Sports Analytics algorithms for performance prediction

Konstantinos Apostolou

SID: 3308170001

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in Data Science

December 2018

THESSALONIKI – GREECE

Sports Analytics algorithms for performance prediction

Konstantinos Apostolou

SID: 3308170001

Supervisor:	Prof. Christos Tjortjis
Supervising Committee Members:	Dr. Christos Berberidis
	Dr. Agamemnon Baltagiannis

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in Data Science

DECEMBER 2018

THESSALONIKI – GREECE

Abstract

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

It is common knowledge, that the term “Sports Analytics” is used more and more nowadays. Sports Analytics can have various applications, that could affect a variety of fields. These could be for example, the prediction of an athlete’s or a team’s performance, the estimation of an athlete’s talent and market value and the prediction of a possible injury. More and more teams and coaches are willing to embed such “tools” in their training sessions, in order to improve their tactics.

This dissertation is divided in two major parts. The first one is a literature review of existing technologies on the subject. The second part focuses on the experiments that were conducted mainly with football data. In these experiments, by using suitable algorithms, a player’s position in the field could be predicted. By accumulating data from past years, we could have an estimation for a player’s goal scoring performance in the next season. Furthermore, in order to make it more specific, the number of a player’s shots can be predicted in each match, something that has a correlation with goal scoring possibility.

At this point, I would like to thank my Supervisor, Professor Christos Tjortjis, for all his help throughout the process, the feedback and suggestions he provided to the issues that occurred and his immediate responses.

Konstantinos Apostolou

7-12-2018

Contents

Table of figures

Chapter 1

Introduction

Sports analytics exist as a concept since many years, but there are a lot of steps that need to be done in order to understand and improve team performance. It is a topic that will gain interest increasingly in the next years. For the purposes of this project we will mostly focus on football (soccer). In order to achieve good accuracy, not only we need to find an accurate database but also the attributes of the database to be relevant to the research.

There are many ways that a team can use data and there are lots of kinds of them. First, they could be related to the way that the game is played (in this case football), not only by each team member but also by the team as a whole. This kind of data have to do with the players average stats such as the number of goals that they score, the number of fouls they commit, with how many red and yellow cards they are booked, how many tackle-ins they do, how many kilometers they run during a match (the use of cameras helps for these stats) and many more. However, it is difficult to compare all those stats in successive matches, because a player's performance depends on his opponents as well. In other words, a striker may score many goals playing versus a bad organized defense, or a bad goalkeeper. Moreover, there are data that give information about how a team manages to score a goal. For, example how many passes they can achieve before they score a goal, what is the average possession of the ball, and for both above facts a major role plays the part of the field that they occur. Obviously, it is different to keep the ball close to the opponent's goalpost. Sometimes, though, there are outliers that show sometimes that other factors play important roles as well, because teams win in football even if they have little ball possession.

So, it is mandatory to use data that are more difficult to collect. These data have to do with players physical condition such as pulse rate when being calm and when sprinting, measure sweatiness and also track a player's sleep. However, most of these data were impossible until recently to collect because these devices were not allowed in football games. It was in March 2015 that the use of EPTS (ELECTRONIC PERFORMANCE AND TRACKING SYSTEMS) was allowed and gave the opportunity for sports analysts to explore other aspects of the game.

There are also other aspects that affect a team’s performance such as weather conditions, the condition of the field and even psychological factors such as the fans support. Another factor is injuries which sometimes could be predicted or prevented.

Chapter 2

Historical Aspects
Football

Sports analytics is becoming more and more challenging nowadays. It is a concept that concerns sports clubs of most sports and even individual athletes. However, the notion of analyzing data and statistics is quite old before even using algorithm and data mining techniques.

The first attempt of applying analytics in a sports game was conducted by Thorold Charles Reep during the 1950 decade. Reep, after graduating from Plymouth Hight School in 1923, was educated as an accountant. He actually managed to win the first prize in a competition that was held in 1928. This competition was organized by Royal Air Force in 1928 for the Accountancy division from which he retired in 1955. [1]

In 1933, an Arsenal football player, Charlie Jones, was giving some lectures in which Reep had participated. Reep also was intrigued by the gaming way of Hebert Chapman which included different playing style of wingers. [2] However, because of the World War II, he was not back in England until 1947, when he realized that all that Chapman style game was not applied. While watching a football game at Swindon Town he was disappointed in the fact that the team could not score. So, in the second half he started to take notes about the game. He concluded that the team should slightly increase the scoring rate in order to be promoted.

A manager at Brentford was fascinated by Reep's work and thus, he was hired as an adviser. His goal there was to help the local team to avoid relegation. After his arrival the team easily gathered the needed points, so they managed to remain in the division [2].

In 1950 Reep introduced the long ball theory (Picture 1). Since he observed that goals were scored when less than three passes are made, he stated that the ball should reach the striker as soon as possible. If done so the more goals the team was going to score improving their rate.

During the years of working until 1967 Reep published a paper about statistical analysis of patterns of play together with Bernard Benjamin. They had data about football games and

specific tactics for 14 years. They concluded that all important moves were achieved with a small number of passes (less than 4).

Even though Reep had been questioned about his results, it is the first actual attempt of applying statistical knowledge to a sports game for improving performance.

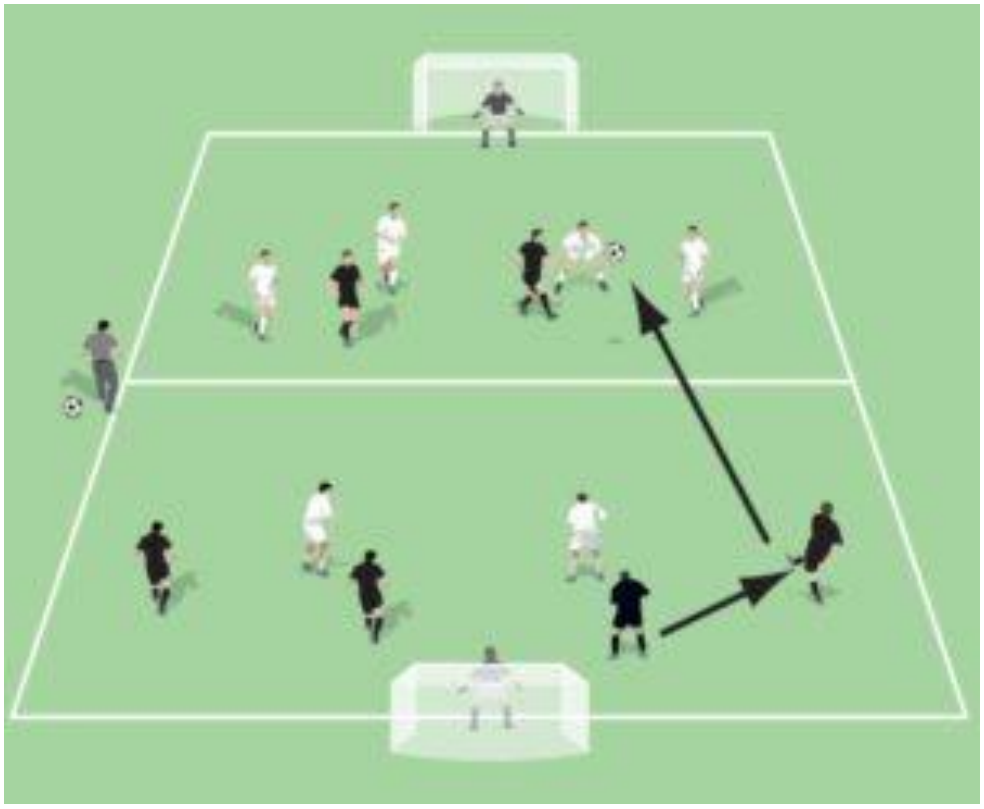


Figure 1 [3]

Nowadays, more sophisticated methods are used in football and in other sports in general, for data mining and decision making based on data. However, we should not forget that Reep was one of the pioneers at this field that was going to become a hot topic in our decade.

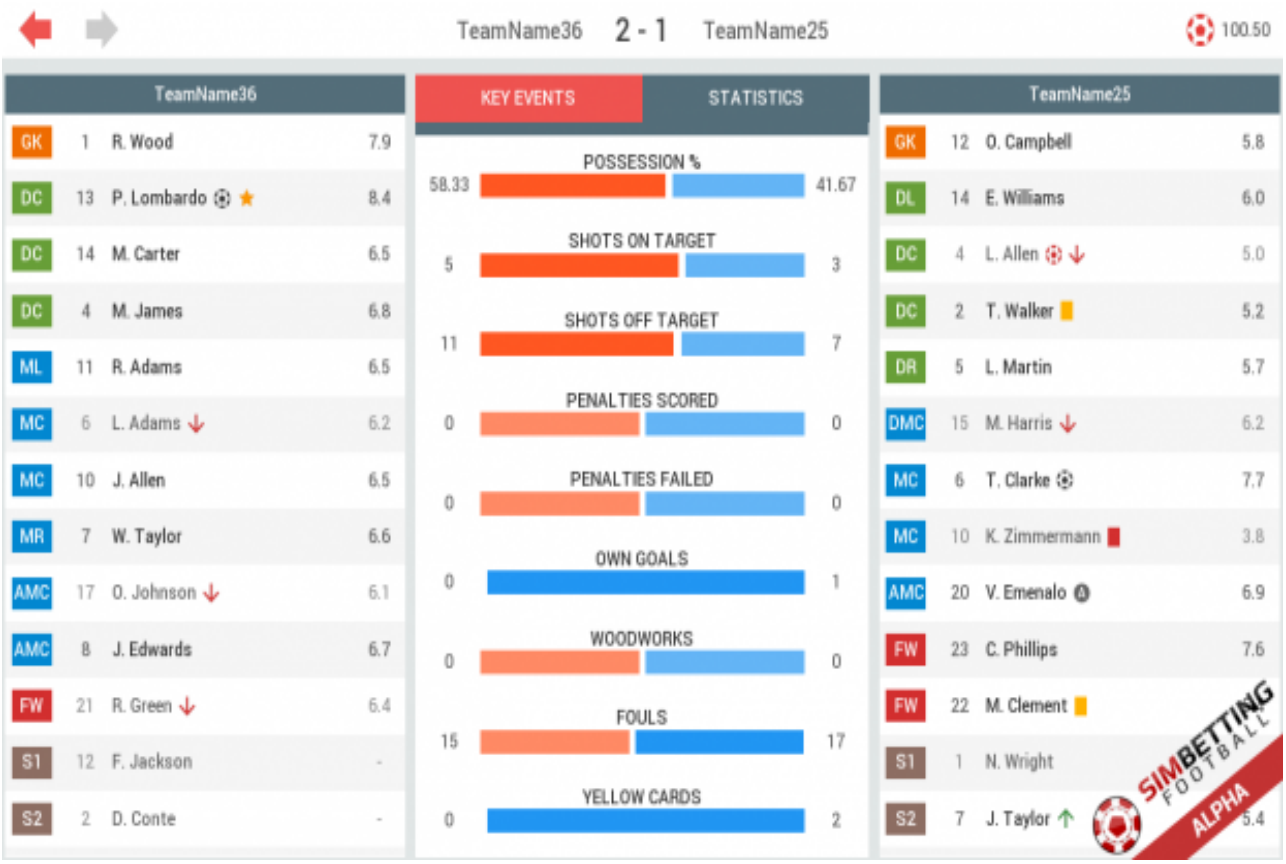


Figure 2 -- statistics in football [22]

German National Football Team

In summer 2014, there was the World Cup organization. In the semifinals, the interest was of course on the match of Germany vs Brazil. Germany managed to win emphatically with 7-1 something that was never achieved before against this opponent.

As it was later pointed out by a German assistant coach named Hansi Flick, he had spent two years studying the Brazilian football players. He managed to gather a lot of data and thus his team was able to achieve such a victory.

It was the first time in the World Cup History that something like that was admitted publicly, however we can't be sure that we would have a different outcome if nothing of this data analysis occurred.

Baseball

Another example of a sports analytics attempt was by George William James who was more involved with baseball. He has written books about baseball history and Statistics. [4]

His goal was to understand and point out why a team could win or lose using statistics. His work was named sabermetrics, something that himself came up with.

At the age of around twenty-five, he decided to stop working at the United States Army and started exploiting his passion for baseball, a sport that he was an avid fan of. He was writing reports about baseball games following a different approach than this that others were, at the time being. At first, he had difficulties to find editors to publish his articles, believing that the audience will not respond. So, he had to reach the audience using a different way, which led him to his decision to publish an annual book titled: "The Bill James Baseball Abstract". As a result, he managed to increase his reputation. [4]

James had tried to convince Major League Baseball to publish stats about each game. However, they refused to provide such information. For this purpose, he managed to organize a 'network' of fans that could gather all these stats that he thought of being important to know, for his work and further analysis. This had a quite successful run, from 1984 to 1991[5]. After publishing two annual essays his network was disbanded something that made James to join other third-party companies such as Stats and Fox Sports [6].

Breakthroughs:

Some of James's statistical breakthroughs were:

- Runs created: quantification of a player's involvement to runs scored.
- Range factor: quantification of a player's defensive involvement.
- Defensive efficiency rating: A statistics about how efficient a defense is by calculating how many offensive balls becoming an "out".
- Win shares
- Pythagorean Winning Percentage
- Game score

- Major League Equivalency
- The Brock2 System
- Similarity scores
- Secondary average
- Approximate Value
- Power/Speed Number
- Temperature gauge [7], [8], [9]

However, James's work was not accepted by professional baseball teams like Charles's work. It was until 2003 when he was hired by Boston Red Sox. [11] Although James does not talk about his work at this team, it is obvious that he has contributed to the team's first Word Series championship.

Stats

Total Runs Leaderboard

All Positions - 2018

2018 ▼ Go

Click on a column heading to sort.

Player	Runs Created	Pitching RunsCreated	Runs Saved	Baserunning Runs	Position	Total Runs
Betts, Mookie	138	0	20	0	16	174
Lindor, Francisco	119	0	14	0	35	168
Trout, Mike	133	0	8	0	22	163
Ramirez, Jose	127	0	3	5	24	159
Yelich, Christian	129	0	4	4	18	155
deGrom, Jacob	1	144	3	0	2	150
Story, Trevor	113	0	1	2	34	150
Baez, Javier	104	0	10	5	30	149
Chapman, Matt	92	0	29	5	22	148
Arenado, Nolan	116	0	5	2	23	146
Machado, Manny	115	0	-10	5	34	144
Freeman, Freddie	117	0	12	1	13	143
Bregman, Alex	121	0	-7	1	26	141
Merrifield, Whit	103	0	9	1	27	140
Cain, Lorenzo	92	0	20	3	24	139
Verlander, Justin	0	133	2	0	2	137
Nola, Aaron	-5	140	-1	0	2	136
Scherzer, Max	5	134	-5	0	2	136
Turner, Trea	98	0	2	1	35	136
Goldschmidt, Paul	119	0	6	-3	12	134

Figure 3-- A Bill James stat matrix [10]

Nowadays there are also other systems that help teams concentrate data such as SportVU as we are going to discuss further below.

A typical image of statistics in a Baseball game is like the following:

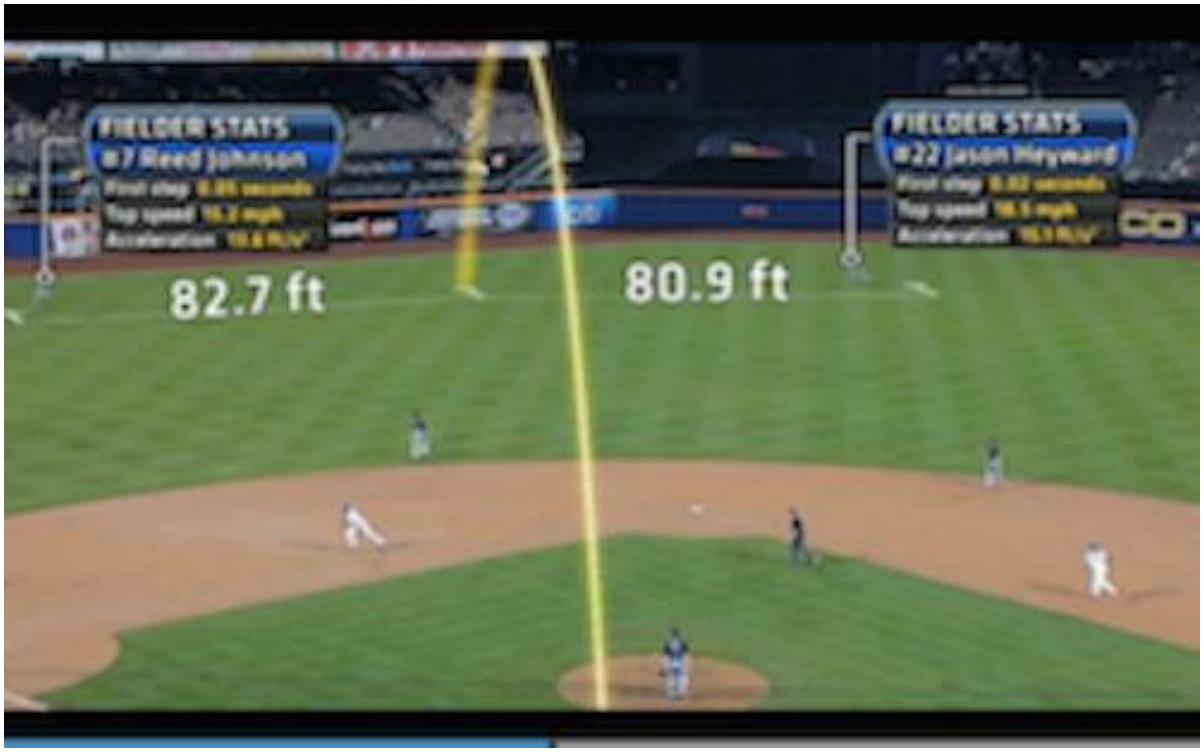


Figure 4 Baseball stats [21]

Basketball

So far, we have seen an introduction about sport analytics historical aspects in football (soccer) and in baseball. However, in basketball the attempts are not so old. It was not until 2005 that Israeli scientists, Gal Oz and Miky Tamir, created SportVU [12] [13].

SportVU is a system that has the ability to track not only the ball but also the athletes, providing data about their positioning and movements in the court. All this data that are gathered can be further analyzed through sophisticated algorithms that the company has created [14].

At first the tracking system was not real time, but this changed in 2011. At the time being not all teams had installed this system. However, in 2014 all teams had installed it in their courts due to the benefits that it was providing [15] [16].

Figure 5 sport vu cameras [17]

Moreover, due to the systems success, in 2016 it was decided to be extended to other sports as well such as football. The Ligue de Football Professionnel's started the adoption followed by other institutions.

All those data that this system provides could be exploited by data scientists and statisticians using machine learning algorithms in order to come to more difficult conclusions, not easily obtained with a quick look at the data [18].

Figure 6 SportVu in football [19]

Figure 7 - Example of sport Vu stats [20]

Chapter 3
Literature Review

In this part of the Thesis, previous work is going to be discussed that has been published mostly during the past decade. The papers that were chosen, belong to this time window, not only because it is important for them to be recent, but also because most of them were written in the past ten years, due to the fact that it is a quite new topic that concerns more and more constitutions. Those could be sports leagues, teams, individual athletes and even betting companies that need an estimation about a game outcome or an athlete's performance. Bellow, the most important implementations are going to be presented.

It is common knowledge that more and more teams are trying to invest in some form of data analytics in order to improve their performance, by gaining even a slight advantage over their opponents. However, there are not so many teams that admit that they have implemented such methods. In addition, it is difficult for a team to reserve the needed funds in order to employ data scientists who will help them draw conclusions [23].

However, the University of Virginia Football team had the will to use their data in order to improve their performance. Due to the lack of funds they asked the University's engineering department for assistance [23]. Their analysis was based on a previous concept of theirs, called the three pillars of data analytics.

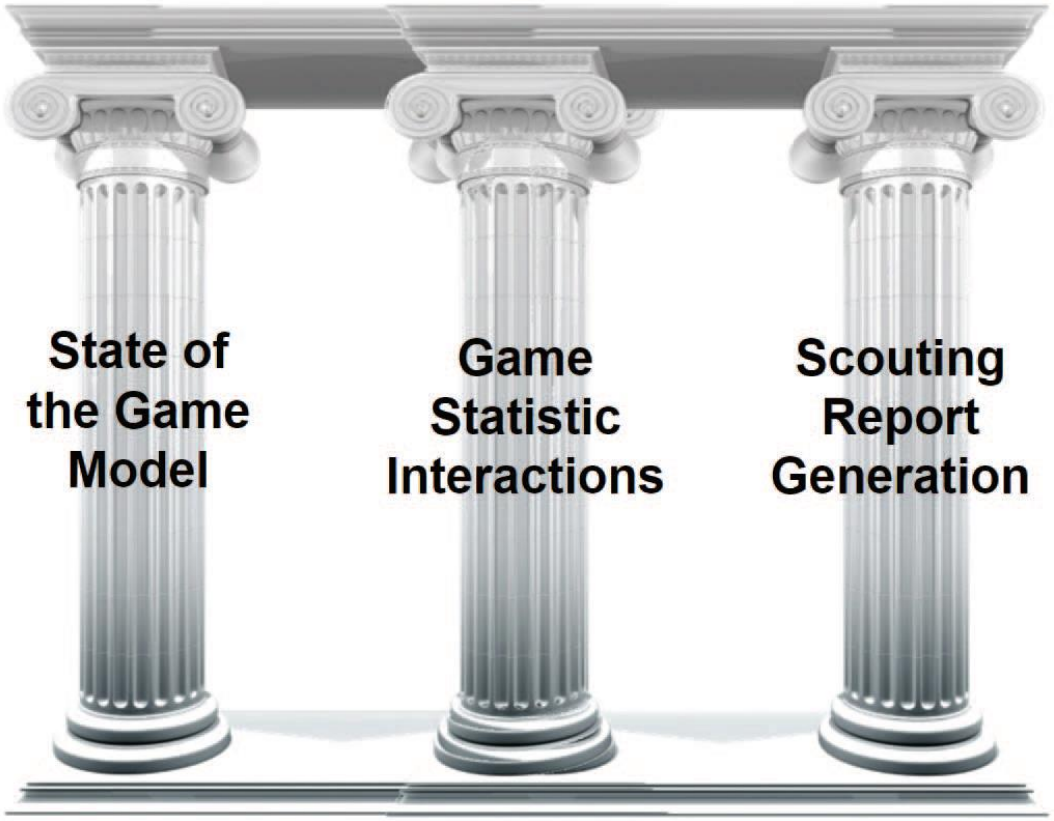


Figure 8--THE THREE PILLARS OF DATA ANALYTICS [23]

These three pillars include:

- State of the game model: Estimation of a team's win probability
- Game Statistics Interactions:
- Scouting Report Generation [23]

The followed process was based at three steps that had to be made.

- Firstly, the analysis of a fourth down decision tool.
- Secondly, a model that can predict the player performance
- And finally, a good exploitation of video analytical tools in order to extract useful conclusions after each game [23] as shown in the picture bellow.

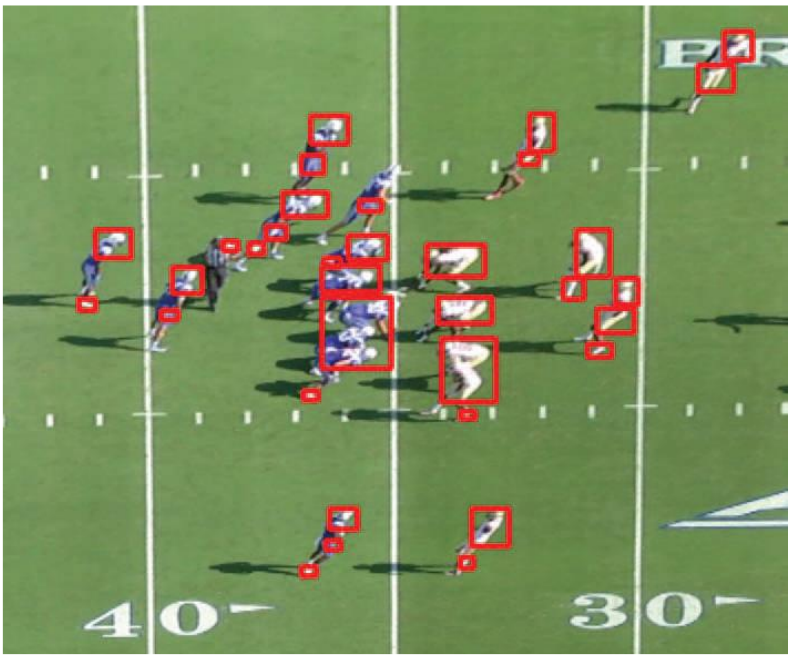


Figure 9-- video analysis [23]

Another approach is based more on data, trying to establish a correlation between the passes of the team and their goal scoring ability. The data are extracted by analyzing each game using the available technologies, for example cameras and sensors. There are also a lot of tools that provide useful help to that cause such as ProZone and Opta, whereas websites like EPLIndex.com and WhoScored.com have great databases [24].

While there is a plethora of data, still a data scientist is needed in order to exploit all this information and provide insights to the teams that are interested. This approach is based on a team's passing ability and accuracy. The data were extracted by four major European leagues, including around 1500 matches and 78 teams [24].

Season 2013/2014		
leagues	4	Germany, England, Spain, Italy
teams	78	20 England, Spain, Italy - 18 Germany
games	1,446	360 games per league in average
events	600,000	450 events per game in average

Figure 10-- data [24]

event	time	player	origin	destination	outcome
pass	17:24	Messi	(65.4, 20.2)	(67.8, 44.1)	successful
attempt	18:12	Messi	(98.4, 15.0)	(118.7, 15.0)	unsuccessful
pass	45:00	Bale	(78.56, 12.2)	(78.5, 36.0)	successful
attempt	55:00	Ronaldo	(89, 45)	(100, 45)	successful
⋮	⋮	⋮	⋮	⋮	⋮

Figure 11--data [24]

In this analysis there is a clear indication that the number of passes performed by a team has an impact on their goal scoring ability (a ratio between the amount of shots that became a goal to the general number of shots) and a more general effect on the team's seasonal performance. However, information about individual players is not taken into account, but their performance as a whole is, while there is not paid too much attention to their defensive skills [24].

In addition to the previous approaches, there is another one more sophisticated that has the advantage of using data that are produced by wearable devices. It was until recently that FIFA allowed the use of such devices during football matches (Electronic Performance and Tracking Systems – EPTS). These systems include accelerometers, gyroscopes, magnetometers [25]. While in sports like Rugby, Hockey and Cricket the use of these devices is widespread in football it is not so much. Of course, the fact that the use of these devices was not allowed played an important role, but even nowadays there are not so many teams that have included their use.

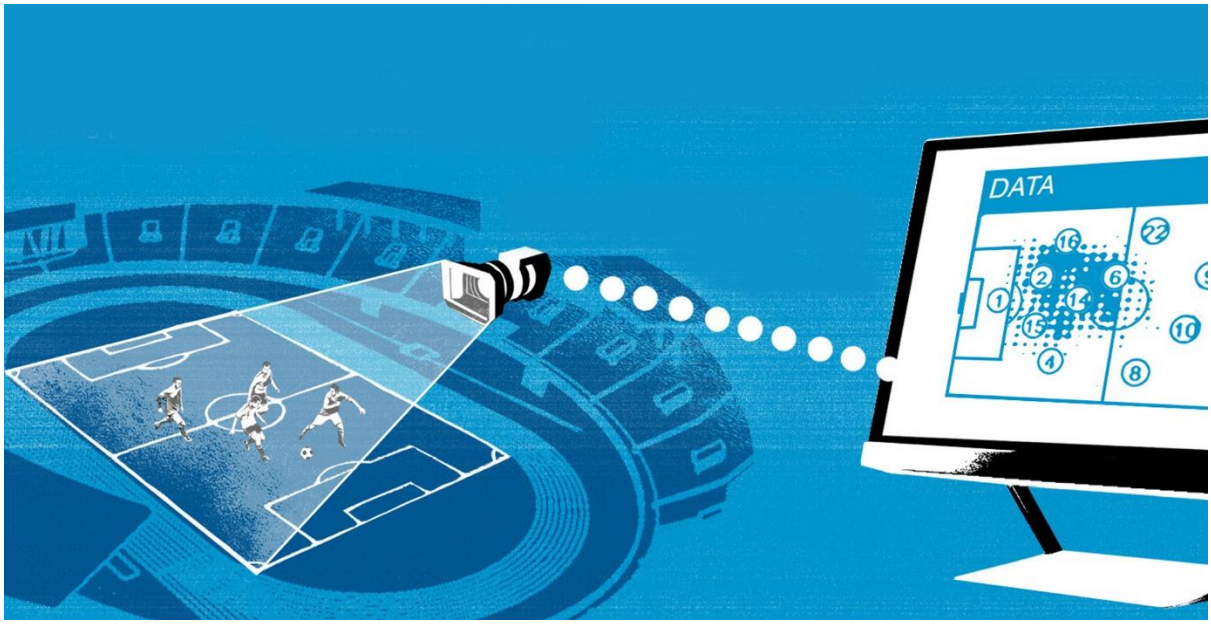


Figure 12 -- The EPTS (FIFA) [26]

In this specific analysis, FC Barcelona is taken as an example. Statistics and data are provided not only for the team, but also for the players. However, while this aims at the improvement of the team's performance, mostly during training sessions, there is a need that the coaches could actually understand and implement this kind of information to their plans [25].

The methodology that is followed is by selecting the most important features that these sensors provide, and then a regression analysis is performed [25]. It concludes that the use of these data both on previous matches and on previous training sessions can actually improve a team's performance significantly and obviously more and more teams are going to implement such approaches to their inventory.

Figure 13-- Wearable Devices [30]

Another analysis relative to the topic is the use of data in order to estimate if an athlete is ready to take part in a competition being either a race or a football match. There are actually some metrics called *markers* which indicate an athlete's "readiness" [27].

However, it is common that this metrics are misinterpreted a lot of times, something that leads coaches to wrong decisions, which may be if an athlete should be purchased to join a team, or if they should release them and sometimes just a simple decision if the athlete should be included in a game [27].

In order to overcome these obstacles, there is a wide use of the available technology that sensors provide and the ability to apply machine learning to the data obtained by those sensors.

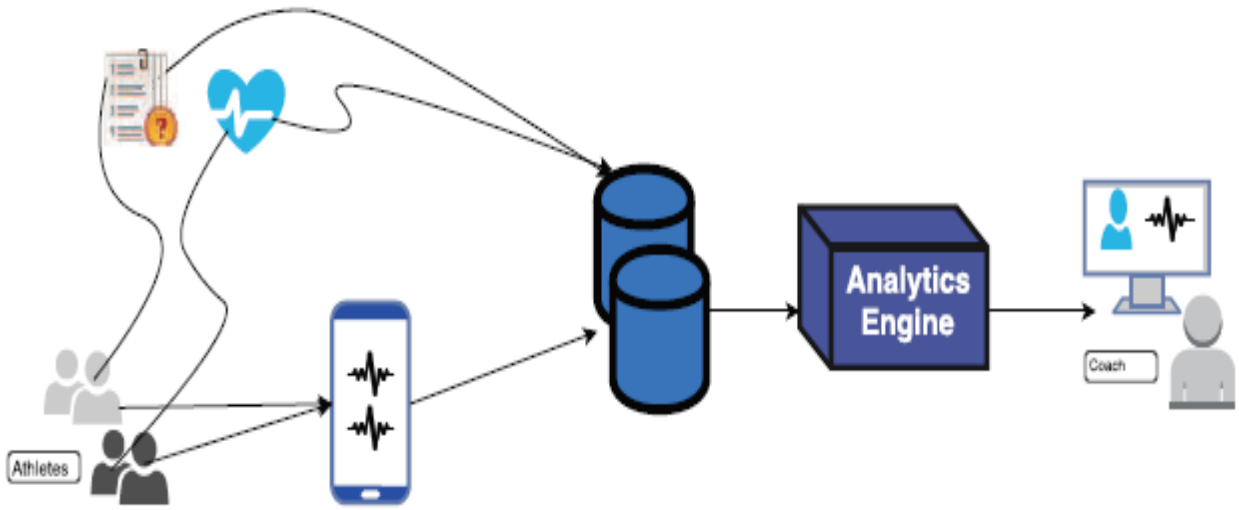


Figure 14 Sensors [27]

The data that the sensors provide at this system are:

- Maximal effort neuromuscular function - Jump test. It is a test of how far an athlete can jump. [27]
- Heart Rate Variability. A metric that shows the different heart rate values of an athlete during a training session. [27]
- Rated Perceived Exertion (RPE). A metric that shows how intense and difficult an exercise is during a training session. [27]

The conclusion of this analysis while having a good accuracy at the test set that was under examination for the purposes of the experiment, it is mentioned that it is difficult to generalize for all kind of athletes.

In addition to the previous approaches, there is another that takes advantage of sensors that can capture the Electroencephalography frequency of brain waves. There is an implementation of data science to those data that of course are gathered from athletes [28].

The experiments were conducted to fifty endurance athletes before and after exercise. The athletes had great experience to their fields, some of them with important achievements to international levels, while their ages varied from twenty to thirty years old. The data collection did not last longer than 180 seconds [28].

10 variables for each athlete were extracted around. The experiment needed to show the difference in the wave frequencies before and after exercise.

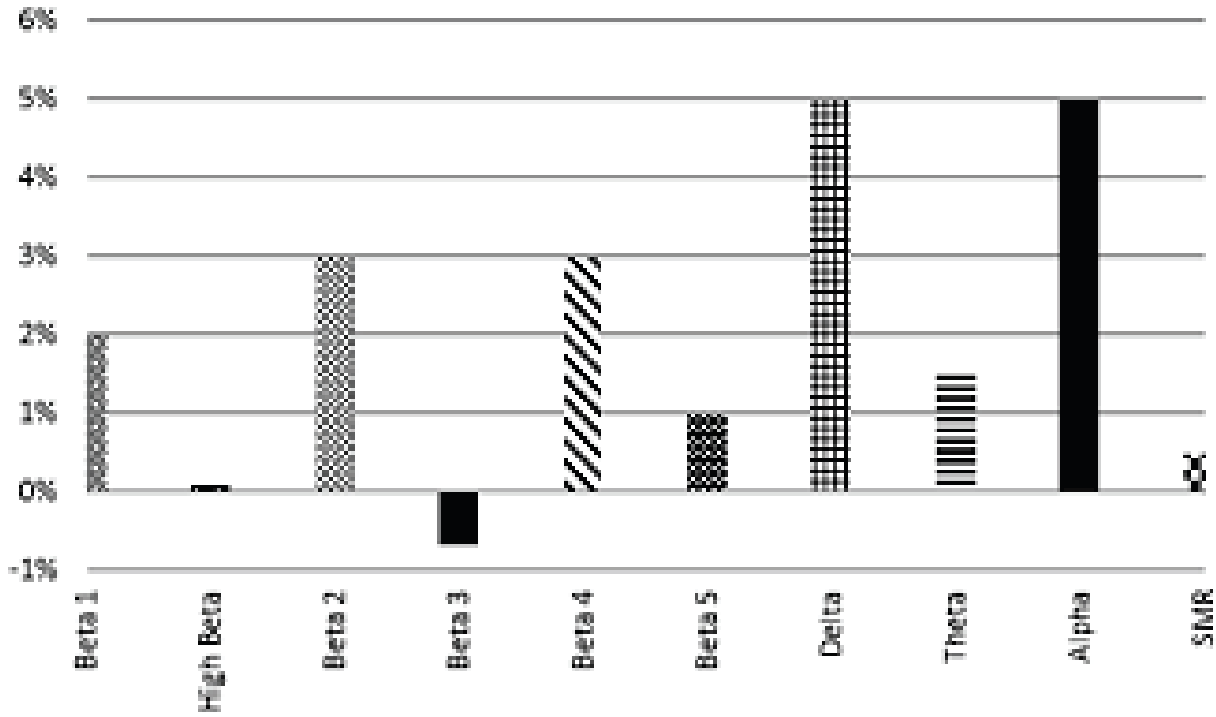


Figure 15 - Difference in wave frequency before and after exercise [28]

This study shows that in order to identify a skilled athlete, we have to look to Beta 4 and Beta 5 waves where there are going to be the biggest differences [28]. It is a not so widespread approach, but it might convince more coaches and teams to adopt it if the results can generalize to a bigger part of the population, and not just at the championship athletes.

In addition to the previous studies, there is another that has the ambition to point out, that for the prediction of a game outcome there is also a quite significant "amount" of chance. For the purposes of this survey, data from leagues of 84 countries were collected [29]. The interesting part is that the focus is not only in one sport but in four:

- Basketball
- Football
- Volleyball
- Handball

There is a try in this survey to actually convert the "amount" of chance to a coefficient and also the "amount" of skill that each team has to another coefficient.

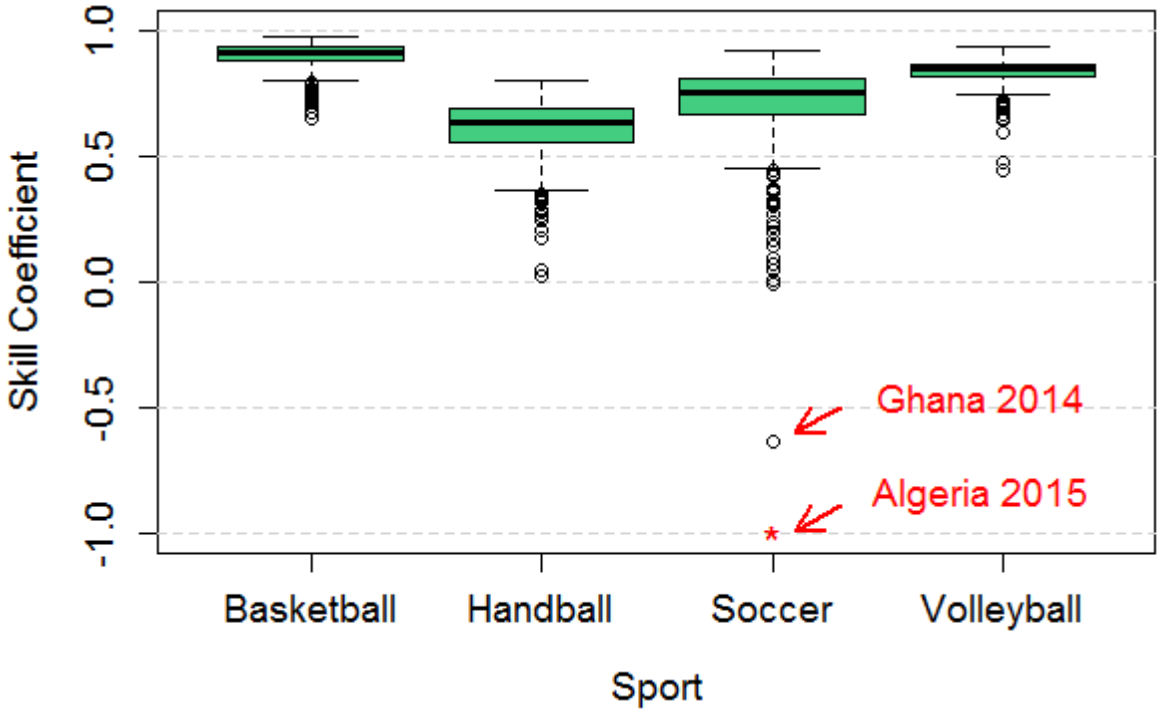


Figure 16 -- skill coeff boxplot [29]

This study concludes to the fact that the most competitive game is basketball followed (surprisingly) by volleyball. It is important to point out that in order to achieve more random leagues in these two sports, we should actually remove 50% and 40% respectively while in football the percentage could be around 20% [29]. In addition, the percentage of luck in each sport is different and sometimes it is different in the same sport, but in different leagues. For NBA for instance, the percentage is around 35%, which is actually a significant number. So, it is important to understand that occasionally chance is difficult to predict, but it has to be taken into account, when a prediction for a game outcome is performed [29].

Furthermore, continuing the literature review, we came across another study which also has to do with data collected by sensors. The difference is that the sensor is not on the athlete. There are also data that are not gathered by cameras or GPS tracking systems.

It is a study about cricket, where a prototype ball is created, and the sensors are on the ball [31].

Figure 17 -- Special Ball for sensor embedding [31]

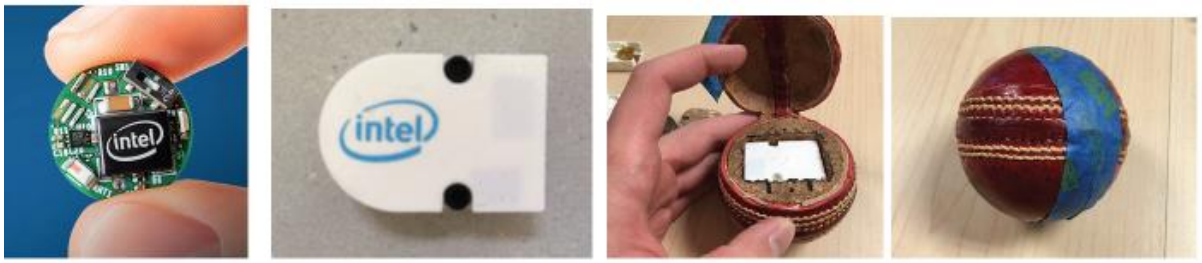


Figure 18 -- Sensor embedded to the ball [31]

The prototype is not perfect; however, the results were encouraging which creates a ground for further exploitation and even adoption by other sports as well [31].

In addition to the previous studies, there is another analysis related to sports analytics, but more attention is paid to the possibility of a player getting injured. It is important to know, given some specific conditions, if a player is going to get injured. Thus, the coach may have the opportunity to take some precautions, in order to avoid such an incident. A serious injury may be a major drawback for a team and the player as an individual. Their form may deteriorate for a long period of time and also it could cost the loss of a title or a prize.

So, in this study there are proposed some methods that could actually predict the injury possibility. The data that were under examination were gathered from 713 Major League Soccer games, and in 548 was at least one injured player [32].

The data that this survey uses, come from weather conditions and the playing surface. The use of data mining and machine learning algorithms helps to the outcome prediction.

Source	Description
MLS Game Data	Game Duration WetBulb Globe Temperature on Field Actual Temperature on Field Surface *(Grass, Turf)*
MLS Injury Data	Temperature Range *(Extreme Cold, Cold, Mild, Warm, Hot, Extreme Heat)* Temperature Score Conditions *(Sunny, Clear, Cloudy, Rain, Snow)* Condition Score
Weather Underground	Temperature Dew Point Humidity Wind Speed Wind Gust Visibility Pressure Wind Chill Heat Index Precipitation Conditions *(Clear, Cloudy, Rain, Thunderstorms, Snow)* Condition Score

Figure 19 -- List of the attributes [32]

The algorithms that were used were:

- C4.5 Decision Tree
- Logistic Regression
- Multilayer Perceptron
- Naive Bayes
- Radial Basis Function Network
- Random Forest
- Support Vector Machine [32]

For all the above algorithms, it was used a 10-fold cross validation in Weka. The results of this study were quite satisfying, however there was not a distinction if an injury was going to happen because of the weather conditions or because of the playing surface [32].

It is thought a first attempt for that kind of prediction with promising results that could generalize to other leagues and maybe other sports as well [32].

In contrast to previous studies that data from wearable devices were obtained and then machine learning algorithms were applied, there is another approach were deep learning is implemented on those data.

Deep learning belongs to the family of machine learning algorithms, while having specific characteristics:

- For feature extraction a lot of layers can be used, and each successive layer uses for input data from the previous one [34].
- It can be performed as supervised but also as unsupervised learning [34].

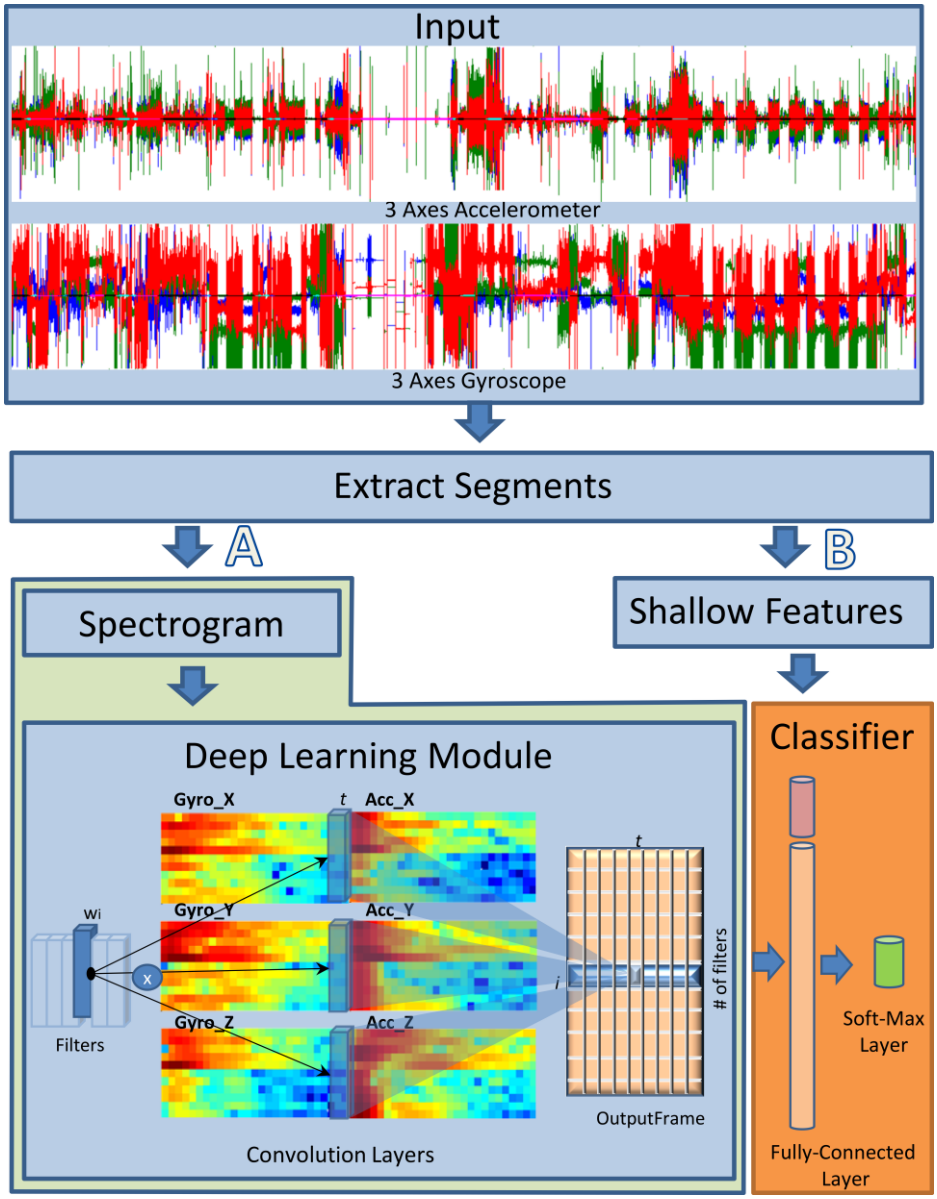


Figure 20 -- The proposed method [33]

The conclusion in this study is that by taking advantage of the deep learning algorithms we can achieve much better results, even from technologies that are being used by professionals to this date.

While this literature review is focused mainly on football (Soccer), there are also proposed studies for other sports that help us make useful conclusions about the way that data mining, machine learning and data from sensors, cameras and wearable devices are going to affect the future of the sports.

As a result, in this next survey the main focus is on tennis which has the advantage of not being a team sport, so it simplifies the study. The study aims to provide information to tennis players on how to improve their serve swing, due to the fact that it is one of the most important movements in tennis. It is an attempt to log the most important details that can make a successful serve swing [35].

The equipment that was used is a Samsung Smart Watch (Gear S2) and a camera to video capture a tennis training session. The smart watch can provide data about the heart rate of the athlete and the positioning of his arm [35].

Figure 21-- Samsung Gear S2 [36]

In order to analyze the arm positioning and connect it to the way that the serve swing is performed, we need to explore how the serve swing is decomposed to stages. For this particular study, the serve swing is divided to 3 stages, Preparation, Acceleration and Follow-Through [35].

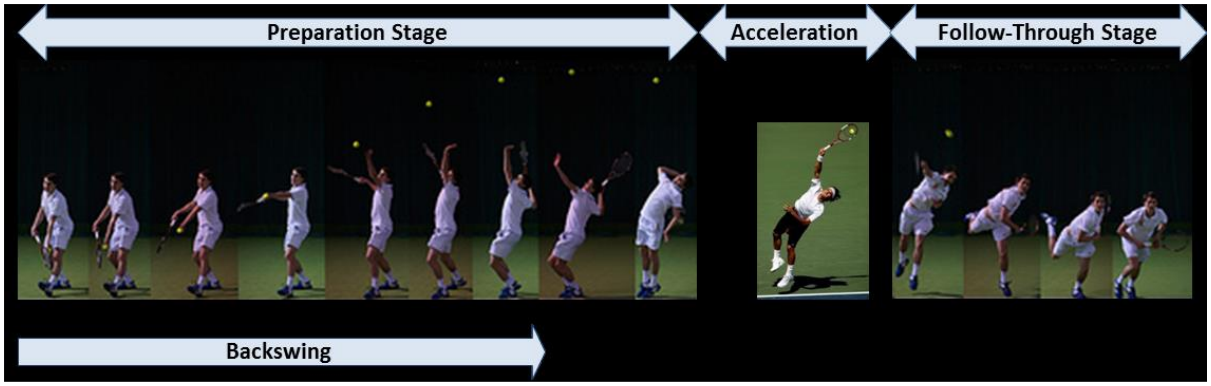


Figure 22 -- Tennis swing serve stages [35]

By accumulating data produced by the sensor and the camera, we can check out how consistent a player's swing is, what is the point that makes the biggest and most important difference, and where an athlete should focus in order to improve their performance [35].

The good news is that this method can be followed to improve also other aspects of tennis players and maybe to be able to generalize it for other similar sports as well.

For the next survey, the focus is on data accumulated strictly by video analysis in football. Computer vision is used to analyze a player's positioning each second of the game, the total distance that they cover, their "favorite" movements and also most of these concepts apply to the ball as well [37].

The first thing that need to be done, is the 3d image of the stadium to be converted to 2 dimensions in order to be easier to edit and gather data. An example of this concept is shown below.

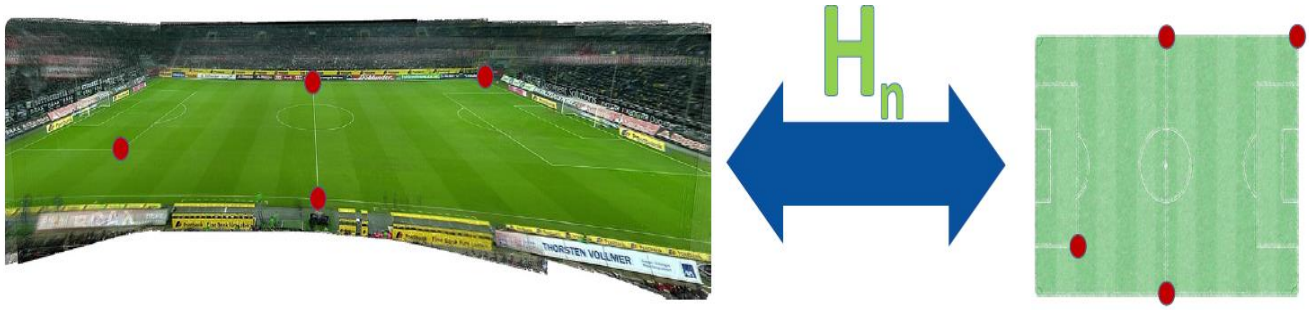


Figure 23 -- Conversion from 3d to 2d [37]

In Figure 23, it is shown that specific points of the image are needed for the conversion, which are pinpointed with red colored dots.

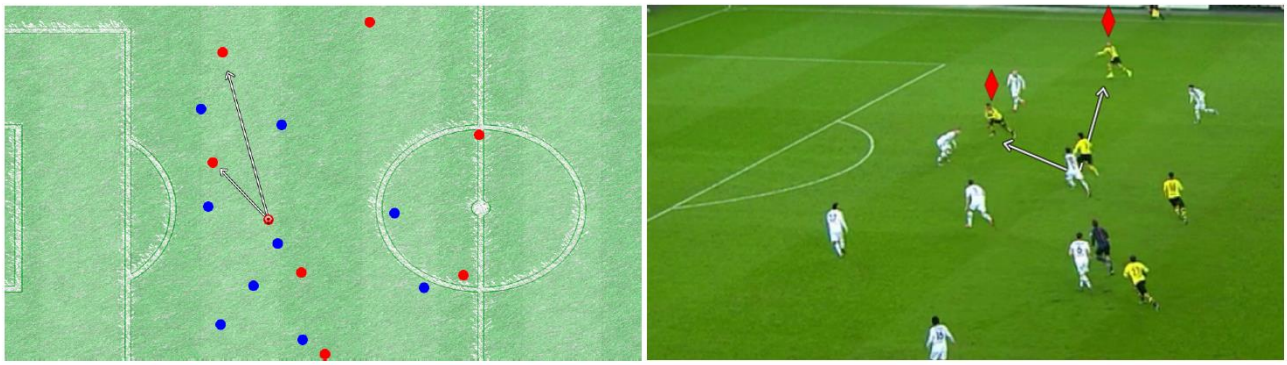


Figure 24 -- players positioning conversion [37]

The figure above shows that it is easier to obtain data from the left picture which is converted by the players in the right one. Then specific algorithms can be applied in order to draw conclusions about a team's performance. This system is quite successful because is used by the most well-known companies in this field [37].

The paper concluded that it is a good technique that is going to be used in the future while mentioning that the fact that video analysis and computer vision can have some disadvantages as well should be taken into consideration [37].

Similarly, to the previous study is this one which is more of a sub category of the way that a video needs to be analyzed. The main concern is the creation of a semi-automatic video image repositioning to vertical axis in order to be easier to obtain the needed data. The accuracy was higher than 85%, while the experiments were conducted on hockey games and players [38].

The process that was followed can be pointed out in the schema below:

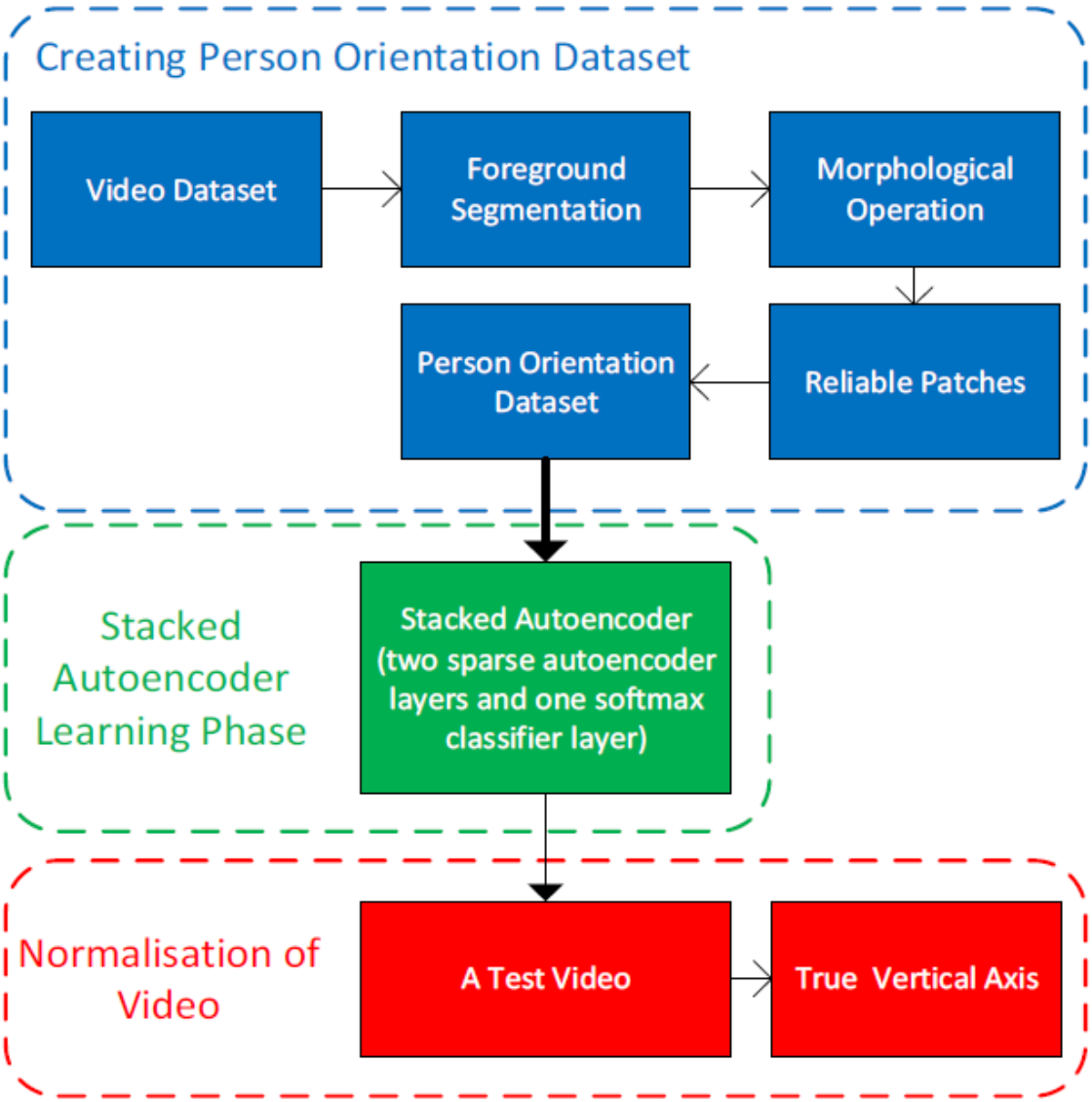


Figure 25-- Followed process [38]

In addition, a possible outcome of the method is shown below:

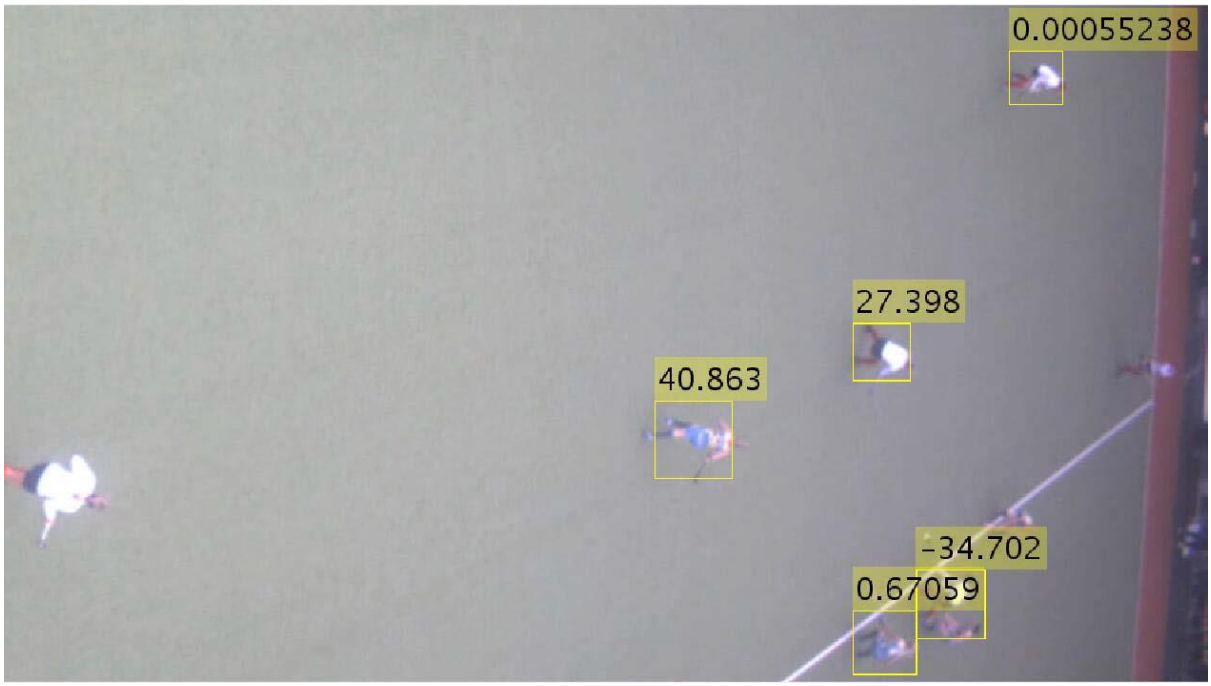


Figure 26 -- outcome of the method [38]

In addition to these studies there are also some others that do not examine the specific field of performance improvement, but are related.

For example, there is a study that is experimenting with NBA players psychology and mood. All those data are obtained by accumulating athletes' activity in social media. For example, all their tweets are gathered alongside with their Facebook statuses. There is a try to study if there is a correlation between such a tweet and the performance of the player in the next day game. The season that was studied was 2012-2013 and there was a sentiment analysis performed on those tweets [39].

We can see below a block diagram of the process:

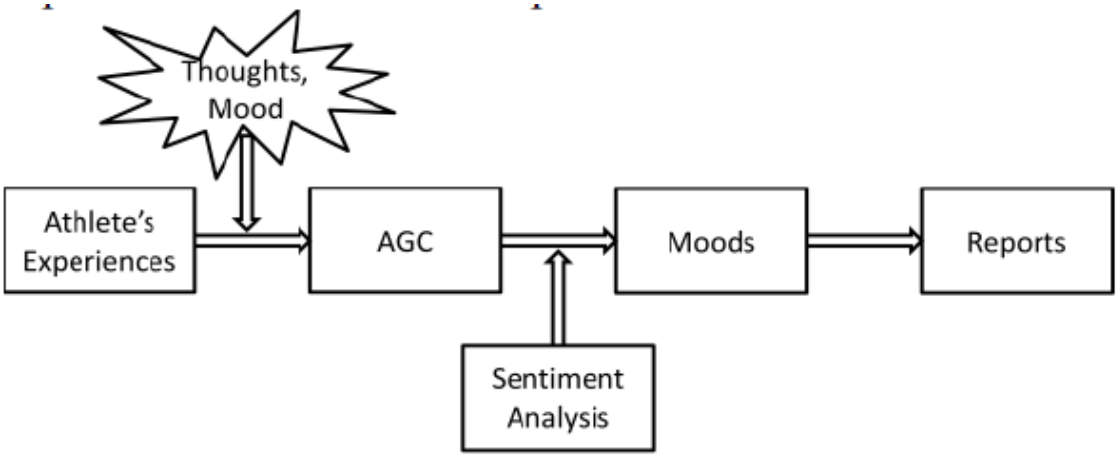


Figure 27 -- block diagram [39]

The study concluded that it is important to perform sentiment analysis on players behavior, because there is a correlation to their performance. It is going to be more widespread in the future, however there are certain limitations such as the fact that the sentiment analysis was performed to individual players and not to the whole team [39].

Chapter 4
General terms

In this part of the Dissertation, some general terms that are relative to the topic are going to be analyzed. It is important though to point them out in order to have a more general view of the subject.

The terms that will be discussed are:

- Data Mining
- Machine Learning
- Sports Analytics

Data Mining

The first term that is going to be explained is *Data Mining,* which is a term that is more likely to come across nowadays, since it has a plethora of applications. With the term Data Mining we mean every procedure and attempt to identify hidden meaning and features in large database systems. Of course, it belongs to the field of Computer Science. The paradox of the term is that we do not try to extract raw data, but patterns that may be implied [40].

The process that needs to be followed in order to complete a Data Mining task is:

- Firstly, a suitable dataset needs to be obtained. We may need to create the dataset on our own.
- Then, the dataset needs to be preprocessed. That means that we have to look for missing values and the data in the same columns to have the same format. For example, a column with years, needs to have integer numbers. If all attributes are not integers, we are most likely going to face inconsistencies in the following steps.
- Then, is the time for the application of the algorithms. These algorithms may have to do with:
    - Classification (Figure 28)

- Clustering (Figure 29)
- Association Rules (Figure 30), etc

- Finally, after we receive a result, we have to evaluateit, if our model is accurate enough and also if it can generalize to other examples or it is just for a unique dataset [40].

It is important to point out that the process that will be followed is not always the same. However, there have been made some attempts to create a general standard that will be followed. The 1999 European Cross Industry Standard Process for Data Mining (CRISP-DM 1.0) and the 2004 Java Data Mining standard (JDM 1.0) are two of the most important. Sadly, it was not a pattern that was easily to adopt so it did not become popular [41].

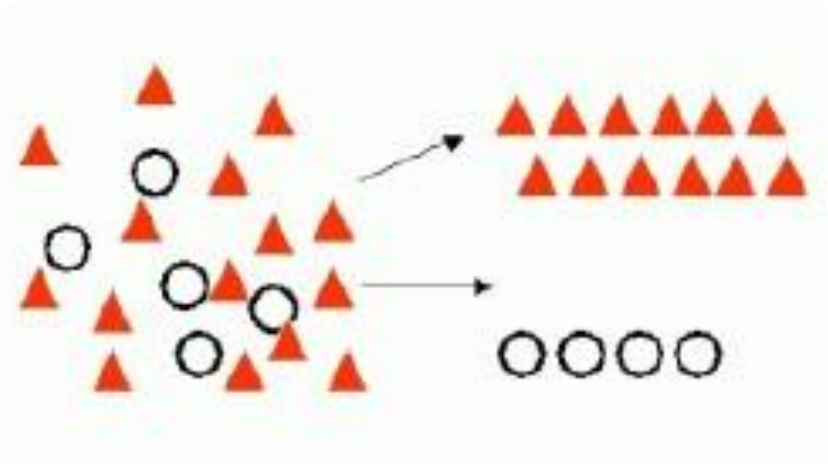


Figure 28 -- classification [43]

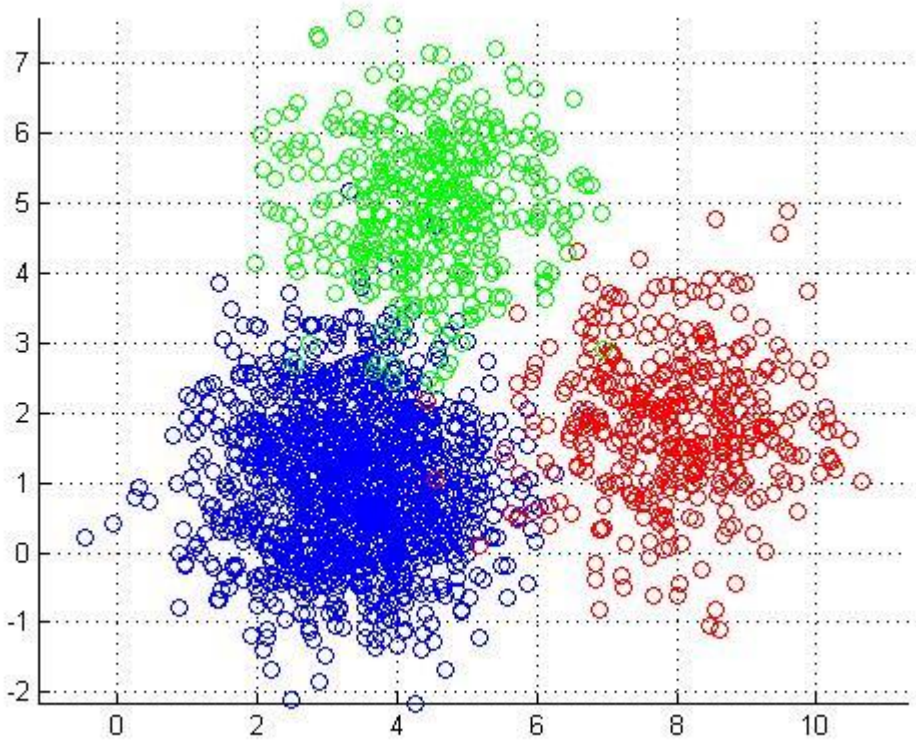


Figure 29-- clustering [42]

Rule	Support	Confidence	Lift
$A \Rightarrow D$	2/5	2/3	10/9
$C \Rightarrow A$	2/5	2/4	5/6
$A \Rightarrow C$	2/5	2/3	5/6
$B \& C \Rightarrow D$	1/5	1/3	5/9

Figure 30 -- association rules [44]

Machine Learning

Since machine learning is a very general concept, there are many definitions for this term. Below we will point out the most common of them, given by the most recognizable and reliable Universities, institutions, professors and organizations.

- “Machine learning is the practice of using algorithms to analyze data, learn from it, and then make a determination or prediction about something in the world. Instead of programming software routines by hand with a particular set of instructions to perform a particular task, the machine is "trained" with large amounts of data and algorithms that allow them to learn how to perform the task” [45].
- “Machine learning is the science of getting computers to act without being explicitly programmed” [46].
- “Machine learning is based on algorithms that can learn from data without relying on rules-based programming.” [47]
- “Machine learning algorithms can figure out how to perform important tasks by generalizing from examples.” [48]
- “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?” [49]

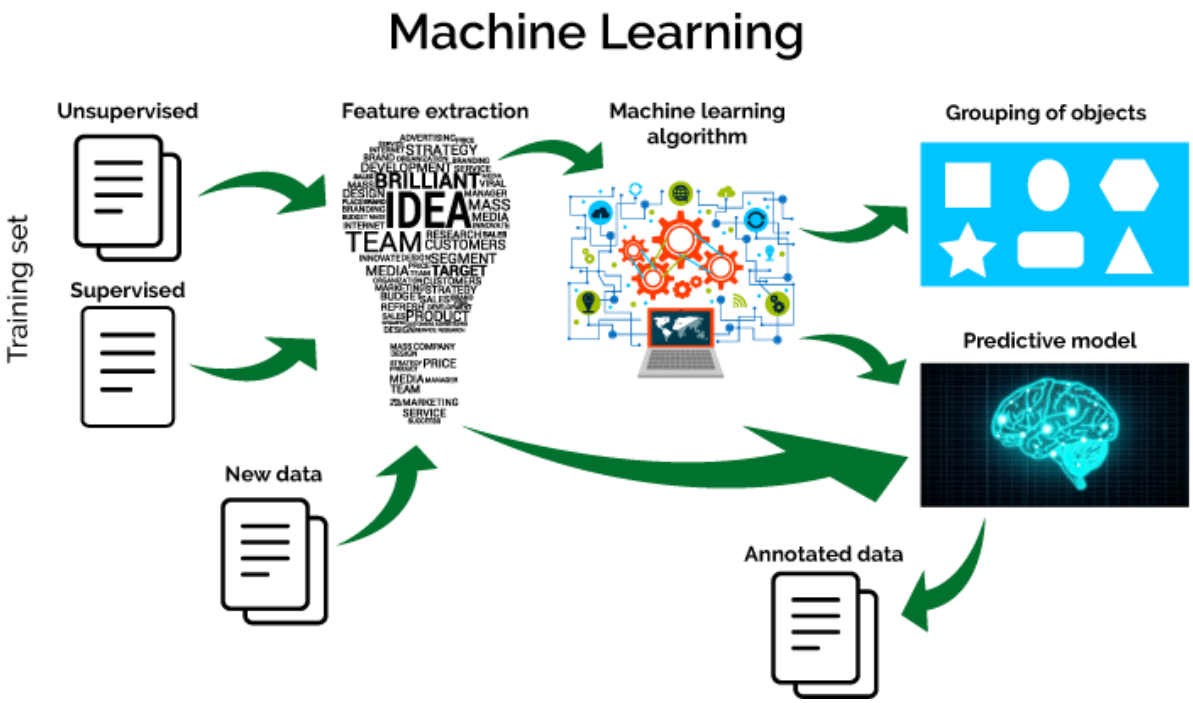


Figure 31 -- Machine Learning Model [50]]

Machine learning can be separated into two major categories:

- Supervised Learning
- Unsupervised Learning

The definitions for the above terms are:

- "In supervised learning, the aim is to learn a mapping from the input to an output whose correct values are provided by a supervisor." [51]
- "In unsupervised learning, there is no such supervisor and we only have input data." [51]

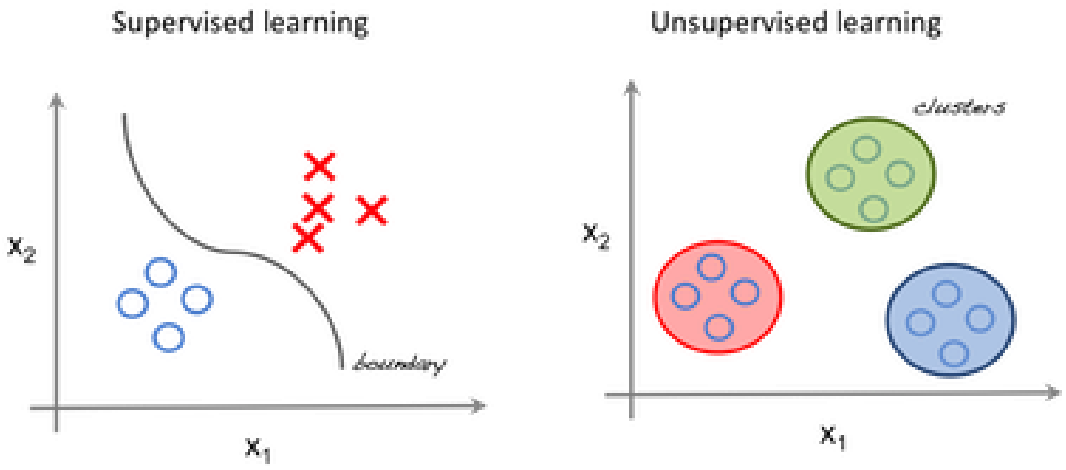


Figure 32 -- Supervised vs Unsupervised learning [52]

Algorithms of supervised learning are amongst others:

- Nearest Neighbor
- Naive Bayes
- Decision Trees
- Linear Regression
- Support Vector Machines (SVM)
- Neural Networks [53]

Algorithms of unsupervised learning are amongst others:

- k-means clustering
- Association Rules [53]

Sports Analytics

Sports analytics is the application of the above methods and implementations to sports in order to draw useful conclusions. Such conclusions may affect the performance of an individual athlete, the team as a whole for a specific game or for the whole season. It can also help teams make predictions for upcoming talents, a player's market value and the possibility of an injury. It is a field that is becoming more and more popular nowadays and it is going to be adopted by a plethora of teams, coaches, individual athletes and companies.

"The analytics split nicely between the front-office and back-office. Front-office analytics include topics like analyzing fan behavior, ranging from predictive models for season ticket renewals and regular ticket sales, to scoring tweets by fans regarding the team, athletes, coaches, and owners. This is very similar to traditional customer relationship management. Financial analysis is also a key area, especially for the pros where salary caps or scholarship limits are part of the equation. Back-office uses include analysis of both individual athletes as well as team play. For individual players, there is a focus on recruitment models and scouting analytics, analytics for strength and fitness as well as development, and predictive models for avoiding overtraining and injuries. Concussion research is a hot field. Team analytics include strategies and tactics, competitive assessments, and optimal roster choices under various on-field or on-court situations." [54]

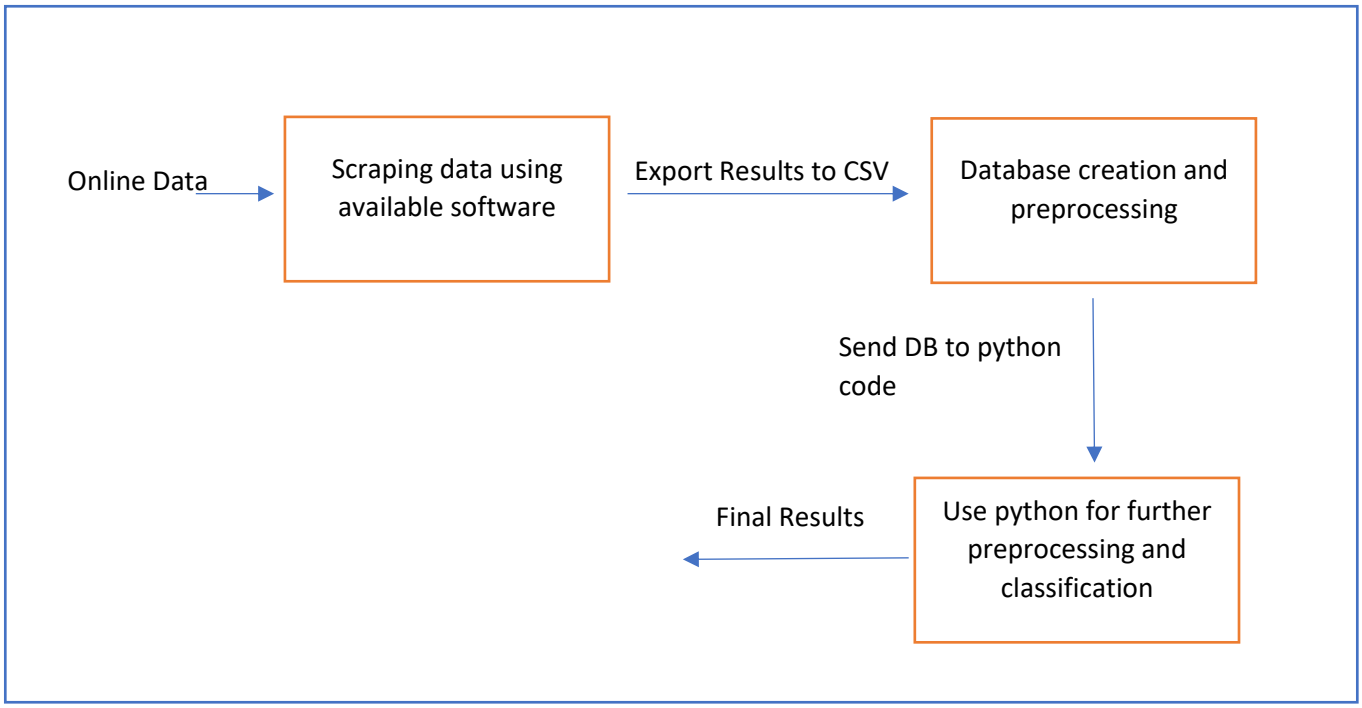
Figure 33-- sports analytics [55]

However, all parts that are interested in this field should be extremely careful. The data should not be accessible to anybody because there are going to occur numerous problems, such as issued with betting companies, or wrong ways of athlete training leading to injuries as a result. [56]

Chapter 5

Followed Process

Block Diagram



We explain the above steps more extensively below:

- Firstly, we need to find data that are relevant to our topic. One of the most complete websites that contains information about football players individually but also about the teams, is whoscored.com. Stats about a player's appearances, how many minutes was in a game for a whole season, the number of goals he scored, his assists and completed passes even how many red and yellow cards he has been booked with, and many more are provided.
- Since we find the data that seem to be related to our topic, we need a software to scrap them and have them in a csv file. There are many scraping tools available online for free.
- Now, we have reached the point that we have our database. However, it is not "clean" and there are many attributes that we do not need. So, the csv file that we obtained from the scraping tools needs to be edited and processed before we use it in python.
- After the preprocessing is completed, we get the database in python, where a final preprocessing may be needed and then we can start searching for classification algorithms that provide best results. Of course, there should be a variety of classifiers that are going to be used, in order to be able to compare the results and chose the classifier that is more accurate. However, the same classifier should be used for all the players, otherwise the "method" will not be fair, and the results will be biased.

Experiments

First Experiment – Players position

1. One of the first experiments that has been conducted, is the prediction of the position of a football player. This was handled by obtaining a database which was crawled from the website https://sofifa.com [60]. It includes data from the football game FIFA 18. The database has attributes for many players. After looking at the database, we can determine that preprocessing has to be done. For this purpose, python will be used. For example, the columns have missing values, which need to be handled. There are names that contain characters not compatible to the UTF-8 system, so they need to be transformed. There are also issues with wrong data types (a column that contains integers is parsed as string).

Below, we can have a look of the database:

```
Name,Age,Nationality,Overall,Potential,Value,Wage,Special,Acceleration,Aggression,Agility,Balance,Ball control,Composure,Crossing,Curve,Dribbling,Finishing,Free
Cristiano Ronaldo,32,Portugal,94,94,95.5M,565K,2228,89,63,89,63,93,95,85,81,91,94,76,7,11,15,14,11,88,29,95,77,92,22,85,95,96,83,94,23,91,92,31,80,85,88,FOR
L. Messi,30,Argentina,93,93,105M,565K,2154,92,48,90,95,95,96,77,89,97,95,90,6,11,15,14,8,71,22,68,87,88,13,74,93,95,88,85,26,87,73,28,59,90,85,FOR
Neymar,25,Brazil,92,94,123M,280K,2100,94,56,96,82,95,92,75,81,96,89,84,9,9,15,15,11,62,36,61,75,77,21,81,90,88,81,80,33,90,78,24,53,80,83,FOR
L. Suarez,30,Uruguay,92,92,97M,510K,2291,88,78,86,60,91,83,77,86,86,94,84,27,25,31,33,37,77,41,69,64,86,30,85,92,93,83,87,38,77,89,45,80,84,88,FOR
M. Neuer,31,Germany,92,92,61M,230K,1493,58,29,52,35,48,70,15,14,30,13,11,91,90,95,91,89,25,30,78,59,16,10,47,12,85,55,25,11,61,44,10,83,70,11,GK
R. Lewandowski,28,Poland,91,91,92M,355K,2143,79,80,78,80,89,87,62,77,85,91,84,15,6,12,8,10,85,39,84,65,83,25,81,91,91,83,88,19,83,79,42,84,78,87,FOR
De Gea,26,Spain,90,92,64.5M,215K,1458,57,38,60,43,42,64,17,21,18,13,19,90,85,87,86,90,21,30,67,51,12,13,40,12,88,50,31,13,58,40,21,64,68,13,GK
E. Hazard,26,Belgium,90,91,90.5M,295K,2096,93,54,93,91,92,87,80,82,93,83,79,11,12,6,8,8,57,41,59,81,82,25,86,85,85,86,79,22,87,79,27,65,86,79,FOR
T. Kroos,27,Germany,90,90,79M,340K,2165,60,60,71,69,89,85,85,85,79,76,84,10,11,13,7,10,54,85,32,93,90,63,73,79,86,90,87,69,52,77,82,74,88,82,MID
G. Higuain,29,Argentina,90,90,77M,275K,1961,78,50,75,69,85,86,68,74,84,91,62,5,12,7,5,10,86,20,79,59,82,12,70,92,88,75,88,18,80,72,22,85,70,88,FOR
Sergio Ramos,31,Spain,90,90,52M,310K,2153,75,84,79,60,84,80,66,73,61,60,67,11,8,9,7,11,91,88,93,72,55,86,68,52,85,78,79,91,77,84,89,81,63,66,DEF
K. De Bruyne,26,Belgium,89,92,83M,285K,2162,76,68,80,75,87,84,90,83,85,83,83,15,13,5,10,13,53,56,65,84,86,30,77,84,88,90,85,40,75,87,51,73,90,82,MID
T. Courtois,25,Belgium,89,92,59M,190K,1282,46,23,61,45,23,52,14,19,13,14,11,85,91,69,86,88,13,15,68,31,17,11,27,13,81,32,36,16,52,38,18,70,44,12,GK
A. Sanchez,28,Chile,89,89,67.5M,265K,2181,88,80,90,87,87,86,80,78,90,85,78,10,10,15,12,13,70,42,85,73,82,30,77,86,87,81,84,35,84,85,39,72,83,83,MID
L. Modric,31,Croatia,89,89,57M,340K,2228,75,62,93,94,92,84,78,79,86,71,77,13,9,7,14,9,55,76,67,83,82,66,80,79,88,92,73,73,71,82,80,58,90,74,MID
G. Bale,27,Wales,89,89,69.5M,370K,2263,93,65,77,65,87,85,87,86,89,87,85,15,15,11,5,6,86,59,85,80,90,51,76,86,87,86,91,52,95,76,55,80,79,76,FOR
S. Aguero,29,Argentina,89,89,66.5M,325K,2074,90,63,86,91,89,90,70,82,89,90,72,13,15,6,11,14,68,24,80,63,83,13,83,91,89,79,88,12,84,74,20,74,83,85,FOR
G. Chiellini,32,Italy,89,89,38M,225K,1867,68,92,59,64,57,82,58,60,58,33,31,3,3,2,4,3,84,88,89,59,49,92,50,28,82,59,78,90,78,68,92,91,50,45,DEF
G. Buffon,39,Italy,89,89,4.5M,110K,1335,49,38,55,49,28,70,13,20,26,15,13,89,88,74,90,84,13,28,75,35,13,10,22,12,80,37,39,11,43,39,11,69,50,17,GK
P. Dybala,23,Argentina,88,93,79M,215K,2063,88,48,91,85,93,84,80,88,92,85,84,5,4,4,5,8,68,24,75,71,88,14,86,84,84,83,82,20,84,83,20,65,84,88,FOR
J. Oblak,24,Slovenia,88,93,57M,82K,1290,43,34,67,49,16,55,13,13,12,11,14,84,90,77,87,84,15,19,76,26,12,14,11,11,84,29,22,18,60,41,12,78,55,13,GK
A. Griezmann,26,France,88,91,75M,150K,2104,87,69,90,80,86,86,82,84,87,88,75,14,8,14,13,14,80,35,87,75,82,23,71,91,90,79,81,11,86,76,22,61,76,87,FOR
Thiago,26,Spain,88,90,70.5M,225K,2185,77,57,90,86,92,83,72,85,90,69,77,6,11,7,9,13,58,78,75,90,82,49,75,83,84,91,75,62,68,75,64,59,86,90,MID
P. Aubameyang,28,Gabon,88,88,61M,165K,2078,95,43,77,70,82,84,77,78,82,88,74,6,9,15,9,9,79,48,77,64,79,28,79,89,87,77,81,36,96,79,25,77,77,86,FOR
L. Bonucci,30,Italy,88,88,44M,210K,1995,62,82,60,52,75,84,44,56,69,39,61,2,2,3,2,4,88,90,85,85,65,84,70,38,86,79,74,88,72,73,88,85,74,58,DEF
J. Boateng,28,Germany,88,88,48M,215K,1989,72,82,58,53,71,86,69,56,67,34,31,7,12,15,6,5,85,83,75,80,58,88,46,47,82,75,79,90,78,74,91,91,76,53,DEF
D. Godin,31,Uruguay,88,88,40M,125K,1930,62,86,63,58,76,82,55,49,53,42,51,6,8,15,5,15,92,88,89,70,43,87,50,48,85,79,67,89,67,67,86,80,52,47,DEF
M. Hummels,28,Germany,88,88,48M,215K,2038,62,66,64,58,77,91,64,65,68,55,53,15,6,10,5,6,89,89,68,85,51,85,68,56,85,80,71,90,65,66,92,85,79,60,DEF
```

Figure 34 -- Players DB

After eliminating those issues, we need to decide where the classification will be done at this primary level. For this step WEKA was used. So, given a player's attributes we try to determine their position on the field. Since there are many positions, and many players have more than one (e.g. striker and left winger) there is a generalization for positions: Forward, Midfielder, Defender and Goalkeeper. By selecting the most important elements, the accuracy achieved is almost 82% with Random Forest and SMO using cross-validation with 10 folds. However, the experiments were done in a small percentage of the players (120 players).

Our database in Weka looks like below:

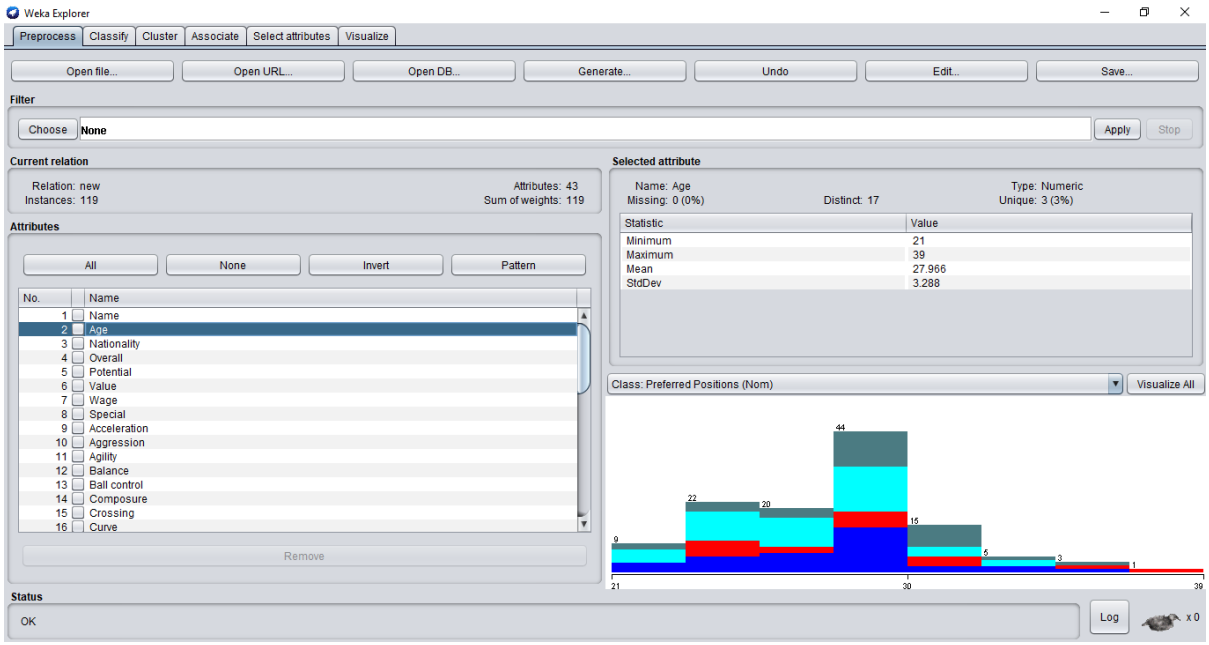


Figure 35 -- DB in Weka

The outcome of random forest can be seen here:

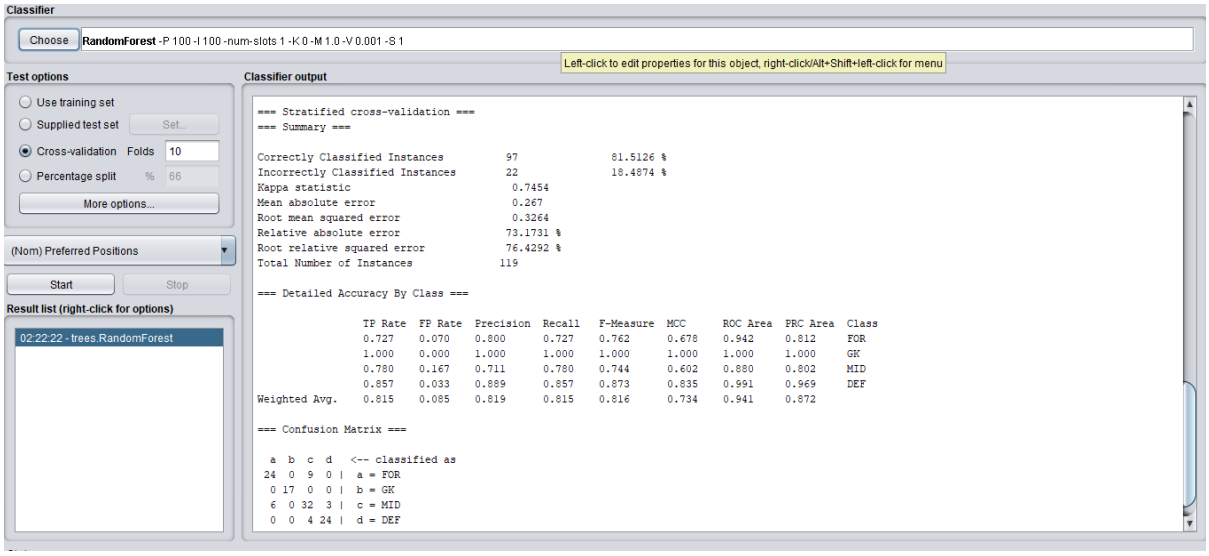


Figure 36 -- Random Forest outcome

The outcome of SMO can be seen here:

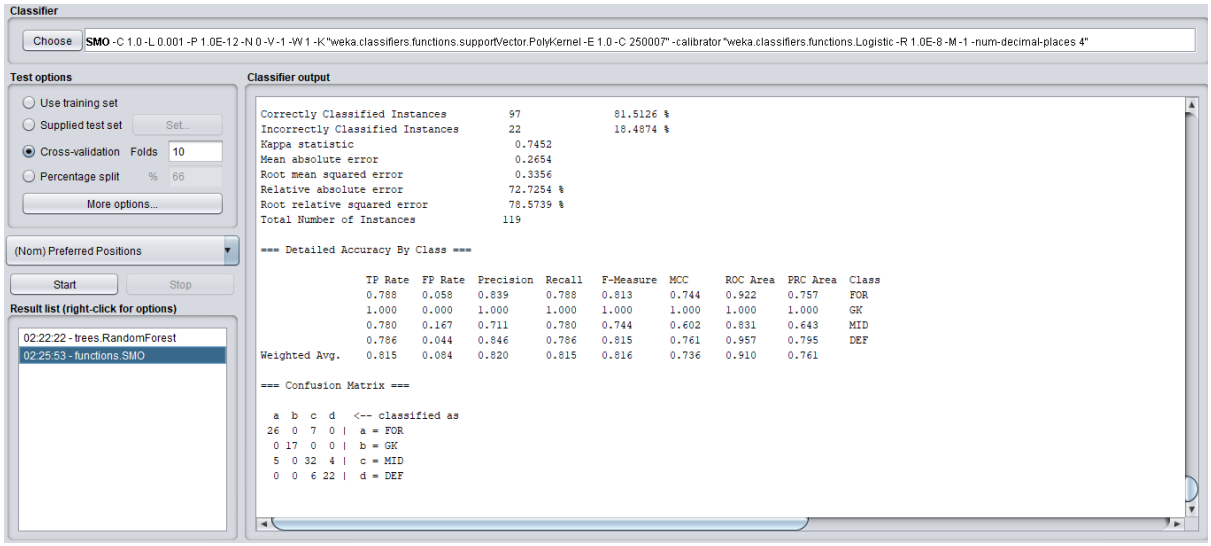


Figure 37 -- SMO outcome

Here is a confusion matrix:

```
=== 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
```

We see that for goalkeepers there aren't any misclassification. The errors occur mostly in the midfielder position due to the generalization that was done. That means that midfielders share attributes with forwards and defenders.

Second Experiment – Number of Goals

2. The next experiment that was conducted, was more related to players performance. Messi and Suarez of Barcelona were used as examples to predict many statistics including the number of goals that the player has scored. In order to be able to examine the process's accuracy, we used data [59] until the season 2016-2017 and then we precited results for season 2017-2018.

The method that was followed was:

Firstly, scrap data from whoscored.com and create a database. We had a csv file that looked like that:

```
Season,Apps,Mins,Goals,Assists,SpG,KeyP,Drb,Fouled,Off,Disp
2016/2017,32,2833,37,9,5.3,2.3,3.7,2.3,0.1,2.6
2015/2016,31,2730,26,16,4.8,2.3,3.5,1.8,0.3,2.2
2014/2015,37,3375,43,18,4.9,2.5,4.6,2.4,0.6,2
2013/2014,29,2508,27,11,5,2.4,4.6,1.5,0.5,2.2
2012/2013,28,2644,46,12,5.1,1.4,3.8,2.3,0.5,2.6
2011/2012,36,3268,50,16,5.5,2.5,4.8,2.2,0.7,2.6
2010/2011,31,2859,31,18,4.5,2,5.6,1.7,0.5,2.9
2009/2010,30,2841,34,10,4.7,1.9,4.1,2.4,0.3,2.5
```

Table of attributes (explanation):

Season	The season for which the player has those stats.
Apps	The number of appearances in a particular season.
Mins	Minutes played
Goals	Total goals
Assists	Total assists
SpG	Shots per game
KeyP	Key passes per game
Drb	Dribbles per game
Fouled	Fouled per game
Off	Offsides per game
Disp	Dispossessed per game

Then, we had to make our test set. Keeping in mind that we do not know anything about season 2017-2018 we created another csv file to be used as a test set that looked like the above. The only difference was that an average value for each attribute was used.

After, getting both csv files in python, we then could run the classification process. However, we should decide for which attribute we need a prediction. Because both Messi and Suarez are key players for Barcelona, with great scoring capabilities the prediction was firstly for the number of goals that they would score in season 2017-2018. So, this attribute had to be dropped in python only for the test set though.

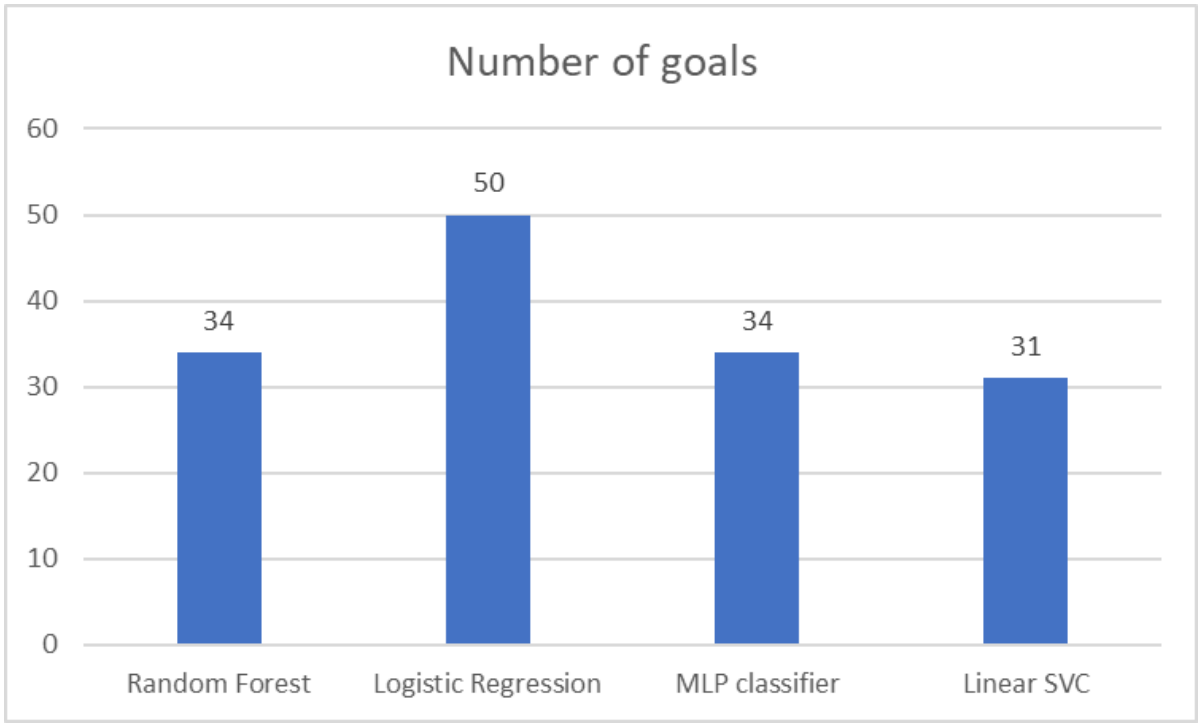
In order to be more accurate about the results the experiments were conducted by using 4 classification algorithms obtained by sklearn libraries.

Those algorithms were:

- Random Forest
- Logistic Regression
- MLP classifier
- Linear SVC

For both players the same method was followed.

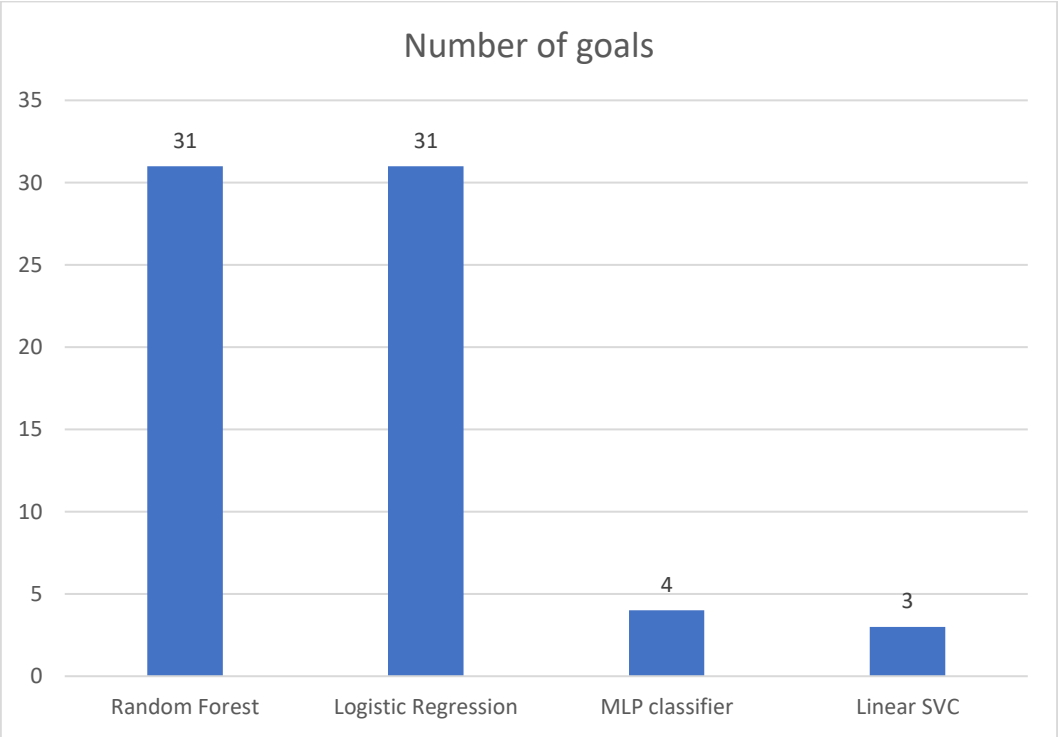
Messi during the season 2017-2018 scored 34 goals. The classifiers we used predicted the following results:



From the above graph, we can determine that the best results were given by Random Forest and MLP classifier followed by Linear SVC. The worst results were provided by Logistic Regression.

The same procedure was followed for Suarez.

Suarez during the season 2017-2018 scored 25 goals. The classifiers provided these results:



We can see that Random Forest and Logistic Regression have the same accuracy. However, MLP classifier and Linear SVC not only provide the worst results, but also their accuracy is terrible.

This experiment shows us that we can predict quite accurately the stats that we are interested in for some players. Thus, for example based on a player's as an individual performance we can predict the team's performance. So, we can predict how many goals a team will score during a season, how many goals will a team concede and so many different statistics. Although, the number of goals that a team will score in a season could be predicted using the team's previous years statistics, but by aggregating each player's statistics we achieve better accuracy. This is actually reasonable, considering that teams change every year and the players that are competing are not the same.

Third Experiment – Next match shot number

For the next experiment, the focus was mainly on the shots that Messi was going to attempt in a match. It was observed that Messi typically scores when he has the chance to shot at least five times in a match.

So, the idea was to predict the number of shots that Messi was going to attempt in a match.

The process was the following:

Firstly, the data that had to be accumulated were from all his previous seasons. A table can be seen below.

Apps	Mins	OutOfBox	SixYardBox	PenaltyArea	Total
8	665	2.8	0.1	1.8	5
36	2997	2.6	0.3	2.6	6
34	2833	2.2	0.2	2.9	5
33	2730	2	0.3	2.5	5
38	3375	1.6	0.4	2.9	5
31	2508	1.9	0.3	2.9	5
32	2644	1.9	0.3	2.8	5
37	3268	1.7	0.5	3.3	6
33	2859	1.2	0.4	3	5
35	2841	1.7	0.3	2.6	5
1	90	5	0	0	5
1	90	0	0	2	2
1	90	4	0	3	7
1	90	2	0	0	2

Figure 38

In the last lines of the above table, we have data about Messi's current season. Then we need to create a test set about the next match.

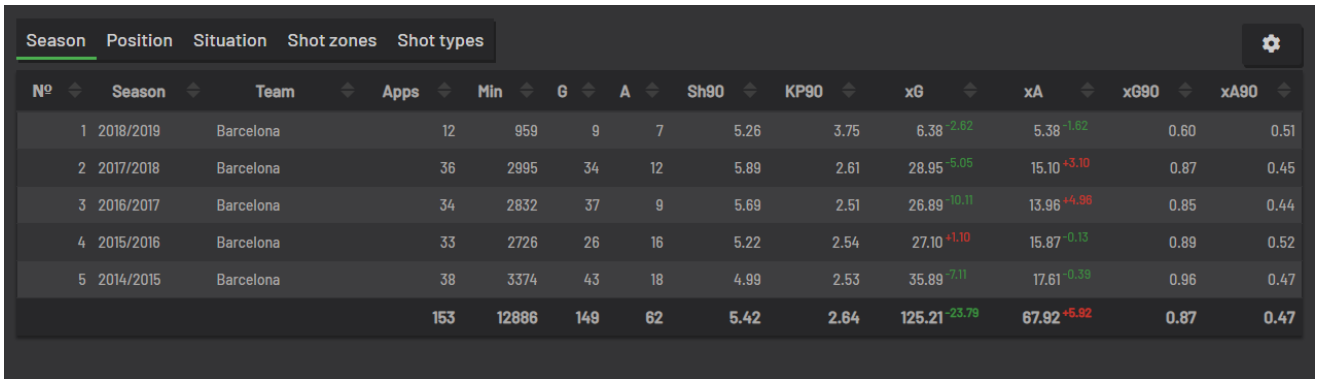
Apps	Mins	OutOfBox	SixYardBox	PenaltyArea
1	90	2,75	0	1,25

Figure 39

For this we had to used data from the website: understat.com.

It has statistics for a plethora of players and also for a lot of their attributes.

Specifically, for Messi we can get:



№	Season	Team	Apps	Min	G	A	Sh90	KP90	xG	xA	xG90	xA90
1	2018/2019	Barcelona	12	959	9	7	5.26	3.75	6.38 -2.62	5.38 -1.62	0.60	0.51
2	2017/2018	Barcelona	36	2995	34	12	5.89	2.61	28.95 -5.05	15.10 +3.10	0.87	0.45
3	2016/2017	Barcelona	34	2832	37	9	5.69	2.51	26.89 -10.11	13.96 +4.96	0.85	0.44
4	2015/2016	Barcelona	33	2726	26	16	5.22	2.54	27.10 +1.10	15.87 -0.13	0.89	0.52
5	2014/2015	Barcelona	38	3374	43	18	4.99	2.53	35.89 -7.11	17.61 -0.39	0.96	0.47
			153	12886	149	62	5.42	2.64	125.21 -23.79	67.92 +5.92	0.87	0.47

Figure 40 – [57]

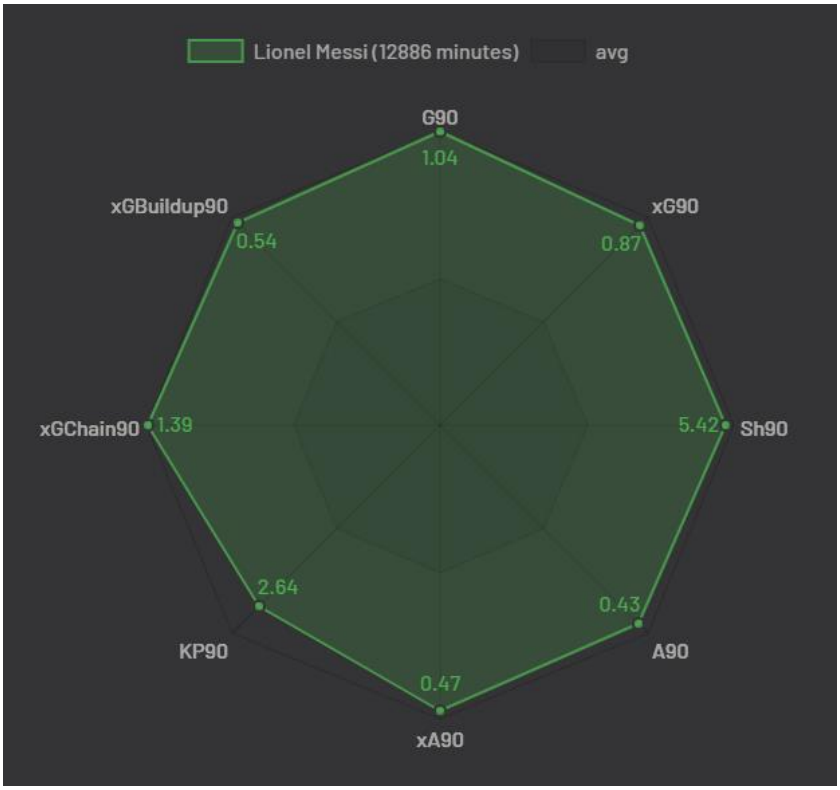


Figure 41-- [57]

The explanation of the above abbreviations can be seen on the table below:

AMC	Attacking Midfield Center
FWR	Forward Right
FW	Forward
FWL	Forward Left
Sub	Substitute
G90	Goals per 90 minutes
xG90	Expected goals per 90 minutes
Sh90	Shots per 90 minutes
A90	Assists per 90 minutes
xA90	Expectesd assists per 90 minutes
KP90	Passes lead to shot per 90 minutes
xGChain90	Total XG of every possession the player is involved in per 90 minutes
xGBuildup90	Total XG of every possession the player is involved in without key passes and shots per 90 minutes

Figure 42 -- [57]

Below there is an image that represents every shot that Messi did. The different colors describe the outcome of the shot.

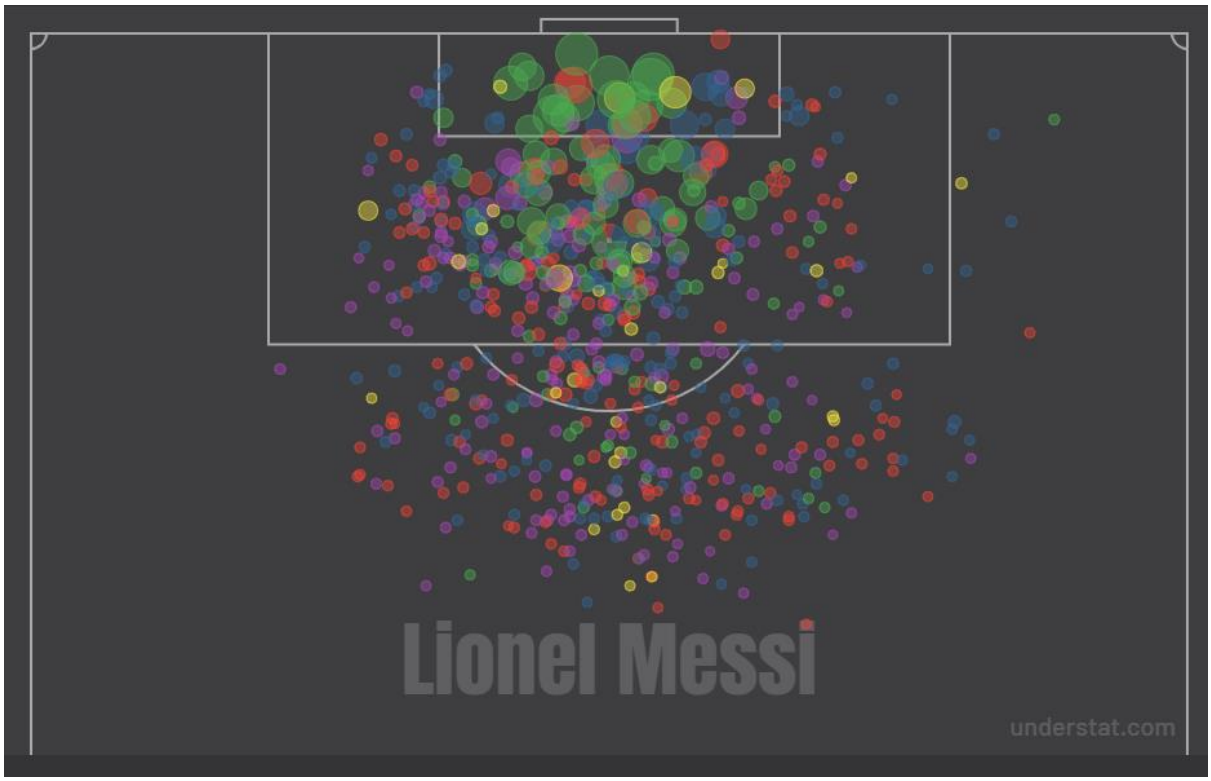


Figure 43 -- [57]

After completing our training and test-set and preprocess the data we can go to python again for experiments. We will show an example for the football match that Barcelona had against Real Sociedad on 15 September 2018.

We run the code python and we get from RandomForest that the attempts on Messi's shot on this match are going to be 2.133. By rounding that number, we can say that Messi is going to shoot twice.

Here are the stats from the actual match:

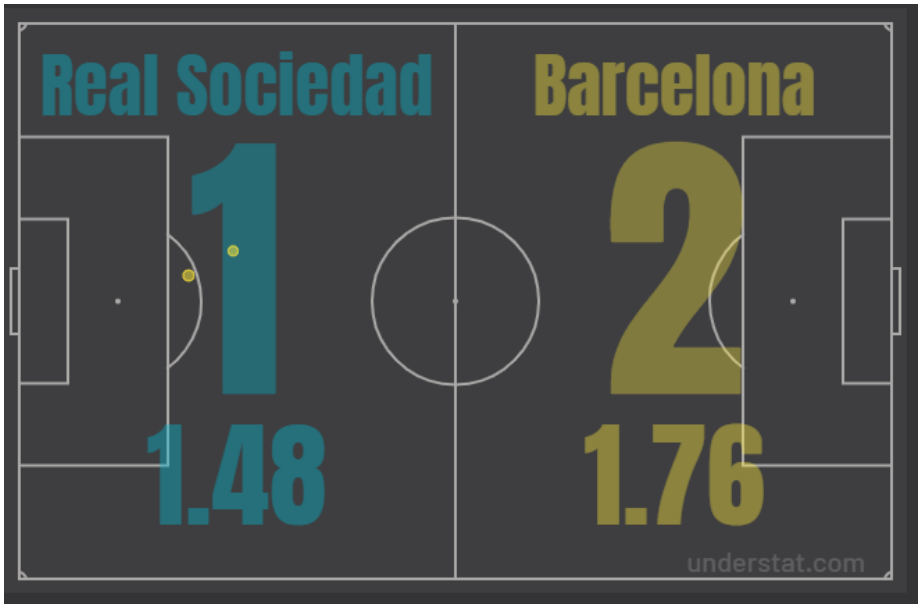


Figure 44 -- [57]

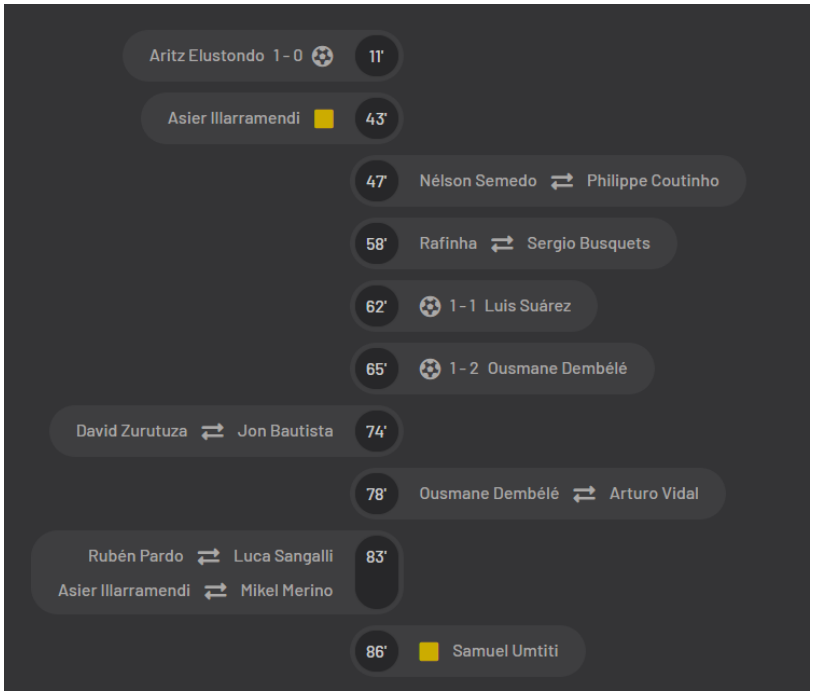


Figure 45 -- [57]

As we can see from the figures above, Messi actually shot twice and also, he didn't score. The same technique can be applied for other Barcelona matches with Messi as an example.

3. After completing the above step the next experiment was to find a relation between goal difference (goal forward and goals against) and the points that a team gathers at the end of the season. However, the results were not encouraging.

During, conducting the experiment and searching for related work, there was already conducted such an experiment for Premier League for season 2016-2017. [58]

The stats were obtained again from whoscored.com and Skysports. The process was predicting a 62% chance of a team being in the top four and an 83.5% accuracy for a team being in the relegation zone [58].

However, the goal difference was an accurate hint of how well a team would perform in the next season. The interesting part was that the number of shots that a team does in a game is inversely correlated on how well the team will do next year. This might have to do with the fact that many shot attempts could mean that the team is incapable of scoring [58].

At the end of the day, despite the fact that this experiment does not provide the desired results there is still space for data analysis in order to determine if there is another key factor that could play an important role for a team's point gathering during a season.

Chapter 6

Tools and technologies used

For the first experiment that was clearly a classification issue, determining what is the position of a football player in the field **Weka** was used.

"The WEKA workbench is a collection of machine learning algorithms and data preprocessing tools that includes virtually all the algorithms described in our book. It is designed so that you can quickly try out existing methods on new datasets in flexible ways. It provides extensive support for the whole process of experimental data mining, including preparing the input data, evaluating learning schemes statistically, and visualizing the input data and the result of learning.

As well as a wide variety of learning algorithms, it includes a wide range of preprocessing tools. This diverse and comprehensive toolkit is accessed through a common interface so that its users can compare different methods and identify those that are most appropriate for the problem at hand. WEKA was developed at the University of Waikato in New Zealand; the name stands for Waikato Environment for Knowledge Analysis. Outside the university the WEKA, pronounced to rhyme with Mecca, is a flightless bird with an inquisitive nature found only on the islands of New Zealand. The system is written in Java and distributed under the terms of the GNU General Public License. It runs on almost any platform and has been tested under Linux, Windows, and Macintosh operating systems." [61]

Algorithms Used

Random Forest

Random forest is an easy to train algorithm and has many advantages. It belongs to the supervised learning category and can be used for both classification and regression issues [62].

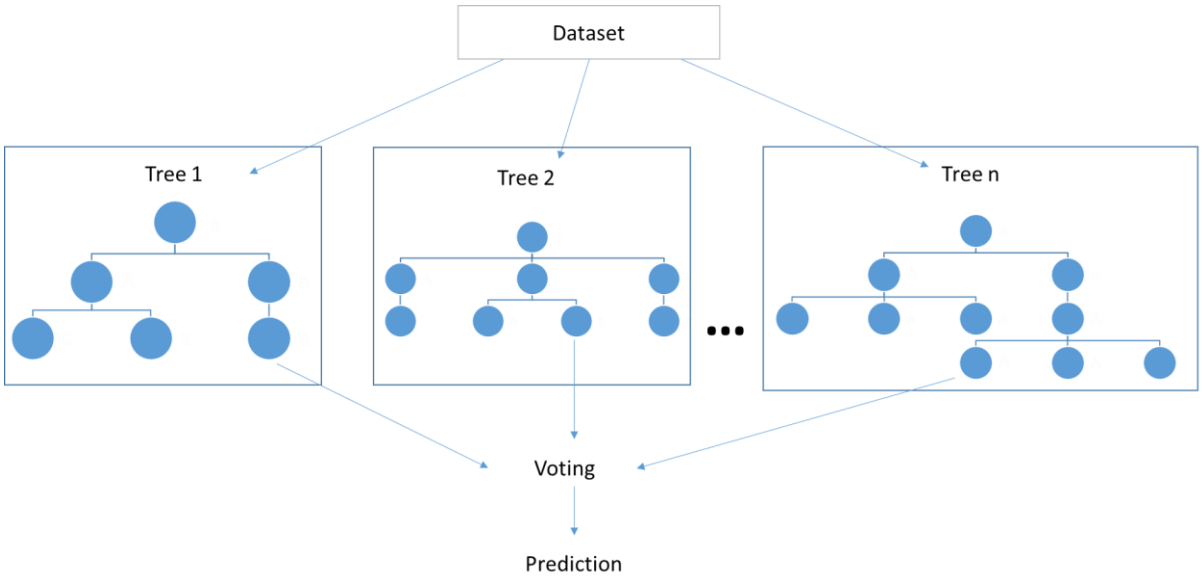


Figure 46 -- [63]

SMO

SMO stands for Sequential minimal optimization. It is used for the solution of the quadratic problem that occurs during the application of support vector machines. It was introduced in 1998 by John Platt. [64]

It belongs to the support vector machines family and it runs sequentially and not in parallel.

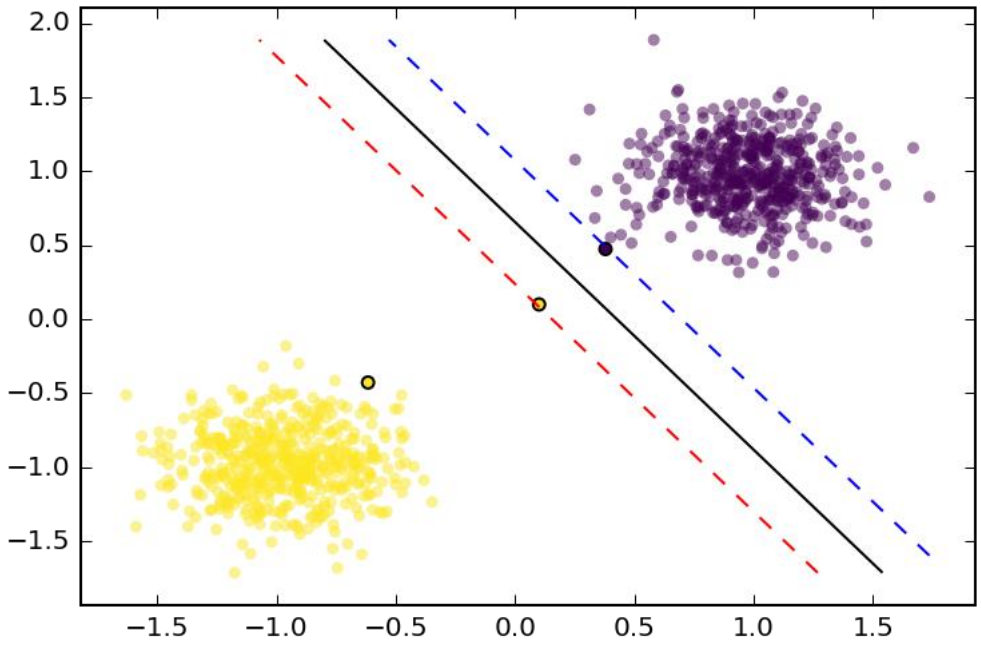


Figure 47 -- SVM [65]

Logistic Regression

"Logistic regression, despite its name, is a linear model for classification rather than regression. Logistic regression is also known in the literature as logit regression, maximum-entropy classification (MaxEnt) or the log-linear classifier. In this model, the probabilities describing the possible outcomes of a single trial are modeled using a logistic function." [67]

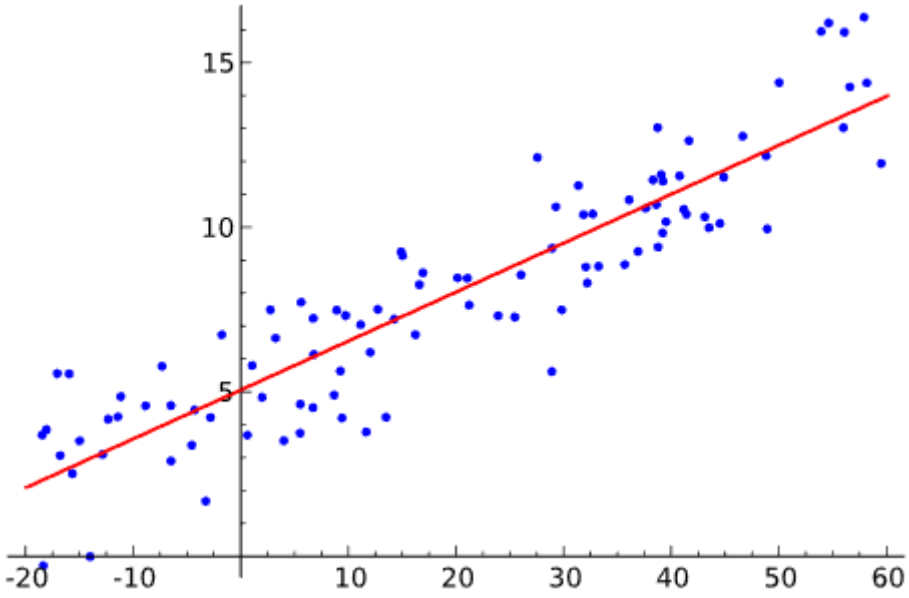


Figure 48 -- Logistic Regression [66]

MultiLayer Perceptron

"Multi-layer Perceptron (MLP) is a supervised learning algorithm that learns a function by training on a dataset, where is the number of dimensions for input and is the number of dimensions for output. Given a set of features and a target, it can learn a non-linear function approximator for either classification or regression. It is different from logistic regression, in that between the input and the output layer, there can be one or more non-linear layers, called hidden layers." [68]

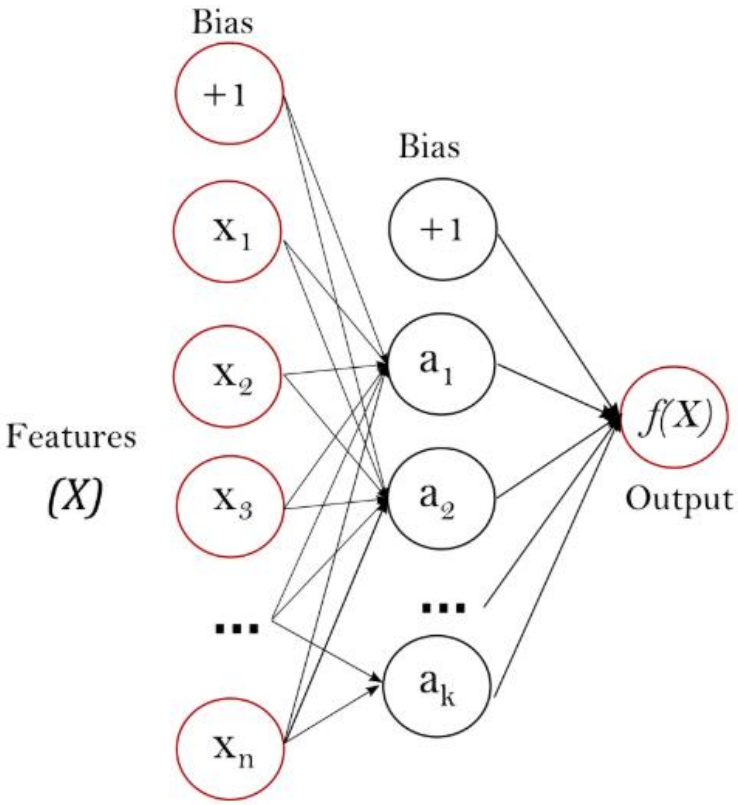


Figure 49 -- Multilayer Perceptron [68]

Linear SVC

SVC is a more general version of the SMO described above. It is actually a support vector machine.

Linear SVC is a subset of SVC.

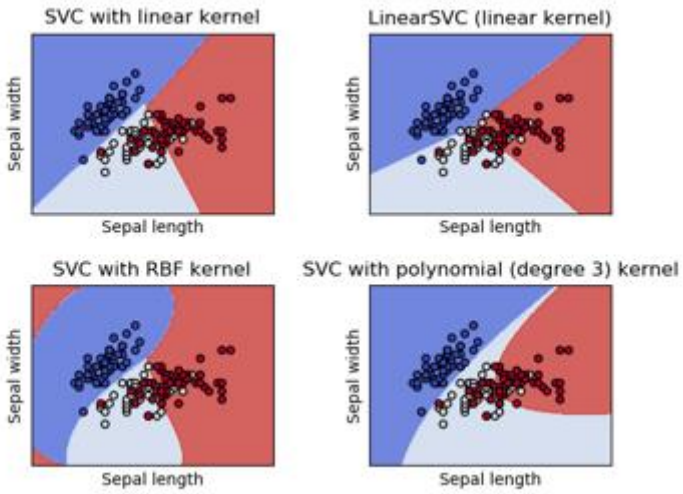


Figure 50 -- SVC [69]

Chapter 7

Conclusion

At the end of the day, we can understand that sports analytics is going to be an important part of a team's performance in the following years. There is going to be a huge amount of data that with the proper exploitation methods will improve the accuracy of the algorithms.

More and more teams are embedding such practices to their inventory for improving their game. Barcelona has a whole team of data scientists for data analytics in order to study their game.

In this study, we explored 3 basic aspects of sports analytics in football.

The first one, was to classify an athlete to the position that he is more suitable, by using statistics that were available about their style of game. We received a good result, taking into consideration that some positions in the field are highly correlated, for example an "extreme" with a midfielder. The good news is, that there were not any misclassifications for the defenders and strikers or goalkeepers.

The second, was the prediction of the number of goals that Messi and Suarez are going to score in a season. By using proper algorithms, we managed to achieve very high accuracy, and the predicted number was really close to the real one.

Finally, the third was the prediction of the number of shots that Messi was going to perform at a specific match of a season. By accumulating proper data and building suitable training and test sets to provide to the algorithms, again we had a really good prediction close to the real number. The importance of the number of shots can be clarified by the fact that there is a correlation to a goal scoring possibility. In this case for example, Messi is possible to score in he attempts at least five shots in a game.

Evaluation and threats to validity

As explained above, we had good accuracy and achieved good results to the experiments that were conducted.

However, for the purposes of this study, we had to focus on specific athletes. It is understandable, that these approaches cannot be applied to all athletes, because they have different characteristics. We have to face every athlete as an individual in order to achieve the results we want.

Future Work

One of the most important aspect for players performance is data collection of wearable devices. Wearable devices can be worn on human bodies either as clothes or as accessories like watches. Thanks to their sensors, they have the ability to provide information about the distances that an athlete covers, the heart rate, the sweat percentage, even data about sleep.

By using those data, an athlete has the opportunity to know his weaknesses and improve. In our topic, which is mostly about football not only the athlete will have access to those data, but also the coach. So, if the football player reaches his limits the coach will know to introduce into the game a reserve. It will improve the team's performance and the coach will adjust the squad with the players that remain available.

Data from wearable devices is a new topic that can provide many key aspects for a team's success. However, it is difficult to obtain those data. Even sport association for football and basketball do not allow the use of these devices at great extent. Probably because they want to protect not only the nature of the sports but also the athletes themselves.

Due to the nature of team sports, such as football, the performance of a team is highly related on how the teammates interact among each other, not only on their individual statistics. In addition, a great role plays the interaction with the opponent.

Although some of those statistics that wearable devices provide, could be obtained by other media, such as cameras that were mentioned above, there is a big difference. And that is that the data are provided real time, while from cameras are at the end of the match.

As a final step, coaches could have all those devices send data to a mobile platform, that will have the ability to make suggestions of what changes should be made. For example, it would be really helpful to advise the coach that a player is tired, and he should make a substitute or that if he throws in the game another player, he may achieve a better result. We can't although rely always on these devices, because sometimes instincts and luck play a really important role for a game outcome.

However, it depends on sport organizations the degree that they will allow the use of those data. We are still at the beginning of exploring this unknown territory.

References

[1] "Goal Scoring in Association Football: Charles Reep". Keithlyons.me. Retrieved 21 November 2013.

[2] Wilson, Jonathan (2009). Inverting the Pyramid: The History of Football Tactics. Orion Books. pp. 138–144

[3] https://www.soccercoachweekly.net/soccer-drills-and-skills/long-ball-game/

[4] Steve Sullivan, State of the Art: The Actuarial Game of Baseball, http://www.contingencies.org/mayjun04/stat.pdf Archived September 27, 2011, at the Wayback Machine.

[5] Glenn Miller (April 28, 1984). "Top Secret: Project Scoresheet to bring hidden facts to the fans". Evening Independent.

[6] Danny Ecker (May 15, 2014). "Stats LLC sold to private-equity firm". Crain's Chicago Business.

[7] Jaffe, Chris (February 4, 2008). "Bill James Interview". The Hardball Times. Archived from the original on February 7, 2008. Retrieved February 16, 2008.

[8] "Bill James Explains New 'Temperature Gauge' Statistic to Determine How Hot or Cold a Hitter Is". NESN. May 7, 2012. Retrieved June 25, 2012.

[9] "Movie Review: "Moneyball," starring Brad Pitt and Jonah Hill". Sports of Boston. September 22, 2011. Archived from the original on September 18, 2012. Retrieved June 25, 2012.

[10] https://www.billjamesonline.com/stats/

[11] Baseball Prospectus 2005, pp.69–70

[12] "Industry Insights: The Evolution of Basketball through Technology | Warsaw Sports Business Club". wsbc.uoregon.edu. Retrieved June 26, 2015.

[13] McCann, Zach (May 19, 2012). "Player tracking transforming NBA analytics". ESPN. Retrieved March 3, 2016.

[14] "Competitive fire helps Kirk Lacob make his own name with Warriors". Retrieved June 26, 2015.

[15] "New age of NBA analytics: Advantage or overload? - The Boston Globe". Retrieved June 26, 2015.

[16] "Bigger than LeBron: How SportVU Will Change Basketball - The Airspace". Retrieved June 26, 2015.

[17] https://www.nba.com/2013/news/features/david_aldridge/11/11/morning-tip-sportvu-cameras-in-arenas-problems-with-nets-qa-with-paul-george/

[18] "Stats LLC and NBA to make STATS SportVU Player Tracking data available to more fans than ever before - NBA.com: NBA Communications". NBA.com: NBA Communications. 2016-01-19. Retrieved 2017-04-02.

[19] https://www.stats.com/player-tracking/

[20] http://sportsvideo.org/main/files/2012/07/Screen-Shot-2012-07-26-at-1.24.31-PM.jpg

[21] https://nyulocal.com/cameras-are-the-future-of-front-office-decisions-in-sports-f26d4f60de2c

[22] https://www.moddb.com/games/sim-betting-football/images/match-summary-screen-statistics

[23] Jack Corscadden, Ross Eastman, Reece Echelberger, Connor Hagan, Clark Kipp, Erik Magnusson, Graham Muller, Stephen Adams, James Valeiras, and William T. Scherer, "Developing Analytical Tools to Impact U.Va.Football Performance"

[24] Paolo Cintia, Luca Pappalardo, Dino Pedreschi, "The harsh rule of the goals: data-driven performance indicators for football teams"

[25] Javier Fernandez, Daniel Medina, Antonio Gomez, Marta Arias, Ricard Gavalda, "From Training to Match Performance: A Predictive and Explanatory Study on Novel Tracking Data"

[26] https://football-technology.fifa.com/en/media-tiles/epts-1/

[27] Babak Moatamed, Sajad Darabi, Migyeong Gwak, Mohammad Kachuee, Casey Metoyer, Mike Linn and Majid Sarrafzadeh, "Sport Analytics Platform for Athletic Readiness Assessment"

[28] Monika Nawrocka, Marcin Łukowski, "Biofeedback EEG data integration and visualization analytics for endurance exercise practices", 2017

[29] Raquel Y.S. Aoki,Renato M. Assunção,Pedro O.S. Vaz de Melo,"Luck is Hard to Beat: The Diiculty of Sports Prediction",2017

[30] https://www.wearable-technologies.com/2016/10/the-hottest-wearables-for-sports-medicine-at-medica-2016/

[31] Mahanth Gowda, Ashutosh Dhekne, Sheng Shen, Romit Roy Choudhury, Sharon Xue Yang, Lei Yang, Suresh Golwalkar, Alexander Essanian, "IoT PLATFORMS FOR SPORTS ANALYTICS"

[32] Sara Landset, Michael F. Bergeron, Taghi M. Khoshgoftaar,"Using Weather and Playing Surface to Predict the Occurrence of Injury in Major League Soccer Games: A Case Study", 2017

[33] Daniele Ravı, Charence Wong, Benny Lo, Guang-Zhong Yang,"A deep learning approach to on-node sensor data analytics for mobile or wearable devices"

[34] Li Deng and Dong Yu,"Deep Learning Methods and Applications"

[35] Manish Sharma, Rupika Srivastava, Akash Anand, Divya Prakash, Lakshmi Kaligounder, "WEARABLE MOTION SENSOR BASED PHASIC ANALYSIS OF TENNIS SERVE FOR PERFORMANCE FEEDBACK"

[36] https://www.samsung.com/gr/wearables/gear-s2-r720/SM-R7200ZKAEUR/

[37] Manuel Stein, Halldor Janetzko, Andreas Lamprecht, Thorsten Breitkreutz,Philipp Zimmermann, Bastian Goldlucke, Tobias Schreck,Gennady Andrienko,Michael Grossniklaus,Daniel A. Keim,"Bring it to the Pitch: Combining Video and Movement Data to Enhance Team Sport Analysis", 2018

[38] Rui Zeng, Ruan Lakemond, Simon Denman, Sridha Sridharan, Clinton Fookes, Stuart Morgan, "Vertical Axis Detection for Sport Video Analytics", 2016

[39] Chenyan Xu, Yang Yu, "Measuring NBA Players' Mood by Mining Athlete-Generated Content",2015

[40] "Data Mining Curriculum". ACM SIGKDD. 2006-04-30. Retrieved 2014-01-27

[41] Hawkins, Douglas M (2004). "The problem of overfitting". Journal of Chemical Information and Computer Sciences

[42] https://www.mathworks.com/matlabcentral/mlc-downloads/downloads/submissions/24616/versions/14/screenshot.jpg

[43] https://gerardnico.com/_media/data_mining/classification.jpg?cache

[44] https://www.saedsayad.com/association_rules.htm

[45] https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/

[46] https://www.coursera.org/learn/machine-learning

[47] https://www.mckinsey.com/industries/high-tech/our-insights/an-executives-guide-to-machine-learning

[48] Pedro Domingos, "A Few Useful Things to Know about Machine Learning"

[49] Tom M. Mitchell, "The Discipline of Machine Learning", 2006

[50] https://towardsdatascience.com/machine-learning-65dbd95f1603

[51] Christos Berberidis,"Machine Learning Principles and Concepts"

[52] https://towardsdatascience.com/unsupervised-learning-with-python-173c51dc7f03

[53] https://towardsdatascience.com/types-of-machine-learning-algorithms-you-should-know-953a08248861

[54] Walter Tichy, "Changing the Game: “Dr. Dave” Schrader on Sports Analytics"

[55] http://thedatascientist.com/why-now-is-the-right-time-for-sports-analytics/

[56] http://tanukamandal.com/2017/12/12/sports-analytics-changed-play/

[57] understat.com

[58] https://theartandscienceofdata.wordpress.com/2016/09/17/predicting-the-english-premier-league-standings/?fbclid=IwAR1YpEtW4Futhk05knAAUyW1so_S8tn7v2LCT7vvqGNfhsmGMiJ-wkIjEE4

[59] www.whoscored.com

[60] sofifa.com

[61] Eibe Frank, Mark A. Hall, and Ian H. Witten, "The WEKA Workbench"

[62] https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd

[63] https://analyticsdefined.com/introduction-random-forests/

[64] Platt, John (1998), Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines

[65] https://jonchar.net/notebooks/SVM/

[66] https://en.wikipedia.org/wiki/Logistic_regression#/media/File:Linear_regression.svg

[67] https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression

[68] https://scikit-learn.org/stable/modules/neural_networks_supervised.html

[69] https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html