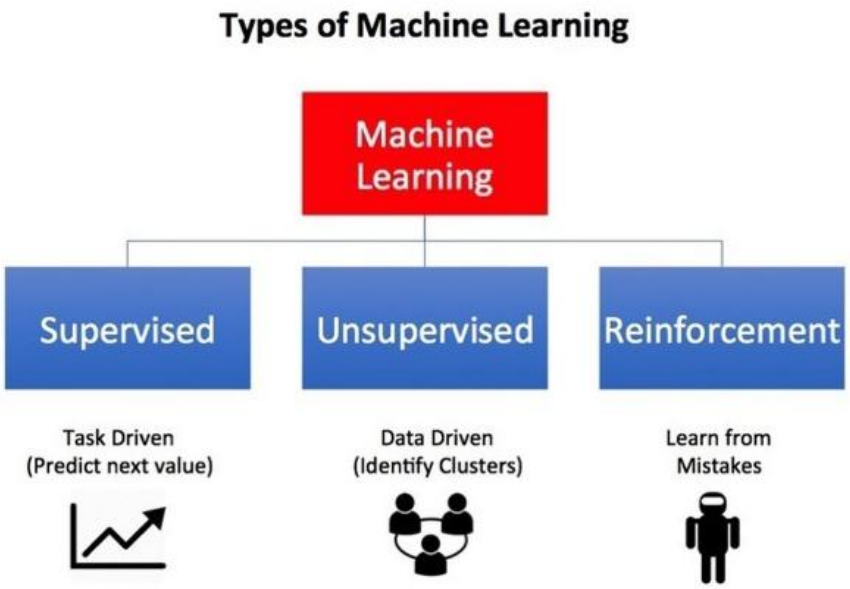

Figure1. Machine Learning approaches [4]

2.2 Supervised Learning

In supervised learning, the algorithms create functions that match the inputs to desired outputs. It trains algorithms with labelled datasets in order to predict outcomes or classify data. Also, supervised learning uses training datasets, which include inputs and correct outputs, to teach models and these models learn over time.

There are two main supervised models, classification models and regression models. Classification models separate data into categories, while regression models try to understand the relationship between independent and dependent variables [5] [6].

Some of the most common supervised learning algorithms are:

1. Naive Bayes
2. Support Vector Machine (SVM)
3. Random Forest
4. K-Nearest Neighbor
5. Decision Trees
6. Neural Networks

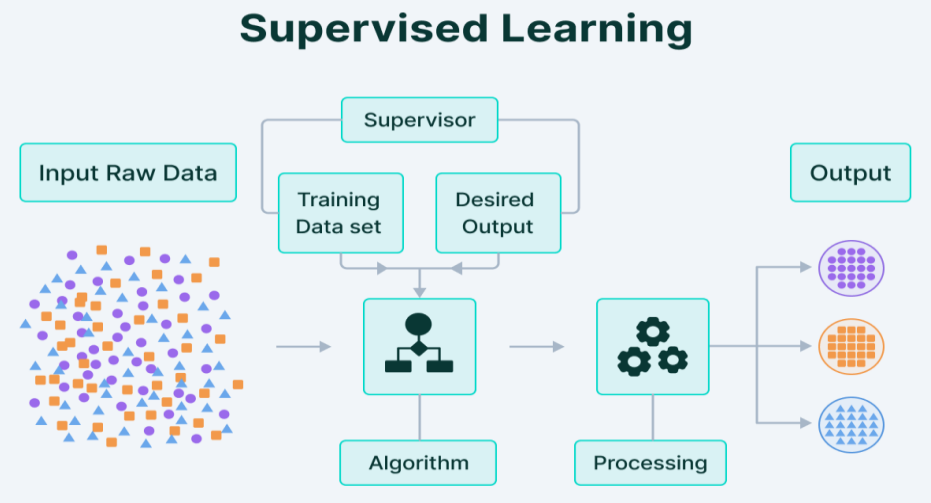

Figure2. Supervised Learning process [7]

2.3 Unsupervised Learning

Besides supervised learning, in unsupervised learning, the models are trained on an unlabelled dataset. To be more specific, the model becomes a self-learner by "studying" the features of the training dataset and then, based on them, the model is ready for predictions on test data. In other words, the model is the "teacher" itself. Unsupervised learning has many applications in many domains like natural language processing, computer vision and speech recognition [8] [9].

Some of the most common unsupervised learning algorithms are:

1. Apriori Algorithms
2. Association Rules
3. Dimensionality Reduction
4. Clustering
5. Principal Component Analysis (PCA)

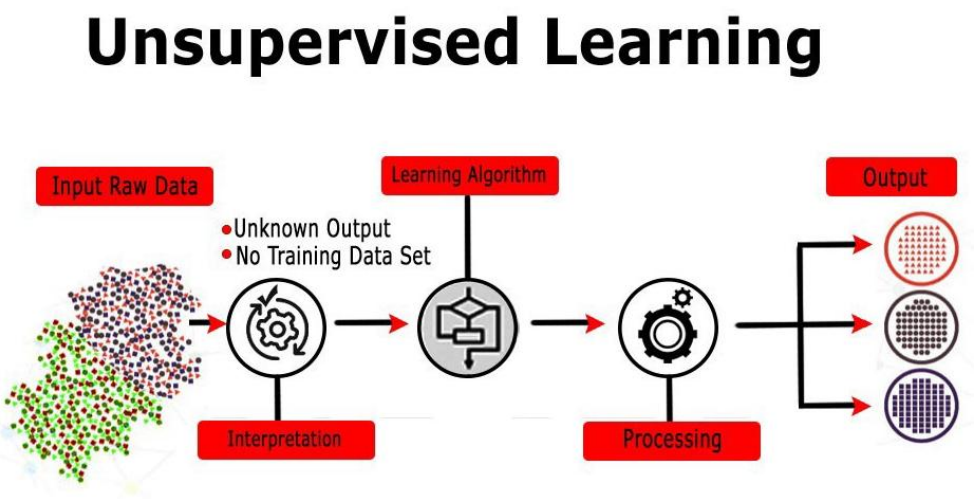

Figure3. Unsupervised Learning process [9]

2.4 Reinforcement Learning

Unlike the other two ML approaches, Reinforcement Learning is a trial-and-error process that learns from the feedback it takes. In Reinforcement Learning, the agent tries to understand what the most appropriate actions are in order to have the best results for the system [10] [11].

Some of the most common reinforcement learning algorithms are:

1. Monte Carlo methods
2. Q-Learning
3. Dynamic Programming
4. TD-Learning

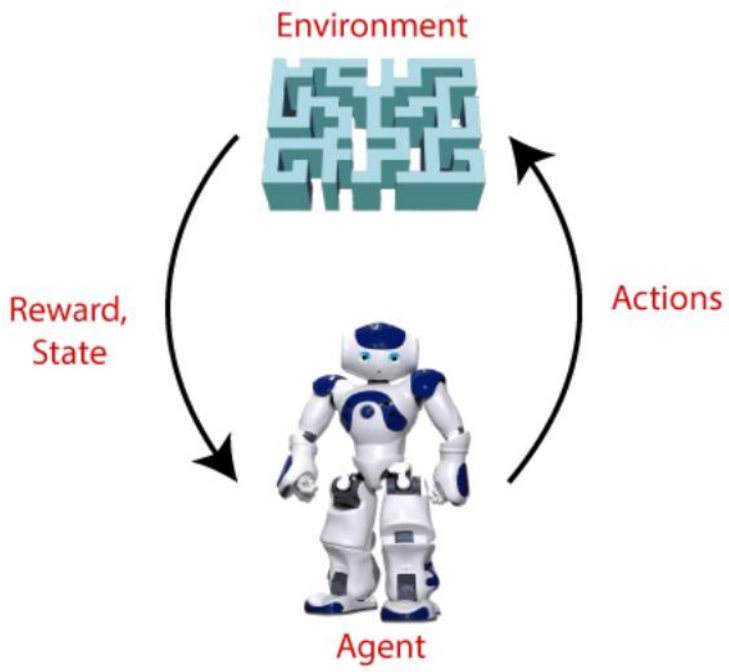

Figure4. Reinforcement Learning process [12]

2.5 Data Mining

The rapid growth of databases has increased the need of tools and techniques which are able to transform data into information. Data mining is the process of extracting information and patterns that were previously undiscovered and then making useful insights from them. Data mining is also known as "knowledge extraction" or "knowledge discovery in databases". Moreover, it is very important because it has many applications in organizations like decision making, query processing and process control [13].

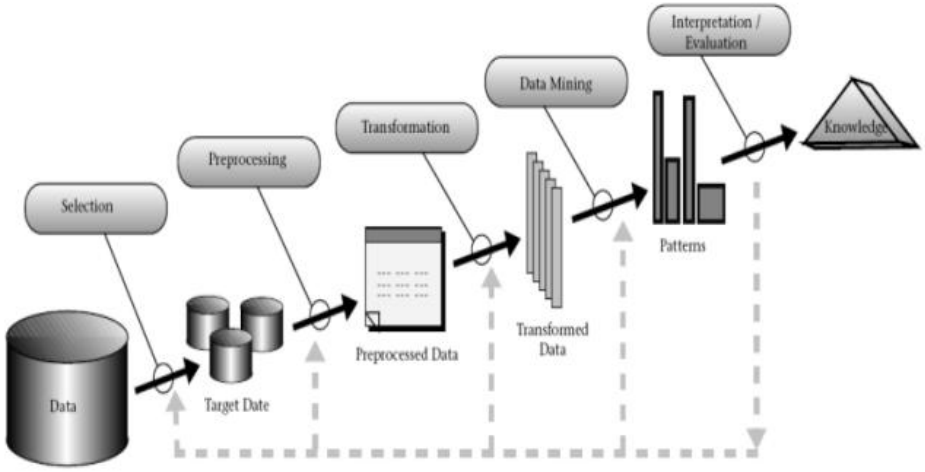

Figure5. Data Mining process [14]

As can be seen from the above figure, there are five main steps during the data mining process. In the beginning, the appropriate data is selected. In the next two steps, data is cleaned and then transformed. After the data's identification and exploration, patterns are exploited in order to extract knowledge from them in the last step [14].

The most famous Data Mining algorithms are:

1. Association Rules
2. Decision Trees
3. Neural Networks

Nowadays, almost every sector is digitalized, and results in huge volume of data. Because of this, Data Mining has many applications. Some of these applications are the following [15]:

1. Market Basket Analysis
2. Educations
3. Bioinformatics
4. Healthcare
5. Criminal Investigation
6. Financial Banking
7. Web Mining

2.6 Sports Analytics

2.6.1 Introduction

Sports analytics is the collection and analysis of historical data that can be used for better decision-making. These data provide an important advantage to coaching staff and athletes generally because their analysis can be very informative and give useful insights about sports.

Generally, it is very vital to understand that Sports analytics is more important than just numbers and goals. Instead, it is something more than that which enables scientists to improve the sport. Today, sports have become so competitive with all this information, which is why they have become enjoyable. Before decades, football was very “slow”. There weren’t tactics and the rules were completely different and that’s why people’s engagement was low.

2.6.2 Sports Analytics’ Applications

Sports analytics has several applications. Some of them are evaluating player performance prediction, ticket prizes, betting and the strategy that a team will use

during a match [16]. All these applications are very important for sports organizations because they can develop their performance and, consequently, increase their value. Moreover, one of the most important sports analytics' applications is the ability to predict players' injuries and the time that they need to heal. Nowadays, we are able to figure out players' vulnerability to injuries and the time that is required for cure [17].

Lastly, predictions about the evolution of the game are the most important application of sports analytics in sports. Big betting companies invest great amount of money in this direction. Knowing the result of a football game or the performance of a player as accurately as possible is a huge achievement. Millions of people bet on sports events, and these companies should be as prepared as possible in order to offer odds and finally win. That's why they pay so much attention to implementing sports analytics.

2.6.3 Sports Analytics' Categories

Sports analytics is divided into two categories, on-field analytics and off-field analytics. On-field analytics has to do with the improvement of players' and team's performance. It is related to player fitness, squad selection or defensive strategy for instance. The main purpose of on-field analytics is to help clubs be better season by season. On the other hand, of-field analytics deals with every aspect of sports analytics except on-field analytics. To be more specific, of-field analytics focuses on growing the profitability of sports organizations, for example, with fan engagement, ticket prices and sales [18].

2.6.4 Wearable Devices

It is obvious that Sports analytics is evolving into sports day by day and year by year. Nowadays, the use of wearable devices is very common among clubs, especially in football, where managers are able to collect incredible information about their players. With all this technology around us, we are able to answer plenty of questions that were unachievable to understand the previous years.

For example, when actually a player needs rest for some days? Are there any chances for a football player to get injured next week? Which is the best type of training that he

should follow to recover after a serious injury? For this reason, many football clubs around the world have adopted the system of wearable devices. These devices collect information for every player individually, and then the decisions are in the coaches' judgment.

Despite that, it is really challenging to get access to information collected by these devices, which have to do with the physical condition of the players. That's because football clubs consider these statistics extremely confidential, so this kind of datasets are not obtained easily for research purposes [19].

Moreover, ML is experiencing great development and gives us the opportunity to analyze videos and images. Many cameras are installed on football grounds, and through them, it is possible to see the whole team's performance and style of playing.

2.7 Modeling

2.7.1 Linear Regression

Linear Regression is a supervised ML algorithm, and it is the most common type of predictive analysis. Its main purpose is to find out the relationship between two or more variables. Moreover, it is used in order to predict the dependent variable's value based on the independent variable's value. With this type of analysis, we are able to calculate the coefficients of an equation which involves many independent variables that are useful for the best prediction of the dependent variable [20] [21]. Also, Linear Regression performs predictions based on numeric or continuous variables like sales, age and price.

Generally, this is the simplest form that can describe the Linear Regression: $y = ax + b$. In this equation, the dependent variable is y, and its value depends on the value of the independent variable x. Also, b represents the constant in the equation. Based on this simple form, Linear Regression can become more complicated with more independent variables. For example, the following equation has three independent variables x, z and w: $y = ax + bz + cw + d$ where d is the constant [20] [21].

The Linear Regression model produces a straight line that depicts the relationship among the variables, as can be seen from the figure below [22]:

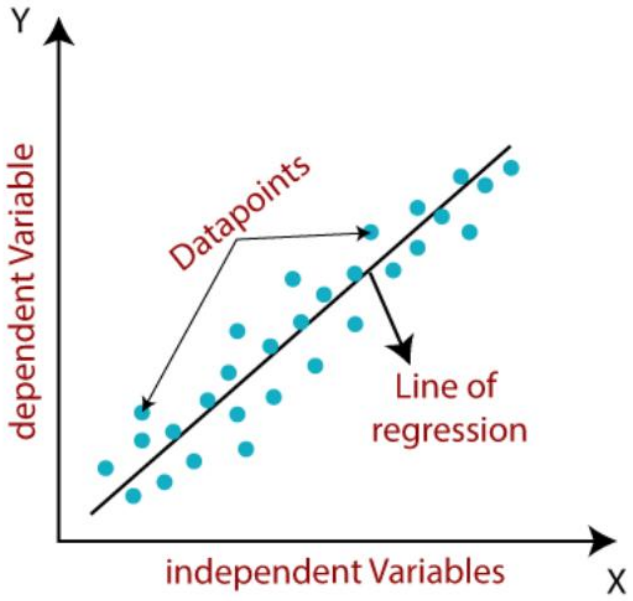

Figure6. Linear Regression representation [22]

The straight line that separates the data points is called regression line. The slope of this line can be increasing or decreasing. To be more specific, there are two types of relationships that depend on the type of the slope, the positive linear relationship and the negative linear relationship. These two relationships can be seen from the figures below. A positive relationship means that when the dependent variable on the y-axis increases then the independent variable on the x-axis increases too. On the contrary, a negative relationship means that when the dependent variable decreases then the independent variable decreases too.

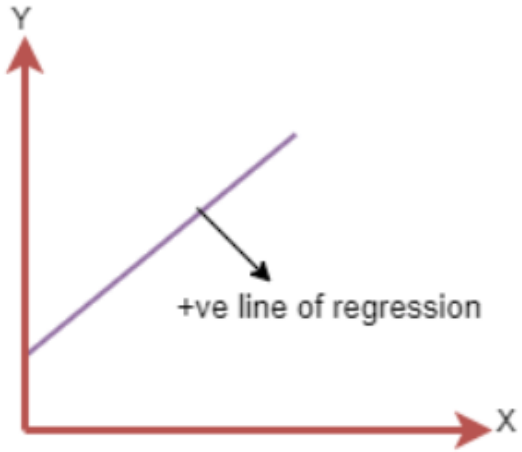

Figure7. Positive Linear relationship [22]

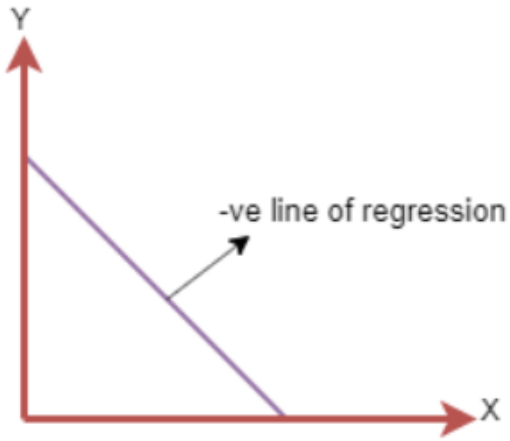

Figure8. Negative Linear Relationship [22]

Moreover, there are two Linear Regression types that are correlated with the number of the independent variables that are used in order to predict the value of the dependent variable. So, if only one independent variable is used for the prediction, then the Linear Regression algorithm is called Simple Linear Regression. On the other side, if more than one independent variable is used for the prediction, then the algorithm is called Multiple Linear Regression [22].

The main goal of Linear Regression is to find the best line that minimizes the error between predicted and actual values. In order to achieve that, we should calculate the cost function which optimizes weights and coefficients and measures how well a Linear Regression model is performing. In Linear Regression, the most common cost function is called Mean Squared Error (MSE). In order to minimize the MSE, we use the Gradient Descent by calculating the slope of the cost function [22].

Overall, Linear Regression models are simple and provide an easy-to-understand mathematical formula which can be used for predictions. But, the most important characteristic of these models is that they can be implemented in many areas, from business to academic studies. Biology and social sciences are also some examples. Businesses collect huge amount of data, and with the help of linear regression they are able to transform them into actionable information. Furthermore, in Sports Analytics the use of Linear Regression can help clubs to predict their team's performance to the next game or the next season [21].

2.7.2 Random Forest

Random Forest is a supervised ML algorithm which is used basically in classification and regression problems. It is based on ensemble learning and it constructs a multitude of decision trees in order to solve a difficult problem and to achieve a better performance of the model. To be more specific, Random Forest has decision trees on different subsets of the dataset, and its goal is to improve the predictive accuracy of this dataset. That is, it aggregates every tree's prediction and then it outputs the optimal result [23]. So, for regression problems, the output of this algorithm is the mean of the individual trees, while in classification problems the output is the class which was selected from the most decision trees. The only problem with this algorithm is the over-fitting, but having a great number of trees, we can prevent it by having a high accuracy [24]. On the other hand, these trees shouldn't be many in number because in that case, the algorithm becomes slow and ineffective.

The following diagram depicts the Random Forest algorithm:

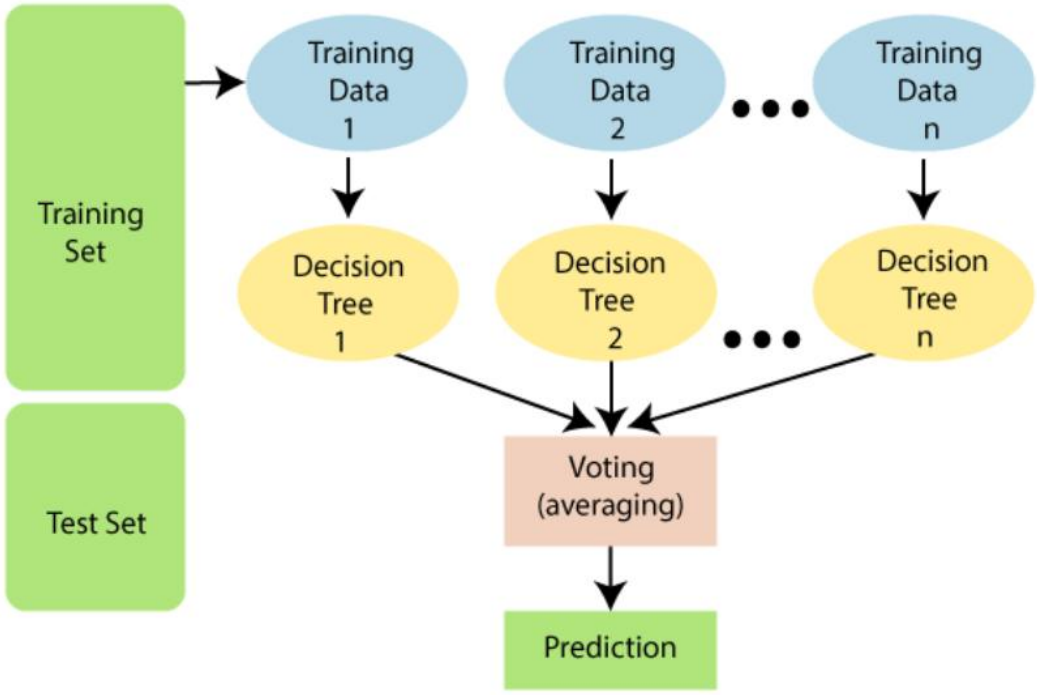

Figure9. Random Forest algorithm [24]

Generally, Random Forest is a very good choice, and it is a highly reliable algorithm. Compared to other algorithms, it takes less time, and the most important thing is that it

can predict the outcome with extremely high accuracy. Moreover, it can preserve the accuracy of the performance when a big part of the data is missing.

Furthermore, Random Forest is a widely used algorithm and it is implemented in different sectors in everyday life. For instance, in the healthcare system, we can identify the risks of diseases, and in the economy we can identify loan risks. Also another application of Random Forest, is in marketing in which we are able to discover trends using this algorithm.

2.7.3 Multilayer Perceptron

Multilayer Perceptron (MLP) is a supervised ML algorithm that makes a function f (): $R^m \rightarrow R^0$ to be trained on a dataset, where m is the number of input's dimensions and 0 is the number of output's dimensions. Given this function a set of features x = x_1, x_2, …, x_m and a target value y, it can learn an approximator for either regression or classification.

The main characteristic that distinguishes it from other ML algorithms is that there can be more than one non-linear-layer between the input and output layers, as can be seen from the figure below [25]:

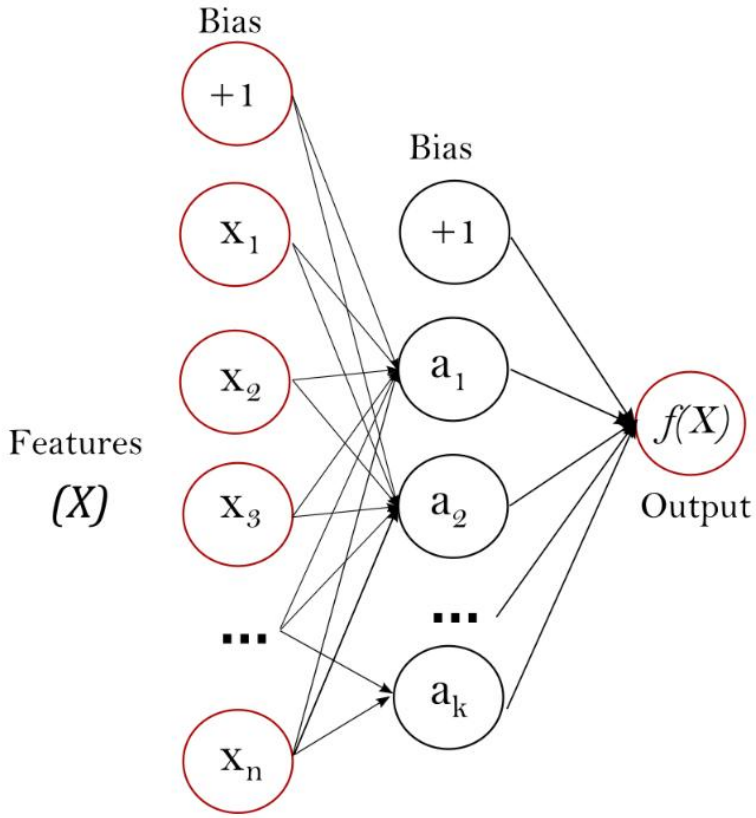

Figure10. MLP with one hidden layer [25]

An MLP consists of more than three layers of nodes whose names are the input layer, hidden layer and output layer consequently. Each node in every layer is connected with a specific weight with all the nodes from the next layer, as can be seen from the above figure. Besides the input node, the others are neurons that use non-linear activation function, which defines the output of that node. The difference between regression and classification is that in the first one, a linear activation function is used for the output layer nodes, while for the second one, a softmax activation function is used. Moreover, backpropagation is used as a technique for training in MLP [26][27].

Each neuron in the hidden layer changes the values from the former layer, and finally, the output layer takes the values from the last layer and converts them into output values. Two very useful advantages of Multilayer Perceptron are the ability to learn non-linear models and the ability to learn real-time models [25].

Furthermore, MLP has many applications in real life. Image recognition, speech recognition and machine translation are some examples.

3 Literature Review

The aim of this section is to present the historical background of Sports Analytics implementations on different previous problems. Similar works and their findings are presented in order to understand the progress until now.

3.1 Historical Background

It is well known that Sports Analytics is used in almost every sport and not only in football. Day by day, more and more clubs and organizations are using different techniques and try to improve their functionality. Into this sector, we present the history of basketball, volleyball and tennis and some works that conducted and had an extremely important role in the evolution of each sport.

3.1.1 Basketball

In 1891, a Canadian teacher of physical education, called James Naismith, attempted to solve a problem that was really hard. To be more specific, during his lessons, students refused to keep up with him. The children were really disruptive, and their behavior was impermissible. In order to solve this unexpected situation, he should invent something ahead of its time. Simultaneously, the need for an indoor game during winter was growing day by day among the students in this college. So, he tried really hard to invent something very interesting that would be played with the ball inside a gymnasium. Finally, he came up with the idea of the sport that it is now known as "basketball" all around the world. Obviously, basketball in Naismith's days wasn't the same as that of today's version. For instance, it was playing with more players, and the rules were different. Then, the game started becoming famous year by year and today is one of the most known sports worldwide [28].

Basketball is professional for many years. National League of Professional Basketball Teams was the first professional basketball competition. This league was invented in 1898, and the teams that participated were offering contracts to their players. The first champions were the Trenton Nationals, with an 18-2 record. Then, it

was a matter of time for basketball to gain extraordinary growth as more and more people became aware of the sport [29].

Nowadays, National Basketball Association (NBA) is the most known basketball league, and it is based in the USA. NBA was formed in 1946, and until now, participation in this league is every basketball player's dream. NBA's clubs buy the best players around the world and that makes the league very competitive and very difficult consequentially. It is admirable the fact that NBA players are the best paid athletes in the world. The league is separated into two conferences (eastern and western) of three divisions, with five teams in each. Totally, there are 30 teams competing each other. Also, NBA includes two "stages", the regular season and play-offs [30]. The number of matches that a team should play depends on how far it will achieve to be in the competition.

Basketball provides a huge amount of data, and that has increased the interest of researchers who seek to investigate the sport further. Analyzing those data can provide incredible insights about the game, the way of playing and the performance of individual players. All this information can be used by coaches, players and managers for better decision-making approaches that will help them in the future. Last years, many works have been conducted on basketball. Following, we present some examples of these works.

Papageorgiou and Tjortjis, in 2022, predicted the daily performance of NBA players who have played at least ten games in the past with the use of regression methods. The final goal of this research was to find the appropriate model which could predict the Fantasy Points for each player participating in a game. Based on these points, they presented the Daily Lineup Optimizer (DLO), which can be used for NBA Fantasy Tournaments. To achieve that, they created a different dataset for every player and each one of them was trained with historical data from season 2010-11 to season 2020-21. They managed to have great accuracy for their models, with an average MAE score of around 7 and an average MAPE score of around 0.3 for all of them. Voting Regressor was the model that was performing better for the majority of basketball players. Finally, they generated a lineup for the 15th May 2021 matchday based on the predicted points, and they achieved a really satisfying score of Fantasy Points (298), which was really close to the actual points on that day [31].

Kirk in 2012 suggested that the most common way to assess a basketball game is missing the spatial analysis, for example, the field goal percentage, known as FG%. Also, during his work, he tried to analyze the way a team plays and the way the team uses the court in order to score. To do that, he separated the field into regions and with the use of visual analytics, he ended up with the following contour, which indicates the point's tendency and the points per attempt [32].

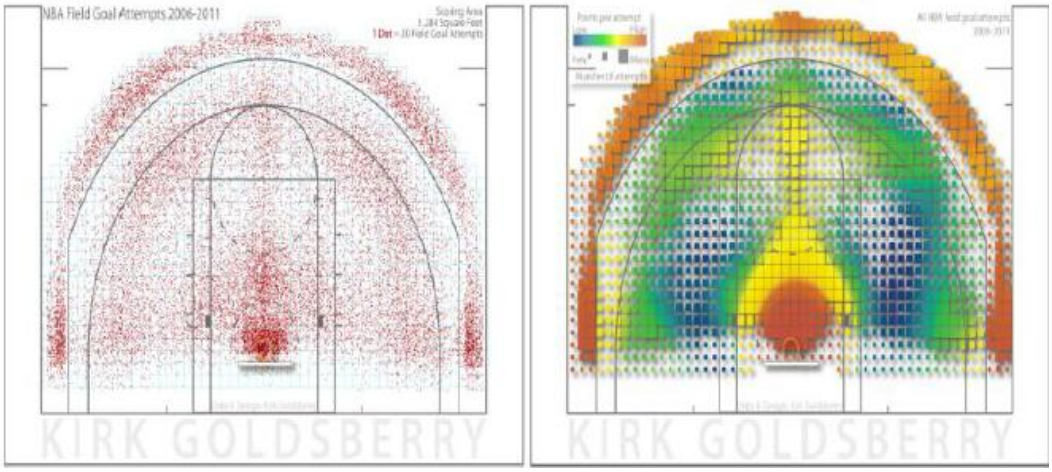

Figure11. Field goal attempts (left), shots and points per attempt (right) [32]

Moreover, one of the thousands of works that have been conducted into the field of basketball is from Shah and Romijnders in 2016. The two researchers applied recurrent neural networks in order to predict if a three-point shot will be successful or not. Their efforts were based on tracking data collected from a tracking system known as SportVu, installed by the NBA organization in every basketball field. They used their models based on the ball's distance from the basket. Totally they used seven different distances, from 2 feet to 8 feet, and the results were impressive as they achieved high accuracy in their predictions [33].

In 2020, Sarlis and Tjortjis, tried to predict the MVP of the NBA seasons from 2017-18 to 2019-20 and the best defender of the same seasons. To do this, they collected valid data from these seasons. For this purpose, they created two different formulas, the Aggregated Performance Indicator (API) and the Defensive Performance Indicator (DPI). The first one was used to predict the MVP while the second one was used to predict the best defender. Both of them were consisted of advanced statistics with regard to players' performances. Finally, the two formulas were able to predict successfully the MVP and the best defender in each case. Moreover, it is really

impressive the fact that the first method is the only one that can predict the MVP with current data [34].

3.1.2 Volleyball

Volleyball is also one of the most famous sports around the world. It is a team sport in which two teams compete with each other. Every team consists of six players, and they play on a court that separates the teams with a net. The goal is to score points by grounding the ball to the opposite team's court.

Obviously, many works have been conducted on the field due to the majority of statistics that it provides during a game. Attributes like blocks, speed of service, jumps and topspin service have attracted scientists' interest in data analysis. Through this process, researchers will be able to understand the characteristics that play the most significant role in the game in order to help coaching staff for better decision making.

In 2015, Chellatamilan et al followed clustering approaches in order to predict the winner of a volleyball game and they achieved high accuracy. To be more specific, Hierarchical clustering, K-Means clustering and Fuzzy C-Means clustering were used for this purpose. The process that the researchers followed can be seen in the table2 [35].

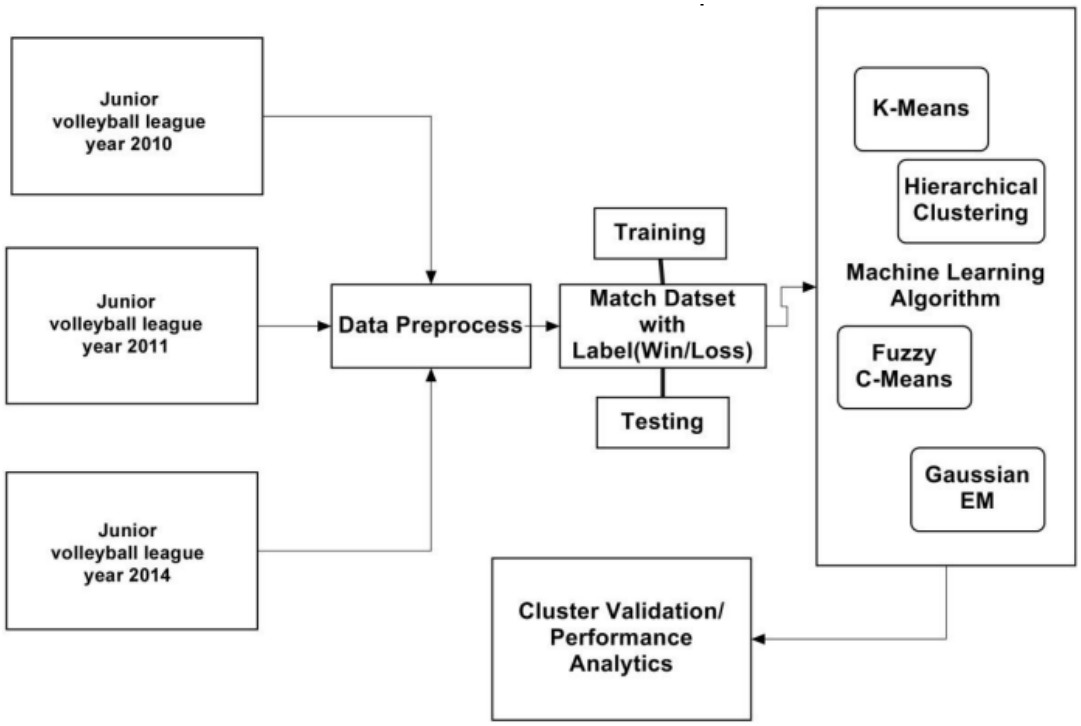

Figure12. Clustering approach for predicting the winner [35]

Moreover, Van Haaren et al, in 2016, attempted to identify team's patterns when a team is attacking using optical tracking data. In other words, they wanted to create an approach which would be able to evaluate an attack as successful or not. To achieve it, they analyzed men's and women's final matches of the 2014 World Championship. During their work, they understood that there were some specific tactics that were followed by the teams and that they ended up with a successful attempt as can be seen from the following table [36].

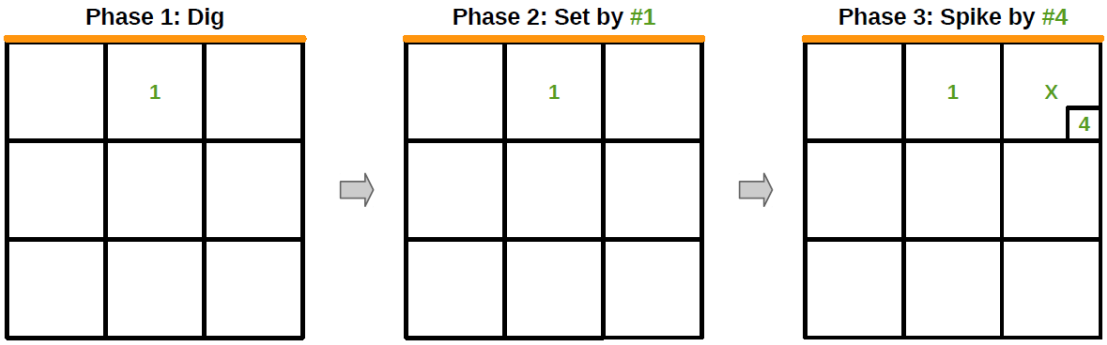

Figure13. Successful pattern from Brazil in men's final [36]

To be more specific, this tactic was used many times by Brazil's national team during the final, and it was very successful. It is obvious that the player with the number 1, known as the setter, sets the ball from the center to the back right corner and then the player with the number 4 hits it direct to the opposite court. X is the location of the ball.

De Leeuw et al, in 2021, conducted an out of the ordinary research with regards to volleyball. In their research, they implemented a ML approach for identifying indicators of training load, which can be used for injury predictions. They collected data for 14 professional volleyball players who fulfilled some conditions regarding their physique (weight, age, height). Then, they applied a supervised ML technique called Subgroup Discovery. The goal of Subgroup Discovery is the detection of subgroups in which the distribution of the target variable is completely different with the distribution of the whole data. Finally, they ended up with the conclusion that the most important step for preventing injuries is the tracking of jump load [37].

Moreover, Akarcesme, in 2021, applied the ML algorithm Logistic Regression in order to predict the results of volleyball matches in the Turkish Men's and Women's

Volleyball League. During his research, he tried to identify which variables play the most important role in this model. The variables were separated into dependent and independent variables. The dependent variable was the match result, while the independent ones were the efficiency of players' positions, like middle blocker efficiency and libero efficiency. The results of the model were pretty good for both men's and women's volleyball leagues. The accuracy for the men's matches was more than 78%, while for women's matches was around 83% which means that the result of a match can be predicted with high accuracy [38].

Finally, in a slightly different research, Tumer and Kocer, in 2017, tried to predict a volleyball team league's rankings using an artificial neural network (ANN). The final model that they used had an accuracy of around 98%, with an MAE of 0.005, which means that they were able to make very good predictions using the previous three seasons' data between 2013 and 2015. Generally, they tested 12 different models of 4, 6, 8 and 10 neurons in logsig, purelin and tansig functions in order to obtain the desired results. Finally, the four-neuron log-sigmoid transfer function had the best performance for the model [39].

3.1.3 Tennis

Tennis is a very famous sport with millions of fans around the world. It is playing with two single players or a pair of players, and the goal is to hit the ball to the opposite side of the court, over the net. All players that participate in this game should hold a racket. Points are collected from a player or a team when the opponent fails to return the ball. It was played for the first time in Birmingham in the 19th century with the name of "lawn tennis" because it was played on grass fields by Victorian gentlemen and ladies. Its origins are dated between 12th and 13th centuries. More specifically, there was a French game known as "jeu de paume", which had very common characteristics with today's tennis. In 1877, it was organized the first Wimbledon Championship, and the winner was Spenser Gore. This tournament was a huge milestone in tennis history because it became the most famous tennis tournament around the world, with millions of fans watching this tournament nowadays [40].

Tennis has increased the interest of scientists because of the great amount of data it provides during every match. Many works have been conducted on this field and the

findings are quite interesting and can be used by people interested in this sport to improve it. Obviously, most of these works focus on estimating the final score or predicting the winner of the game. There are many reasons that lead scientists in this direction. Gambling is one of them because tennis is one of the most famous sports for betting around the world. Tennis' value in the market is increasing day by day. Today, its value is around 59 billion dollars, while it is expected that in 2023 it will be around 92 billion dollars [41].

In 2017, Cornman et al. tried to predict the outcome of tennis games using historical data. They wanted to beat the betting odds. Several models were used for this task, like SVM, Neural Networks, Random Forests and Logistic Regression. Finally, they achieved a prediction score of around 70% for the season 2016-17 [42].

In 2021, similar research was conducted by Gao and Kowalczyk, who were able to predict match outcomes using three different ML models. To be more specific, Logistic Regression, Random Forest Classifier and Support Vector Machine were these models. Generally, the Random Forest model was the model with the highest accuracy, almost 84%, and the model with the most sensitivity to the features that were selected. The spectacular outcome of this research is that this accuracy exceeds the accuracy based on betting odds. Moreover, this research underscores the importance of the serving strength, especially the first serve, as the number one characteristic that plays a crucial role in the match outcome prediction [43].

3.2 Related Work

3.2.1 Introduction

In this part of the dissertation, we present former works with regards to football. More specifically, we will pay attention to relative researche that conducted the previous years and we will present their procedure. So, at the end of this section, we will be able to understand the evolution of Sports Analytics in football.

3.2.2 Performance Prediction in Football

Researchers are more likely to conduct their research on football than the previous years due to the increasing data on this field. This huge amount of data gives the opportunity to provide better decision-making capabilities to the coaching staff [44]. Decisions with regards to football have to be more reliable and the huge amount of data which are available helps in this direction. We live in an era which is experiencing a great number of available sources and previous research on this field, so we can use all this information to expand our knowledge of football.

In 2018, Pariath et al. during their research, they created a Regression model in order to predict the overall performance of a player and a Linear Regression model to predict the market value of a player based on the first model's results. This kind of information would be very useful for coaching and scouting purposes. They collected their data from a football game called FIFA, which has reliable and accurate statistics for thousands of football players around the world. The first model, which has to do with overall performance, achieved an accuracy of around 84% with an RMSE of 2.67 and an MAE of 2.01. On the other hand, the second one, which has to do with market value, was a Linear Regression model and it was more accurate with accuracy around 91% and an RMSE of 0.410 and an MAE of 0.304. The second model's performance was better because it was based on the overall performance value predicted from the first model. Also, they created different models for every player's position, but the results were very similar for all positions. Lastly, they ended up that the best indicator of performance is the overall value of the player [45].

Next, in 2018, Apostolou and Tjortjis conducted research with 3 different experiments. First of all, they tried to predict a player's position on the field based on his characteristics. Throughout this experiment, they achieved an incredible accuracy of around 81.5% with Random Forest and Sequential Minimal Optimization (SMO) [46].

```
=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 24  0  9  0 |  a = FOR
  0 17  0  0 |  b = GK
  6  0 32  3 |  c = MID
  0  0  4 24 |  d = DEF
```

Figure14. Confusion Matrix [46]

It is obvious that most of the players are classified correctly, especially goalkeepers, due to their distinguished attributes. The second experiment had to do with the number of goals per season of two very famous football players, Leonel Messi and Luis Suarez. Random Forest, Logistic Regression, MLP Classifier and Linear SVC were the algorithms that were used for this experiment. As can be seen from the tables below, Random Forest was the algorithm with the best results because it was very close to the actual goals of the two football players (in season 2017-18, Messi achieved 34 goals while Suarez 25) [46].

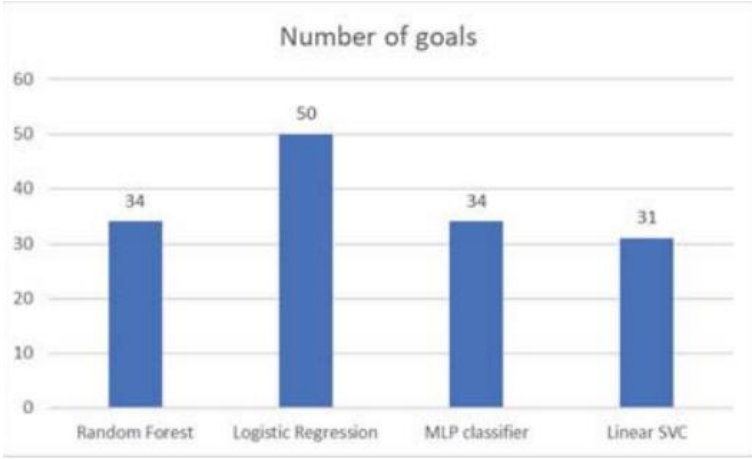

Figure15. Messi's number of goals [46]

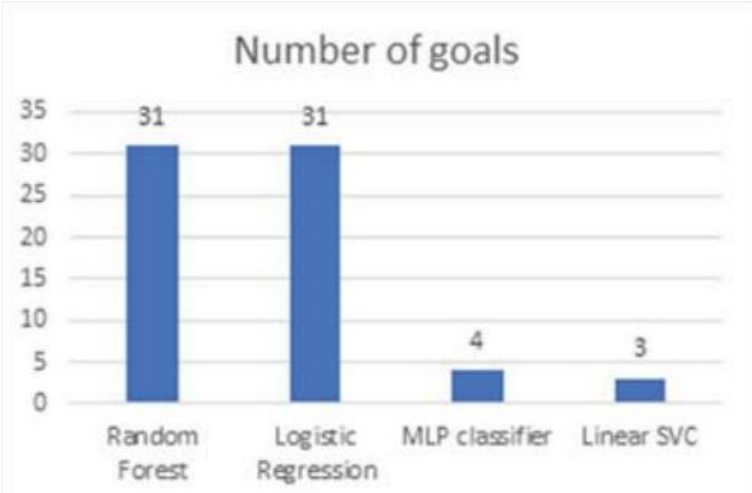

Figure16. Suarez's number of goals [46]

The last experiment was focused on the prediction of the number of shots of a player during a match. Again, Random Forest is the algorithm that provides the highest

accuracy. On a specific match, the algorithm predicted Messi's shots to be 2.133 while actually, they were 2 [46].

Shrinivasan, in 2017, in his research, predicted the number of passes among players in a football game. The impressive characteristic of his work is the fact that he achieved a very small Means Squared Error, around 6.88, which wasn't achieved before. He based his study on networks. To be more specific, he used the passing distribution of the teams in order to create networks which are called passing networks of a team. These networks consist of nodes and edges which represent players and the passes among them correspondingly. Passing distribution records how successfully the players pass the ball between them before a goal. SVM, Random Forest Regression and Gradient Boosting Regressor were the algorithms that he used to end up to predict the number of passes between two players in different games. Simply, the models try to predict new passing networks [47].

In 2017, Ruiz, H., Power, P., Wei, X. & Lucey, P. conducted a research based on the Premier League 2015-16 champions Leicester City. To be more specific, they analyzed this season and based on this analysis they stated how the predicted features could be used in order to predict feature performance. Their goal was to understand what made Leicester City champions from being an average mid-table team. They used different kinds of tools, such as expected goals, plots of strategy and plots of passing quality, and they compared their results with the results of other Premier League teams this season. In the end, they concluded that defensive effectiveness was the major factor that made Leicester win the title. Thinking that their goalkeeper Peter Schmeichel was the second most efficient goalkeeper in the league, it's obvious that Leicester's defense was really crucial [44].

Further research in this field has been done by Manish et al in 2021. Into their research, they predicted the statistics of football players based on their position in previous seasons (forwards, midfielders, defenders and goalkeepers) through different algorithms. They mention that the position of a player has a high dependence on his performance. Various ML models were used by the authors to make predictions. Finally, they concluded that the Multiple Regression algorithm is the most efficient algorithm for predicting a player's performance due to its high accuracy. In more detail, Multiple Regression algorithm has 0.88 MAE, 1.16 MSE and 1.077 RMSE [48].

During a football game, many events happen on the field that affects the outcome of the match. In 2020 Tiwary Ekansh, Sardar Pransajit and Jain Sarika considered the whole environment of a football game and predicted the result of the game with the use of Recurrent Neural Networks (RNNs) and Long Short Term Memory (LSTM). During this study, the researchers considered the current form of the teams that were playing in a football game, i.e. how many matches in a row they have won. Also, they used data from English Premier League from season 2010-11 to season 2017-18 for their experiments. With the appropriate feature extraction, they created features that were very useful to continue. Finally, they ended up that these methods are better than Artificial Neural Networks (ANN) and traditional ML algorithms [49].

In 2019, Pantzalis and Tjortjis conducted two experiments. The first one was team performance prediction and the second one was player performance prediction. For the first experiment, the goal was to predict how a football club would perform during season 2018-19 based on historical data. To do that, two approaches were followed. For the first one, they separated teams in those that their final position would be better than last seasons' and those whose final position would be worse than their previous one. The final accuracy for this approach was around 70%. During the second one, the main idea was to simulate every football match and classify them as home/away win and draw. After that, each club's points were gathered together, and the final league tables were produced. In that case, the accuracy was better and more promising than the first approach. For the second experiment, they tried to figure out which attributes and movements during a game can influence a defender's rating. For this purpose they used Multiple Linear Regression and the results were remarkable as R-squared was 0.9 and adjusted R-squared was 0.88. Finally, they concluded that 30 features were influencing defenders' performance, like interceptions and clearances [50].

Nowadays, it is very common for football clubs to use wearable devices for their players during training session or football game. There are many works with these devices because they help data analysts to collect plenty of data with regards to football player's physique. Besides that, many professional firms conduct statistical analysis and provide their findings to football clubs to extract information from this analysis and monitor their players [51]**.** Wearable devices are very important in decision-making in football. For instance, knowing that a player is exhausted, the coach can decide to let him rest for a couple of days instead of making him play a tough game.

In 2019, Frey et al conducted research into this new field of wearable devices. In more detail, the authors predicted the player's position on the football field with the use of three different ML algorithms (Random Forest, Gradient Boosting and Convolutional Neural Network). The whole data for this research was collected through wearable devices, which measure two important things, player movement (IMU) and player position (GPS). Finally, they achieved good performance for all the three algorithms (around 80% for all of them). Also, they found out that there was a small misclassification in positions that the detection of them are difficult anyway because of the common characteristics that players share in similar positions. For instance, there is a huge similarity between the movements of a winger and a midfielder [52].

Also, Cintia et al. in 2015, made a team performance prediction during a game based on data that was collected from wearable devices. To achieve their goal, the researchers used a player-passing network and a zone-passing network. Working on these network indicators, they were able to describe the performance of a team [51].

Moreover, Al-Asadi and Tasdemir in 2022 estimated players' market value through four ML algorithms. To be more specific, they used Random Forests, Linear Regression, Multiple Linear Regression and Decision Trees. Random Forest was the algorithm which provided the best results for predicting the market value of a player, and it has the lowest error ratio correspondingly with an RMSE of 1.64 and an R^2 of 0.95. In their research, they tried to analyze the collected data in order to understand and identify the most significant features that determine this value. During their analysis, they found out that a player's value on the market is not related only with his skills or his performance generally. In contrast, there are many factors that affect this value like his potential and his international reputation [53].

Similarly, Stanojevic and Gyarmati in 2016 created a simple model for evaluating the market value of football players. They achieved a small improvement in their estimates than the estimates from the specific website famous on its market value specification, called transfer market. Also, they achieved an RMSE of around 0.200 [54].

In a similar research in 2016, Fernandez et al collected data from electronic performance and tracking systems (EPTS). This technology includes microsensor technology and a GPS system that collects mainly physical information from matches and training. Their research indicates that it is possible to use many ML algorithms and predict the physical attributes from the collected data [55].

4 Methodology

This chapter analyzes the process followed in order to predict how many goals a player will score. More specifically, this part of the dissertation will present the whole procedure. Starting with data collection via scraping, how data are cleaned and continuing with feature engineering which was the final part before the ML algorithms implementation. Finally, we will present and compare the evaluation of the results with the use of three different metrics, MAE, MSE and RMSE.

4.1 Process Description

The main purpose of this dissertation is player performance prediction, as already mentioned. To be more specific, the aim of this dissertation is to predict the number of goals that a player will achieve based on his previous seasons. Different machine learning algorithms have been implemented. Linear Regression, Random Forest and Multilayer Perceptron were used.

To begin with, the appropriate dataset should be created. To do this, the data should be scraped from a valid source in order to have results as accurate as possible. For this reason, the data was scraped from Sports Reference, a well-known American organization which is extremely highly active on the field of sports data [1]. On this organization's website, someone can find a huge quantity of statistics about players from many sports. Our goal was to extract information for football players from the last 5 seasons, from season 2017-18 to season 2021-22. So, we ended up with a final database with more than 2,500 football players and more than 30 characteristics about them.

During our experiments, the algorithms weren't trained with all the dataset. Instead, they were trained with the data of the seasons from 2017-18 to 2020-21. So, the season 2021-22 was used only for testing and validation, i.e. to see how good the results were. For this purpose, we should concentrate all statistics of seasons 2017-18, 2018-19, 2019-20 and 2020-21 together into one csv file.

The next step was the pre-processing of those data. Duplicates, null values and noise should be removed from the dataset in order to be "clean". Moreover, some

characteristics of those players weren't important to the whole process, so they were removed from the dataset. Also, some others that appeared as strings in python were converted into floats with the appropriate commands.

The dataset that was used to train the algorithms was consisted of football players from all positions, nationalities, football clubs, values and even status. The only restriction was the league that they participate. More specifically, we chose football players from the five best leagues in the world. These leagues are Premier League (England), La Liga (Spain), Bundesliga (Germany), Ligue 1 (France) and Serie A (Italy). The reason was practical. The statistics about players from different leagues are less. In any case, it is more interesting when we analyze and make conclusions about famous football players.

Then, having a dataset without null values, duplicates, missing values and generally outliers, it was time to search for regression algorithms in order to make the appropriate predictions about the performance of a player. For the whole process, python and, more specifically, PyCharm was used. The following diagram summarizes the process:

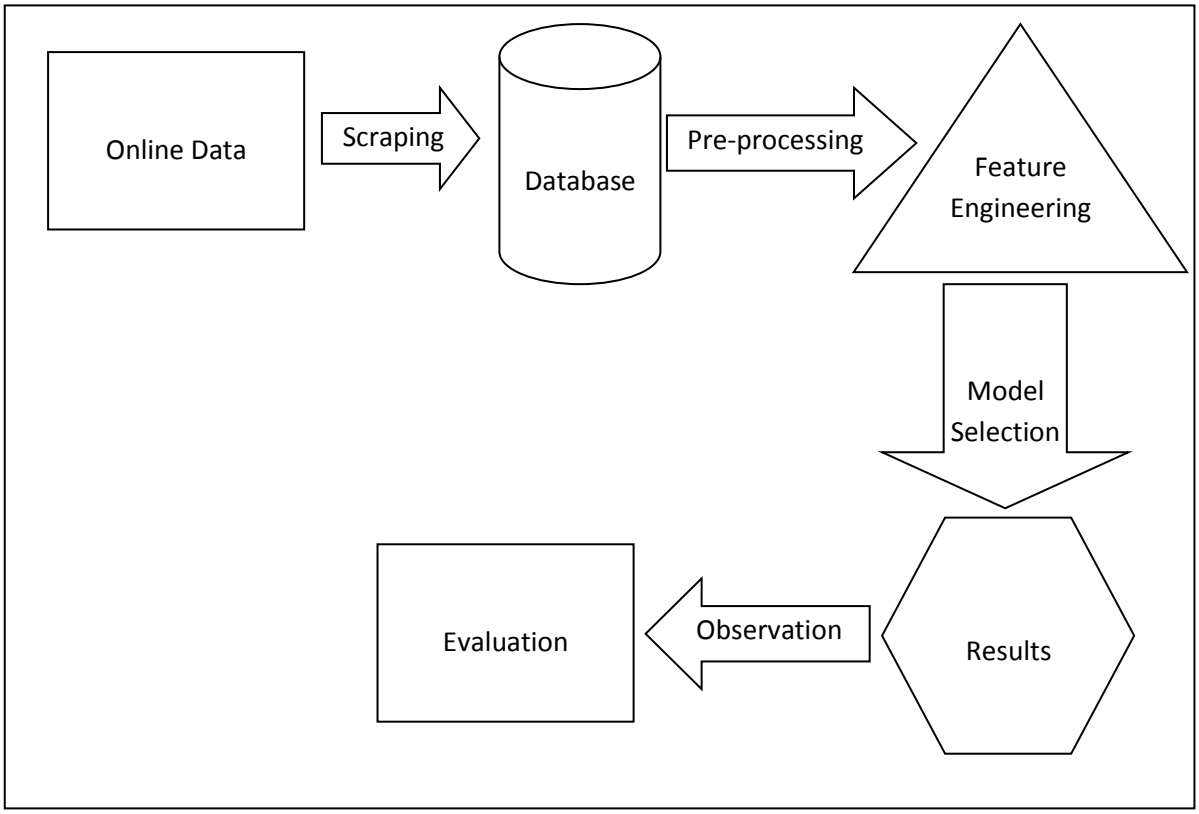

Table1. Process followed

Furthermore, it was necessary to compare the results between the models. This would give us priceless information about the reliability of the models that were used for the predictions. So, the comparison between the methods that were followed was inevitable. This is the only way to find out which model provides the best results. Moreover, the results of the algorithms were assessed in terms of MAE, RMSE and MSE.

Mean Absolute Error (MAE) represents the mean of the absolute difference between the expected and actual value. It is estimated as the sum of absolute errors divided by the size of the data i.e. the number of predictions. Generally, it is very common metric of evaluation in regression models. Its formula is described below [56]:

$$\text{MAE} = \frac{\sum |\, actual\ value - expected\ value \,|}{n}$$

Root Mean Square Error (RMSE) is a very common measure of the differences between predicted and actual values and it is the square root of the mean of the square of the errors. Also, it is described by the formula below [57]:

$$\text{RMSE} = \sqrt{\frac{\sum |\, actual\ value - expected\ value \,|^2}{n}}$$

Mean Square Error (MSE) measures the average of the squares of the errors which is the average squared difference among expected values and actual values. MSE is described by the formula below [58]:

$$\text{MSE} = \frac{\sum |\, actual\ value - expected\ value \,|^2}{n}$$

Generally, these metrics are the most common metrics in regression problems, and they are very helpful throughout the evaluation of the models. MAE, MSE and RMSE should be as low as possible to have a good performance.

4.2 Data Collection

The most important part of the dissertation is the data collection. There are many websites that provide plenty of football statistics from clubs and players. So, we should pay high attention to the validity of the data that we would use because the results wouldn't be accurate otherwise.

The data was collected from Sports Reference. Sports Reference is a big organization that collects huge amounts of data on a daily basis for many sports like football, basketball, baseball and hockey. Different kind of information like scores, statistics and historical background for almost 100,000 players from over 100 competitions are provided and updated day by day.

In more detail, the data that was scraped from this website was data from football players for the seasons 2017-18, 2018-19, 2019-20, 2020-21 and 2021-22. In every season, thousands of players were included. Finally, we ended up with five different csv files. In every one of them there are data for more than 2,500 football players. Having 5 different csv files it was wise to concentrate them into one csv file in order to have the final database. The features of the dataset are presented below:

Player	Name of the player
Nation	Nationality of the player
Pos	Position most commonly played by the player
Squad	Club the player is currently playing
Comp	Competition
Age	Age of the player
Born	Year of the player's birth
MP	Games played by the player
Starts	Games started by the player
Min	Minutes played by the player
90s	Minutes played divided by 90
Gls	Goals scored

Ast	Assists
G-PK	Non-penalty goals
PK	Penalty kicks made
PKatt	Penalty kicks attempted
CrdY	Yellow cards
CrdR	Red cards
Gls per 90	Goals scored per 90 minutes
Ast per 90	Assists per 90 minutes
G+A per 90	Goals and Assists per 90 minutes
G-PK per 90	Goals minus Penalty Kicks made per 90 minutes
G+A-PK per 90	Goals plus Assists minus Penalty Kicks made per 90 minutes
xG	Expected Goals
npxG	Non-Penalty Expected Goals
xAG	Expected Assisted Goals
npxG+xAG	Non-Penalty Expected Goals plus Expected Assisted Goals
xG per 90	Expected Goals per 90 minutes
xAG per 90	Expected Assisted Goals per 90 minutes
xG+xAG per 90	Expected Goals plus Expected Assisted Goals per 90 minutes
npxG per 90	Non-Penalty Expected Goals per 90 minutes
npxG+xAG per 90	Non-Penalty Expected Goals plus Expected Assisted Goals per 90 minutes

Table2. Dataset's Advanced and Basic Metrics

As can be seen from the above table, the majority of these features are considered advanced features because there isn't only the basic information like goals, assists and minutes played for every player. Instead, there are features that combine information and that give us a detailed description of every player's performance throughout the

previous seasons. This is exactly what makes our research more interesting. For instance, the majority of them provide us thorough reports even about the statistics per 90 minutes, i.e. per game.

Moreover, the shape of the dataset looks like this (some of the features):

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Rk	Player	Nation	Pos	Squad	Comp	Age	Born	MP	Starts	Min	90s	Gls	Ast	G-PK	PK	PKatt	CrdY	CrdR
2	1	Max Aarons	engENG	DF	Norwich C	engPremi	21	2000	34	32	2,881	32.0	0	2	0	0	0	8	0
3	2	Yunis Abdelhamid	maMAR	DF	Reims	frLigue 1	33	1987	34	34	2,983	33.1	2	0	2	0	0	5	1
4	3	Salis Abdul Samed	ghGHA	MF	Clermont	frLigue 1	21	2000	31	29	2,462	27.4	1	0	1	0	0	12	3
5	4	Laurent Abergel	frFRA	MF	Lorient	frLigue 1	28	1993	34	34	2,956	32.8	0	2	0	0	0	9	0
6	5	Charles Abi	frFRA	FW	Saint-Ã‰t	frLigue 1	21	2000	1	1	45	0.5	0	0	0	0	0	0	0
7	6	Dickson Abiama	ngNGA	FW	Greuther	deBundes	22	1998	24	5	726	8.1	0	1	0	0	0	3	0
8	7	Matthis Abline	frFRA	FW	Rennes	frLigue 1	18	2003	7	1	103	1.1	0	0	0	0	0	0	0
9	8	Tammy Abraham	engENG	FW	Roma	itSerie A	23	1997	37	36	3,084	34.3	17	4	14	3	3	9	0
10	9	Luis Abram	pePER	DF	Granada	esLa Liga	25	1996	8	6	560	6.2	0	0	0	0	0	2	0
11	10	Francesco Acerbi	itITA	DF	Lazio	itSerie A	33	1988	30	29	2,536	28.2	4	0	4	0	0	2	1
12	11	Ragnar Ache	deGER	MF,FW	Eint Frank	deBundes	23	1998	13	1	259	2.9	0	1	0	0	0	0	0
13	12	Mohamed Achi	frFRA	MF	Nantes	frLigue 1	19	2002	1	0	10	0.1	0	0	0	0	0	0	0
14	13	Marcos AcuÍ±a	arARG	DF	Sevilla	esLa Liga	29	1991	31	26	2,26	25.1	1	3	1	0	0	10	0
15	14	Che Adams	sctSCO	FW	Southamp	engPremi	25	1996	30	23	2,039	22.7	7	3	7	0	0	0	0
16	15	Tyler Adams	usUSA	MF	RB Leipzig	deBundes	22	1999	24	12	1,351	15.0	0	1	0	0	0	5	0
17	16	Sargis Adamyan	amARM	FW,MF	Hoffenhei	deBundes	28	1993	13	2	331	3.7	1	0	1	0	0	0	0
18	17	Martin Adeline	frFRA	MF,FW	Reims	frLigue 1	17	2003	8	2	352	3.9	0	0	0	0	0	1	0
19	18	Amine Adli	frFRA	FW,MF	Leverkuse	deBundes	21	2000	25	13	1,256	14.0	3	2	3	0	0	3	0
20	19	Yacine Adli	frFRA	MF,FW	Bordeaux	frLigue 1	21	2000	36	25	2,26	25.1	1	7	1	0	0	5	0

Table3. The final database

Obviously, this isn't the final shape of the dataset that we used for the algorithms' implementation, but the initial one. Many changes took place to convert this dataset to the appropriate form. These changes include pre-processing and feature engineering and will be presented in the following chapters.

4.3 Pre-processing

The first thing to do in the pre-processing stage was data cleaning. For this reason, we should pay high attention on the form of the data that we collect. To be more specific, there shouldn't be a different kind of "noise" on the dataset that was used for the final predictions. The term "noise" refers to duplicates, null values, missing values or generally outliers. There were some values that appeared as NAN, which means "Not A Number" and it refers to missing values. Besides that, the dataset that was scraped didn't have any of the above characteristics at all, so it wasn't necessary to do more actions with regard to this. That's why we were extremely focused on the website from that we would scrape data. It was vital to scrape as much clean data as possible.

Next, we should do the appropriate changes on the features of the dataset. To do that, we removed some of them that were strings and could not be used to our experiments. For our experiments, we wanted to use information like the number of goals and assists

a player achieved or the number of yellow and red cards he received during the previous seasons and not statistics like his nationality or the league that he participates.

For this purpose, the next step in the pre-processing phase was to convert some data to the appropriate form. Some of the features which were integers appeared like strings and this was a huge problem because we couldn't implement the ML techniques. So, we converted these features with the proper commands from strings to float numbers.

Last but not least, an essential detail was the fact that in our dataset, we should include players that have played for all of the seasons. So, we had to remove players that started their career after season 2017-18 and they exist to our dataset, or players that finished their career before season 2021-22 and they exist to the dataset too.

All these players were removed, as mentioned before, because they had statistics for less than 5 seasons into our data. Therefore, the final dataset consists of more than 800 players, which is less than the initial number of 2.500 players. Obviously, it would be a huge problem if someone had statistics for less than five seasons because the algorithm would not be trained with many seasons for specific players and their predictions would be inaccurate. After that, the dataset was clean and ready for the next phase, which was feature engineering.

4.4 Feature Engineering

Having our final dataset, we had to deal with an important problem. The goal was to train the algorithms with 4 out of 5 seasons (from 2017-18 to 2020-21) of statistics and then test them with the last season (2021-22). The predictions will have to do with the last season (2021-22). Nevertheless, by training the algorithms with the current form of their statistics, the results would not be fair and valid.

To be more specific, if we want to predict the number of goals that a player is going to achieve in season 2021-22, it isn't right to train the algorithm with statistics from this season, like assists, cards etc. It is desirable to convert the dataset into a form that each row contains past statistics. In other words, we had to convert the data to historical data.

Our goal is to transfer the goals from a season to the previous one. Now, the algorithm will learn for example how many goals a player achieved in season 2018-19 based on his performance in season 2017-18, and so on. That's the correct way to train

the algorithm because our purpose is to find out how many goals a player will achieve in season 2021-22 based on his performance in season 2020-21. In other words, we create season-lag features. Generally, lag-features are values at former time step which are considered helpful because they generated with the assumption that what happened in the past can affect or contain useful information for the future [59].

For this purpose, we created four different datasets from the initial one.

1. 2018-2021_3 → Includes statistics for the seasons 2017-18, 2018-19 and 2019-20
2. 2018-2021_Goals_3 → Includes statistics for seasons 2018-19, 2019-20 and 2020-21
3. 2021-2022_3 → Includes statistics from season 2020-21
4. 2021-2022_Goals_3 → Includes statistics from season 2021-22

So, when we wanted to implement our algorithms, we invoked the features that we wanted. To be more specific, for the training, we invoked all the features from dataset "2018-2021_3" except from the goals, because the goals were invoked from the dataset "2018-2021_Goals_3". During the testing part, dataset "2021-2022_3" was invoked respectively. Moreover, the split of the dataset that was followed was 80% of data for training while 20% of data for testing.

5 Results

In this chapter, the algorithms' results will be presented. After that, we will be able to compare them and decide which one performs better. For this purpose, an example with two football players will be presented too.

5.1 Linear Regression Results

By implementing Linear Regression, we are able to get the following results. It is obvious that all metrics have very good results because their values are very close to zero.

Metrics	Linear Regression
MAE	1.60
MSE	7.06
RMSE	2.66

Table4. Linear Regression Evaluation

5.2 Random Forest Results

By implementing Random Forest we are able to get the following results. As can be seen from the table below, the results are decent because all of them are very close to zero too.

Metrics	Random Forest
MAE	1.69
MSE	7.75
RMSE	2.78

Table5. Random Forest Evaluation

5.3 Multilayer Perceptron Results

By implementing Multilayer Perceptron we are able to get the following results. As can be seen from the table below, MLP has good results considering that all of them are really close to zero as the previous two models.

Metrics	Multilayer Perceptron
MAE	1.63
MSE	7.68
RMSE	2.77

Table6. Multilayer Perceptron Evaluation

It is obvious that Linear Regression performs better than the other two algorithms. All its metrics are small and very close to zero. In order to understand their difference better, we will present an indicative example of two football players and their predictions. For this reason, we chose to illustrate the performance prediction of David Silva and Danny Welbeck.

The selection of them isn't random. They both play as midfielders and strikers. If we chose to illustrate center backs, i.e. defenders, for example, the predictions would be very accurate, because a defender usually scores zero goals during a season. So, it isn't challenging for the algorithm to predict that a defender will score zero goals for instance. For the example below, it is important to mention that their actual goals for season 2021-22 was 2 for David Silva and 6 for Danny Welbeck.

Moreover, we chose these players to underline the differences between the models. Obviously, there are many examples of players whose actual goals were predicted accurately by all models. Also, there are many players whose goals' predictions were far away from the actual ones.

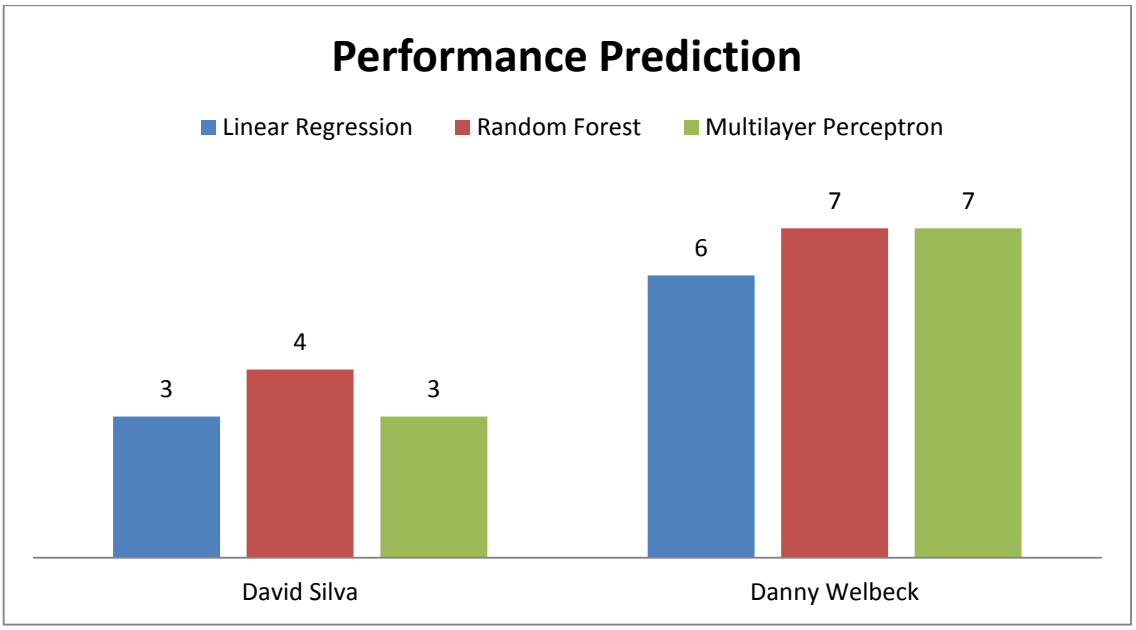

Table7. Example of Players' Performance Prediction

The results of the above table confirm the initial assumptions that Linear Regression performs better to our problem. In the first case, prediction for David Silva, Linear Regression was very close to the actual goals (1 goal difference) and in the second case, prediction for Danny Welbeck, it was accurate. The next better model was MLP because the first prediction for David Silva was 3 (1 goal difference) and the second one for Danny Welbeck was 7 (1 goal difference). Finally, Random Forest results were good but worst than the other two models. In more detail, the prediction for David Silva was 4 (2 goals difference) and the prediction for Danny Welbeck was 7 (1 goal difference).

If we sum up the differences between actual and predicted goals for every model, we will conclude that Linear Regression's performance was the best with 1 goal difference in total. Then MLP follows with 2 goals difference in total and last is Random Forest with 3 goals difference in total.

6 Discussion

As mentioned before, one of the most important parts of this dissertation is the comparison between the algorithms. We evaluated them with the use of three metrics, and now we compare these evaluations and see which algorithm performs better. That's the only way to understand their performance and how well we deliver the problem.

Therefore, we will compare the results of every metric of each algorithm. The results are gathered and represented in the below tables:

Algorithms	MAE
Linear Regression	1.60
Random Forest	1.69
Multilayer Perceptron	1.63

Table8. MAE comparison

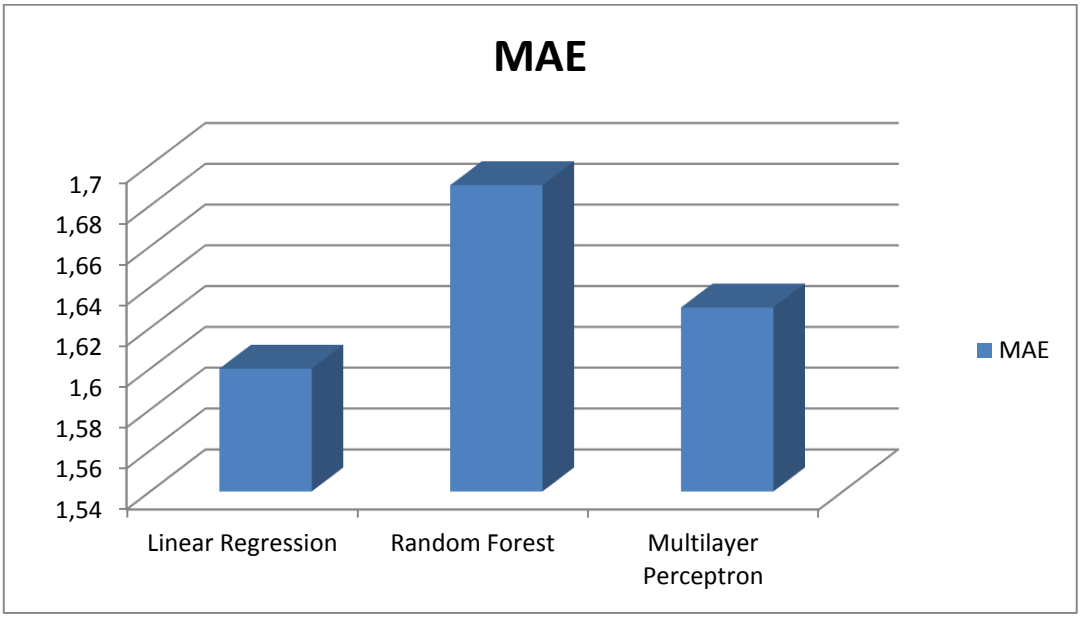

Figure17. MAE Bar Chart

Algorithms	MSE
Linear Regression	7.06
Random Forest	7.75
Multilayer Perceptron	7.68

Table9. MSE comparison

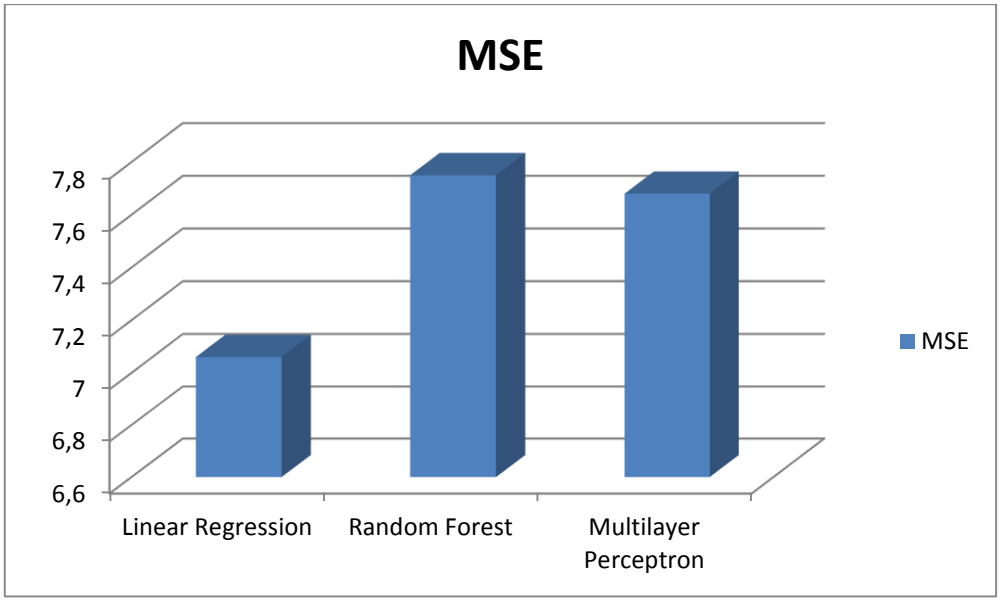

Figure18. MSE Bar Chart

Algorithms	RMSE
Linear Regression	2.66
Random Forest	2.78
Multilayer Perceptron	2.77

Table10. RMSE comparison

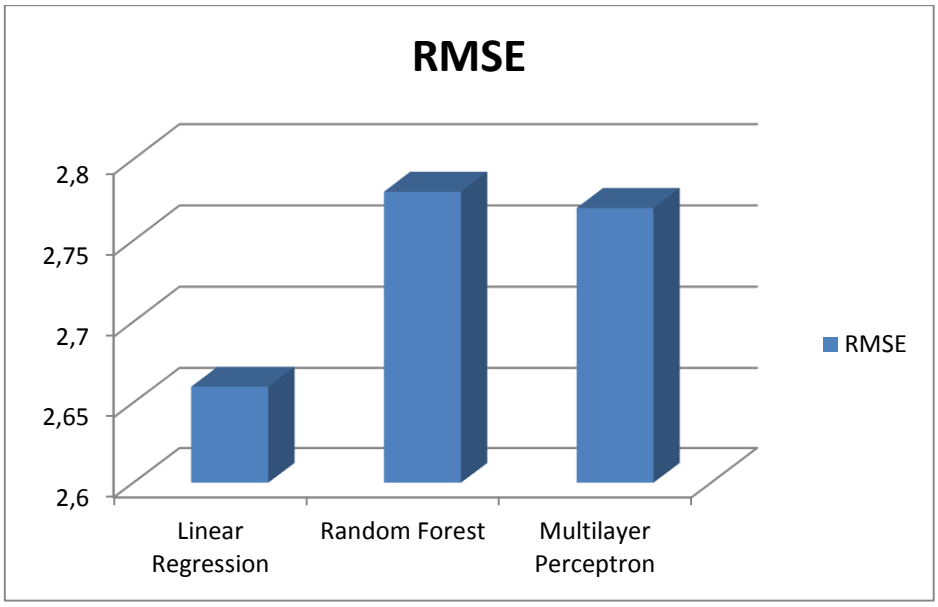

Figure19. RMSE Bar Chart

As can be easily observed from the above tables and representations, it is obvious that Linear Regression performs better. More specifically, Linear Regression has the lowest MAE, MSE and RMSE of the other two algorithms.

MAE is the most important metric because it directly depicts how close the predictions are to the real values. We observe that all the algorithms have MAE very close to 2 with small differences. That means that we achieved a really good accuracy for all of them. For instance, Linear Regression is able to predict the goals of a player with a small deviation of two goals. So, players who actually scored five goals, for example, will be predicted to score between three or seven goals. This two goals deviation is true for the most of the predictions but not for all of them. For instance, there are players that their goal predictions are far away from the actual ones. Generally, the average predictions are very good, and that's why we see that MAE is very close to zero.

Also, we should mention that MSE and RMSE are very important too. They are two metrics that are extremely helpful because they are very reliable and show us the error of the regression prediction each one of them with its own way.

Generally, there isn't good value for these metrics. Their values depend only on the dataset and the reason that we compare them has to do with the fact that we work on the same dataset throughout the research. Otherwise, it would be wrong to compare

different models on completely different datasets. But, as a general rule, we know that if we want to be sure about our predictions and their reliability, we should calculate these metrics as closer to zero as possible [56][57][58].

To sum up, it seems the other two algorithms, Random Forest and Multilayer Perceptron are reliable, because their metrics are also very close to zero. However, if we look closer to the above tables and representations, we will observe that Multilayer Perceptron performs better than Random Forest. To be more specific, all its metrics are smaller than those of Random Forest.

Now that we have completed the evaluation of our models and their comparison, we end up to the conclusion that Linear Regression is the best model for predicting how many goals a player will score the next season. So, these models can be used by coaches or data scientists in football clubs to take crucial decisions about their squad.

Furthermore, it is very important for every researcher to understand his work's value and its contribution to the scientific community. Taking into consideration previous researches and related work that was presented in chapter three, we can see that our results our promising. In more detail, our results are better than those of the most previous works. On the other hand, there are some researches with better results with regards to MAE, MSE and RMSE.

It is important to mention that this comparison has threats because many researches were focused on different sport's datasets. To be more specific, it is completely different to work on a basketball dataset and predict the number of points a player scores on a season than working on a football dataset. Most of the statistics have fluctuations which don't happen in football datasets. Consequently, algorithms perform completely different on different datasets.

To conclude, the lessons that we gained throughout the process are countless. First of all, the most important part is to find and work on a valid dataset. Differently, there will be problems with the validity of the results. Then, the algorithms selection is highly importance too. Their selection happened based on the bibliography, i.e. we used the algorithms that were most used in other researches. Also, the evaluation of them is very important. For this purpose, we should decide which metrics are the most appropriate. Finally, we understood that the algorithm's comparison should happen very carefully due to its importance. Based on this comparison we are able to decide which algorithm performs better and make our conclusions.

7 Conclusions

In this chapter, we present a summary of the methodology that was followed and the final results. Then, we highlight the threats to validity with regard to our research and finally, we suggest future work.

7.1 Conclusion

It is important to understand that sports analytics is going to play an important role in sports in the future. More and more teams around the world hire experts in this field in order to improve their performance. This trend is observed especially in football which is the most famous sport globally. To achieve this improvement, football clubs pay great attention to improving their players individually.

In this dissertation, we tried to predict the performance of football players who are currently active. The aim of this dissertation was to predict the goals that a player scores as accurately as possible, taking into consideration his previous season's statistics.

For this purpose, three approaches were followed to deal with this issue. To be more specific, we implemented three different ML algorithms in order to generate goal predictions. These algorithms are Linear Regression, Random Forest and Multilayer Perceptron.

We trained the algorithms with a dataset which includes more than 2.500 football players that are currently playing football at a professional level. To train the algorithms, we used statistics from seasons 2017-18 to 2020-21, except for the last season, 2021-22, which was used for testing because we aimed to predict the performances for this season.

Our findings were very interesting. Applying these algorithms, we concluded that Linear Regression, Random Forest and Multilayer Perceptron had very similar results, but Linear Regression had the most accurate ones.

Generally, the results were very good as the final predictions were very close to the goals actually scored by the players. Also, the effectiveness of the models was measured with three metrics: MAE, MSE and RMSE.

MAE, MSE and RMSE were smaller for Linear Regression predictions than Multilayer Perceptron and Random Forest consecutively. Moreover, we should take into account that there are many different factors that can affect a football match, so this kind of predictions are very difficult and they cannot be very accurate.

To conclude, nowadays, sports analytics is a hot topic, and it is taken into consideration by many companies, clubs and players. A huge amount of data is generated day by day, and ML and Data Mining are suitable tools to collect and then analyze them.

Throughout the next years, every sports club might have its own team of data scientists, not only in football but also in other sports. Consecutively, there will be a need of the appropriate innovative tools and software from every organization in order to handle all this information. It is obvious that the whole industry is growing fast, and everyone in this field should be prepared.

In conclusion, this dissertation can be very helpful to football clubs, managers and coaching staff generally because they can use the findings to their own advantage by improving their decision-making. Knowing how well a player will perform next season is invaluable information that will improve the whole squad collectively.

7.2 Threats to Validity

As explained earlier, we achieved very good accuracy and admirable results during our experiments. Besides that, there are some threats with regard to the research that should be mentioned.

The example of Messi shows us that there are many out-field factors that can affect the performance of a player. Messi was scoring on average 30 to 36 goals every single year, but in the season that we made the predictions (2021-22) he achieved only 6 goals.

This instance gives us to understand that it may not possible to predict with extremely high accuracy the performance of a football player.

Playing environment (hot or cold weather, natural or synthetic grass), emotions (psychological situation) and team environment are some examples of these out-field factors that can highly affect the individual performance of a player [60].

Additionally, it is vital to talk about the final results with regard to the database. Into our dataset, there are many football players whose actual goals were 0 for the season 2021-22 (the season that we made the predictions). Probably, these players are defenders, goalkeepers or even midfielders with no scoring capabilities. Therefore, it is easy for the algorithms to predict their goals because throughout all the previous seasons, their actual goals were 0 too.

On the other hand, it is extremely difficult for the algorithms to predict future goals of a striker. That's because a striker has fluctuations to his scoring record. Maybe one season, he is in a perfect condition, which means that he scores many goals, while the previous seasons were at a really bad tempo.

7.3 Future Work

The results from our research indicate that we can predict the performance of a player for the next season with sufficient accuracy based on historical data. However, new works should have occurred in this field in the same way for further improvement.

Obviously, those results can improve with the addition of more features. Training the algorithms with as many statistics as possible can increase the accuracy of the models. In addition to that, another way to increase the accuracy is to use different ML techniques.

Moreover, as mentioned in the dissertation, nowadays, more and more clubs are trying to collect their data from wearable devices, cameras and magnetic jackets [51]. Analyzing the information that those devices provide, future research on this field can focus not only on statistics, like goals, assists and cards, that were used in this

dissertation but on some very important information like the shooting and position habits of a player, player's physique, heart condition and player movements into the pitch, i.e. statistics that it was extremely hard to collect throughout our research because they are considered confidential and it is really challenging to have access to them [19]. Training ML algorithms with all this information can be proven vital for future predictions.

Furthermore, it would be very interesting to take into account only center forwards, i.e. strikers, instead of all players, as we did in our dissertation. That would generate a difficult environment for the algorithms, which would have to deal with predictions of players who score a different number of goals every season. That happens to our dissertation partially because there are many players with 0 goals every season on average, as mentioned before.

Lastly, all the information gained from this dissertation can be used for further analysis by computing the goals that a team will score in total, not individually. By making goal predictions for the whole squad of a football club, we can concentrate on this information and predict the final position of the club. Will the team achieve more goals than in previous years? If yes, the team is probably going to finish in a better position than in previous years.

References

[1] Sports Reference. (2022). [Data set]. https://www.sports-reference.com/

[2] Burns, E. (2021, March 30). *Machine Learning*. Enterprise AI. https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML

[3] Jordan, M. I. & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), pp. 255-260. doi:10.1126/science.aaa8415

[4] Bhatt, S. (2019, April 19). *Reinforcement Learning 101 – Towards Data Science*. Medium. https://towardsdatascience.com/reinforcement-learning-101-e24b50e1d292

[5] IBM. (2020, August). *Supervised Learning*. https://www.ibm.com/cloud/learn/supervised-learning

[6] Nasteski, V. (2017). An overview of the supervised machine learning methods. *Horizons.B*. 4, pp. 51-62.doi: 10.20544/horizons.b.04.1.17.p05

[7] Baheti, P. (2022, October 21). *Supervised and Unsupervised Learning [Differences & Examples]*. V7. https://www.v7labs.com/blog/supervised-vs-unsupervised-learning

[8] Dridi, S. (2021). Unsupervised Learning - A Systematic Literature Review. *ResearchGate*, doi: 10.13140/RG.2.2.16963.12323

[9] Mahesh, B. (2020). Machine Learning Algorithms – A Review. *International Journal of Science and Research (IJSR)*, 9(1), pp.381-386. doi:10.21275/ART20203995

[10] Jia, J. & Wang, W. (2020). Review of reinforcement learning research. *2020 35th Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, pp. 186-191, doi: 10.1109/YAC51587.2020.9337653

[11] Glorennec, P. Y. (2000, September). Reinforcement Learning: An overview. *Proceedings European Symposium on Intelligent Techniques (ESIT-00)*, pp.17-35

[12]JavaTPoint. *Reinforcement Learning Tutorial*. https://www.javatpoint.com/reinforcement-learning

[13] Chen, M.S., Han, J. & Yu, P.S. (1996). Data Mining: An Overview from a Database Perspective. *IEEE Transactions on Knowledge and Data Engineering*, 8(6), pp. 866-883. doi: 10.1109/69.553155

[14] Bharati, M. & Bharati, R. (2010). Data Mining Techniques and Applications. *Indian Journal of Computer Science and Engineering*, 1(4), pp. 301-305

[15] Neha, K. & Maramreddy, Y.R. (2020). A Study on Applications of Data Mining. *International Journal of Scientific & Technology Research*, 9(2)

[16] Ahmed, M. (2022, February 4th). *Understanding Sports Analytics Simplified*. HEVO. https://hevodata.com/learn/sports-data-analytics/

[17] Sarlis, V., Chatziilias, V., Tjortjis, C. & Mandalidis, D. (2021). A Data Science Approach Analysing the Impact of Injuries on Basketball Player and Team Performance. *Information Systems*, 99, 101750. doi: 10.1016/j.is.2021.101770

[18] Proem Sports. (2021). https://www.proemsports.com/single-post/2017/06/22/The-Evolution-and-Future-of-Analytics-in-Sport

[19] Gyarmati, L. & Hefeeda, M. (2016). Competition-wide Evaluation of Individual and Team Movements in Soccer. *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, doi: 10.1109/icdmw.2016.0028

[20] *What is Linear Regression?*. IBM. https://www.ibm.com/topics/linear-regression

[21] Complete dissertation. (2022). *What is Linear Regression?*. https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/what-is-linear-regression/

[22] JavaTPoint. *Linear Regression in Machine Learning*. https://www.javatpoint.com/linear-regression-in-machine-learning

[23] Raj, A. (2021, December 14). A Quick and Dirty Guide to Random Forest Regression. *Medium.* *https://towardsdatascience.com/a-quick-and-dirty-guide-to-random-forest-regression-52ca0af157f8*

[24] JavaTPoint. *Random Forest Algorithm*. https://www.javatpoint.com/machine-learning-random-forest-algorithm

[25] 1.17. *Neural network models (supervised)*. Scikit-learn. https://scikit-learn.org/stable/modules/neural_networks_supervised.html

[26] Wikipedia. (2022, September 8). *Multilayer perceptron*. https://en.wikipedia.org/wiki/Multilayer_perceptron

[27] Subramanya, T., Harutyunyan, D., & Riggio, R. (2019). Machine Learning-Driven Service Function Chain Placement and Scaling in MEC-enabled 5G Networks. *Computer Networks*. 106980, doi: 10.1016/j.comnet.2019.106980

[28] Myerscough, K. (1995). The game with no name: the invention of basketball. Routledge. 12(1), pp.137-152, doi: 10.1080/09523369508713887

[29] Staffo, D. F. (1998). The Development of Professional Basketball in the United States, with an Emphasis on the History of the NBA to its 50th Anniversary Season in 1996-97. *Physical Educator*. 55(1), pp.9, https://www.proquest.com/docview/1437929073/fulltextPDF/686FE0E6F1DB4A0FPQ/1?accountid=135021

[30] Louis. (2022, March). *How does the NBA Conference System Work*. Basketball Noise. https://basketballnoise.com/how-does-the-nba-conference-system-work/

[31] Papageorgiou, G. & Tjortjis, C. (2022). *Data Mining in Sports: Daily NBA Player Performance Prediction* [Master's thesis, International Hellenic University]. International Hellenic University Repository. https://repository.ihu.edu.gr/xmlui/handle/11544/29991

[32] Goldsberry, K. (2012). CourtVision: New Visual and Spatial Analytics for the NBA. *2012 MIT Sloan sports analytics conference*. 9, pp. 12-15

[33] Shah, R. & Romijnders, R. (2016). Applying Deep Learning to Basketball Trajectories. *ArXiv: Neural and Evolutionary Computing.*

[34] Sarlis, V. & Tjortjis, C. (2020). Sports Analytics – Evaluation of Basketball Players and Team Performance. *Information Systems*, 93, 101562. doi: https://doi.org/10.1016/j.is.2020.101562.

[35] Chellatamilan, T., Ravichandran, M. & Kamalakkannan, K. (2015). Modern Machine Learning Approach for Volleyball Winning Outcome Prediction. *Global Journal of Multidisciplinary Studies*, 4(12), pp. 63-71.

[36] Van Haaren, J., Davis, J., Ben Shitrit, H. & Fua, P. (2016). Analyzing Volleyball Match Data from the 2014 World Championships Using Machine Learning Techniques. *The 22nd ACM SIGKDD International Conference*, doi: 10.1145/2939672.2939725

[37] De Leeuw, A.-W., Van Der Zwaard, R. & Knobbe, A. (2021). Personalized Machine Learning Approach to Injury Monitoring in Elite Volleyball Players. *European Journal of Sport Science*, pp. 1-10. doi: 10.1080/17461391.2021.1887369

[38] Akarcesme, C. (2017). Is It Possible to Estimate Match Result in Volleyball: A New Prediction Model. *Central European Journal of Sports Science and Medicine*, 17, pp. 5-17. doi: 10.18276/cej.2017.3-01

[39] Tumer, A. E. & Kocer, S. (2017). Prediction of team league's rankings in volleyball by artificial neural network method. *International Journal of Performance Analysis in Sport*, 17(3), pp. 202-211. doi: 10.1080/24748668.2017.1331570

[40] Bruce, M. (2022, June 30). *tennis*. Encyclopedia Britannica. https://www.britannica.com/sports/tennis

[41] Sarcevic, A., Vranic, M., Pintar, D. & Krajna, A. (2022). Predictive modeling of tennis matches: a review. *2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO)*. p1099-1104, doi: 10.23919/MIPRO55190.2022.9803645.

[42] Cornman, A., Spellman, G. & Wright, D. (2017). Machine Learning for Professional Tennis Match Prediction and Betting.

[43] Gao, Z. & Kowalczyk, A. (2021). Random Forest Model Identifies Serve Strength as a Key Predictor of Tennis Match Outcome. *IOS Press*, 7(4), pp. 255-262, doi: 10.3233/JSA-200515

[44] Babbar, M. & Rakshit, (2019). A systematic Review of Sports Analytics. *ResearchGate*.

[45] Pariath, R., Shah, S., Surve, A. & Mittal, J. (2018). Player Performance Prediction in Football Game. *Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pp. 1148-1153. doi: 10.1109/ICECA.2018.8474750.

[46] Apostolou, K. & Tjortjis, C. (2018). Sports Analytics Algorithms for Performance Prediction. *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*. pp. 1-4, doi: 10.1109/IISA.2019.8900754.

[47] Srinivasan, B. (2017). A Social Network Analysis of Football – Evaluating Player and Team Performance. *2017 Ninth International Conference on Advanced Computing (ICoAC)*. pp. 242-246, doi: 10.1109/ICoAC.2017.8441301.

[48] Manish, S., Bhagat, V. & Pramila, R. (2021). Prediction of Football Players Performance using Machine Learning and Deep Learning Algorithms. *2021 2nd International Conference for Emerging Technology (INCET)*. pp. 1-5, doi: 10.1109/INCET51464.2021.9456424.

[49] Tiwari, E., Sardar, P. & Jain, S. 2020. Football Match Result Prediction Using Neural Networks and Deep Learning. *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*. pp. 229-231, doi: 10.1109/ICRITO48877.2020.9197811.

[50] Pantzalis, V. C. & Tjortjis, C. (2020). Sports Analytics for Football League Table and Player Performance Prediction. *2020 11th International Conference on Information, Intelligence, Systems and Applications (IISA. doi:10.1109/iisa50023.2020.9284352)*.

[51] Cintia, P., Pappalardo, L. & Rinzivillo, S. (2015). A network-based approach to evaluate the performance of football teams. *ResearchGate*.

[52] Frey, M., Murina, E., Rohrabach, J., Walser, M., Haas, P. & Dettling, M. (2019). Machine Learning for Position Detection in Football. *2019 6th Swiss Conference on Data Science (SDS)*. pp. 111-112, doi: 10.1109/SDS.2019.00009.

[53] A. Al-Asadi, M. & Tasdemir, S. (2022). Predict the Value of Football Players Using FIFA Video Game Data and Machine Learning Techniques. *IEEE Access*, 10, pp. 22631-22645, doi: 10.1109/ACCESS.2022.3154767.

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

[55] Fernandez, J., Medina, D., Gomez, A., Arias, M. & Gavalda, A. (2016). From Training To Match Performance: A Predictive and Explanatory Study on Novel Tracking Data. *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*. pp. 136-143, doi: 10.1109/ICDMW.2016.0027.

[56] S. AllWright. (2022). *What is a good MAE Score? (simply explained)*. https://stephenallwright.com/good-mae-score/

[57] S. AllWright. (2022). *What is a good RMSE value? Simply explained*. https://stephenallwright.com/good-rmse-value/

[58] S. AllWright. (2022). *What is a good MSE Score? (simply explained)*. https://stephenallwright.com/good-mse-value/

[59] Lazzeri, F. (2020). *Machine Learning for Time Series Forecasting with Python.* (1st ed.). John Wiley & Sons, Inc.

[60] Dahl, K.D. (2013). *External Factors and Athletic Performance* [Senior Honors Thesis, Liberty University]. Liberty University Repository https://digitalcommons.liberty.edu/honors/347/