INTERNATIONAL
HELLENIC
UNIVERSITY

# Sports Analytics Algorithms for Performance Prediction

**Paschalis Koudoumas**

SID: 3308190012

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

*Master of Science (MSc) in Data Science*

JANUARY 2021

THESSALONIKI – GREECE

INTERNATIONAL
HELLENIC
UNIVERSITY

# Sports Analytics Algorithms for Performance Prediction

**Paschalis Koudoumas**

SID: 3308190012

| | |
|---|---|
| Supervisor: | Assoc. Prof. Christos Tjortjis |
| Supervising Committee Members: | Assoc. Prof. Maria Drakaki<br>Dr. Leonidas Akritidis |

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

*Master of Science (MSc) in Data Science*

JANUARY 2021

THESSALONIKI – GREECE

# Abstract

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

Sports Analytics exist as a term and concept for many years, but nowadays, it is implemented in a different way that affects how teams, players, managers, executives, betting companies and fans perceive statistics and sports.

Machine Learning can have various applications in Sports Analytics. The most widely used are for prediction of match outcome, player or team performance, market value of a player and injuries prevention. This dissertation focuses on the quintessence of football, which is match outcome prediction.

The main objective of this dissertation is to explore, develop and evaluate machine learning predictive models for English Premier League matches’ outcome prediction. A comparison was made between XGBoost Classifier, Logistic Regression and Support Vector Classifier. The results show that the XGBoost model can outperform the other models in terms of accuracy and prove that it is possible to achieve quite high accuracy using Extreme Gradient Boosting.

# Acknowledgements

At this point, I would like to thank my Supervisor, Professor Christos Tjortjis, for offering his help throughout the process and providing me with essential feedback and valuable suggestions to the issues that occurred.

Paschalis Koudoumas

4-1-2021

# Contents

# List of Figures

# List of Tables

# 1 Introduction

The dissertation is structured in 7 chapters as follows: the first chapter is the introduction; the second chapter presents the historical background for sports analytics in certain sports and the previous work on football; the third chapter offers a short introduction to some general terms that are used in the dissertation; in the fourth chapter, the whole methodology process is analyzed; the fifth chapter describes the procedure of creating prediction models; in the sixth chapter, the results of the models and their evaluation are presented; and finally, in chapter seven, conclusions are presented and future work is recommended.

The dissertation focuses on football and more specifically on the English Premier League. The aim of the dissertation is to predict match outcomes for English Premier League. Football was selected because it is considered the world's most popular sport regarding its fanbase. In 2019, English Premier League had 39 thousand tickets per game. On the other hand, there are specific difficulties in predicting football match outcomes because these can be affected by both internal and external factors. The abundance of free online data regarding football can be considered as an asset, but the proper selection of data suitable for outcome prediction is needed.

Even though it is quite difficult to predict a football match outcome, this dissertation proves that it is possible to generate predictions with relatively good results for a single football season. More specifically, creating features related to each team's form and strength from the traditional football statistics of a game, was crucial for the model to achieve a satisfactory performance. XGBoost model scored the highest accuracy score compared to Logistic Regression and Support Vector Classifier.

Generally, sports analytics is a topic that increasingly gains interest. Sports clubs have already started using software, censors, cameras, and specific techniques to collect and analyze data that help managers, analysts, trainers, and executives to make short-term and long-term decisions.

Moreover, sports analytics become indispensable for betting reasons. Both clients and betting companies are very interested in sports data. Betting companies want to please the clients offering good betting odds as well as make profit, and clients search for statistics, especially advanced ones to make a betting decision.

# 2 Background

This chapter is split into 2 subsections. The first one refers to the historical background of baseball, basketball, and football and the second one to the past research in football.

## 2.1 Historical Background

Sports analytics is a new scientific area that grows rapidly as it is used by more and more professional teams and professional athletes in every sport. Nevertheless, the idea of collecting and analyzing data that may help the improvement of the team or individual performance was presented many years ago before using algorithms and machine learning techniques.

Several decades ago, even traditional statistics were not monitored for most sports. For instance, there is no information about the assists that were given at the World Cup football tournament of 1954 because the “key” pass that leads to a goal was not statistically significant during that period.

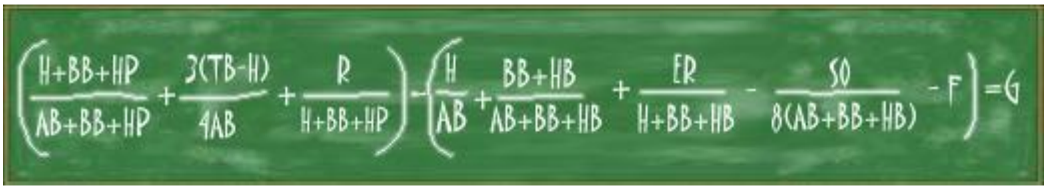
Surprisingly, whilst football is the most famous sport, baseball was the first sport that sports analytics were applied.

### 2.1.1 Baseball

In 1906, the Chicago Cubs achieved the best record with 116 wins in 154 games in regular season and reached to the finals. The competitor of the title was the Chicago White Sox, known as the “hitless wonders”. Sportswriter Hugh Fullerton implemented his own baseball analysis and according to that analysis, he predicted that the weaker Chicago

White Sox would win the championship. Despite the odds, Chicago White Sox shocked everyone and won the championship. Because of this prediction, Hugh Fullerton became famous and he published the book *“Touching Second: The Science of Baseball”* in 1910 [1].

In 1947, baseball executive Branch Rickey thought innovative and he hired statistical analyst Allan Roth. His main idea was to convert Roth’s observations from the team data into game insights. Rickey was one of the first executives that realized the worth of statistics and sports analytics. In 1954, Life magazine published an article based on Rickey decision, entitled “Goodbye to Some old Baseball Ideas”, which focus on the suggestion that a team’s performance might be precisely clarified by a statistical formula [2].



$$\left(\frac{H+BB+HP}{AB+BB+HP}+\frac{3(TB-H)}{4AB}+\frac{R}{H+BB+HP}\right)-\left(\frac{H}{AB}+\frac{BB+HB}{AB+BB+HB}+\frac{ER}{H+BB+HB}-\frac{SO}{8(AB+BB+HB)}-F\right)=G$$

Figure 1: Branch Rickey's famous formula [2].

In 1971, baseball researcher L. Robert Davis called 30 people who are interested in baseball and statistical research for a meeting in New York. That meeting leads to the creation of the Society for American Baseball Research. The main target of this society is to endorse and develop the historical and statistical research in baseball.

In 1977, William (Bill) James, maybe the most prominent person in baseball history, published his annual book *“The Bill James Baseball Abstract” from 1977 to 1980.* That period was crucial for sports analytics because that was the first time that a book about sports analytics became widespread. Furthermore, Bill James is the “father” of sabermetrics. Sabermetrics is the statistical analysis of baseball data designed to measure baseball players’ performances. He published almost 30 books about baseball history and statistics until he was hired as senior consultant of co-owner and general manager of Boston Red Sox in 2002. In 2004, Boston Red Sox won its first World Series since 1918. The team won again the series in 2007 and in 2013 with Bill James on the front office [5].

In 2003, Michael Lewis published the book named *“Moneyball”*. This book refers to the Oakland Athletics and their general manager Billy Beane. Beane utilized sabermetric analysis to construct teams that qualified for five postseasons from 2000 to 2006, whilst had one of the lowest wage budgets in baseball [3]. Figure 2 depicts team salaries in 2002 for Oakland Athletes, who had the third-lowest wage budget in the MLB [4].

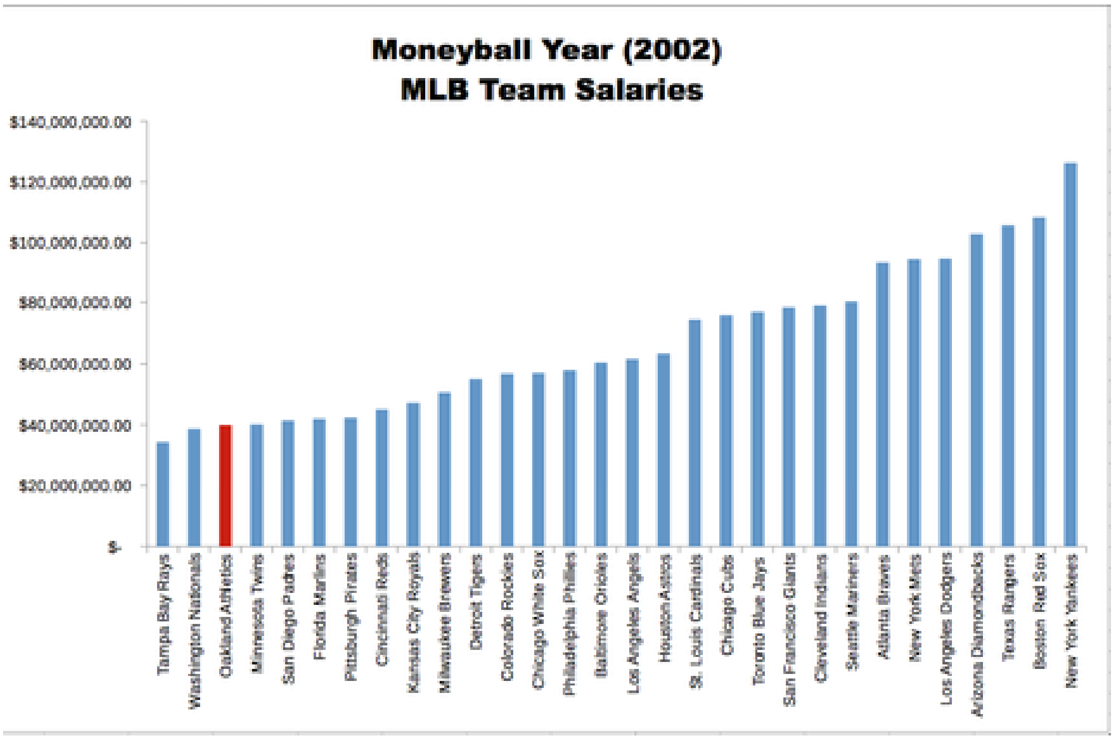


Figure 2: Team salaries in 2002 [4].

In 2017, all baseball teams which participated in Major League Baseball, employed at least one sabermetrician.

### 2.1.2 Basketball

Generally, at basketball, the statistically better team usually wins the game. Despite basketball dependence on statistics, sports analytics came very late compared to other sports. In 2004, a book, titled “Basketball on Paper”, was published by Dean Oliver.

Nowadays, Dean Oliver referred as the godfather of advanced basketball statistics. He divided the basketball game points into efficiency and pace [6].

In 2005, two Israeli scientists, Gal Oz and Miky Tamir, created the famous SportVU. It is a system that tracks the movements of every player and the ball 25 times per second. Since 2014, all NBA teams use the SportVU system. The six cameras that were installed in every arena, provides every team with advanced statistics depends on speed, distance, player separation and ball possession. Moreover, SportVU has been expanded in football [7] [8] [9].

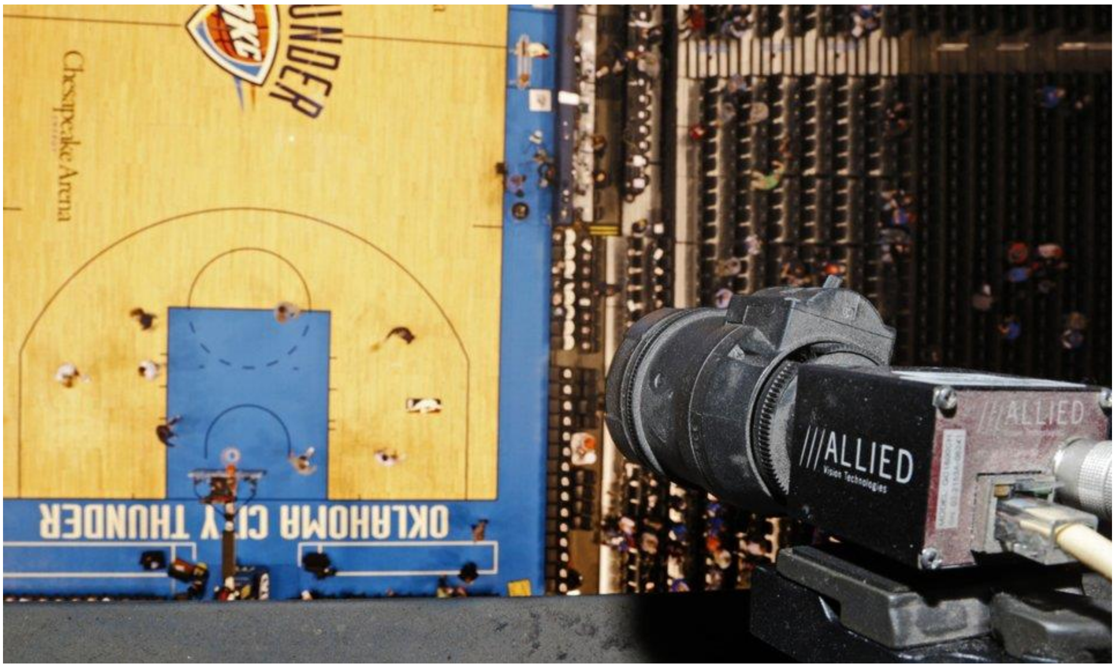


Figure 3: SportVU camera in NBA [10].

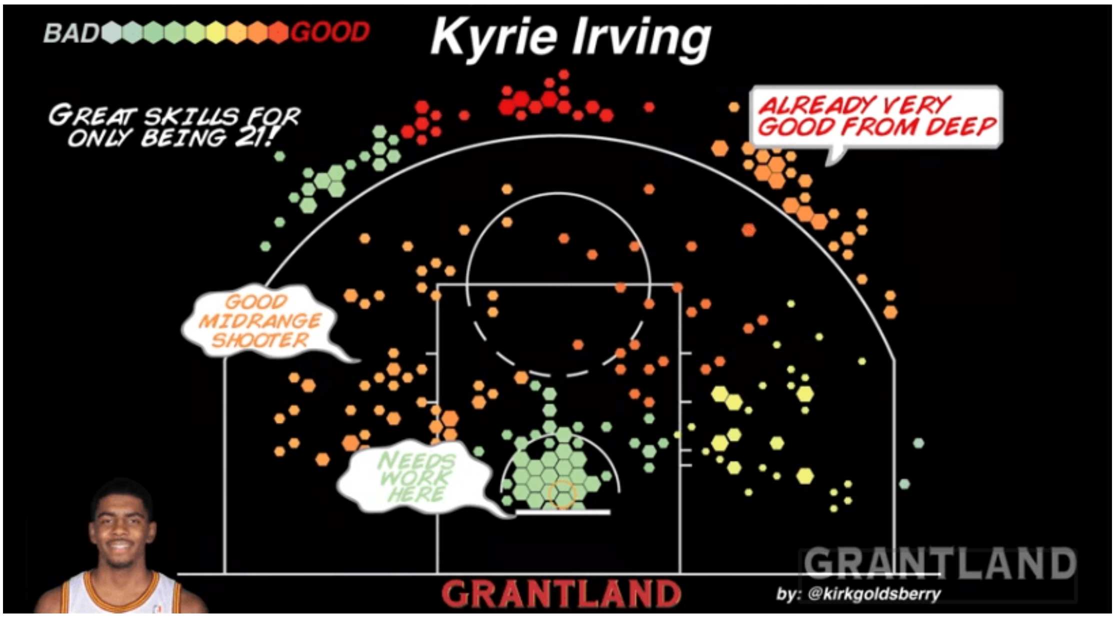


Figure 4: Kyrie Irving's report from SportVU [11].

### 2.1.3 Football

Analysts have been applying mathematical models in sports for a long time. Charles Reep was the first who applied mathematics in football. During 1950, he utilized statistics and demonstrated the long-ball theory in order to improve the probability of winning [12]. However, Charles Reep approaches were controversial. A sports journalist Jonathan Wilson argued that the long-ball theory was totally wrong [13]. Those days, the mathematics used were uncomplicated and done by hand. The rapid growth of mathematics and computer science gives the ability to the sports teams to use data science in order to analyze the data about the sport [14].

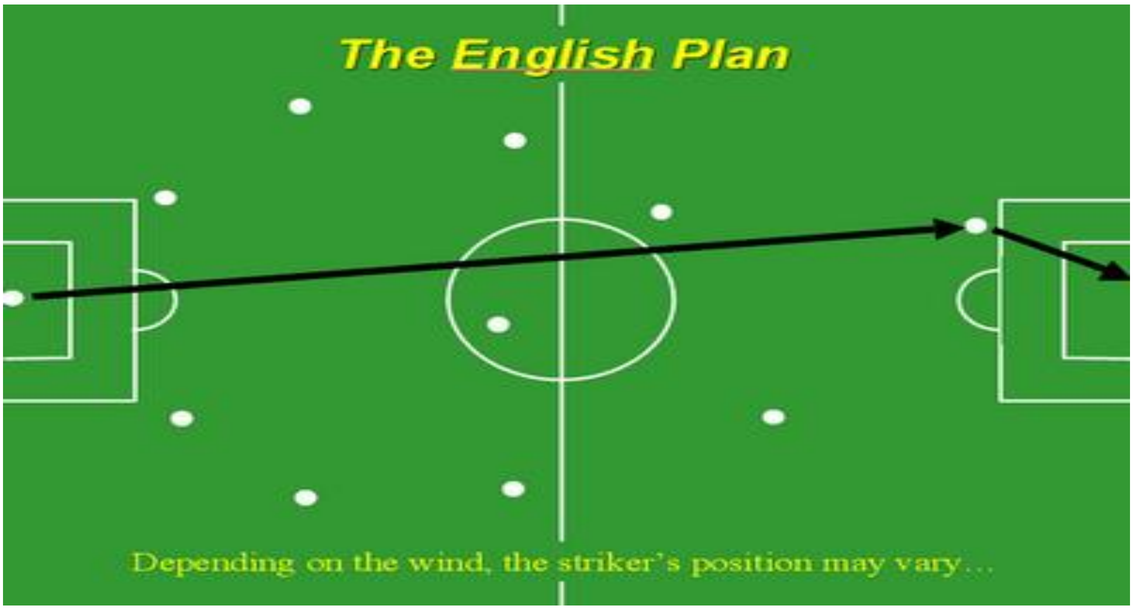


Figure 5: Long-ball theory [15]

Nowadays, Liverpool FC have the football analytics dream team. More specific, the research team comprises of Ian Graham (PhD in Physics), Bill Spearman (PhD in Physics and ex-CERN), Tim Waskett (PhD in Astronomy) and Dafydd Steele (Statistical researcher). This group helps the team sign the most suitable players for coach's style. Their contribution to winning the Champions League (2019) and English Premier League (2020), was significant [14]

During the past decade, there was a substantial change of football analytics. Advanced statistics have appeared. The most important ones are Expected Goals, Expected Assists, Defensive Coverage and Sequences. Expected Goal may be the most innovative. It measures the quality of chances created and conceded. In 2015, Midtjylland FC won the first championship of their history, signing players according to the Expected Goal model [16].

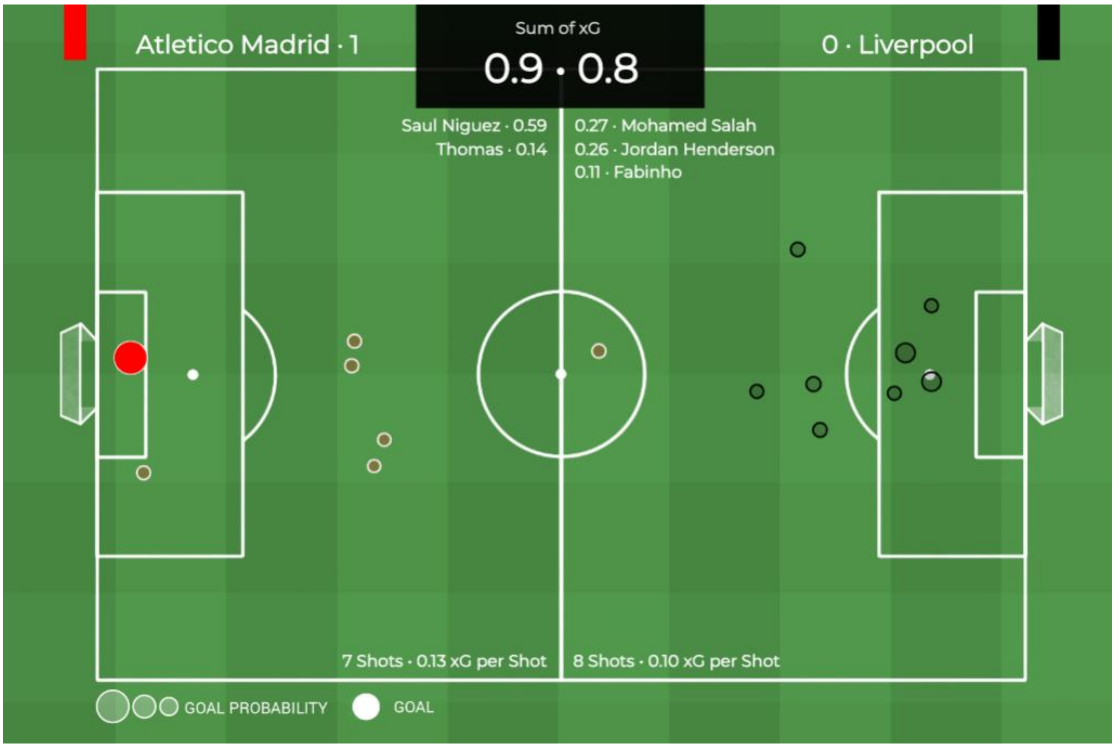


Figure 6: Score and the Expected Goals [17].

## 2.2 Literature Review

### 2.2.1 Introduction

In this part of the dissertation, previous works and the remarkable researches will be presented and discussed. We have separated the past research in four categories. Firstly, we will inspect at the general overview of football predictions and the machine learning techniques that have been utilized. Secondly, we will focus on research related to using team ratings to enhance predictions. Thirdly, we will talk about the factors that affect the outcome of football matches. Finally, we will present the papers that demonstrate models estimating the expected goals that a team is expected to have achieved.

### 2.2.2 Football Prediction Overview

Generating predictions for football outcomes has been a significant research subject since the middle of the 20th century. Moroney (1956) [18] and Reep (1971) [19] utilized both the negative binomial distribution and Poisson distribution to model the total goals scored in a football match, depending on historical team results.

In 1982, Maher [20] pioneered using Poisson distributions to model home and away team defensive and attacking abilities. His main purpose was to predict the mean number of goals for each team. Following Maher's approach, in 1997, Dixon and Coles [21] generated a model that produced probabilities for game results and scores by following Poisson distribution. Dixon and Coles model was very popular and used as a benchmark for other models. This model developed on a Poisson regression model, which indicates that the expected number of goals for each team converted to goal probabilities and finally to match result probabilities.

Furthermore, this model measures an attacking and defending rating for each team by calculating the maximum likelihood estimates of these ratings on previous outcomes and utilizes a weighted function to exponentially down weight previous results depending on the length of time that divides an outcome from the actual prediction time.

In 2000, Rue and Salveson [22] used Monte Carlo simulation to make predictions.

In the early 2000s, researchers were trying to directly predict games outcomes (win or draw or loss) instead of predicting the match scores and then generating match outcome probabilities. More specific, in 2000, Forrest and Simmons [23] applied a classification model to predict the match result rather than generating predictions for the goals scored. This helped them to prevent the challenge of interdependence on the two teams scores.

In the same year, Kuypers [24] constructed a model that predicts future match outcomes, using variables extracted from a season's match outcomes. Moreover, he was among the first that tried to make profitable betting strategies by using his model.

The past research tried to forecast both match score and match outcome. Subsequently, we will now explore the recent research that has been done on the topic, by applying modern Machine Learning algorithms.

In 2005, Goddard [25] built a regression model that took recent performance, team quality, game significance and geographical distance into account. Using geographical distance as a variable, he tried to detect a rivalry between teams. Goddard published one of the first papers that took into consideration other variables than actual match results. He identified that there was the likelihood of a profitable betting return comparing his model predictions with the betting odds. He inferred that models based on scores and models based on outcomes have similar accuracy, and the best method might be a hybrid model.

In 2006, Joseph at al. [26] evolved four machine learning models: A Naïve Bayes, a Bayes learner, a KNN learner and a MC4 learner. He deduced that MC4 learner was the best in detecting the attributes with the greatest influence on the result.

In 2011, Hucaljuk et al. [27] tried to predict football scores using different Machine Learning techniques: Naïve Bayes, Bayesian Networks, LogitBoost, KNN, Random Forest, and Artificial Neural Networks (ANN). They concluded that Artificial Neural Networks achieved the best results.

More recently, in 2016, Tavakol [28] used a linear model. He took into consideration historical player data and the historical outcomes between the two teams. In order to train the model, he used feature extraction and feature aggregation techniques because of the big number of features.

In 2019, Gu at al. [29] gathered high volume of data about teams and players. They applied big data analytical techniques and developed a Support Vector Machine and an Ensemble Machine Learning algorithm. The Ensemble algorithm could improve completely the Support Vector Machine’s results. In the same year, Apostolou and Tjortjis

[78] achieved to predict the goals that a famous player will score in the next season using Logistic Regression.

In 2020, Chazan and Tjortjis [79] tried to predict English Premier League using different Machine Learning techniques. SVM with polynomial kernel was the best in terms of accuracy.

By exploring the past research that have been done on the topic of constructing models to predict football game scores and outcomes, we gained valuable information on the techniques we should use and the potential issues we should avoid.

### 2.2.3 Rating Systems

Some researchers followed another approach in order to perform match predictions. They used rating systems for theirs purposes. One of the most popular rating systems that is used nowadays, is the “ELO rating” system. In 1978, Elo [30], a Hungarian physics professor, invented this rating system to rate chess players. Its usage has been broadened for football clubs. In 2003, Buchdahl [31] used ELO ratings to update teams’ relative strengths.

In 2010, Hvattuma [32] utilized an ELO system and applied regression models to obtain covariates for football match predictions. He used two forms of ELO ratings. The first one took into consideration the match outcome and the second one took into account the actual match score. To evaluate his predictions, he tested various benchmarks and he inferred that he acquired better results utilizing ELO ratings.

To generate predictions for football Euro 2016 tournament, several approaches have been tested. In 2016, Lasek [33] used ordinal regression ratings and least squares ratings in combination with Poisson regression, to provide predictions about the tournament and simulate it many times utilizing Monte Carlo simulations.

In 2017, Viswanadha et al. [34] utilized player ratings instead of using team ratings to predict cricket match results applying a Random Forest classifier.

In 2020, Sarlis and Tjortjis [80] used advanced basketball analytics including ELO ratings to predict the MVP and the defensive player of the year.

In 2013, Fenton [35] invented “pi-rating”, a new team rating type. It dynamically evaluates teams by exploring the relative differences in score over time. The same year, Lasek et al. [36] proposed that an ensemble model from various rating systems can have substantially better results than a model using only one rating system.

### 2.2.4 Factors that affect the result of a football game

Surely, there are many factors that decide a football match. Vecer at al. [37], in 2009, detected red card as a factor and Catteeuw et al. [38], in 2010, identified referee decisions as a factor that affects a football game. These factors are in-court parameters. However, there are plenty of out-court factors which are more complicated than the previous ones.

The most significant factor may be the players. Even though, it is possible to evaluate the quality of players by utilizing rating systems, there are various factors to take into account. An important one, is the interrelations among teammates. Generally, it is based on the strengths and the weaknesses of each player in combination with the team’s culture and mentality and it appears to be an important parameter of a team’s performance according to Gréhaigne et al. [39], in 2005.

Possibly the most used performance metric in football is the team form. Team form is the overview of the result of five games. This number is acknowledged by experts and adopted by many football statistics sites like “statsfc.com”. In 2014, Arabzad et al. [40] followed a different interesting approach. They used more than five matches. For example, they used the outcomes of last four home and away games.

Another factor that can affect team performance is players' psychology. It is a non-measured parameter. In 2012, Constantinou et al. [41] utilized knowledge about motivation of winning the game and team spirit to measure psychological difference between two teams located in the same region.

There are players in a team like the captain who plays a more important role than other players. These players called key players. Messi is a great example. He contributes to the winning mentality of his team with his special football abilities. Even though experts have accepted this approach, there has not been many researches. In 2006, Joseph et al. [42] performed good results by constructing a model based on key players.

### 2.2.5 Expected Goals (xG) Models

Expected Goals models is a quite modern approach that focus on analyzing game data to understand the number of goals a team should have scored based on the statistics that have been noted during the match.

In 2012, MacDonald [43] constructed an Expected Goals model to estimate the performance of National Hockey League (NHL) matches. He used two metrics. The first took into consideration the shots and the missed shots. The second one based on shots, missed shots, and blocked shots. This model helped teams to understand if, for instance, they were missed too many goal opportunities or if their gameplay did not lead to enough goal opportunities. He achieved promising results for this Expected Goals model.

In 2015, Lucey [44] utilized features from spatiotemporal data to find the possibility of each shot of turning into a goal. Features were shot location, defender proximity, game phase etc. as extracted from spatiotemporal data. This enables the calculation of a team's effectiveness in a match and the estimation of the number of goals a team would have been expected to achieve.

More recently, in 2016, Eggels at al. [45] used geospatial data for the shot and the part of the body that a player used for it. They applied classification techniques like logistic regression, decision trees and random forest, to classify the opportunities into probabilities of actually scoring a goal.

# 3 General Terms

In this part of the dissertation, the following general terms that are related to the topic will be explained:

- Data Mining
- Machine Learning
- Machine Learning Algorithms
- Sports Analytics

## 3.1 Data Mining

Data Mining is every process and attempt to detect hidden patterns, trends and meaning in large database systems. [46] Data mining belongs to the computer science field. It combines statistics and AI (Artificial Intelligence) with database management. [47]

The completion of a data mining task requires the following procedure (Figure7):

- Data extraction, collection and selection and load in data warehouses.
- Dataset needs to be preprocessed. Data should be cleaned and simplified.
- The cleaned data should be transformed into a supported structure for Data Mining algorithm.
- Mining is the major step of learning and knowledge extraction. Mining can be split into six stages:
    1. Anomaly detection.
    2. Association rule learning.
    3. Clustering.
    4. Classification.
    5. Regression.
    6. Summarization.
- Evaluation of the significance for the extracted knowledge.

- Visualization. [48] [49]

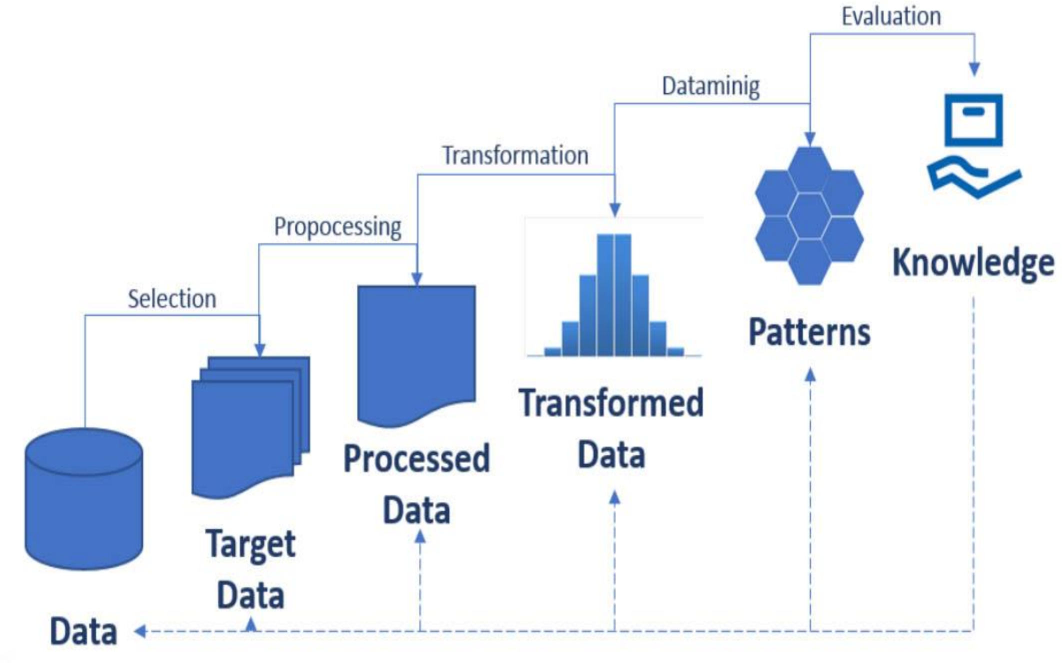


Figure 7: Data Mining process. [49]

## 3.2 Machine Learning

There are many definitions for the term “Machine Learning”. The most recognizable and reliable ones will be mentioned below:

- “Machine Learning is the science of getting computers to act without being explicitly programmed”. [50]
- “The field of Machine Learning seeks to answer the question “How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes” [51]?
- “A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E” [52].

- “Machine learning is based on algorithms that can learn from data without relying on rules-based programming” [53].

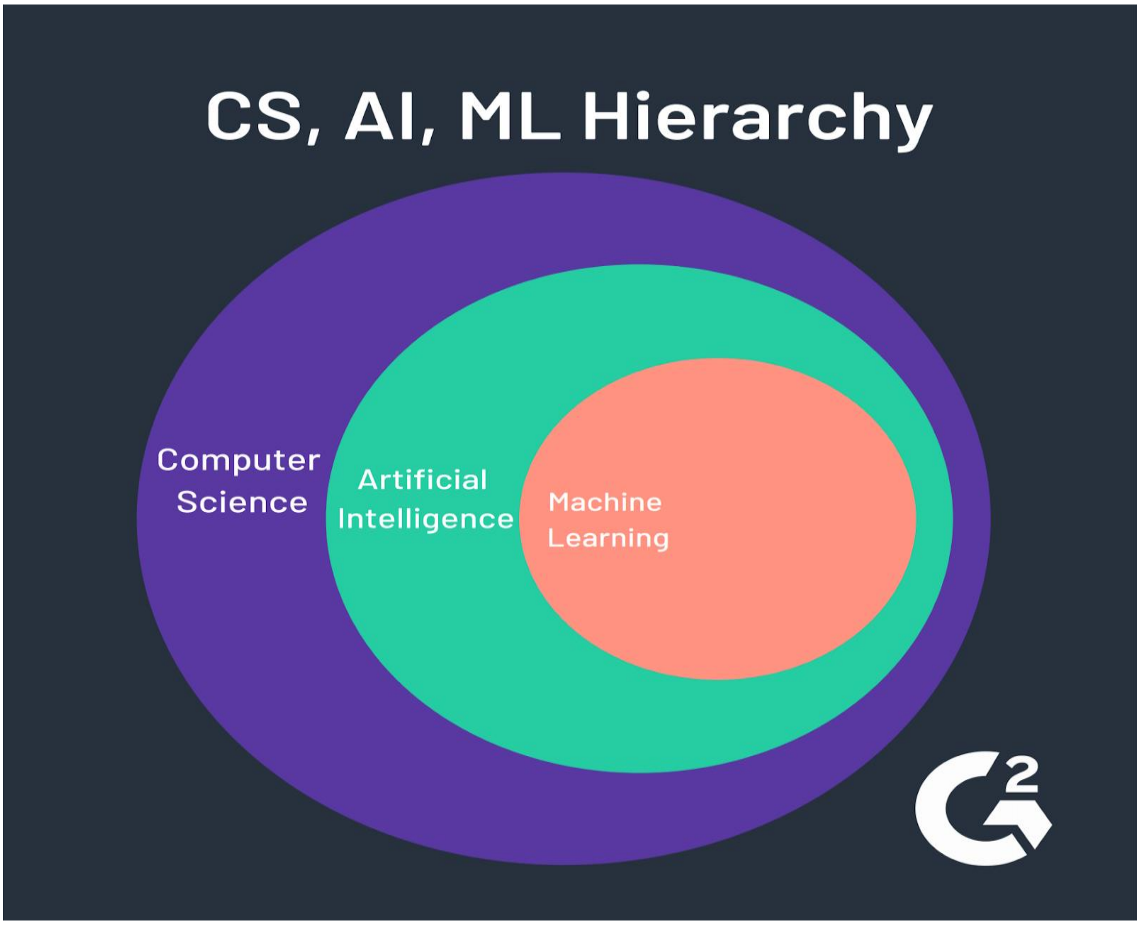


Figure 8: The hierarchy of Machine Learning [54].

Machine Learning can be split in three main categories:

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

Below, in figure 9, it demonstrated a typical Machine Learning structure with the most significant parts.

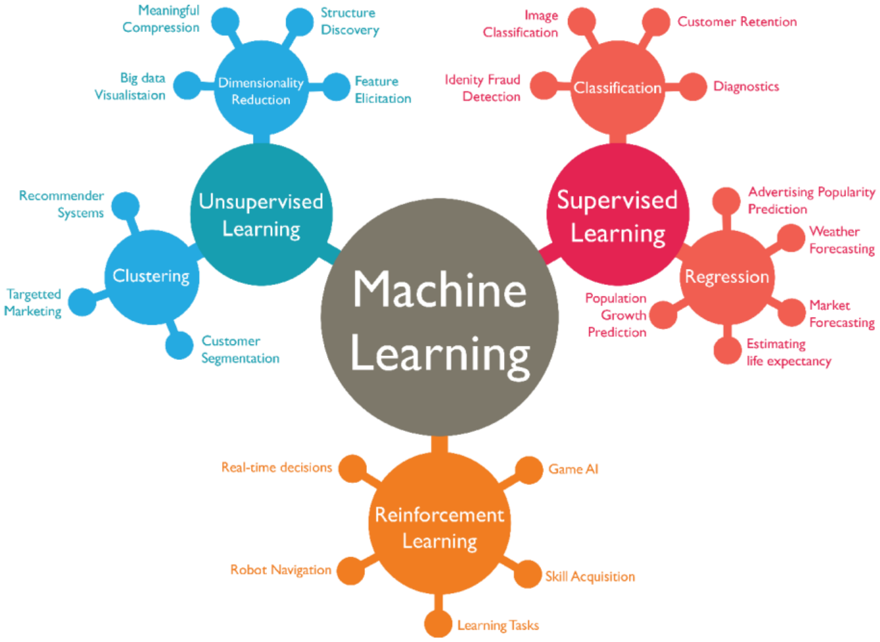


Figure 9: Machine Learning structure and the most important tasks [55].

Supervised Learning is the Machine Learning task of learning a function that maps an input to an output based on examples input-output pairs. It infers a function labeled training data consisting of a set of training examples. The optimal result is when the algorithm correctly determines the class labels for unseen cases [56] [57].

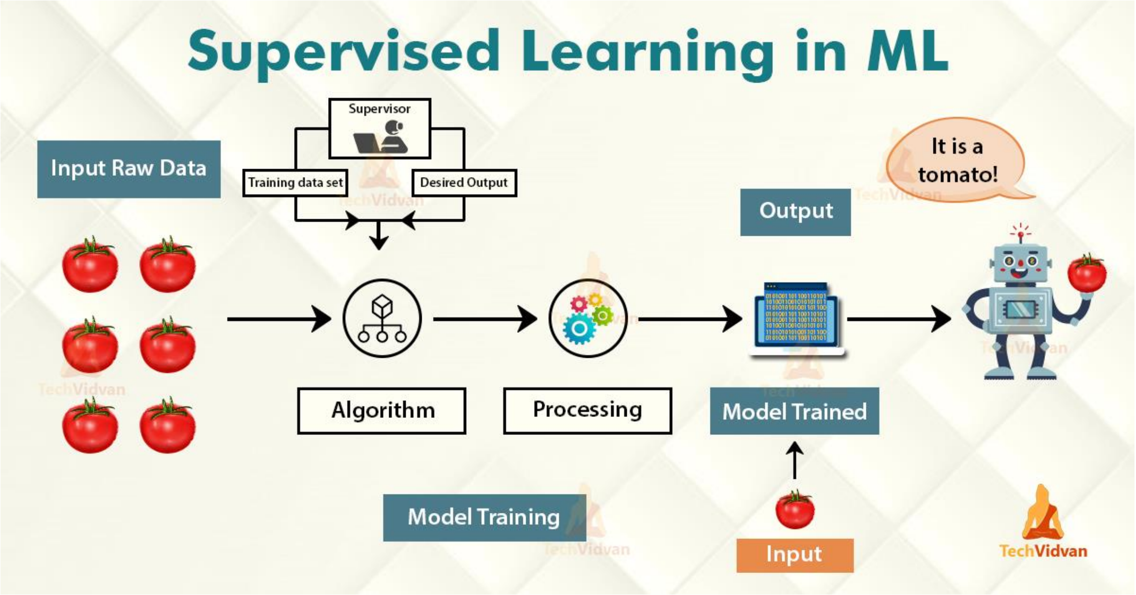


Figure 10: Supervised Learning [58].

Unsupervised learning is a type of machine learning that looks for previously undetected patterns in a data set with no pre-existing labels and with a minimum of human supervision. In contrast to supervised learning that usually makes use of human-labeled data, unsupervised learning, also known as self-organization allows for modeling of probability densities over inputs [59]

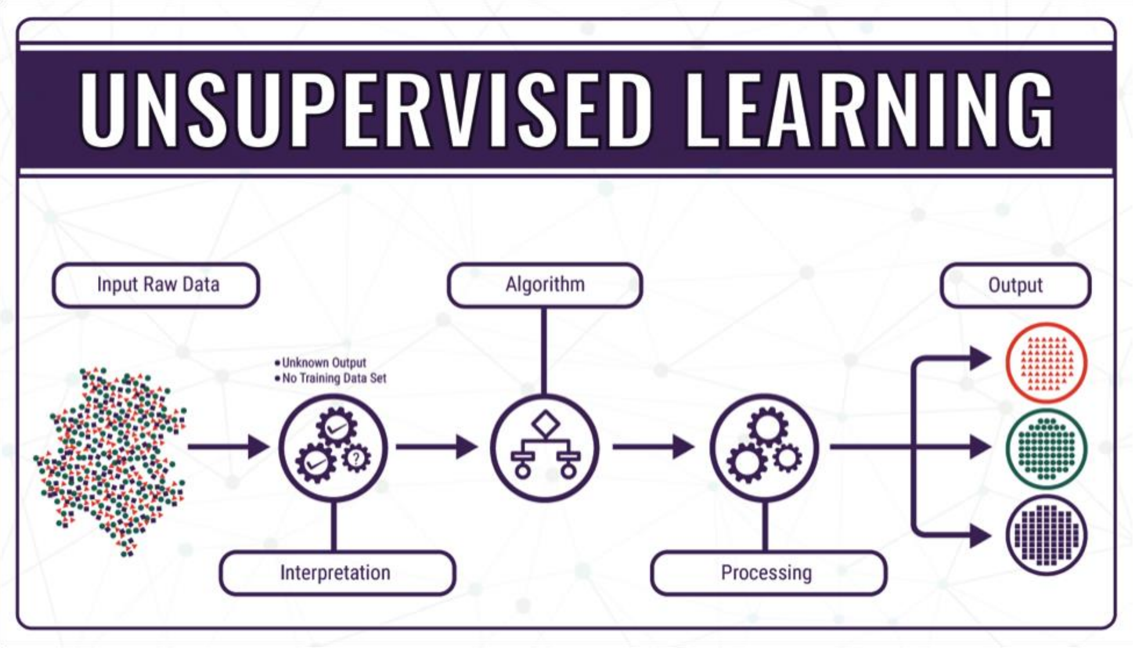


Figure 11:Unsupervised Learning [60].

Reinforcement learning differs from supervised learning in not needing labelled input/output pairs be presented, and in not needing sub-optimal actions to be explicitly corrected. Instead, the focus is on finding a balance between exploration (of uncharted territory) and exploitation (of current knowledge) [61].

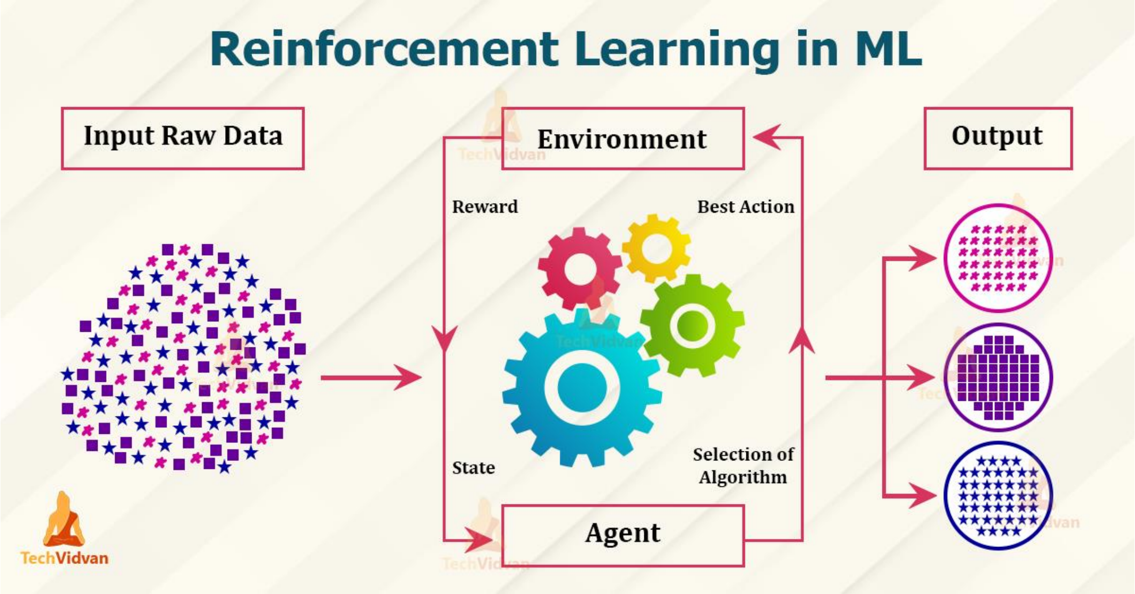


Figure 12: Reinforcement Learning [62].

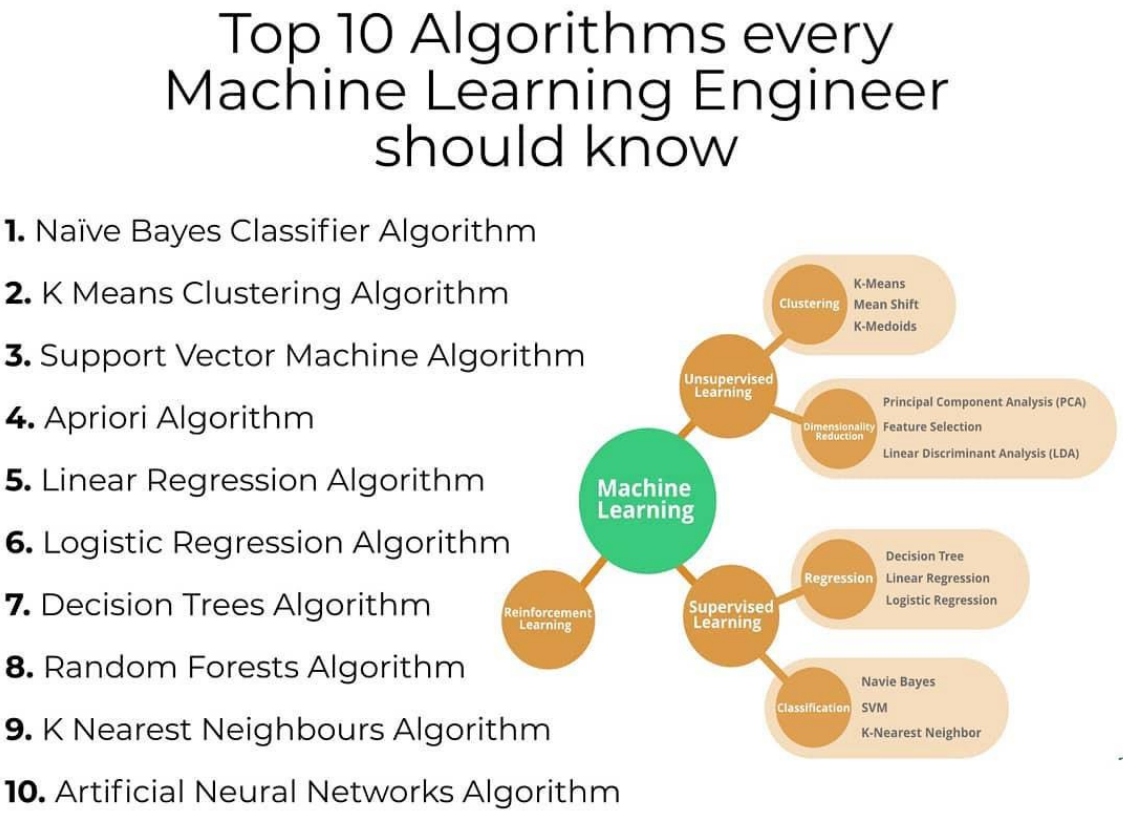
### 3.2.1 Machine Learning Algorithms

Machine learning algorithms are programs that adjust themselves to perform better as they are exposed to more data. The “learning” part of machine learning means that these programs alter how they process data over time, much as humans change how they process data by learning.A machine-learning algorithm is a program with a explicit way to adjusting its own parameters, given feedback on its last performance in making predictions about a dataset [63].

Some of the most widely used Machine Learning algorithms are the following:

- Linear Regression
- Logistic Regression
- Support Vector Machine
- Naïve Bayes
- Random Forest
- K-Nearest Neighbor
- Decision Trees

- Gradient Boosting algorithms
- Neural Networks
- Principal Component Analysis
- Apriori
- K-mean



Top 10 Algorithms every Machine Learning Engineer should know

1. Naïve Bayes Classifier Algorithm
2. K Means Clustering Algorithm
3. Support Vector Machine Algorithm
4. Apriori Algorithm
5. Linear Regression Algorithm
6. Logistic Regression Algorithm
7. Decision Trees Algorithm
8. Random Forests Algorithm
9. K Nearest Neighbours Algorithm
10. Artificial Neural Networks Algorithm

Figure 13: Machine Learning algorithms [64].

## 3.3 Sports Analytics

Sports Analytics is the usage of past historical data and advanced statistics in order to measure and enhance performance, make data-driven decisions and make predictions about the performance and results. The purpose of that use is to provide a competitive advantage to a team or individual [65].

There are two main aspects of Sports Analytics. On-field and off-field analytics. On-field analytics deals with improving the on-field performance of teams and players. It copes with aspects such as game tactics and player fitness. Off-field analytics deals with the business side of sports. Off-field analytics focuses on supporting a sport club or body surface patterns and insights through data that would help boost ticket and merchandise sales, improve fan engagement, etc [66].

Figure 14: Sports Analytics [67].

However, the data should not be accessible to anyone due to lack of expertise and may lead to problems, such as issues with betting companies [68].

# 4 Methodology

Jupyter Notebook was the software that was used for the whole process. All models were developed and evaluated using Python 3.7.3. During data collection procedure, some data were typed manually, and Microsoft Excel was utilized for that reason.

```
#DataFrames in Python
In [  ] :  This is pandas DataFrame
In [ 6 ] : import pandas as pd
           import numpy as np
In [ 7 ] : df = pd.DataFrame(data*np.array([[1,2,3],
           [4,5,6]], dtype=int), columns=['A', 'B', 'C'])
           df
```

Figure 15: Jupyter notebook [69].

## 4.1 Process Description

Firstly, an appropriate dataset needed to be found. There are several websites with football statistics both for teams and players. We chose a set of data from website “datahub.io” with English Premier League matches.

After the data acquisition, it was necessary to clean and transform the dataset in our desirable form. For our purposes, we wanted to split each football season into a separate csv file. Then, the csv files were uploaded to Jupyter Notebook.

The data needed to be preprocessed. They were tested for null values, duplicate values, and noise. Python language was utilized to clean the data and build the machine learning models. Then, feature engineering and extraction took place to keep and create the suitable features.

Finally, the results that were acquired, were evaluated in terms of accuracy and F-score. More information and in-depth description are presented and explained below.

### 4.1.1 Data Collection

In Europe, English Premier League is probably the most famous football league. Consequently, detecting data regarding English Premier League matches is not so challenging, as there are plenty of sources that can provide historical data for collection and using.

The data were downloaded from "datahub.io". The data contain English Premier League fixtures and results from season 2009/2010 to season 2018/2019. The last season was used to perform predictions. The rest ones were used for the training phase. After collecting the data, various actions needed to take place to the Machine Learning models' variables and features.

The features of the dataset are presented below:

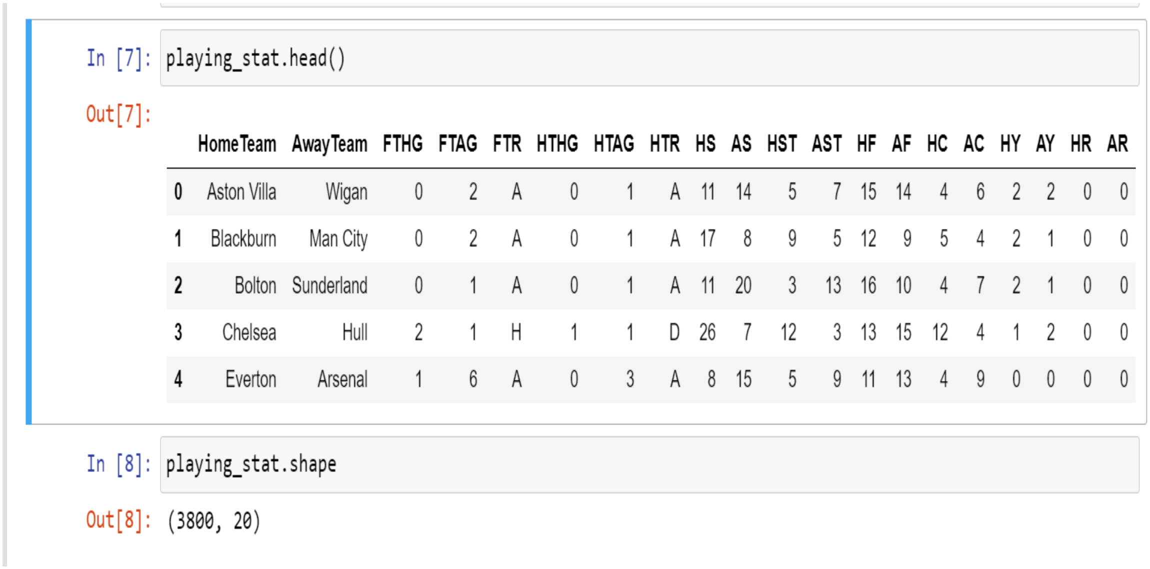
- Div: League Division
- Date: Match date
- HomeTeam: Home team
- AwayTeam: Away team
- FTHG: Full time home team goals
- FTAG: Full time away team goals
- FTR: Full time result (H=Home Win, D=Draw, A=Away Win)
- HTHG: Home team half time goals

- HTAG: Away team half time goals
- HTR: Half time result
- Referee: The referee of each match
- HS: Home team shots
- AS: Away team shots
- HST: Home shots on target
- AST: Away shots on target
- HF: Home team fouls committed
- AF: Away team fouls committed
- HC: Home team corners
- AC: Away team corners
- HY: Home team yellow cards
- AY: Away team yellow cards
- HR: Home team red cards
- AR: Away team red cards

Each csv file contains 23 columns and 380 rows.

### 4.1.2 Data Cleaning

The first step of the pre-processing phase was data cleansing. The dataset did not need many actions, because there were not many missing values. Every season was a separate dataset in csv format. Each file contains 23 columns and 380 rows. In season 2014-2015, there was one match more than it should have. It had an invalid row, and this specific row was deleted. Some rows were deleted because they were not important for our purpose. Specifically, the columns "Div", "Date" and "Referee" were erased. Then, the 10 csv files were concatenated into a dataframe. The new dataset contained 20 columns and 3800 rows, as it is demonstrated in figure 16.



```
In [7]: playing_stat.head()
Out[7]:
```

| | HomeTeam | AwayTeam | FTHG | FTAG | FTR | HTHG | HTAG | HTR | HS | AS | HST | AST | HF | AF | HC | AC | HY | AY | HR | AR |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Aston Villa | Wigan | 0 | 2 | A | 0 | 1 | A | 11 | 14 | 5 | 7 | 15 | 14 | 4 | 6 | 2 | 2 | 0 | 0 |
| 1 | Blackburn | Man City | 0 | 2 | A | 0 | 1 | A | 17 | 8 | 9 | 5 | 12 | 9 | 5 | 4 | 2 | 1 | 0 | 0 |
| 2 | Bolton | Sunderland | 0 | 1 | A | 0 | 1 | A | 11 | 20 | 3 | 13 | 16 | 10 | 4 | 7 | 2 | 1 | 0 | 0 |
| 3 | Chelsea | Hull | 2 | 1 | H | 1 | 1 | D | 26 | 7 | 12 | 3 | 13 | 15 | 12 | 4 | 1 | 2 | 0 | 0 |
| 4 | Everton | Arsenal | 1 | 6 | A | 0 | 3 | A | 8 | 15 | 5 | 9 | 11 | 13 | 4 | 9 | 0 | 0 | 0 | 0 |

```
In [8]: playing_stat.shape
Out[8]: (3800, 20)
```

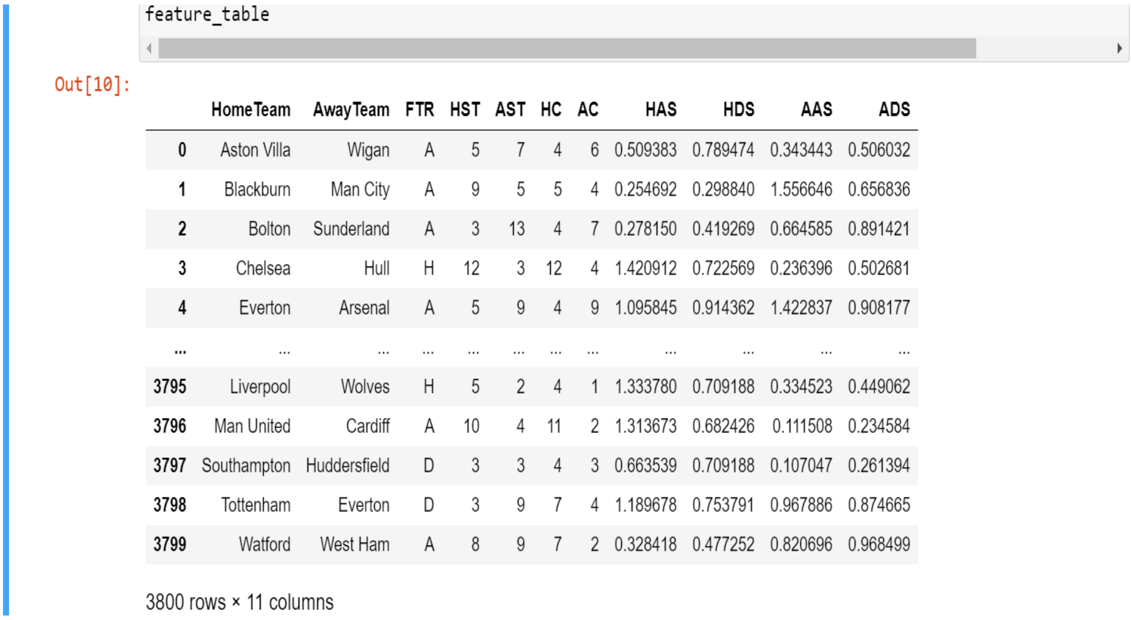
Figure 16: Dataset example and dataset shape.

### 4.1.3 Feature Extraction and Feature Engineering

The table was not very informative, so we needed to extract some features from it, relating to the offensive and the defensive capabilities of the home and the away team.

Firstly, we extracted some features for the teams per season. The new features were:

- HAS: Home attacking strength
- HDS: Home defensive strength
- AAS: Away attacking strength
- ADS: Away defensive strength



```
feature_table
```

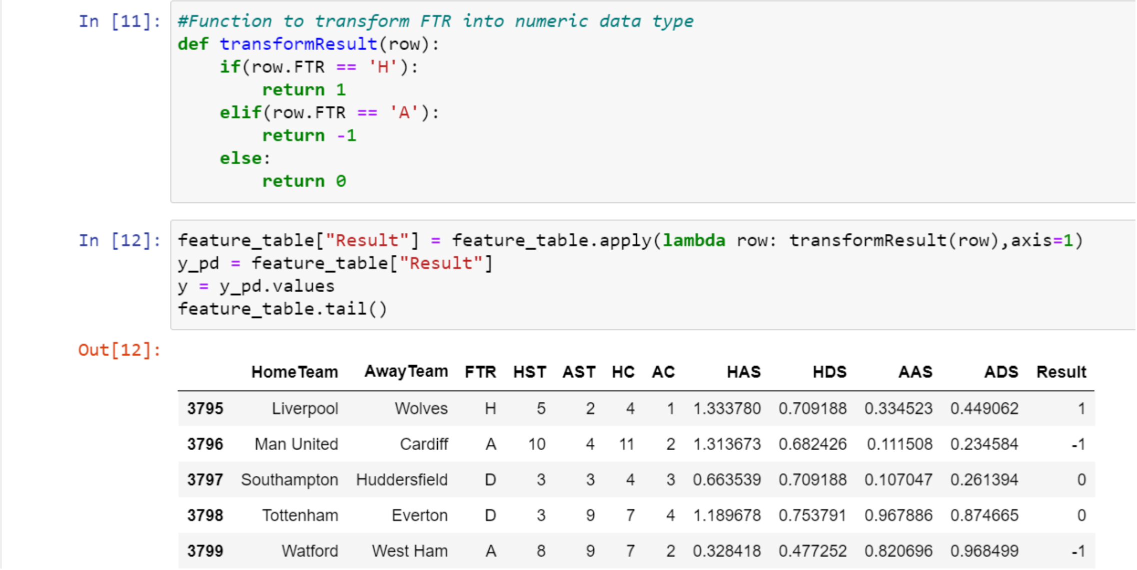
Out[10]:

| | HomeTeam | AwayTeam | FTR | HST | AST | HC | AC | HAS | HDS | AAS | ADS |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | Aston Villa | Wigan | A | 5 | 7 | 4 | 6 | 0.509383 | 0.789474 | 0.343443 | 0.506032 |
| **1** | Blackburn | Man City | A | 9 | 5 | 5 | 4 | 0.254692 | 0.298840 | 1.556646 | 0.656836 |
| **2** | Bolton | Sunderland | A | 3 | 13 | 4 | 7 | 0.278150 | 0.419269 | 0.664585 | 0.891421 |
| **3** | Chelsea | Hull | H | 12 | 3 | 12 | 4 | 1.420912 | 0.722569 | 0.236396 | 0.502681 |
| **4** | Everton | Arsenal | A | 5 | 9 | 4 | 9 | 1.095845 | 0.914362 | 1.422837 | 0.908177 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **3795** | Liverpool | Wolves | H | 5 | 2 | 4 | 1 | 1.333780 | 0.709188 | 0.334523 | 0.449062 |
| **3796** | Man United | Cardiff | A | 10 | 4 | 11 | 2 | 1.313673 | 0.682426 | 0.111508 | 0.234584 |
| **3797** | Southampton | Huddersfield | D | 3 | 3 | 4 | 3 | 0.663539 | 0.709188 | 0.107047 | 0.261394 |
| **3798** | Tottenham | Everton | D | 3 | 9 | 7 | 4 | 1.189678 | 0.753791 | 0.967886 | 0.874665 |
| **3799** | Watford | West Ham | A | 8 | 9 | 7 | 2 | 0.328418 | 0.477252 | 0.820696 | 0.968499 |

3800 rows × 11 columns

Figure 17: Table example after initial feature extraction.

Then, the full-time result that was our classification target, was transformed into numeric data type. Specifically, we transformed a home win into a 1, a home defeat into a -1 and a draw into a 0.



In [11]:
```
#Function to transform FTR into numeric data type
def transformResult(row):
    if(row.FTR == 'H'):
        return 1
    elif(row.FTR == 'A'):
        return -1
    else:
        return 0
```

In [12]:
```
feature_table["Result"] = feature_table.apply(lambda row: transformResult(row),axis=1)
y_pd = feature_table["Result"]
y = y_pd.values
feature_table.tail()
```

Out[12]:

| | HomeTeam | AwayTeam | FTR | HST | AST | HC | AC | HAS | HDS | AAS | ADS | Result |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **3795** | Liverpool | Wolves | H | 5 | 2 | 4 | 1 | 1.333780 | 0.709188 | 0.334523 | 0.449062 | 1 |
| **3796** | Man United | Cardiff | A | 10 | 4 | 11 | 2 | 1.313673 | 0.682426 | 0.111508 | 0.234584 | -1 |
| **3797** | Southampton | Huddersfield | D | 3 | 3 | 4 | 3 | 0.663539 | 0.709188 | 0.107047 | 0.261394 | 0 |
| **3798** | Tottenham | Everton | D | 3 | 9 | 7 | 4 | 1.189678 | 0.753791 | 0.967886 | 0.874665 | 0 |
| **3799** | Watford | West Ham | A | 8 | 9 | 7 | 2 | 0.328418 | 0.477252 | 0.820696 | 0.968499 | -1 |

Figure 18: Transforming full-time results into numeric data.

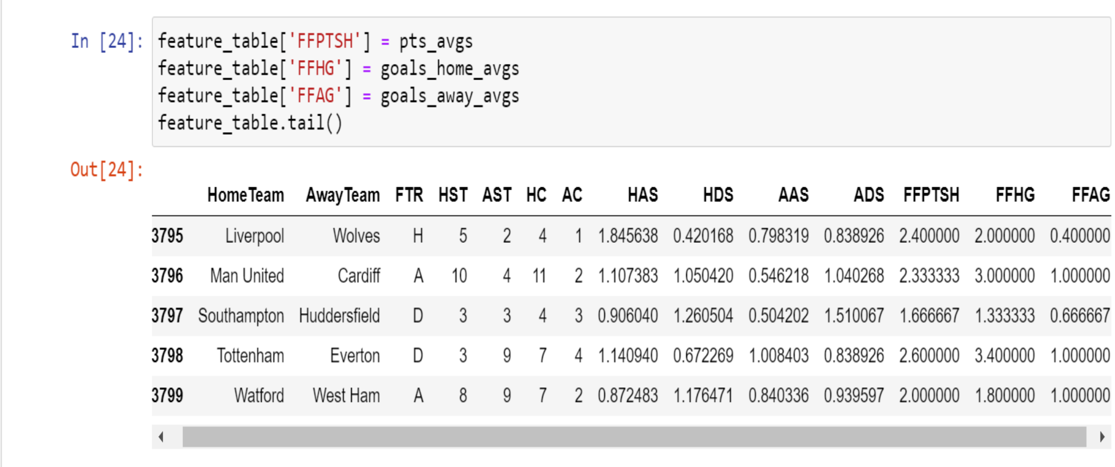
Because these features considered a long time of period (season long) and did not take into consideration the head-to-head results between teams, it needed to extract some features that were related to the head-to-head results. For example, Manchester City was a better team than Sunderland, but Sunderland used to win Manchester City in the English Premier League. The features that we created are the following:

- FFHG: Average goals of the home team in the last n matches against the away team.
- FFAG: Average goals of the away team in the last n matches against the home team.
- FFPTSH: Average points of the home team in the past n matches against the away team.

At this phase, we had to deal with a big issue. The problem was the case that there were not n matches between the teams prior to a given match. The problem was split into two cases.

The first case was that there are no prior matches between the two teams. This usually happened because teams are relegated or promoted each season. So, it is possible a team which has never been promoted again, to be promoted and play matches with an English Premier League stable team, so no previous matches will be available. Furthermore, the first matches of the season that are included in this dissertation (season 2009/2010) have no prior matches between teams. Therefore, we solved this problem by filling each of the features with mean values. More specifically, instead of utilizing average goals of home team in the past n matches against the away team, we used the average goals of home teams across all seasons. We followed the same procedure with average away goals and average home points.

The second case was that there are m prior matches between two teams, but m < n. To solve this problem, we used the average across these m matches.



In [24]:

```
feature_table['FFPTSH'] = pts_avgs
feature_table['FFHG'] = goals_home_avgs
feature_table['FFAG'] = goals_away_avgs
feature_table.tail()
```

Out[24]:

| | HomeTeam | AwayTeam | FTR | HST | AST | HC | AC | HAS | HDS | AAS | ADS | FFPTSH | FFHG | FFAG |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **3795** | Liverpool | Wolves | H | 5 | 2 | 4 | 1 | 1.845638 | 0.420168 | 0.798319 | 0.838926 | 2.400000 | 2.000000 | 0.400000 |
| **3796** | Man United | Cardiff | A | 10 | 4 | 11 | 2 | 1.107383 | 1.050420 | 0.546218 | 1.040268 | 2.333333 | 3.000000 | 1.000000 |
| **3797** | Southampton | Huddersfield | D | 3 | 3 | 4 | 3 | 0.906040 | 1.260504 | 0.504202 | 1.510067 | 1.666667 | 1.333333 | 0.666667 |
| **3798** | Tottenham | Everton | D | 3 | 9 | 7 | 4 | 1.140940 | 0.672269 | 1.008403 | 0.838926 | 2.600000 | 3.400000 | 1.000000 |
| **3799** | Watford | West Ham | A | 8 | 9 | 7 | 2 | 0.872483 | 1.176471 | 0.840336 | 0.939597 | 2.000000 | 1.800000 | 1.000000 |

Figure 19: Table with averages.

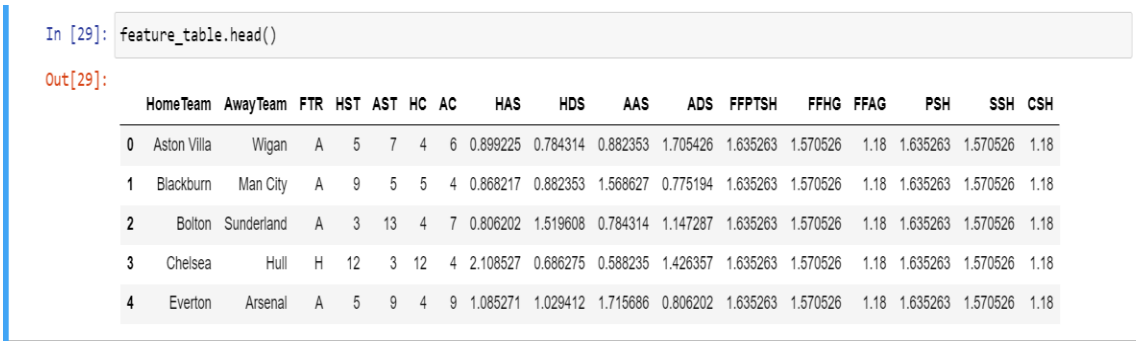
Then, the next step was extracting other features relating to form. Team form is one of the most famous metrics used to measure a team’s recent condition and performance. Taking the last 5 league matches into account is the most widely used and accepted method to measure a team’s form. Actually, the official statistics provided by English Premier League use this way of defining form [70]. Good form is that a team has been winning many matches lately. If a team has been getting good results lately, it is possible to have a good result in the upcoming match.

The features were related to home team form and away team form. In order to create them, we took into account the average points and goals of home team in the last n games and average points and goals of away team in the last n games. Missing or incomplete data were handled in the same way as missing or incomplete data in the past features.

The feature relating to form that we created were:

- PSH: Points streak for the home team
- SSH: Goals scored streak for the home team
- CSH: Goals conceded streak for the home team
- PSA: Points streak for the away team
- SSA: Goals scored streak for the away team

- CSA: Goals conceded streak for the away team

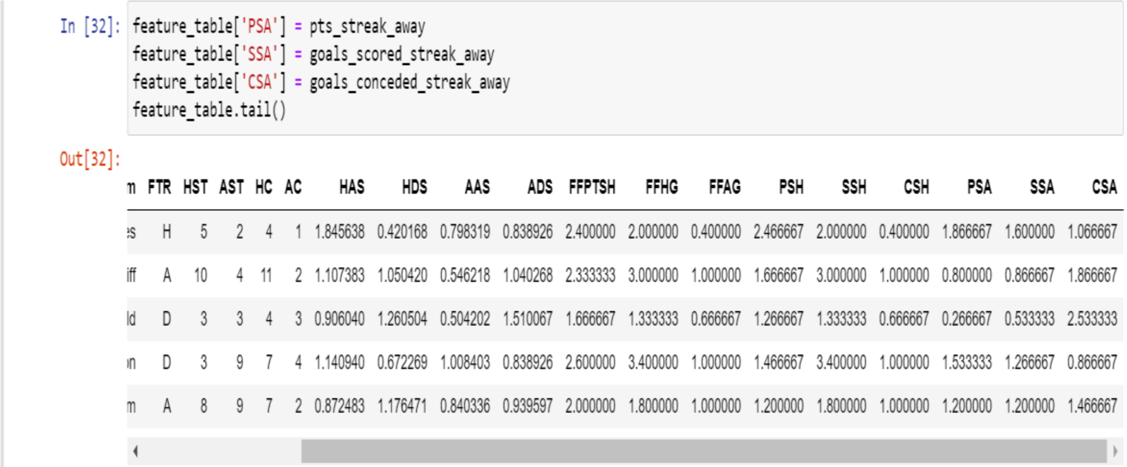


```
In [29]: feature_table.head()
Out[29]:
```

| | HomeTeam | AwayTeam | FTR | HST | AST | HC | AC | HAS | HDS | AAS | ADS | FFPTSH | FFHG | FFAG | PSH | SSH | CSH |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Aston Villa | Wigan | A | 5 | 7 | 4 | 6 | 0.899225 | 0.784314 | 0.882353 | 1.705426 | 1.635263 | 1.570526 | 1.18 | 1.635263 | 1.570526 | 1.18 |
| 1 | Blackburn | Man City | A | 9 | 5 | 5 | 4 | 0.868217 | 0.882353 | 1.568627 | 0.775194 | 1.635263 | 1.570526 | 1.18 | 1.635263 | 1.570526 | 1.18 |
| 2 | Bolton | Sunderland | A | 3 | 13 | 4 | 7 | 0.806202 | 1.519608 | 0.784314 | 1.147287 | 1.635263 | 1.570526 | 1.18 | 1.635263 | 1.570526 | 1.18 |
| 3 | Chelsea | Hull | H | 12 | 3 | 12 | 4 | 2.108527 | 0.686275 | 0.588235 | 1.426357 | 1.635263 | 1.570526 | 1.18 | 1.635263 | 1.570526 | 1.18 |
| 4 | Everton | Arsenal | A | 5 | 9 | 4 | 9 | 1.085271 | 1.029412 | 1.715686 | 0.806202 | 1.635263 | 1.570526 | 1.18 | 1.635263 | 1.570526 | 1.18 |

Figure 20: First 5 rows of the table with the home team form features.

In Figure 20, the first 5 rows with the home team form features are demonstrated. The features contain mean values as it was described above for missing or incomplete previous matches data.



```
In [32]: feature_table['PSA'] = pts_streak_away
         feature_table['SSA'] = goals_scored_streak_away
         feature_table['CSA'] = goals_conceded_streak_away
         feature_table.tail()
Out[32]:
```

| n | FTR | HST | AST | HC | AC | HAS | HDS | AAS | ADS | FFPTSH | FFHG | FFAG | PSH | SSH | CSH | PSA | SSA | CSA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| :s | H | 5 | 2 | 4 | 1 | 1.845638 | 0.420168 | 0.798319 | 0.838926 | 2.400000 | 2.000000 | 0.400000 | 2.466667 | 2.000000 | 0.400000 | 1.866667 | 1.600000 | 1.066667 |
| iff | A | 10 | 4 | 11 | 2 | 1.107383 | 1.050420 | 0.546218 | 1.040268 | 2.333333 | 3.000000 | 1.000000 | 1.666667 | 3.000000 | 1.000000 | 0.800000 | 0.866667 | 1.866667 |
| ld | D | 3 | 3 | 4 | 3 | 0.906040 | 1.260504 | 0.504202 | 1.510067 | 1.666667 | 1.333333 | 0.666667 | 1.266667 | 1.333333 | 0.666667 | 0.266667 | 0.533333 | 2.533333 |
| n | D | 3 | 9 | 7 | 4 | 1.140940 | 0.672269 | 1.008403 | 0.838926 | 2.600000 | 3.400000 | 1.000000 | 1.466667 | 3.400000 | 1.000000 | 1.533333 | 1.266667 | 0.866667 |
| m | A | 8 | 9 | 7 | 2 | 0.872483 | 1.176471 | 0.840336 | 0.939597 | 2.000000 | 1.800000 | 1.000000 | 1.200000 | 1.800000 | 1.000000 | 1.200000 | 1.200000 | 1.466667 |

Figure 21: Table with features related to form both for home and away team.

# 5 Prediction Models

After pre-processing phase, the data were ready to be used in predictive models. Three Machine Learning models were developed and evaluated. More specific, the Machine Learning models were Logistic Regression, XGBoost Classifier and Support Vector Classifier.

## 5.1 Explanatory Variables

All models developed were utilizing the same variables to train and make predictions. Table 1 demonstrates these variables below:

| Variable Name | Description |
|---|---|
| **HAS** | Home team attacking strength |
| **HDS** | Home team defensive strength |
| **AAS** | Away team attacking strength |
| **ADS** | Away team defensive strength |
| **FFPTSH** | Average points of the home team in the past n matches against the away team |
| **FFHG** | Average goals of the home team in the last n matches against the away team |
| **FFAG** | Average goals of the away team in the last n matches against the home team |
| **PSH** | Points streak for the home team |
| **SSH** | Goals scored streak for the home team |
| **CSH** | Goals conceded streak for the home team |

| | |
|---|---|
| **PSA** | Points streak for the away team |
| **SSA** | Goals scored streak for the away team |
| **CSA** | Goals conceded streak for the away team |

Table 1: Explanatory Variables.

## 5.2 Classification Models

XGBoost Classifier, Logistic Regression, and Support Vector Classifier were the 3 classification models that were used to predict the matches outcomes.

### 5.2.1 Boosting with XGBoost

In Machine Learning, boosting is one of the most famous terms. Boosting operates by sequentially applying a classification algorithm to reweighted parts of the training data and taking a weighted majority vote of the sequence of classifiers therefore produced. Gradient Boosting algorithms focus on the difference between predictions and ground truth. They use a differential loss function [71].

The XGBoost Classifier gets as an input of independent variables and provides an outcome in the form of a class. In our case, predicting the result of the match, there are 3 classes which represent all possible outcomes. We trained the algorithm by splitting the data into training and validation set. The validation set was the matches in season 2017/2018 and the training set was the matches from season 2009/2010 to 2016/2017.

### 5.2.2 Hyperparameter Tuning

XGBoost API gives the ability to the user to adjust a specific number of parameters. These parameters are defined prior to the training process and their values used to handle the learning process. The are several options for tuning these parameters. Grid Search is one of the most widely used methods.

### 5.2.3 Grid Search Cross Validation

This method is quite simple to be understood and interpreted. Let us assume that the parameters that need tuning are P = ($P_1$, $P_2$, …, $P_n$). An easy way to execute Grid Search is to create a vector of lower bounds l and one of upper bounds u for each P. Grid Search will produce $n^m$ grid points to check in the form of ($a_i$, $b_i$). Then, the grid point with the best outcomes in terms of measures defined by the user is kept [72].

Grid Search validates the performance of the grid points using k-fold Cross Validation. It is a method that evaluates Machine Learning models completing the following procedure:

- Randomly shuffles the dataset.
- Divides the data in k folders. K is defined by the user.
- Each unique group is going to set to be the test set, and the rest groups used as training set.

The number of folds is usually 10 as it has been able to produce models with low bias [73].

The test set that was used for predictions was 2018/2019. The results of the grid search are shown in the table 2 below:

| Hyperparameter | Season 2018/2019 results |
|---|---|
| **N Estimators** | 30 |
| **Max Depth** | 3 |
| **Min Child Weight** | 1 |
| **Colsample by tree** | 0.9 |
| **Learning rate (eta)** | 0.1 |

Table 2: Grid Search results for XGBoost.

### 5.2.4 Logistic Regression

Logistic regression is a statistical classification technique. This method gives the capability to model the relationship between dependent and independent variables. It can be also treated as a special issue of linear regression models. Nonetheless, the binary response variable contravenes regression model’s normal assumptions. A logistic regression model indicates that a suitable function of the fitted probability of the occurrence is a linear function of the detected values of the available variables. The leading benefit of this method is that it can generate an uncomplicated probabilistic blueprint of classification. The disadvantages are that logistic regression does not have the ability to solve issues and problems of nonlinear and interactive aftermaths of explanatory variables. It is a regression technique for forecasting a dichotomous dependent variable. Moreover, it is suitable for circumstances in which you desire to manage to predict the presence or absence of a result depend on values of set of predictor variables. Furthermore, this model can comprise the major effects and interaction terms. A significant phase in the procedure of modeling a dataset is deciding if confounder and interaction term appear in the data. The word confounder interprets a covariate that is related to the dependent variable of significance and the primary independent variable. In case of both variables are related then the relationship between them is called to be confounded [74] [75] [76].

### 5.2.5 Hyperparameter Tuning

The results of the Grid Search for the Logistic Regression classifier are shown in the table 3 below:

| Hyperparameter | Season 2018/2019 results |
|---|---|
| **C** | 10 |
| **Penalty** | l2 |

Table 3: Grid Search results for Logistic Regression.

### 5.2.6 Support Vector Machines

In Machine Learning, Support Vector Machines are supervised learning models with associated learning algorithms that analyze data both for classification and regression. For example, given a set of training samples, each belongs to one of two categories, a Support Vector Machine algorithm constructs a model that assigns new data to one or the other category. Furthermore, Support Vector Machines can perform a nonlinear classification using “kernel” which mapping the inputs into high dimensional feature spaces [77].

### 5.2.7 Hyperparameter Tuning

The results of the Grid Search for Support Vector Classifier are presented in the table 4 below:

| Hyperparameter | Season 2018/2019 results |
|---|---|
| **C** | 10 |
| **Gamma** | 0.01 |

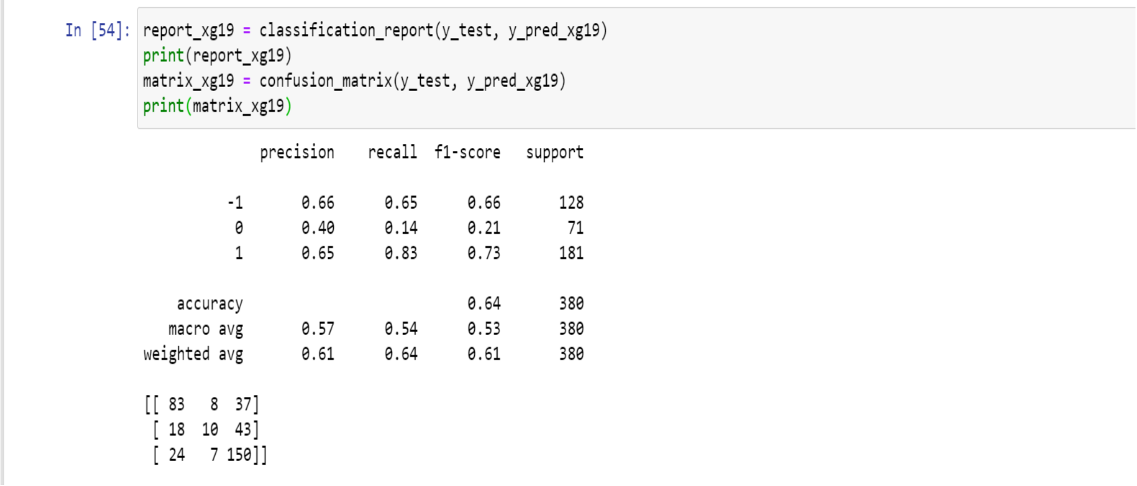
Table 4: Grid Search results for Support Vector Classifier.

# 6 Results and Discussion

Each model was run and evaluated based on its performance on the English Premier League season 2018/2019, which was the test set. Accuracy was the metric that is used for the performance comparison. Accuracy means the ratio of correct to incorrect predictions in 380 football matches' outcomes.

## 6.1 XGBoost Classifier

In this part of the dissertation, the performance of the XGBoost Classifier is demonstrated. As we can see in figure 22, this classifier achieved 64% in terms of accuracy in season 2018/2019. In addition, in class "-1", which was the "away win" class, it has 66% in terms of F1-Score, in class "0" which was the "draw" class, it has 21% and in class "1" which was the "home win" class, it has 73%. We can also infer both from the F1-score metric and the confusion matrix that the XGBoost classifier was less powerful in predicting the draws than the other two classes. However, it was quite foreseeable because the draw outcome is considered as the most difficult outcome to predict in a football game and for this reason, this outcome constitutes usually a high-reward odd for betting players. The model was quite good in "home teams wins" class by correctly classifying 150 from 181 wins.



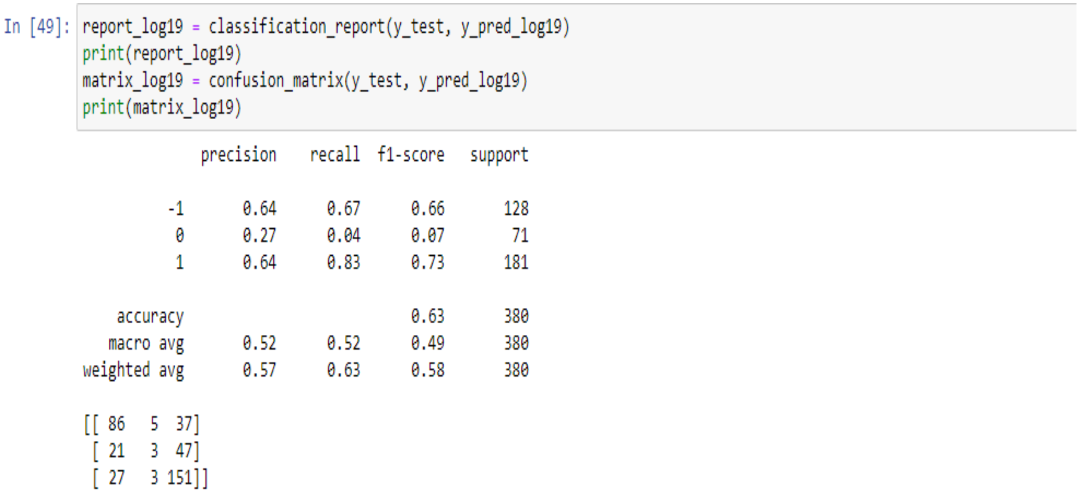
```
In [54]: report_xg19 = classification_report(y_test, y_pred_xg19)
         print(report_xg19)
         matrix_xg19 = confusion_matrix(y_test, y_pred_xg19)
         print(matrix_xg19)
```

```
              precision    recall  f1-score   support

          -1       0.66      0.65      0.66       128
           0       0.40      0.14      0.21        71
           1       0.65      0.83      0.73       181

    accuracy                           0.64       380
   macro avg       0.57      0.54      0.53       380
weighted avg       0.61      0.64      0.61       380

[[ 83   8  37]
 [ 18  10  43]
 [ 24   7 150]]
```

Figure 22: Main metrics results and confusion matrix for XGBoost Classifier.

## 6.2 Logistic Regression

As for XGBoost Classifier, after tuning the parameters for Logistic Regression classifier using Grid Search, the model tested in 2018/2019 season. In figure 23, it is illustrated the performance of this classifier. Specifically, this model achieved 63% in terms of accuracy in 380 matches. Furthermore, classification report shows a difficulty in predicting draws as the XGBoost Classifier. It is also a bit more accurate predicting the home teams' wins than the away teams' wins. In terms of F1-Score, it shows 66% in "away win" class, 73% in "home team" class and 7% in "draw" class.

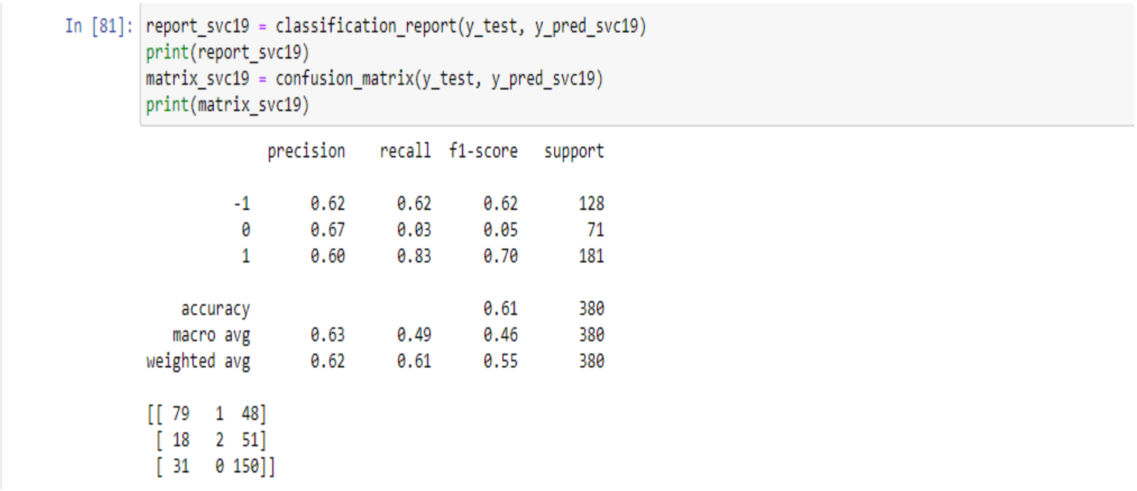


```
In [49]: report_log19 = classification_report(y_test, y_pred_log19)
         print(report_log19)
         matrix_log19 = confusion_matrix(y_test, y_pred_log19)
         print(matrix_log19)

                       precision    recall  f1-score   support

                  -1       0.64      0.67      0.66       128
                   0       0.27      0.04      0.07        71
                   1       0.64      0.83      0.73       181

            accuracy                           0.63       380
           macro avg       0.52      0.52      0.49       380
        weighted avg       0.57      0.63      0.58       380

        [[ 86   5  37]
         [ 21   3  47]
         [ 27   3 151]]
```

Figure 23: Main metrics results and confusion matrix for Logistic Regression.

## 6.3 Support Vector Classifier

In figure 24, it is demonstrated the performance of the Support Vector Classifier. The model achieved 61% in terms of accuracy. Unfortunately, even after hyperparameter tuning, there was no improvement. Even though, the overall accuracy score is adequate, the model is quite bad when it tried to classify the “draw” class. More specific, it only predicted 2 over 71 draw outcomes. Moreover, the model is satisfactory at home team wins and away team wins predictions, with 70% and 62% in terms of F1-Score, respectively.



```
report_svc19 = classification_report(y_test, y_pred_svc19)
print(report_svc19)
matrix_svc19 = confusion_matrix(y_test, y_pred_svc19)
print(matrix_svc19)
```

```
              precision    recall  f1-score   support

          -1       0.62      0.62      0.62       128
           0       0.67      0.03      0.05        71
           1       0.60      0.83      0.70       181

    accuracy                           0.61       380
   macro avg       0.63      0.49      0.46       380
weighted avg       0.62      0.61      0.55       380

[[ 79   1  48]
 [ 18   2  51]
 [ 31   0 150]]
```

Figure 24: Main metrics results and confusion matrix for Support Vector Classifier.

## 6.4 Comparative Results for Classifiers

It can be inferred that the XGBoost Classifier slightly outperforms the other two models. It achieved 64% accuracy, whilst SVC and Logistic Regression had 61% and 63% accuracy, respectively. Additionally, XGBoost outplayed the other two classifiers in predicting the draws, which was the major issue that we had to deal with. It is worth mentioned that after Grid Search Cross Validation, the classifiers had better performance except for Support Vector Classifier. Tables 5 and 6 are representative of the comparative results in outcome predictions of the XGBoost Classifier, the Logistic Regression model, and the Support Vector Classifier.

| **Classifier** | **Accuracy (%)** |
|---|---|
| **Logistic Regression** | 63 |
| **XGBoost Classifier** | 64 |
| **Support Vector Classifier** | 61 |

Table 5: Accuracy of classifiers

| Classifier/F1-Score (%) | Class -1 (Away) | Class 0 (Draw) | Class 1 (Home) |
|---|---|---|---|
| **XGBoost Classifier** | 66 | 21 | 73 |
| **Logistic Regression** | 66 | 7 | 73 |
| **Support Vector Classifier** | 62 | 5 | 70 |

Table 6: F-1 Score of models for each class

In table 6, it is demonstrated that XGBoost Classifier is the best classifier for match outcomes predictions. More specifically, in Class "-1" and Class "1", XGBoost Classifier and Logistic Regression have the same results in terms of F1-Score. Support Vector Classifier has slightly worse results than the other models. In Class "0", even though the XGBoost Classifier outperforms the other models, it achieves quite poor performance.

XGBoost model is the best of the three models in predicting the season 2018/2019 English Premier League outcomes.

## 6.5 Evaluation of Results

In this section, problems that came up during the whole process and solutions given are discussed. The project result is evaluated and threats to validity as well.

One of the main challenges encountered during the project was to select appropriate data to use to build a predictive model. There is abundance of public data regarding football. The selection of the dataset was crucial because it was impossible to utilize every free online data due to the computational power that would be needed. Fortunately, the dataset contained reliable and useful information about English Premier League games for 10 seasons. During feature selection step, some insignificant and useless for the models attributes were eliminated such as the referee name, the division, and the half-time result.

On the other hand, gathering advanced football statistics and data about players from wearable devices, which might improve the model performance, is very difficult because there are not available online. Furthermore, the financial data about teams and players have inaccuracies because teams do not confirm them.

Handling newly promoted teams was another problem. Regarding these teams, we assigned mean values to some variables. Even though, it was necessary for the project, some teams were underestimated from these average values.

Another problem was the difficulty in predicting draws. XGBoost was the best after tuning, but it still had a poor performance in predicting draws. The solution usually is to use class imbalance techniques. This solution was tested with slightly worse results. An extensive customization of classifiers regarding the dataset requirements, would probably give slightly better results. Nevertheless, this process might provide bad generalization to our model and surpass the scope of this dissertation.

There are several factors that decide a football game result. A result is not affected only by team's and player's ability. Unfortunately, we cannot predict some external factors. When a team plays a significant European match during the same week, the coach may perform a rotation which means that he will rest his best players. Another factor is luck because the best team in the pitch does not always win, especially in football. The manager of a team plays a significant role in the whole team processes and performance. In addition, injuries of important players is a factor that affects a game result. These factors can be considered as threats to validity.

The result of this project is quite promising. Accuracy level can be considered as more than satisfactory. In fact, model's accuracy score is higher than Chazan's best model for English Premier League [79]. Our model is able to predict 243 of 380 matches in English Premier League season 2018/2019. It can predict correctly 64% of the matches' outcomes in the whole season. Under certain circumstances, it can be used for betting reasons.

# 7 Conclusion and Future Work

## 7.1 Conclusions

In this research, possibly the most fundamental aspect of sports analytics was explored. In football, match outcome and by extension overall team performance, is the desideratum of football clubs, betting companies and fans.

The main objective of building a model by exploring several Machine Learning techniques has been accomplished. We used Machine Learning algorithms such as XGBoost Classifier, Logistic Regression and Support Vector Classifier in order to generate match outcome predictions. XGBoost Classifier produced good results, taking into consideration that a football game can be affected by many internal and external factors.

Concluding, sports analytics has already become a major part of a team’s performance as teams collect and analyze more efficient the huge amount of data generated by players through censors and other tools. More and more teams hire sport analysts and data scientist to handle and get useful insights from the data. For example, two of the biggest football clubs in the world, such as Barcelona and Liverpool, have hired a complete team comprising data scientists and sports analysts. Sports analytics has already been considered as a fast-growth industry.

## 7.2 Future Work

There are many directions in which this project can be led with more time and resources.

The model can be improved with more data regarding match events. For instance, possession, specific shots data, player formation and ratings, referee ratings, even manager ability, could be crucial for the model.

An interesting future extension is to apply Monte Carlo simulations to predict future events, such as the probability of a team winning the championship. Monte Carlo simulations run many times and depend on random sampling to produce predictions.

Additionally, team data, such as possession, number of crosses, number of headers, etc., could be utilized to categorize each team into separate categories of playing styles. Then, models could be constructed to understand the difference between the playing styles of two different teams and help the prediction of the outcome of a match between these teams.

For researchers who want to expand this research in the future, it is recommended to go deeper into the difficulty in predicting a draw, using common techniques, such as tuning of the classes' weights or following a different strategy.

Finally, a potential future extension could be the investigation of betting odds and whether our model could suggest good value bets and profitable betting strategies that are profitable over time.

# References

[1] Neyer, R., Sabermetrics | Statistics. [online] Encyclopedia Britannica. Available at: <https://www.britannica.com/sports/sabermetrics>.

[2] B. Rickey. Goodby to some old baseball ideas. *LIFE*, vol. 37, 2 August 1954.

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

[4] D. Leewood, "Moneyball," 2 October 2011. [Online]. Available: https://en.wikipedia.org/wiki/Moneyball

[5] James, Bill. The bill james baseball abstract 1987. Ballantine Books, 1987.

[6] NABC., Timeout Feature: The Early Days Of Basketball Analytics. [online] Available at: <https://www.nabc.com/nabc_releases/timeout_features/2016/timeout-analytics>.

[7] Cohen, B., 2017. How An Israeli Tech Startup Changed The NBA. [online] WSJ. Available at: <https://www.wsj.com/articles/the-israeli-tech-startup-that-changed-the-nba-1510068323>.

[8] Steinberg, L., 2015. CHANGING THE GAME: The Rise Of Sports Analytics. [online] Forbes. Available at: <https://www.forbes.com/sites/leighsteinberg/2015/08/18/changing-the-game-the-rise-of-sports-analytics/?sh=2d307f54c1fd>.

[9] Holmes, B., 2014. New Age Of NBA Analytics: Advantage Or Overload? - The Boston Globe. [online] BostonGlobe.com. Available at: <https://www3.bostonglobe.com/sports/2014/03/29/new-age-nba-analytics-advantage-overload/1gAim4yKYXGUQ2CTAe7iCO/story.html?arc404=true>.

[10] Slater, A., 2014. Outlook 2014: Innovative Sportvu Cameras Changing Statistical Analysis In The NBA. [online] Oklahoman.com. Available at: <https://oklahoman.com/article/4084024/outlook-2014-innovative-sportvu-cameras-changing-statistical-analysis-in-the-nba>.

[11] Douglass, B., 2013. How Sportvu Came To Be The NBA Analytics Game Changer. [online] Sporttechie.com. Available at: <https://www.sporttechie.com/sportvu-pulls-nba-analytics-forward/>.

[12] Larson, O., 2001. Charles Reep: A major influence on British and Norwegian football. Soccer & Society, 2(3), pp.58-78.

[13] Wilson, J., 2013. Inverting the pyramid: the history of soccer tactics. Bold Type Books.

[14] Soccerment.com. 2020. The Growing Importance Of Football Analytics | Soccerment. [online] Available at: <https://soccerment.com/the-importance-of-football-analytics/>.

[15] Betts, E., 2011. Formation Renovation: Rethinking The Long Ball. [online] The Other 87. Available at: <https://theother87.wordpress.com/2011/10/19/formation-renovation-rethinking-the-long-ball/>.

[16] Rathke, A., 2017. An examination of expected goals and shot efficiency in soccer. Journal of Human Sport and Exercise, 12(2), pp.514-529.

[17] Driblab.com. 2020. [online] Available at: <https://www.driblab.com/analysis-team/expected-goals-xg-what-it-is-and-how-it-works/>.

[18] M. J. Moroney. Facts from figures, 3rd edn. Penguin: London, 1956.

[19] C. Reep. Skill and chance in ball games. Journal of the Royal Statistical Society Series A 131: 581-585, 1971.

[20] M. J. Maher. Modelling association football scores. Statistica Neerlandica, 1982.

[21] M.J. Dixon, S.C. Coles. Modelling association football scores and inefficiencies in the football betting market. Applied Statistics, 1997.

[22] H. Rue, O. Salvesen. Prediction and retrospective analysis of soccer matches in a league. Statistician, 2000

[23] Forrest, D. and Simmons, R., 2000. Forecasting sport: the behaviour and performance of football tipsters. International Journal of Forecasting, 16(3), pp.317-331.

[24] Kuypers, T., 2000. Information and efficiency: an empirical study of a fixed odds betting market. Applied Economics, 32(11), pp.1353-1363.

[25] Goddard, J., 2005. Regression models for forecasting goals and match results in association football. International Journal of forecasting, 21(2), pp.331-340.

[26] Joseph, A., Fenton, N. and Neil, M., 2006. Predicting football results using Bayesian nets and other machine learning techniques. Knowledge-Based Systems, 19(7), pp.544-553

[27] Hucaljuk, J. and Rakipović, A., 2011, May. Predicting football scores using machine learning techniques. In 2011 Proceedings of the 34th International Convention MIPRO (pp. 1623-1627). IEEE.

[28] Tavakol, M., Zafartavanaelmi, H. and Brefeld, U., 2016. Feature Extraction and Aggregation for Predicting the EURO 2016. In MLSA@ PKDD/ECML.

[29] Gu, W., Foster, K., Shang, J. and Wei, L., 2019. A game-predicting expert system using big data and machine learning. Expert Systems with Applications, 130, pp.293-305

[30] Elo, A.E., 1978. The rating of chessplayers, past and present. Arco Pub..

[31] Buchdahl, J., 2003. Fixed odds sports betting: Statistical forecasting and risk management. Summersdale Publishers LTD-ROW.

[32] Hvattum, L.M. and Arntzen, H., 2010. Using ELO ratings for match result prediction in association football. International Journal of forecasting, 26(3), pp.460-470.

[33] Lasek, J., 2016. EURO 2016 Predictions Using Team Rating Systems. In MLSA@ PKDD/ECML.

[34] Viswanadha, S., Sivalenka, K., Jhawar, M.G. and Pudi, V., 2017. Dynamic Winner Prediction in Twenty20 Cricket: Based on Relative Team Strengths. In MLSA@ PKDD/ECML (pp. 41-50).

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

[36] Lasek, J., Szlávik, Z. and Bhulai, S., 2013. The predictive power of ranking systems in association football. International Journal of Applied Pattern Recognition, 1(1), pp.27-46.

[37] Vecer, J., Kopriva, F. and Ichiba, T., 2009. Estimating the effect of the red card in soccer: when to commit an offense in exchange for preventing a goal opportunity. Journal of Quantitative Analysis in Sports, 5(1).

[38] Catteeuw, P., Gilis, B., Wagemans, J. and Helsen, W., 2010. Offside decision making of assistant referees in the English Premier League: Impact of physical and perceptual-cognitive factors on match performance. Journal of Sports Sciences, 28(5), pp.471-481.

[39] Gréhaigne, J.F., Grehaigne, J.F., Griffin, L.L. and Richard, J.F., 2005. Teaching and learning team sports and games. Psychology Press.

[40] Arabzad, S.M., Tayebi Araghi, M.E., Sadi-Nezhad, S. and Ghofrani, N., 2014. Football match results prediction using artificial neural networks; the case of Iran Pro League. Journal of Applied Research on Industrial Engineering, 1(3), pp.159-179.

[41] Constantinou, A.C., Fenton, N.E. and Neil, M., 2012. pi-football: A Bayesian network model for forecasting Association Football match outcomes. Knowledge-Based Systems, 36, pp.322-339.

[42] Joseph, A., Fenton, N.E. and Neil, M., 2006. Predicting football results using Bayesian nets and other machine learning techniques. Knowledge-Based Systems, 19(7), pp.544-553.

[43] Macdonald, B., 2012, March. An expected goals model for evaluating NHL teams and players. In Proceedings of the 2012 MIT Sloan Sports Analytics Conference, http://www.sloansportsconference. com.

[44] Lucey, P., Bialkowski, A., Monfort, M., Carr, P. and Matthews, I., 2014, February. quality vs quantity: Improved shot prediction in soccer using strategic features from spatiotemporal data. In Proc. 8th annual mit sloan sports analytics conference (pp. 1-9).

[45] Eggels, H., van Elk, R. and Pechenizkiy, M., 2016, September. Explaining Soccer Match Outcomes with Goal Scoring Opportunities Predictive Analytics. In MLSA@ PKDD/ECML.

[46] Chakrabarti, S., Ester, M., Fayyad, U., Gehrke, J., Han, J., Morishita, S., Piatetsky-Shapiro, G. and Wang, W., 2006. Data mining curriculum: A proposal (Version 1.0). Intensive Working Group of ACM SIGKDD Curriculum Committee, 140, pp.1-10.

[47] Clifton, C., Data Mining - Pattern Mining. [online] Encyclopedia Britannica. Available at: <https://www.britannica.com/technology/data-mining/Pattern-mining>.

[48] TWIN, A., 2020. Data Mining: How Companies Use Data To Find Useful Patterns And Trends. [online] Investopedia. Available at: <https://www.investopedia.com/terms/d/datamining.asp>.

[49] Katya. networthis. 2020. [online] Available at: <https://www.networthis.com/data-mining-process-tools-text-mining/?doing_wp_cron=1608216817.8412499427795410156250>.

[50] Ng, A., 2020. Machine Learning. [online] Coursera. Available at: <https://www.coursera.org/learn/machine-learning>.

[51] Mitchell, T.M., 2006. The discipline of machine learning. School of Computer Science.

[52] Mitchell, T.M., 1997. Machine Learning, volume 1 of 1.

[53] Pyle, D. and San José, C., 2015. [online] mckinsey.com. Available at: <https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/an-executives-guide-to-machine-learning#>.

[54] Reynoso, R., 2019. Understanding Artificial Intelligence And Machine Learning. [online] Learn.g2.com. Available at: <https://learn.g2.com/artificial-intelligence-and-machine-learning>.

[55] Shewan, D., 2019. 10 Companies Using Machine Learning In Cool Ways. [online] Wordstream.com. Available at: <https://www.wordstream.com/blog/ws/2017/07/28/machine-learning-applications>.

[56] Russell, S.J. and Norvig, P., 2010. Artificial Intelligence-A Modern Approach, Third International Edition.

[57] Mohri, M., Rostamizadeh, A. and Talwalkar, A., 2012. Foundations of machine learning.[Sl].

[58] TechVidvan. n.d. Supervised Learning Algorithm In Machine Learning - Techvidvan. [online] Available at: <https://techvidvan.com/tutorials/supervised-learning/>.

[59] Hinton, G.E., Sejnowski, T.J. and Poggio, T.A. eds., 1999. Unsupervised learning: foundations of neural computation. MIT press.

[60] TechNative. 2020. Why Unsupervised Machine Learning Is The Future Of Cybersecurity. [online] Available at: <https://www.technative.io/why-unsupervised-machine-learning-is-the-future-of-cybersecurity/>.

[61] Littman, M.L. and Moore, A.W., 1996. Reinforcement Learning: A Survey, Journal of Artificial Intelligence Research 4.

[62] TechVidvan. 2019. Reinforcement Learning Algorithms And Applications - Techvidvan. [online] Available at: <https://techvidvan.com/tutorials/reinforcement-learning/>.

[63] Nicholson, C., n.d. Machine Learning Algorithms. [online] Pathmind. Available at: <https://wiki.pathmind.com/machine-learning-algorithms>.

[64] Prashanth, M., 2020. Top 10 Best Machine Learning Algorithms. [online] Medium. Available at: <https://mahiprashanth866.medium.com/top-10-best-machine-learning-algorithms-acd9fd2c0471>.

[65] Holman, V., 2018. What Is Sports Analytics? Victor Holman Explains - Agile Sports Analytics. [online] Agile Sports Analytics. Available at: <https://www.agilesportsanalytics.com/what-is-sports-analytics/>.

[66] Ray, Sugato (June 22, 2017). "The Evolution and Future of Analytics in Sport". Proem Sports | Sports Analytics | Singapore & India.

[67] Information Management and Analytics Club, IMI New Delhi. 2018. Sports Analytics: Growing Prevalence In 2018. [online] Available at: <https://imacimi.wordpress.com/2018/10/30/sports-analytics-growing-prevalence-in-2018/>.

[68] Tanuka's Blog. n.d. Sports Analytics Have Changed The Way Sports Are Played - Tanuka's Blog. [online] Available at: <https://tanukamandal.com/2017/12/12/sports-analytics-changed-play/>.

[69] Willems, K., 2019. Jupyter Notebook Tutorial: The Definitive Guide. [online] datacamp.com. Available at: <https://www.datacamp.com/community/tutorials/tutorial-jupyter-notebook>.

[70] Premierleague.com. 2020. Premier League Table, Form Guide & Season Archives. [online] Available at: <https://www.premierleague.com/tables>.

[71] Zhang, Z., 2020. Boosting Algorithms Explained. [online] Medium. Available at: <https://towardsdatascience.com/boosting-algorithms-explained-d38f56ef3f30>.

[72] Dufour, J.M. and Neves, J., 2019. Finite-sample inference and nonstandard asymptotics with Monte Carlo tests and R. In Handbook of Statistics (Vol. 41, pp. 3-31). Elsevier.

[73] Brownlee, J., 2020. A Gentle Introduction To K-Fold Cross-Validation. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/k-fold-cross-validation/>.

[74] Hajmeer, M. and Basheer, I., 2003. Comparison of logistic regression and neural network-based classifiers for bacterial growth. Food Microbiology, 20(1), pp.43-55.

[75] Khemphila, A. and Boonjing, V., 2010, October. Comparing performances of logistic regression, decision trees, and neural networks for classifying heart disease patients. In 2010 international conference on computer information systems and industrial management applications (CISIM) (pp. 193-198). IEEE.

[76] Tsangaratos, P. and Ilia, I., 2016. Comparison of a logistic regression and Naïve Bayes classifier in landslide susceptibility assessments: The influence of models complexity and training dataset size. Catena, 145, pp.164-179.

[77] Cortes, C. and Vapnik, V., 1995. Support-vector networks. Machine learning, 20(3), pp.273-297.

[78] Apostolou K., Tjortjis C., 'Sports Analytics algorithms for performance prediction', IEEE 10th Int'l Conf. on Information, Intelligence, Systems and Applications (IISA 2019), pp. 469-472, 2019.

[79] V. Chazan–Pantzalis, C. Tjortjis, ' Sports Analytics for Football League Table and Player Performance Prediction', Proc. 11th IEEE Int’l Conf. on Information, Intelligence, Systems and Applications (IISA 20) 2020.

[80] V. Sarlis, C. Tjortjis, “Sports Analytics – Evaluation of Basketball Players and Team Performance”, Information Systems, 2020.