INTERNATIONAL
HELLENIC
UNIVERSITY

Sports Analytics Algorithms for Performance Prediction

Chazan – Pantzalis Victor

SID: 3308170004

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in Data Science

DECEMBER 2019

THESSALONIKI – GREECE

INTERNATIONAL
HELLENIC
UNIVERSITY

Sports Analytics Algorithms for Performance Prediction

Chazan – Pantzalis Victor

SID: 3308170004

Supervisor:	Prof. Christos Tjortjis
Supervising Committee Members:	Dr. Stavros Stavrinides
	Dr. Dimitris Baltatzis

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in Data Science

DECEMBER 2019

THESSALONIKI – GREECE

Abstract

Sports Analytics is not a new idea, but the way it is implemented nowadays have brought a revolution in the way teams, players, coaches, general managers but also reporters, betting agents and simple fans look at statistics and at sports.

Machine Learning is also dominating business and even society with its technological innovation during the past years. Various applications with machine learning algorithms on core have offered implementations that make the world go round.

Inevitably, Machine Learning is also used in Sports Analytics. Most common applications of machine learning in sports analytics refer to injuries prediction and prevention, player evaluation regarding their potential skills or their market value and team or player performance prediction. The last one is the issue that the present dissertation tries to resolve.

This dissertation is the final part of the MSc in Data Science, offered by International Hellenic University.

Acknowledgements

I would like to thank my Supervisor, Professor Christos Tjortjis, for offering his valuable help, by establishing the guidelines of the project, making essential comments and providing efficient suggestions to issues that emerged. I would also like to thank him for his promptness and his cooperativeness throughout the whole process.

I would like to thank my family for their support and their patience during the past few months. They have done everything possible for me and I could not have made it without them.

This dissertation is devoted to the little guy I do everything for, my greatest inspiration, my son, Vasilis.

"Research is to see what everybody else has seen
and to think what nobody else has thought."
Albert Szent–Gyorgyi

Chazan–Pantzalis Victor
25–11–2019

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This dissertation is split in 6 chapters; the first chapter is the present introduction part. In the second chapter, historical background of sports analytics in certain games, as well as outstanding previous work in sports analytics are presented. In chapter three, general terms, algorithms and tools that are mentioned or used in the dissertation are described. A short definition to the problem this thesis deals with, along with the approach followed is the subject of the fourth chapter. Experiments and their evaluation are conducted on chapter five. Finally, in chapter six, conclusions are drawn and future work is suggested.

The dissertation is focused on football (soccer). The scope of the dissertation is team and player long–term performance prediction. Football was selected because of the uncountable statistical categories and historical data that provides, because of its fame and because of the simplicity of its rules and of national championships formats. On the other hand, there are special difficulties that are extensively explained, which make long–term prediction in football a difficult task.

There is an abundance of online data regarding football, which is an asset, but also creates the need of filtering and usage of proper data for team and player performance prediction. Unfortunately, this is not always easy, as sometimes there is an intersection of data for team performance prediction and data for player performance prediction. Thus, it is very important to find the correct proportion of data used for the experiments in each case.

Additionally, team and player performance can be affected by incidents that are not depicted in the data collected; a team is rated higher than should be when their opponents play really bad. A player might have a low rating when coming into action after a serious injury.

Finally, the nature of football makes statistical recording of game events and the player and team rating, an ambiguous process. Following the same pattern, performance prediction is not easy either and long–term performance prediction is even tougher but also not sufficiently studied.

Nevertheless, as it is proved in this dissertation, it is possible, up to a certain level, to make some long–term predictions, especially for team performance. Specifically, in this thesis, a reliable prediction of the final league table for certain leagues is presented. Prediction is relatively good, mainly with regard to the champion and the teams that win European qualification. The points each team gathers are calculated after simulating every match of the season. What makes this research interesting is that the prediction is conducted *before* the

beginning of the season, with no official games played, only with historical data and the information gathered during summer break. Another novelty of this dissertation is that *advanced statistics* from previous seasons (e.g. expected goals) are used for prediction.

Other predictions for team performance in the thesis refer to whether a team is going to have a better season than last one. Furthermore, another issue in the dissertation is the detection and recording of personal skills and statistical categories that separate an excellent central defender from an average central defender.

There is an increasing interest in sports analytics and in performance prediction. Clubs have started using sophisticated devices and software in order to gather and analyze data generated by players during training sessions and matches. Data processing is a valuable aid to managers, coaches, football analysts and trainers in short–term decision making and in long–term organization development.

Sports analytics are also a hot trend for betting reasons. Bets are becoming more complex, gamblers are searching for uncommon betting choices and companies have to satisfy clients' needs, but on the same time they have to offer reasonable stakes. Therefore, an extensive analysis of all data available is a prerequisite for betting companies.

Finally, fans are also very interested nowadays in advanced statistics and how they affect a game of football. Until recently, newspapers and web sites in their articles offered only a simple game recap with few simple statistics recorded. During the last few years, web sites with high–level of knowledge in football tactics have emerged. They are gaining popularity very fast and use sports analytics to support their view on football.

For all the aforementioned reasons, the use of sports analytics has increased during the last few years and will continue to increase; it was only recently that Fédération Internationale de Football Association (FIFA) allowed the usage of optical and GPS tracking systems, known as Electronic Performance & Tracking Systems (EPTS), by teams. Those devices will bring a revolution in the way data are collected, as well as in the amount of collected data. For sports analytics, there is a golden era ahead of us.

Chapter 2

2.1 Historical Background

Sports Analytics (SA) is a relatively new science field that is now being used in every sport by almost every professional team or every professional athlete. However, many years before scientists actually dealt with SA, individual coaches, sportsmen and sport reporters grasped the importance of collecting and analyzing sports data and used them alongside traditional training.

Even the statistics that we now consider "traditional" were not being recorded at most sports some decades ago. For example, we do not know who gave the most assists at the World Cup football tournament of 1950, as no one had realized that a pass leading to a goal is a significant statistical category and should be recorded.

Unsurprisingly, football was not the first sport that SA were implemented. The first attempts by sportsmen to use statistics and analysis based on collected data was made to older sports, like baseball, tennis or even boxing.

2.1.1 Baseball

A lot of people have watched the movie "*Moneyball*" (2011) by director Bennett Miller and have read the homonym book by Michael Lewis back from 2003. It's the real story of how Billy Beane, the General Manager (GM) of the Oakland Athletics, a team playing in the Major League Baseball (MLB) of the United States, managed to create a successful team, despite of the small budget available. Beane and his assistant, Paul DePodesta, used a sabermetric approach to sign players and employ tactics. Sabermetrics (term derived by the acronym SABR, meaning Society for American Baseball Research) is the detailed and empirical analysis of baseball data and statistics. Some people know the story behind the book and are aware of sabermetrics, but few of them actually know that sabermetrics were not invented or first used by Billy Beane in 2002.

Some early forms of baseball were being played in England even before 1750. The immigration brought those games at the other side of the Atlantic Ocean and the first recorded baseball game was held in 1838, in Ontario, Canada. The first recorded game in USA took place, in Hoboken, New Jersey in 1846, a year after the first set of rules for a baseball game was formed in New York. National Association of Baseball Players was formed in 1857 and National League was founded in 1876. Another league, the American League was founded in

1893 and was conducted separately form National League for ten years. Starting from 1903, the winners of each league compete to each other to win the World Series Champion. So, baseball is an old traditional American sport. Therefore, it is not by luck that it is the game with the most statistics recorded than any other game in history.

It was the 1861 when a newspaper reporter and amateur statistician, Henry Chadwick claimed that an analysis on the athletes' play at the bat and in the field should be made in order their skills to be assessed. He also invented a scorekeeping template, so that games would be recorded consistently. Additionally, he listed totals of games played, outs, runs, home runs, and strikeouts, the first baseball database. Chadwick presented his work by publishing a rulebook called "*Beadle's Dime Base Ball Player*", which nowadays is sought after by collectors in its original form [1].

In 1906, the Chicago Cubs, having dominated the regular season with the best record in both leagues (116 wins in 154 games) reached to the World Series final. The opponent was their fellow citizens, Chicago White Sox, known as the "Hitless Wonders". The nickname was created because the Sox had the worst batting average in the American league (.230). Sportswriter Hugh Fullerton, however, after analyzing the two opponents predicted that the ostensibly weaker White Sox side would win the series and actually White Sox did win the title by 4 games to 2, creating one of the biggest upsets in baseball history. Having gained fame because of his prediction, Fullerton published the book "*Touching Second: The Science of Baseball*" in 1910. Despite the fact that the book had some inaccuracies, it is remarkably interesting in respect to statistics and analysis. Even complex diagrams exist in its pages [2].

In 1913, "Elias Sports Bureau" was founded by Munro brothers. It was the first sports data company. Its objective was to keep and sell baseball data. The success was immense and the firm is still in the business under the same activity.

A very innovative man, Branch Rickey served as a GM in multiple teams for almost 40 years. Rickey made some very pioneering efforts during the first half of the 20th century, namely the development of farm system (i.e. a minor league system) at the 30's and the break of the color barrier in 1945, signing African American player Jackie Robinson.

While Rickey was the President and GM of Brooklyn Dodgers during the 40's, he was the first to hire a full–time statistician, named Allan Roth in 1947. The objective was to reform Roth's observations regarding to team data into game strategies. Rickey was one of the first executives to grasp the value of statistics and SA. Actually, many regard him as the father of sabermetrics, before even baseball analytics were called that way. Along with Roth they made revolutionary observations. For example, they concluded that on–base percentage was more meaningful statistic than batting average or that a player performance was impacted by left and right hand splits.

More important, in August of 1954, LIFE magazine published an article by Branch Rickey called "*Goodby to Some Old Baseball Ideas*". In the article, as seen in Figure 1, Rickey presented his own formula for evaluating team efficiency, a very fascinating attempt in order to measure and quantify what it takes to win a baseball game. In the second part of his formula, related to team defense, Rickey used metrics that even nowadays are considered modern sabermetrics. At the beginning of the article, Rickey described his equation as "*a device for measuring baseball*" and noted that "*It is most disconcerting and at the same time the most constructive thing to come into baseball*" [3].

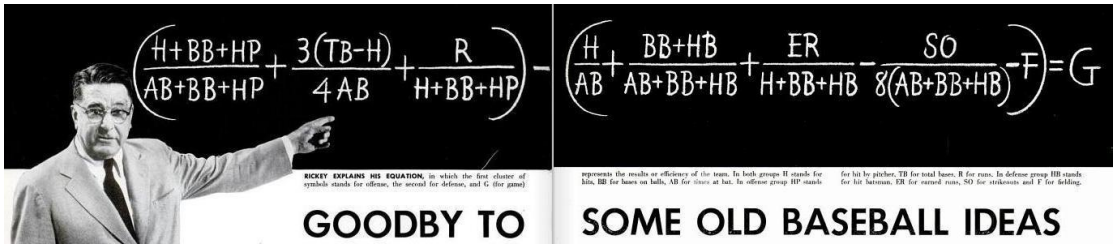


Figure 1: Branch Rickey explains his famous formula for evaluating the efficiency of a baseball team **[3]**.

As technology evolved and computers emerged, it was only a matter of time before they were used for sports analysis. The Mills brothers rented an IBM 1620 and used it to investigate and measure what makes a clutch player. In 1970 they presented their conclusions in a book called "*Player Win Averages: A Computer Guide to Winning Baseball Players*".

The next year, baseball researcher and writer L. Robert Davids took the initiative to call circa 30 "statistorians" as he named them (i.e. people interested in baseball history and statistical research) for an organizational meeting in Cooperstown, New York. Those operations lead to the foundation of the Society for American Baseball Research (SABR). SABR's main objective is to promote the research of history and statistics in baseball.

However, SA still had not gained the attention of many team executives and the fans. Bill James, who now is considered as maybe the most influential person in the history of baseball, was obliged to self–publish his annual book "*The Bill James Baseball Abstract*" from 1977 up to 1980. The first edition sold less than 80 copies, but James was passionate and obsessive. Very soon he became popular among baseball fans and a media conglomerate agreed to publish and promote future editions of his abstracts. That was a very decisive moment for the sport in general, as it was the first time that a book about SA reached a mass audience. His abstracts were written in a very stimulating way and contained in–depth statistics, but also study and analysis of the previous season in an effort to decide what is the line that separates winning from losing teams. Bill James was also the first man that used the term sabermetrics. After publishing almost 30 books devoted to baseball history and statistics, he was hired and still now serves as a senior advisor on Baseball Operations for the Boston Red Sox [4].

In 1981, STATS Inc. was founded by John Dewan. It is a company that provides content to multimedia platforms, television broadcasters, leagues, teams and players. It covers more than

300 leagues and 83,000 events on annual basis. Following STATS, similar companies like Opta and Prozone were created during the 90's. Websites devoted to sabermetrics, like BaseballProspectus.com emerged with the spread of the internet. STATS is being noted here because Bill James has worked for the company in the past and because STATS' first customer was a baseball team, the Oakland Athletics.

SA were gaining ground but still did not dominate the baseball industry. That was about to change with the dawn of the 21st century.

At the begging of the 2001–2002 Major League Baseball (MLB) season, the Oakland Athletics GM, Billy Beane had 44 million USD available for team salaries, which ranked Athletics in the 28th position of the MLB team salaries between 30 teams, as seen in Figure 2. Correspondingly, New York Yankees which ranked 1st in team salaries had 125 million USD to their disposal. Beane and his assistant, Paul DePodesta decided to follow a sabermetric approach. They had noticed that a team with a high on–base percentage had a higher probability to score runs, thus to win a game. So, the general idea was that they would draft and trade players only with high on–base percentage, indifferently from their rest statistics and player skills. The result was that Athletics finished the season with the second best record (.636), only behind Yankees and that they became the first team in the history of the American League to record 20 consecutive wins [5] [6].

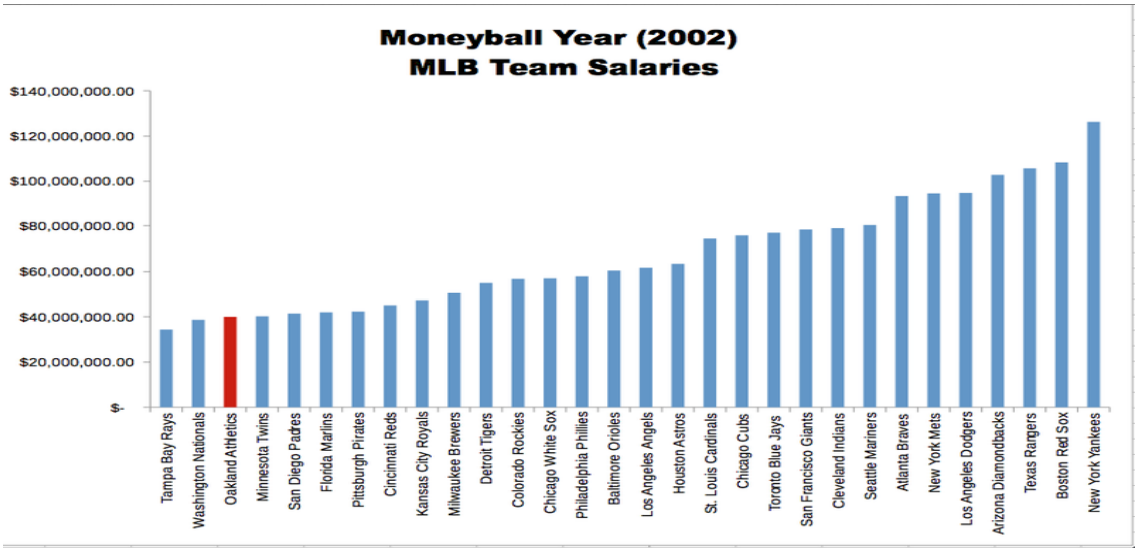


Figure 2: Distribution of team salaries in 2002. Team salaries ranged from about 35 million USD to about 120 million USD. The Oakland Athletics had the third-lowest team payroll in the league **[7]**.

Despite being eliminated in the first round of the postseason, Beane's reputation had grown significantly and his ideas that had even been mocked, looked more appealing after the end of the season. Boston Red Sox, inspired by his approach, offered 12.5 million USD to make him their new GM. In a shocking decision, Beane declined the offer to stay at Oakland, but the Red Sox did not change their aspiration, which was to hire a GM who would focus on sabermetrics; 29–year–old Theo Epstein took over and became the youngest GM in the history of MLB. Less

than two years later, the Red Sox won their first World Series Championship in 86 years. They repeated their success in 2007. Epstein, then, moved to Chicago Cubs in 2011, rebuilt the whole organization from scratch and in 2016 the team won their first World Series Championship since 1908! A remarkable success [8].

During those 15 years (2002 – 2016) the MLB teams that hired full–time sabermetric analysts and started to use SA were dramatically increased and enjoyed great success; the New York Mets managed to reach to the 2015 World Series Final, despite having dropped their payroll by more than 200 million USD in two years [8].

On the contrary, teams that were adjusted too late to the new status, faced failure, both in an athletic and in financial terms; the Philadelphia Phillies, the last team to hire a full–time analyst, won the World Series in 2008, but faced a tailspin in the years to follow, that ended with an embarrassing winning percentage of .389 in 2015, the worst in the whole MLB [8].

By the end of 2016, it was obvious to everyone that sabermetric analysis had become as important as traditional scouting and traditional tactics and maybe even more.

2.1.2 Tennis

Tennis is also an old game, as it has an ancient origin from the 12th century. Of course, the game evolved and was formed to what we know today during the 19th century. Tennis analysis emerged really fast. Manuals and articles –mainly unsigned– that provided statistics on played balls, aces, rallies, faults, double faults, total points won etc. existed even from 1885. Also, players ranking system was first established in 1973, but the prime idea was presented by William Strunk Jr. in 1890. According to him, players should be given some points in relation to the tournament games they participated. Thus, every player would end up with his own rank.

2.1.3 American Football

The first American Football game took place in November of 1869, between two college teams. In the years to follow, associations relevant to the sport were launched and the rules of the game brought a dispute between the athletes, before finally the sport was divided in American and Canadian Football.

In January 1900, a statistical analysis was performed and published about two college games. The analysis measured the key events of the games and also recorded the estimated distances for rushing and punts.

Almost forty years before Moneyball, an American Football team, founded in 1960 and participating in National Football League (NFL), the Dallas Cowboys, tried to exploit the power of statistics. An intelligent, aggressive salesman, Tex Schramm, the GM of Cowboys at the time, had previously been in charge of a large television network during the broadcast of Winter Olympics. He was inspired by the usage of computers in figuring out scores and standings for the Olympic Games and talked to IBM experts about his idea; football players evaluation using computers. When he took over in Dallas, IBM sent Salam Qureishi, an Indian computer programmer and statistician, a modest young man with absolute no knowledge of the game to talk with Schramm. They were two totally different characters, but their coexistence at the organization turned out to be very successful.

Schramm wanted to find a way to efficiently draft the right college players for Dallas and expected from Qureishi to optimize the player selection task. The initial problem that forced Schramm to consider computers was that the team had too much data for too many players. Qureishi's problem was that computing power in 1962 was very limited. Qureishi had to throw away most of candidate players' attributes and eventually cut down the database into five dimensions (character, quickness & body control, competitiveness, mental alertness and strength & explosiveness). The Cowboys, based on Qureishi's analysis drafted a lot of good players, even if their choices initially seemed bizarre to other teams' GMs (Bobby Hayes who later entered the sport's hall of fame is just one such case). The result was that Dallas Cowboys made it to five Super Bowl appearances and two titles during the 70's.

More than that, except from the sport success, economic success was also achieved. Dallas Cowboys are the most valuable NFL team. They worth over 2 billion USD according to Forbes, which makes them the fifth most valuable sport organization in the world, every sport included. Their annual revenues are approximately 270 million USD and they own the world's largest domed stadium with a capacity of 100,000 seats. Unsurprisingly, Cowboys are considered the most iconic team of NFL, something like Real Madrid in European Football.

The methodology for analyzing sports data is not necessarily limited into sports. In 1967, Cowboys owner, Texas oil multimillionaire Clint Murchison Jr., founded a new company, the "Optimum Systems Inc." which would hold rights to the software developed by Qureishi. The firm was equally owned by the Cowboys, some other NFL teams and Qureishi himself. Optimum exceeded the narrow frame of football and collaborated with corporations, municipal governments and other entities with data selection problems [9].

This holds also in the case of FiveThirtyEight.com. This website started in March 2008 by a baseball analytics background statistician, named Nate Silver. But FiveThirtyEight does not provide only baseball–based data. It covers every famous sport competition and not only that; the website has separate sections about politics, science & health, economics and more. In 2008,

Silver's analytics system correctly predicted how 49 out of the 50 states would vote in the USA Presidential Election [6].

2.1.4 Basketball

Basketball's main difference from other team sports like Baseball, Football or Soccer is that Basketball does not have any predecessors. Canadian physical education professor at Young Men's Christian Association Training School (YMCA), at Springfield, Massachusetts, James Naismith was trying to keep his students occupied. In December of 1891, the days were rainy and the sports activities should be carried out indoors. Naismith created, wrote down and explained to his students the rules of a new game he had thought. His students loved it and basketball was officially invented.

Basketball is a game depending heavily on statistics. Unlike football, predicting the winner team is not that difficult in basketball, since the better team (i.e. the one that is statistically superior) wins the game most of the times. National Basketball Association (NBA) championship is now full of innovative and substantial stats that describe the way a team attacks or defends with high accuracy. Moreover, individual player statistics indicate not only how good a player is, but also how important he is for his team.

Before reaching the present state, however, basketball data had been untapped for too many years. University coach Lloyd Messersmith was one of the first pioneers of SA in basketball. He published many scientific papers while being at DePauw University from 1930 to 1945. His doctoral dissertation was titled "*The Development of a Measurement Technique for Determining the Distances Traversed by Players in Basketball*". Messersmith was very interested in notational analysis for sports (i.e. the study of players' movement patterns and techniques and team tactics and strategy). He tried to communicate his work to wide audience, but without much success. His notational analysis on basketball unfortunately did not receive much attention [10].

Elsewhere, however, during those years, another revolutionary man, Howard Hobson, also a University coach was studying and statistically analyzing shooting percentages mainly, but also other aspects of the game, for the last 13 years. In 1949, he published his findings in a book named "*Scientific Basketball*". Hobson's main concern was that the game had become too physical and that players were all gathered in a close range from the basket. The solution that he proposed was the adoption of a different type of shot; the three points shot. He also suggested that the free throws lane should be widened and that a shot clock should be used. Hobson was successful as a coach too, as he won a National Collegiate Athletic Association (NCAA) champion with University of Oregon and five conference titles with University of Yale [11].

Legendary Dean Smith was the coach of North Carolina University for 36 consecutive years, recorded 879 wins, appeared in 11 NCAA final fours, winning two national champions. He is also known for recruiting Charlie Scott, the university's first African American scholarship basketball player, for having a high graduation rate, with 96.6% of his athletes receiving their degrees and for having the privilege to coach Michael Jordan at his youth.

Before working as a first coach at North Carolina, Smith was the assistant coach at Air Force Falcons. While being there, in 1955, he developed a possession evaluation system. Smith was convinced that the absolute number of points scored or allowed could be a misleading statistic as in that way, the number of possessions for each team were not taken into consideration. In order to evaluate team effectiveness more accurately, Smith rather measured the average number of points scored for each possession. In his own words: "*Our goals are to exceed .85 points per possession on offense and keep our opponents below .75 points per possession through our defensive efforts*" [11].

At the early 60's, Paul Keller, a math teacher and high school basketball coach in Delaware, Ohio used a system similar to Smith's. He named it "Offensive Efficiency Rating System" but incorporated more statistics to that, like defense efficiency rating, turnovers, field goal percentage, free throw percentage, and rebound percentage. His rating system was widely used by colleges [11].

In 2004, Dean Oliver published a book, titled "*Basketball on Paper*". It soon became very popular and is now regarded as the holy bible of basketball analytics. Oliver specified four key categories that assess the real ability of a team: shooting efficiency, rebounding percentage, turnovers per possession and getting to the free throw lane [11].

One year later, two Israeli scientists, Gal Oz and Miky Tamir introduced SportVU to the world. SportVU is a system of cameras which collects data with a rate of 25 frames per second. As seen in Figure 3, it provides real time statistics, like the position of a player and the ball which can be leveraged for in–depth analysis by teams. SportVU was acquired by STATS and started its official collaboration with NBA in 2010. Not only the association provides the fans with advanced statistics, but also every NBA team has installed SportVU and has access to its contents. SportVU has also been expanding in football during the last years [6].

Figure 3: SportVU is using cameras to locate and record position and movement of players and ball **[12] [13]**.

But numbers are just numbers and have nothing to offer if no one perceives their essence and tries to convents them into meaningful game strategies. Daryl Morey, the GM of Houston Rockets and a former statistical consultant, conducted an analysis that proved that two–point shot attempts (and mainly the long two–point shots) were outweighed by the 50 percent gain in points from a successful three–point attempt. Morey's conclusion was that the Rockets should have as an offensive goal what is depicted in Figure 4; to either take a short–range shot from the key or attempt a three–point shot. The first category of shots offers only two points, but has a high probability of success. The second category of shots has a lower success probability, but it's a risk worth taking, since a good shot would reward the team with an extra point. The only inefficient shots are the long two–point shots; low probability of success and only two points for every successful attempt [14].

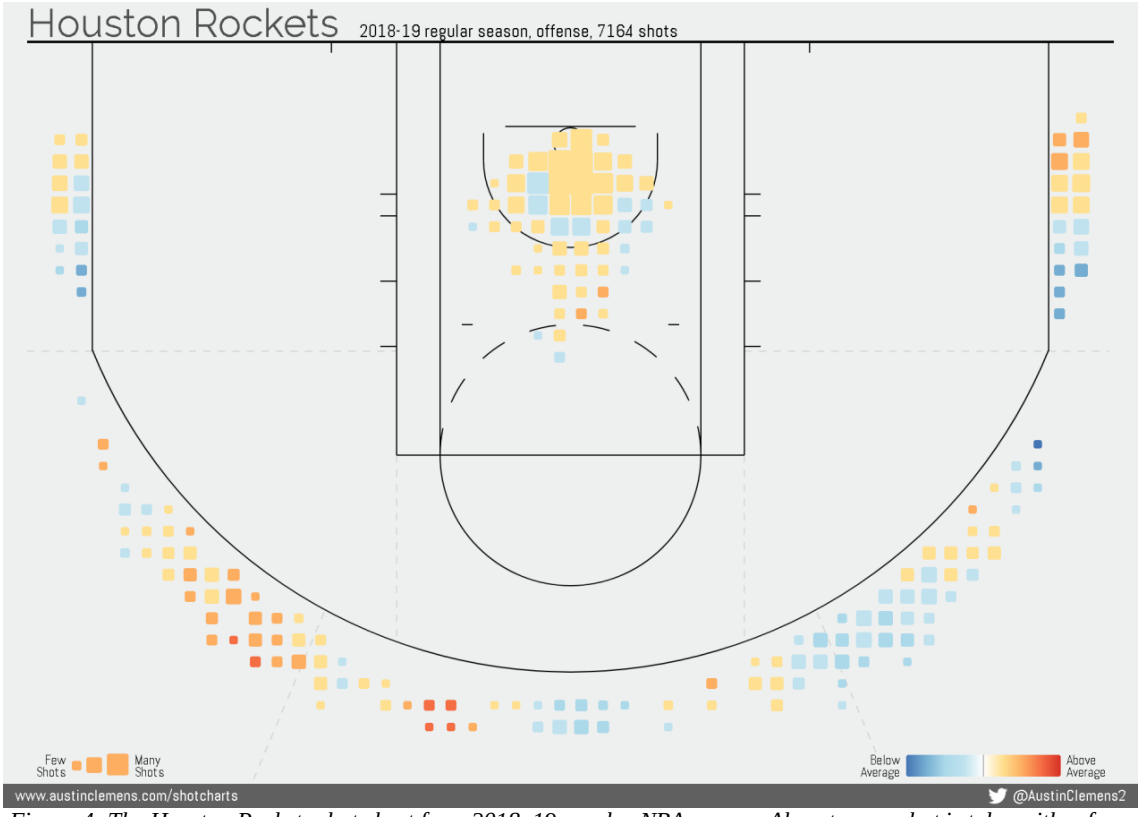


Figure 4: The Houston Rockets shot chart from 2018–19 regular NBA season. Almost every shot is taken either from inside the key or from the 3–point territory **[15]**.

Houston Rockets changed dramatically the way they were shooting, but consequently a lot of NBA teams decided to follow their example. As seen in Figure 5, the NBA champion of 2018–2019 was looking like a three point contest at times. In the previous season, an average NBA team took almost 29 three–point shots per game, while the equivalent number for the 2018–2019 season was 32 shots, almost a 10% raise, in just one year. To emphasize on that, last year San Antonio Spurs shot fewer three–point shots than any other team with 24.8 per game. In season 2006–2007, Golden State Warriors also recorded 24.8 three point attempts per game, only that by this performance they ranked 1st in the whole NBA in three–point shots. Milwaukee

Bucks, the NBA team with the best winning percentage for the 2019 campaign (.732) achieved that because they boosted their three–point shots per game from 24.8 to 38.2 [16].

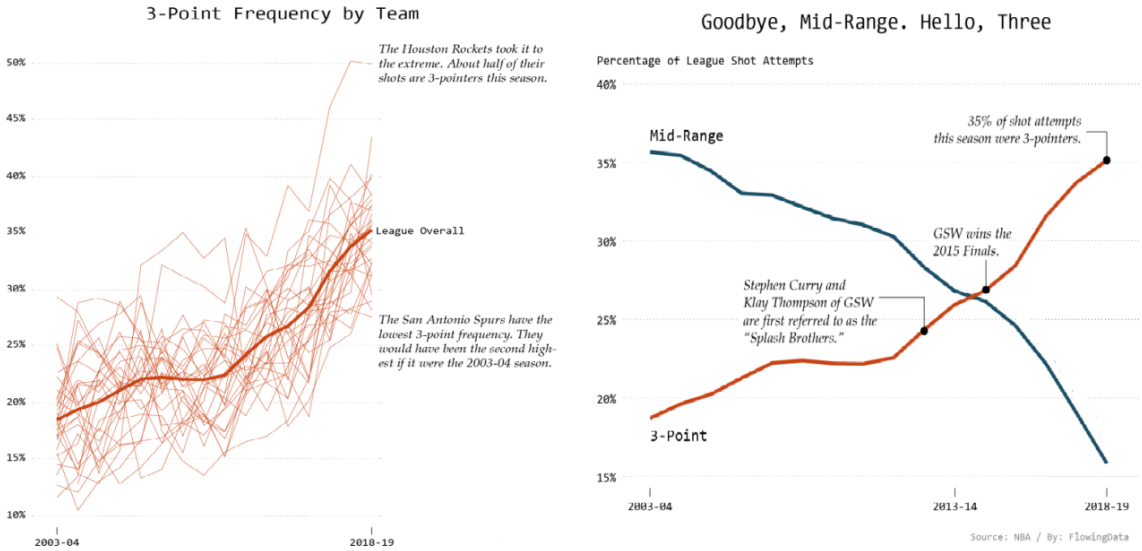


Figure 5: On the left, 3–point frequency by team and team overall for the last 15 years. Houston Rockets on top of the league. On the right, mid–range vs 3–point shot frequency on the same time period **[17]**.

Houston Rockets still have not won the champion, but are among the contenders every year. Maybe the only reason why they have not succeeded is because of the Golden State Warriors. Warriors, another team specialized in three–point shots, with three of the best shooters in the league, namely Steph Curry, Klay Thompson and Kevin Durand have eliminated the Rockets in the postseason three times during the last four years.

In 2020, with scouting and SA at their best, it is most certain that a good idea will be emulated by almost everyone. So, the need for new, innovative ideas is more essential than ever.

2.1.5 Motorsports (Formula 1)

In most of the sports, data generated from players is a vital part of the sport. In Formula 1 (F1), data are the sport itself. F1 cars have more than 200 sensors attached on them. Williams Grand Prix Engineering team estimated in 2011 that for each race, a total amount of 30 terabytes (TB) of streaming data are being generated. There are 10 teams in the championship, so we are talking about almost 300 TB of data just for a single weekend. That is more than 6000 TB or 6 petabytes (PB) of data for all the races of a season! And the data that are being generated in the races is just a small portion of the data that are being generated in winter tests, when F1 cars cover thousands of kilometers testing mechanical parts, tires, strategies, aerodynamics, engine performance etc. in order to achieve the perfect set up for the races. Those are clearly big data as they contain volume, variety, velocity, veracity and value [18].

First thing is how to manage all this data. Every F1 team has its own hardware equipment and use customized software solutions. They have on–board storage systems and cloud computing is important too; powerful internet connections provided from multinational conglomerate communication companies upload data really fast so the employees on team factory have immediate access; it takes no more than an hour in order to transfer the 30 TB of data that are produced during the race. There are race analysts both on track and on team headquarters. Their number varies from 50 up to 200 employees just for data analysis. They exploit state–of–the–art visualizations, search for patterns and anomalies from the streaming data and offer their advices to strategists and team principals [18] [19] [20].

Before big data analytics and cloud computing era, the critical point of racing was driver's capability. But F1 is the capstone of motor technology, so it was certain that data would finally conquer the sport. During the 70's the telemetry systems had already been sophisticated and small enough to fit to a racing car. By the 80's, the on–board electronic systems were multiplied and telemetry could send data to the garage on real time. Initially there was a data storage problem, so the telemetry could not be active during the whole race, but temporarily. Today, real time streams of big data have boosted the sport [19]. Data from the car sensors are fed into simulators and different scenarios are emerged to give real time insights to strategists, as seen in Figure 6.

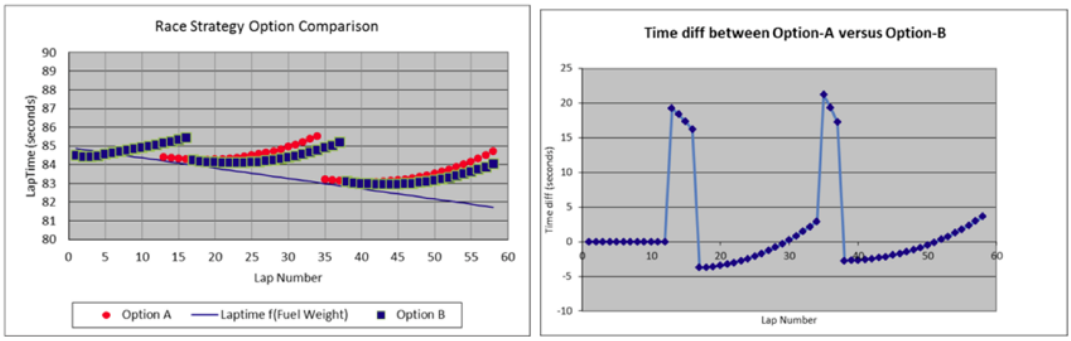


Figure 6: Race strategy example in Formula 1, analyzing multiple pit stop choices **[21]**.

Television broadcast provides the fans with real–time data of high interest; g–force measurements, heat maps that show the temperature of each part of the car and –lately– probability estimations about a possible overtake. But, generally, pre and post racing analysis is not provided by teams and is not approachable in any way. Even among team personnel there are certain grades of accessibility privileges to data. F1 is not unaware of scandals of technology spying. The one thing spectators can have a glimpse at is the strategy that a team embraces, taking in consideration every piece of information concerning the race (i.e. car data, weather forecast, track condition, tires condition, competition's acts, data from previous races etc.) by just paying attention to race incidents.

A very good example is the 2005 Monaco Grand Prix. Finnish McLaren Mercedes driver, Kimi Raikkonen was the fastest driver on Saturday's qualifying and started the race from the

first place. He was followed by Spanish title–contender Fernando Alonso with Renault in second place. In lap 23, Minardi's driver Christijan Albers crashed into a wall and the Safety Car was deployed as the road was blocked by the destroyed Minardi. It is a ground truth in F1 that when the Safety Car is on track, you have to pit. In that way a driver loses relatively less time on pits because the drivers on the track have to slow down, following the pace of the Safety Car. Alonso pitted immediately as it was expected, but surprisingly Raikkonen stayed on the track. This decision was made by McLaren chief strategist Neil Martin, after assessing all the data available. He and his partners had to make a difficult decision in less than a minute. It would not be possible if SA had not supplied them with every possible scenario. It was an unprecedented move and it was also risky; if Raikkonen had pitted, he would certainly be in front of Alonso. But without pitting under Safety Car, maybe Alonso would overtake him when Raikkonen would finally make his pit stop and McLaren would lose a certain victory. The Finnish seemed to be in a bad spot, but he fought back with some very fast laps, which gave him a 35–seconds lead from Alonso. Raikkonen then pitted and when he exited the pits he was still in front of the Spaniard by 13 seconds! Also, Raikkonen had a fresh set of tires, while Alonso was struggling on the track with his rear tires wearing rapidly. Raikkonen won the race, Alonso lost two places finishing fourth and their time difference rise up to 36 seconds. It was a brilliant data–driven decision making from McLaren that gave Kimi Raikkonen, nicknamed "The Iceman", his first and only victory in the famous Grand Prix [22].

2.1.6 Football (soccer)

Football or Association Football or soccer in USA (a jocular abbreviation of the word association, with the suffix -er appended to it) is maybe the most famous game in the world. Ancient games like cuju in China, episkyros in Greece, harpastum in Roman Empire, kemari in Japan and chuk–guk in Korea are considered the ancestors of football and rugby. During the mid–19th century, various types of mixed football and rugby games existed that were played in English schools and Universities, with different sets of rules. The Football Association (FA) was formed in October, 1863 in London with the unification of the rules to be its highest priority. Teams that preferred rugby rules seceded from FA and formed their own union in 1871. The oldest football competition is the FA Cup that was first held in 1872 and London club Wanderers were the first winners. The first international game took place in the same year, between Scotland and England in Glasgow.

Despite that football is considered a European or even an English game, the first preview of SA in football was given by an American newspaper reporter, David F. Barrett. In 1910, Barret published a chart with players' game actions, apart from goals, as he had noticed that everyone

was paying attention at goals and goal scorers and not to other equally interesting points of the game.

Today, a statistical review of a game is taken for granted, especially for simple statistics like shots taken by each team. This was not always the case. It is believed that the first time that opponent's attempts for goal were recorded was in 1937, for a friendly game between a mixed Central European team and a mixed Western European team.

Maybe the first British notational analyst was Charles Reep. He was a huge fan of the systems that manager Herbert Chapman was using. Chapman had introduced the W–M formation and he liked his teams bringing the ball forward as quickly as possible. Reep was disappointed when after the World War II found out that the ideas of –the now deceased– Chapman were abandoned. He considered the game to be fairly slow and the teams to lack in attacking skills. In 1950, Reep claimed that with minor changes, Swindon Town would improve the goals scored per game from 2 to 3. Brentford manager, Jackie Gibbons was impressed and offered him a job as his advisor in February of 1951. Under Reep's and Gibbons' guidance, Brentford not only managed to achieve its goal of avoiding relegation but also collected 20 out of the possible 28 points remaining. Reep was also working for a newspaper during the 50's and in 1968 he published a statistical analysis of patterns of play in football, along with Bernard Benjamin. The dataset that was used for this paper included 578 games that took place from 1953 up to 1967. In his articles Reep criticized possession football and the findings of his paper confirmed that most goals are scored after the ball is passed 3 or less times between players. So, according to Reep, a more direct approach should be adopted, where the goalkeepers or the defending players should make long aerial balls, searching for the attacking players, skipping the midfield as much as possible. This approach was called "*long ball*" and was the trademark of England national football team and of most English clubs as well for many years. It also influenced other national teams, such as Norway [23].

Despite being innovative and influential, Reep is also controversial. He influenced many managers of his time, but his work has been criticized lately. Sports journalist and writer Jonathan Wilson highlighted that the 91.5% of the passing sequences that Reep studied had 3 passes or less, but only the 80% of goals came from sequences with equal or less than 3 passes. Therefore, the conclusion should be that goals are scored more often when there are sequences with 4 passes or more, rather than the opposite. According to Wilson, either Reep was misled by his own findings or he himself successfully tried to mislead others [24].

In August, 2001, PA Sport, a multimedia sport news agency created "Football Live", a computer program able to collect and distribute statistics from English and Scottish football matches in real time. Its usage led to a massive boost of displayed statistics. Television, internet and newspapers were getting their stats via Football Live. Opta acquired Football Live in 2014.

Nowadays, we are living the era of advanced metrics in football. Some statistics that looked advanced ten years ago, now they are obsolete. For example, how significant is to know how many shots were taken by a team? A shot that was made under bad circumstances (i.e. a marked player shots outside the box) should have the same value with a shot that was made from a perfect position inside the area from an unmarked striker? A complete pass from one defender to the goalkeeper should have the same value with a key pass that leads to a shot or with an assist that leads to a goal? Working in that direction, many remarkable advanced statistics have emerged during the last decade, such as Expected Goals, Expected Assists, Expected Points, Packing, Defensive Coverage, Sequences and more.

Maybe the most revolutionary metric is the Expected Goals (xG). It is a statistical measure of the quality of chances created and conceded (Expected Goals Against or xGA). xG probabilistically assign a score from 0 to 1 to each chance based on several variables such as assist type, shot angle and distance from goal, whether the player that took the shot was marked or not, etc. Shot quality evaluation is usually being achieved by training neural network prediction algorithms over large datasets of shots. xG are calculated both for every individual player, but also cumulatively for the whole team, as seen in Figure 7. The model is applicable in other sports too, like ice hockey. In football there is a difficulty to accurately predict the final outcome of a game, because of the small number of goals achieved. xG eliminate some of the randomness of the actual goals scored and give us a better picture of team performance. The model of xG has not avoided criticism, but there have been certain cases that the method was implemented with great success, such as the champion that FC Midtjylland won in 2015, signing players who matched this model. Midtjylland was just founded in 1999 and this was their first title ever. It was followed by another champion three years later and a Danish cup in 2019 [25].

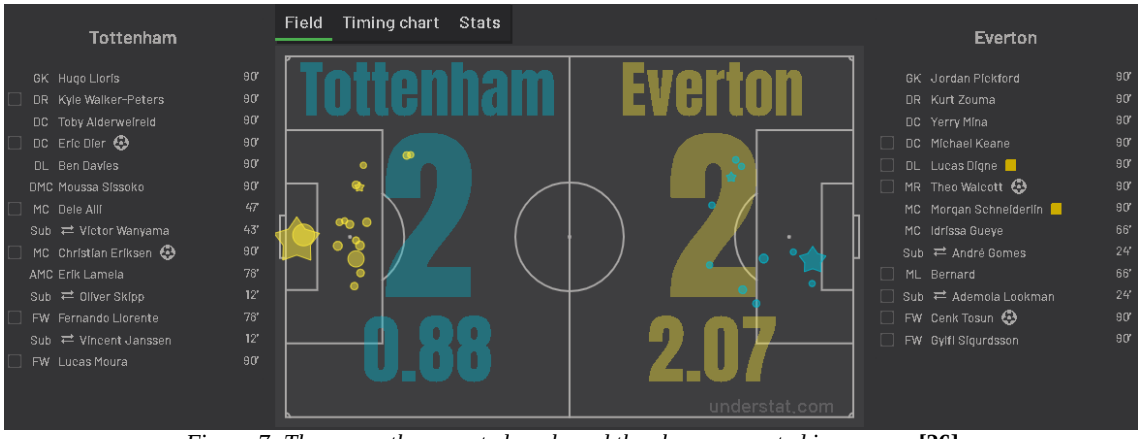


Figure 7: The score, the expected goals and the chances created in a game. **[26]**

It is worth noting FiveThirtyEight.com again. The web site combines xG and team ratings in order to predict football matches for 36 football leagues. Then, they run Monte Carlo simulations, with an objective to forecast the final league table, which team is going to win the

league, which teams are going to achieve European qualification and which teams are going to be relegated. They make similar predictions for a variety of sports too [27].

Today, there are several web pages referring to football statistics, like whoscored.com, understats.com, eplindex.com or statsbomb.com. Some of them provide live coverage of football games. As seen in Figure 8, there are actually various statistics that every web page focuses on different categories, so one can choose from many sources in order to find the suitable statistic. Some of the pages even contain remarkable articles about football analysis as well.

Figure 8: Typical live coverage of a game, with several statistics **[28]**.

The need for analytics arose with the nascence of sports, but back then very few people had realized the power of analysis. Revolutionary pioneers had small or bigger contribution to the elaboration of SA; they did the tough job, until computers and machines came to their rescue. Data are now collected easily and technology has broadened the scope of SA substantially. Statistics and algorithms took over and scientists are getting more involved into sports and guide SA to a world of unlimited opportunities.

2.2 Literature Review

As discussed in the Historical Background section, some forms of SA existed in the past, but they were mainly nonrecurring actions of individual pioneers. Today, in scientific and research level, there already are many innovative SA applications. Leagues, professional teams, athletes and betting companies are getting more and more interested into those methods, as they seek ways to achieve their goals. It is true that not every sport organization has implemented SA or has hired data scientists so far, but it is expected that in the years to follow, a notable amount of money will be invested on analysis. During the last 20 years, mainly, modern techniques, sophisticated algorithms and dynamic tools for sports analysis are developed. As can be seen in Figure 9, articles, reports, papers and dissertations related to SA are being constantly published. This revolution has created a huge ocean of tools that can be used in sports for analysis, prediction, evaluation or improvement and organizations just have to choose what matches their objectives and their ambitions.

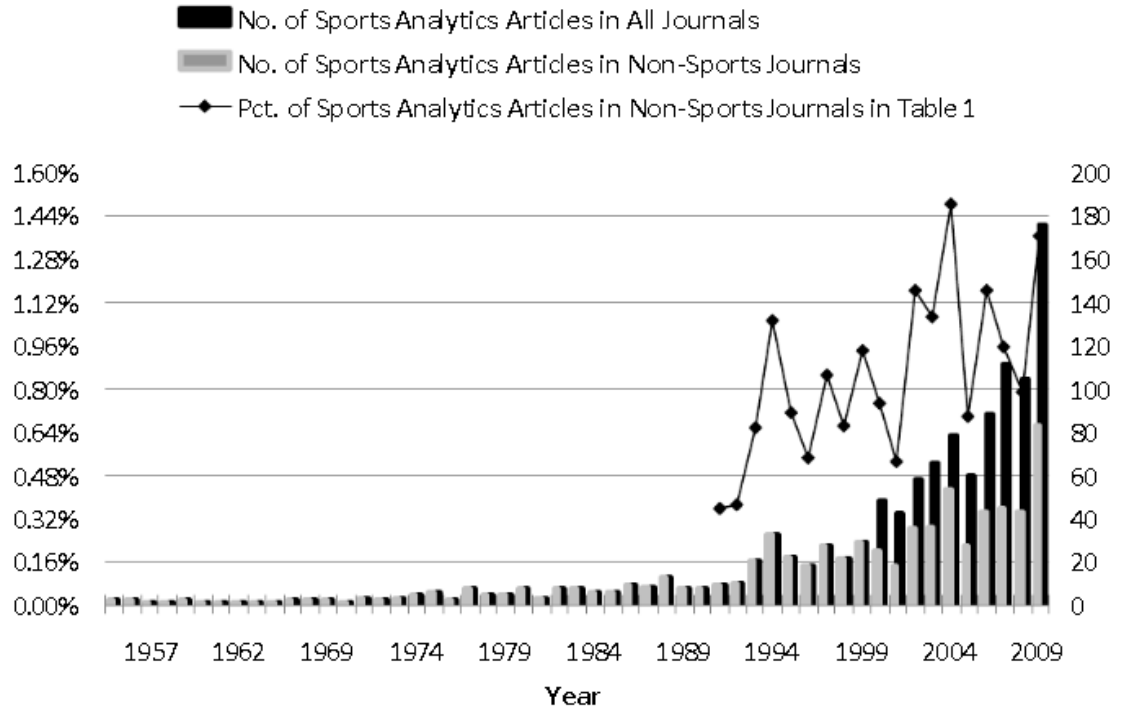


Figure 9: The number and proportion of journal articles on sports analytics has risen substantially since 1990 **[29]**.

In this section of the Dissertation, the most significant researches and previous works are briefly presented. The proposed methods are highlighted and pros and cons are discussed. The criteria used to evaluate the papers that were finally used for the Literature Review were the following:

I. Relevance with the experiments conducted for the Dissertation.

II. Relevance with football analytics.

III. Innovative ideas concerning other sports.

IV. Non–obsolescence of researches.

2.2.1 Definitions and Data Composition

To begin with, it would be intriguing to see how some researchers and sport writers perceive the concept of SA:

Morgulev et al. consider SA as the examination and modeling of sporting performance using scientific techniques, while Passfield and al. describe SA as an emerging discipline, highlighting that it is a mix of a wide domain of specialties from human physiology and kinetics to sport science, big data, data science, data mining, mathematics and statistical analysis. They both agree that thorough analysis of large datasets can expand our knowledge in the sport science, in players and their behavior and has the power to reform the sport [30] [31].

Nate Silver, in his book, while analyzing several types of predictions on a probabilistic level, he notes that it is important to be able to transform the quantitative information to qualitative and being rigorous when analyzing the volume of data available. He is also pointing out the fact that technology will add value to SA and will be a priceless asset in measuring things that were not possible to be measured, but does not underestimate the merits of traditional scouting. He, finally, states his belief that innovative ideas will dampen the inevitable deviation of the SA's predictions from the real facts due to the unpredictable nature of sports [32].

But what sports data consist of? Data can be quantitative or qualitative. They can be structured, semi–structured or unstructured. It is collected from various sources; pitch actions of footballers, recorded statistics, spatiotemporal data, biographical data, medical history of players, scouting reports, real–time streaming data, data from cameras or wearable devices and more. When collected, must be standardized, centralized, integrated and analyzed and assumptions must be presented in a simple but stimulating manner [33].

2.2.2 Game Result Predictive Models

One of the most interesting topics in SA is the prediction of the outcome of a game. There is a plethora of scientific papers referring to this subject and researchers approach the prediction problem from a different angle. A simple prediction strategy is to predict the goals that the two teams are going to score in a game. This approach was mainly implemented during the late 90's and the early 00's.

The first time that a sufficient model for predicting the result of a game was created was in 1997 by Dixon and Coles. The model was able to extract probabilities for the goals scored in a game, following a Poisson distribution. It is considered a classic and is used as benchmark against every new model created. The calculated probability of opponent teams scoring a certain number of goals is converted into score probabilities and then into match outcome probabilities. Finally, each team is assigned with an attacking and a defensive rating which depends on the past results. A weighted function is used to down–weight past results according to the time space between the game and the time of prediction [34].

Rue and Salvesen first and then Crowder at al. used Monte Carlo simulation to predict results and they implemented a dynamic structure in their models, as they were updating the parameters of their algorithms as more game results were included into their dataset [35] [36].

Of course, some sophisticated methods had been implemented even before 2000. Pollard and Reep introduced "yield" in 1997. It is a metric, defined as the probability of a goal being scored minus the probability of a goal being conceded. They proved that open play has higher yield than set plays (i.e. free kicks, corners, throws–in) [37].

In the years to follow, researchers, instead of trying to predict goals or points scored, mainly focused on directly predicting wins, draws and losses (and in a variant, wins and no–wins) or wins and losses for sports that draw is not an option (e.g. basketball).

Goddard, in 2005, compared the two methods (i.e. modeling the goals scored vs modeling win–draw–lose match result) and concluded that a hybrid model is achieving the best prediction performance. He also was one of the first to use other variables than previous match results. He leveraged features like the importance of individual games, geographical distance between the two opponents and more. For the win–draw–lose method Goddard used an ordered probit regression model and apart from the aforementioned attributes, he exploited a database of English game results from the last 25 years. Goddard also included in his work a comparison of his predictions with the betting odds of the games and came to the conclusion that achieving a positive betting return over time is not impossible [38].

Lago-Penas and al. performed an one–way ANOVA and discrimination analysis and concluded that the most discriminating factors that separate the winning and non–winning sides are the shots on goal, the crosses, the match location, the ball possession and the ability of the opponent team (based on ranking). Similarly, Harrop and Nevill, analyzing data using a regression model, supported that the best predictor is the pass accuracy, followed by the number of shots, the number of passes and dribbles (the fewer the better) and the venue of the match [39] [40].

In 2016, Mao et al. collected and analyzed 480 matches of the Chinese Super League regarding 21 performance–related variables and the match outcome (win, draw, loss). They used cumulative logistic regression in a generalized linear model, taking the value of each

performance–related variable as an independent variable to predict the logarithm of the odds of winning. They classified the teams as upper and lower ranked teams and made a cluster analysis for the goal difference of each game. In that way, they separated their dataset into close and unbalanced matches. They claimed that the features that provide the most positive effects are shots on goal, shot accuracy, tackles and aerials won [41].

In the same year, using a linear model too, but implementing a different approach, Tavakol et al. exploited the past results between the two teams, but also historical player data for their prediction on the EURO 2016 competition. They had a large number of attributes, so they clarified that they had to use dimensionality reduction methods in order to extract the best attributes and use them for prediction [42].

Tax and Joustra employed a set of factors from public data and also performed dimensionality reduction techniques (principal component analysis – PCA) along with machine learning (ML) algorithms (Naive Bayes and Multilayer Perceptron – MLP) in order to predict the Dutch football championship. They achieved an accuracy of almost 55% in their predictions and claimed that building a profitable betting strategy is possible. Nevertheless, they proved that a hybrid model, combining public data and betting odds could improve accuracy [43].

Gomez et al. used a spatial approach, as they divided the pitch into five primary zones with subzones inside them. They analyzed 1900 matches of the Spanish League and performed another data reduction method; factor analysis with principal components. After that they concluded that only four factors were worthwhile to focus on:

i) Turnovers in zone 5.2 (i.e. offensive small area) and Crosses in zone 4 (i.e. between the midfield circle and offensive semi–circle area),
ii) Goals and shots in zone 5.1 (i.e. offensive goal area), Turnovers in zone 4, and ball recovery in zone 2 (i.e. between the defensive semi-circle area and midfield circle),
iii) Goals and shots in zone 5.2 and
iv) Turnovers in zone 5.1.

All the factors above are highest for winning teams and teams tend to increase those features when playing at home [44].

Bayesian networks (BN) have also been vastly used for predicting the outcome of a football game. Back in 2006, Joseph et al. implemented a variety of ML algorithms (Decision Tree, Naive Bayesian Network, Statistics–based Bayesian Network and K–nearest neighbors) and compared them to BN built by a reliable domain expert. They claimed that while BN are easy to be constructed, their performance is impressive. The only drawbacks they detected is that expert knowledge is necessary and that players often change clubs or even retire from football, so the model quickly becomes out of date [45].

Artificial neural networks (ANN) have gained popularity during the last decade and have also been used for prediction in football. McCabe and Trevathan coped with four different

sports and four different leagues respectively, namely NFL for American football, AFL for Australian football, Super Rugby for rugby and EPL for football (soccer). Using data from 2002 up to 2008 and a Multilayer Perceptron (MLP), trained with Back Propagation (BP) and equipped with conjugative–gradient algorithms, they tried to predict match results. The ANN had a structure of 20–10–1, i.e. 20 input layer nodes, 10 hidden layer nodes and 1 output layer node. The same features were used for every sport. The best average prediction performance concerned to rugby with 67.5%, while soccer was proved to be hard to predict, with an average prediction performance of 54.6% (still, over 50%) [46]. Then, Hucaljuk and Rakipovic came to the conclusion that ANN performed better than any other ML technique they used (Naive Bayes, Bayesian Networks, LogitBoost boosting algorithm, K–nearest neighbors and Random Forest) [47].

In an alternative approach, Kampakis and Adamides used social media data (i.e. twitter posts) to predict the outcome of football matches. They built a model based on EPL games and compared it to predictive models based on historical data and football statistics. The authors ensured that the Twitter–based model was better than historical–based model. Then they used mixed models and as a result, the prediction accuracy rose. Thus, it was proved that Twitter contains useful information and can be a helpful source for predicting the outcome of a game. A limitation of their study was that the experiment was conducted on EPL games from a period of only 3 months [48].

2.2.3 Game Result Comparative Models

Apart from papers about predictive analysis, papers referring to comparative analysis are discussed in this thesis. The main components that are compared are wins and losses. The authors of the following papers attempted to discover the attributes that made the difference between winning and losing sides. It appears that a noteworthy attribute that most researchers point out is the efficiency. Efficiency is defined as the number of goals divided by the number of shots. Shots on goal, pass accuracy, quality of the opponent team, venue of the match and ball possession also seem to be significant variables [49].

Broich et al. analyzed 153 games from the German Bundesliga (i.e. the first division championship) and concluded that efficiency is by far the most significant parameter for the match outcome. Other important variables, according to them, are number of shots, number of passes and number of ball contacts. They also discovered that efficiency is mostly affected by the shooting condition (i.e. location and situation of shooting) and by the last few passes just before the shooting rather than the whole sequence of passes. Finally, an interesting conclusion

was that despite what is generally believed, the distance coverage of the team is not a statistically significant winning attribute [50].

Kapidzic et al. used data derived from two different sources; the Bosnian Premier League and the European Championship of 2008. They did not evaluate efficiency, as discussed above, but their results indicated that the losing teams of the league mainly lacked interaction between players and co-operation during the transition from defense to offense. The losing teams of the European Championship had problems at creating chances that ended in shots on goal. They indicated that the number of shots within 16 meters from the goalpost and the number of accurate passes are the attributes that separate winning from losing teams [51].

Three more researches on win/lose indicators are highlighting some very interesting variables that contribute to the teams' success: Ruiz-Ruiz et al. analyzed every match of the 2006 World Cup and stated that winning teams have more entries into the penalty area of the opponent team [52]. Jankovic et al., using as a dataset the matches of the 2010 World Cup, concluded that in big competitions, such as the World Cup, the success is heavily depending on the number of successful attacks and passes. It was also found that winning teams were better on the long passes specifically [53]. Finally, Armatas et al. discovered that when a team scores the first goal of a game, has a 71.43% chances of winning the game and only 12.38% chances of losing. They also noted that more goals are scored in the second half and that probability of scoring a goal increases as the game time proceeds. There is a probability of just 12% for a goal to be scored in the first 15 minutes of a game, while the corresponding probability for the last 15 minutes of the game is 23.3% [54].

There have also been comparative analyses where the researchers used teams' different positions of tournament rankings as a compassion measure. Armatas et al. after analyzing the top and the last teams of Greek League for ten years, concluded that better–ranked teams are more efficient in shooting and passing. In particular, he found that top teams need less attempts to score a goal than other teams and that they score more goals from inside the penalty area. He inferred that this was the result of their tactical and technical superiority, as they can enter the area more often than other teams. In terms of passing, what discriminates a top team is the higher number of key–passes (i.e. passes that lead to a shot, regardless if the shot is successful or not) [55].

Clemente, in his article about 2010 World Cup, confirmed the findings of the previous search, adding that successful teams in tournaments also score more goals in open play. Additionally, he included attacking zones or zones of goals conceded as indicators that designate the most successful teams [56].

Luhtanen et al. studied selected offensive and defensive variables of players and teams in the European Championships (EURO) of 1996, held in England and of 2000, held in Netherlands and Belgium. Their research gives prominence to the dynamic nature of football, as

it seems that in the EURO 1996 case, success to the tournament is being predicted mostly using the defensive variables, namely number of interceptions and success rate of all the defensive actions. Contrary, the percentage of the successful passes and the percentage of successful goal scoring trials, hence attacking variables were found to better predict success in EURO 2000 [57].

Bekris et al. used a different approach; they compared matches with at most one–goal difference (i.e. short range) to matches with at least three–goals difference (i.e. wide range). They found out that wide range winners outplayed their opponents in ball possession percentage, number of passes, "one vs one" duels won, number of shots, number of shots on target and shooting accuracy percentage. Contrariwise, those findings do not stand for short range matches, which are more sensitive to luck [58].

Another evaluation criterion involves possession, passes and passes sequences. Later, in the Literature Review, pass effectiveness is discussed thoroughly, but here, another comparative analysis is presented. The authors, Hughes and Franks analyzed the national teams that participated in 1990 World Cup and divided them into two groups: successful teams were the teams that reached at least to the quarterfinals. All other teams were considered unsuccessful. Then, they used an interesting approach. They measured "goals/shots per 1000 possessions" in order to extract knowledge for the ball possession. The successful teams were better in converting possession into shots on goal. Additionally, when the passing sequence was relatively long (i.e. 8 passes or more), the probabilities for a successful team to create a chance for goal were significantly higher. Conversely, it was proven that when a passing sequence becomes very large, the team needed to attempt more shots to score a goal [59].

2.2.4 Rating Systems

In this section of the Literature Review, the concept of rating is presented. Rating is a single number which is used to describe the strength of a team in comparison to other teams in a particular time. Most famous rating system is the ELO Ratings. It was first used in 1978 by Arpad Emmerich Elo, a Hungarian–origin Physics professor and chess grandmaster. Elo used it to rate chess players, but ELO Ratings was later used in other sports too.

The rating is changing based on game results. After each match, points are transferred from the losing team to the winning team. The points being transferred depend on the position that teams hold before the game is played. When a low–ranked team beats a high–ranked team, more points are being transferred than when the opposite occurs. This efficient tuning adjusts the rankings, as underrated teams win points more easily than overrated teams. The rating also takes into consideration the importance of the match, the goal difference of the teams in the final

outcome and the venue of the game. The formula that describes these parameters is the following:

$$R_{new} = R_{cur} + I \times G \times \left(R_{act} - R_{\exp} \right), \text{ where:}$$

- R_{new} is the ELO Rating after the match is played.
- R_{cur} is the ELO Rating before the match is played.
- I is a positive number that denotes the match importance.
- G is a positive number that denotes the effect of the goal difference (GD). Typically is defined as follows:

$$G = \begin{cases} 1, & if \quad GD \le 1 \\ 1.5, & if \quad GD = 2 \\ \dfrac{GD+11}{8}, & if \quad GD \ge 3 \end{cases}$$

- R_{act} is the actual outcome of the match, which is defined as followed:

$$R_{act} = \begin{cases} 1, & for \quad win \\ 0.5, & for \quad draw \\ 0, & for \quad loss \end{cases}$$

- $R_{\exp}$ is the expected outcome of the match, takes values in the range (0,1) and is calculated as follows:

$$R_{\exp} = \frac{1}{1+10^{\frac{R_{away} - R_{home}}{400}}}, \text{ where}$$

  - R_{away} is the rating of the away team.
  - R_{home} is the rating of the home team [60] [61].

There are other rating systems, such as Probabilistic Intelligence Ratings, Split–ELO Ratings, (a transformation of ELO Ratings), the rating system that was proposed from Knorr-Held [62] and the pi-rating system, which was proposed by Constantinou and Fenton in 2012.

The pi-rating is developed dynamically, as can be seen in Figure 10. It takes into consideration the inconsistency between predicted and actual goal difference. Home and away ratings are used for the calculation of the overall team rating, via the simple following formula:

$$R_{\tau} = \frac{R_{\tau H} + R_{\tau A}}{2}, \text{ where:}$$

- R_{τ} is the overall rating of team τ.

- $R_{\tau H}$ is the rating of team τ when team is playing at home.
- $R_{\tau A}$ is the rating of team τ when team is playing away.

Denoting the home team as α end the away team as β, home and away ratings are being updated in a cumulative way:

1. Update the home rating of the home team: $\hat{R}_{\alpha H} = R_{\alpha H} + \psi_H(e) \times \lambda$.
2. Update the away rating of the home team: $\hat{R}_{\alpha A} = R_{\alpha A} + \left(\hat{R}_{\alpha H} - R_{\alpha H}\right) \times \gamma$.
3. Update the home rating of the away team: $\hat{R}_{\beta A} = R_{\beta A} + \psi_A(e) \times \lambda$.
4. Update the away rating of the away team: $\hat{R}_{\beta H} = R_{\beta H} + \left(\hat{R}_{\beta A} - R_{\beta A}\right) \times \gamma$.

where:

- $R_{\alpha H}$ and $R_{\alpha A}$ are the current home and away rating of home team.
- $R_{\beta H}$ and $R_{\beta A}$ are the current home and away rating of away team.
- $\hat{R}_{\alpha H}$ and $\hat{R}_{\alpha A}$ are the revised home and away rating of home team.
- $\hat{R}_{\beta H}$ and $\hat{R}_{\beta A}$ are the revised home and away rating of away team.
- e is error between predicted and actual goal difference.
- $\psi(e)$ is a logarithmic function of e.
- λ and γ are learning rates.

The measuring error e is calculated as follows:

g_D is defined as the difference between the goals of the home and the goals of the away teams:

$$g_D = g_H - g_A.$$

$\hat{g}_D$ is the expected goal difference and is calculated based on the following formula:

$$\hat{g}_D = \hat{g}_{DH} - \hat{g}_{DA}.$$

$\hat{g}_{DH}$ is the expected goal difference for a team, playing on home against the average opponent, while $\hat{g}_{DA}$ is the expected goal difference for a team, playing away against the average opponent. The following formulas are used for the estimation of those values:

$$\hat{g}_{DH} = 10^{\frac{|R_{\tau H}|}{3}} - 1 \text{ and } \hat{g}_{DA} = 10^{\frac{|R_{\tau A}|}{3}} - 1.$$

Given all these, the error e is simply calculated as: $e = \left| g_D - \hat{g}_D \right|$.

Finally, we have the function $\psi(e) = 3 \times \log(1+e)$.

It is proven by the authors that pi-ratings outperform every ELO–based ranking system in terms of accuracy, but also that pi-ratings are capable of generating profits against published market odds, which is a substantial feature [63].

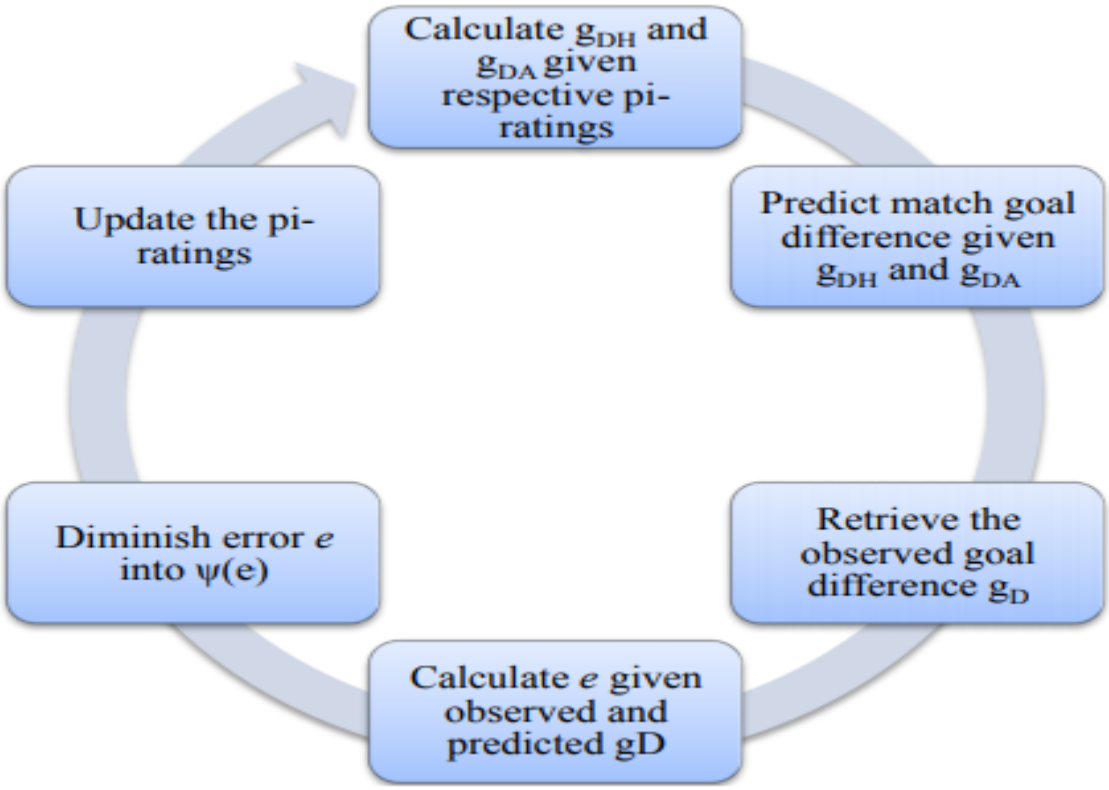


Figure 10: The process of updating pi-ratings **[63]**.

The aforementioned ranking systems were used in several papers. Hvattum and Arntzen used ELO Rating differences between teams as covariates in ordered logit regression models. The ELO–based models were comparable to other typical prediction methods, but still could not beat the bookmakers' odds [64]. Constantinou and Fenton used the pi-ratings that they had earlier invented for model validation, trying to make long–term prediction over team performance [65]. Long–term predictions in football still have not been extensively explored and that is an issue which is further analyzed in this thesis.

2.2.5 Expected Goals (xG) Models

The basic principles of xG have been analyzed in section 2.1.6 of the thesis. It would be useful to cite some papers that implemented the metric, as xG are used later for the experiments of this thesis.

Wright et al., in 2011, tried to predict which factors make a goal scoring opportunity better than others and concluded that those are position of attempt, goalkeeper's position and type of shot. They did not mention xG on their paper, but certainly worked in that direction, as the variables for goal scoring they investigated are the same with those that are used for the evaluation of the xG derived from a scoring opportunity [66].

In 2015, Lucey et al. proposed a method for the estimation of the likelihood of opportunities in football, using spatiotemporal data. They used an estimate called Expected Goal Value and found that more important variables for evaluating the likelihood of a goal are the game state, the proximity of a defender, the interaction of surrounding players, the speed of play and of course the shot location [67].

Next year, Eggels et al. used the xG trying to build a classification model to classify each scoring opportunity into a scoring probability. They leveraged geospatial data and implemented various classification techniques. They also indicated that xG could be further used for evaluating players and whole seasons, but they warned that probability estimates of goal scoring opportunities may suffer from high standard deviation [68].

2.2.6 Long–Term Prediction Models

Being able to predict the final outcome of a certain game is important, but maybe not important as the prediction of a team performance for the whole season. It is obvious that it is very hard to predict the long–term performance of a team and it is much harder to predict its performance by comparison with the performance of other teams. Not much work has been done upon this challenging task so far, but this thesis deals with it. Therefore, in this section, some very stimulating researches over that issue are discussed.

One of the most intriguing but also almost unexplored scopes is the prediction of the final table of a championship. Van Haaren and Davis emphasize on the difficulty to predict the exact position of a team in the final table, because it depends on the final position of every other team. Another obstacle in their method was the games that ended in a draw. Ranking systems they used for simulating match results have difficulties in predicting draws. The result is a high variance on the predicted number of points for each team. However, they indicated two substantial metrics needed for evaluating the quality of the predicted final tables: the percentage of correctly predicted relative positions and the mean squared error regarding positions [60].

Oberstone developed a multiple regression model, ending up with 6 independent variables which he assessed to be sufficient for predicting the final league table of EPL in terms of points, instead of accurate positions. The six variables were percentage of goals to shots (i.e. goals divided by shots), percentage of goals scored outside box, ratio of short/long passes, total

number of crosses, average goals conceded per game and number of yellow cards. He also used one–way ANOVA to investigate which pitch actions differentiate the four best teams from all the others in the league. He used data from 2007–2008 season of EPL and managed to achieve outstanding results [69].

Three more papers focused on the financial strand of a football club, rather than pitch performance. Hall et al. studied the correlation between payroll and team performance, deploying 25 year–long data from English football. They concluded that –unsurprisingly– there is a higher success probability for the teams that spend more in payroll and that top teams are more likely to spend big amount of money [70]. Kringstad and Olsen used data from Norwegian league and focused on the relationship between financial strength and sporting outcome. They presented some mixed results; evidence suggested that budgeted revenue was a success indicator, but only for bottom–half teams, while static and dynamic regression models they implemented supported the notion of budgeted revenues being a driver of sporting outcome. Their final conclusion was that money is a significant factor of success, but only to a certain extent and that focus on athletics is still vital [71]. Coates et al. used data from every team that participated in Major League Soccer (MLS) of USA during the years 2005 and 2013. They examined the relationship between salary level and dispersion with football success. They revealed that while the wage bill of team has a positive effect on success, salary inequality has a negative effect on success. In that way they proved that cohesion is essential in football [72].

Moreover, regarding success in professional football, Gerhards and Mutz considered market value of team players as by far the most important feature, even compared to athletic features. They analyzed data for 12 leagues over a period of 5 years, i.e. 60 different competitions. They confirmed that in 63.3% of those competitions, the championship was won by the team with the biggest market value. Furthermore, only once the league was won by a team that was outside top–three in terms of market value; it was Leicester City FC in 2016. They also inferred that fluctuation has a negative effect on team performance and that cultural diversity has a limited impact. In general, they claimed that is relatively easy to predict the outcome of a competition but raised their concerns over competitive imbalance (see section 2.2.11) [73].

Cintia et al. used pass–based performance indicators (more on that on section 2.2.7) and other efficient metrics, like the Pezzali score. The signification of this metric lies on the fact that reward the teams that are effective on both sides of the pitch, i.e. in offensive skills and in defensive duties. It is formulated as follows:

$$Pezzali\ score(team) = \frac{|goals(team)|}{|attempts(team)|} \times \frac{|attempts(opponent)|}{|goals(opponent)|}.$$

They simulated games from four major leagues and claimed that they achieved superb results, as they predicted game outcome with an accuracy of almost 60%. They also found that the final rankings in the simulated championships were very close to the true rankings; the difference between actual points and simulated points (i.e. the ranking error) was close to zero. Nevertheless, some teams had a considerable ranking error, which was explained by very high or very low Pezzali score (i.e. some teams had been very effective, while others had not). Finally, they marked the simplicity of their models and encouraged researchers to work with more complex models as they reckoned that there is room for improvement in accuracy [74].

Constantinou and Fenton, studying predictive accuracy in long–term team performance, were skeptical about pure ML as they regarded that it comes with a cost in accuracy. They proposed a method which they called smart–data. It is a knowledge engineering and data engineering approach by which they chose the proper data, rather than the available data. Instead of taking individual match results into consideration, they preferred to exploit external factors which may influence the strength of a team (i.e. managerial changes, European qualification, newly promoted teams etc). With those factors they built new, such as "true team strength", "expected performance" and more. Their goal was to predict the final table in terms of point won by each team. They achieved great results, managed to single out certain external factors that boost or worsen a team performance and focused on the quality of their data, not on the quantity [65].

2.2.7 Pass Effectiveness Models, Networks of Passes and Spatiotemporal Data

Football pass is one of the most important actions on field, along with shots and defensive plays. There has been a long lasting debate amongst football experts about the style of passes that a team should adopt. There are the fans of long passing, like Charles Reep who was discussed in section 2.1.6 and the whole movement consisting of coaches, sportsmen or simple fans influenced by him. Then there are the short–pass enthusiasts, like coach Pep Guardiola that created the most successful Barcelona FC side during the previous decade, embracing the "tiki–taka" approach (i.e. a big number of short passes).

Actually, Gyarmati et al. used passing sequences and pass networks to compare and differentiate the styles of different teams. They used Barcelona as a case study and concluded that their game does not consist of uncountable random passes but rather has a precise, finely constructed structure [75]. While, Bialkowski et al. leveraged a combination of match statistics, event data and player tracking data to identify the two opposing teams in a game. Their model had 70% accuracy [76].

Therefore, it soon emerged the need of evaluating the gain of a pass or a sequence of passes, whatever the passing approach of a team might be. Unsurprisingly, several papers coped with two fundamental issues:

I. The evaluation of pass effectiveness in football.

II. The usage of pass–based ratings in predictive models.

Cakmak et al. conducted an outstanding research, introducing a metric, named Pass Effectiveness. They based pass evaluation upon mathematic grounds. Pass effectiveness is being extracted from the combination of five other measurable pass metrics; gain of a pass, pass advantage, goal chance, decision time and pass effectiveness of the next pass. Their revolutionary work is briefly presented.

First, they defined the risk of a pass, based on the probability of an intervention from an opponent player. So, if P_1 pass the ball to his teammate P_2 and P_3 is an opponent player, it holds that $Risk\left(P_3, pass\left(P_1, P_2\right)\right) = intervention\ probability\left(P_3, pass\left(P_1, P_2\right)\right)$.

The intervention probability is learned from past game data and depends on the proximity of the opponent player to the passing line, the minimum speed that opponent player should run to intervene the pass and the time that the ball takes to travel from P_1 to P_2.

Thus, there is a risk area, i.e. a region in which any opponent team player might intervene in the passing route, which depends on pass distance. Regions around players are modeled as circles which have the players for their center and their radius is learned from training data.

The overall risk of a pass is the cumulative risk of the pass with respect to all the opponent players that are located in the risk area of a pass. So, there is the following formula:

$$Overall\ risk\left(pass\left(P_1, P_2\right)\right) = \sum_{P_i} Risk\left(P_i, pass\left(P_1, P_2\right)\right).$$

Then, they defined the sense of the Threat–Posing Player. Let's assume that a player has the ball and wants to reach to his closest spot at the shooting line (i.e. an imaginary line from which a player can shoot with good probabilities for goal) A threat–posing player is an opponent that can reach that spot no later than the player with the ball. Denoting with $T(P)$ the number of threat–posing players for a player P, then the formula

$$Gain\left(pass\left(P_1, P_2\right)\right) = T\left(P_1\right) - T\left(P_2\right)$$

describes the gain of a pass in terms of moving the ball away from opponents.

Next is the pass advantage, which is an attribute of a player position in relation to his teammates and is defined as follows:

$$Pass\ advantage\left(P\right) = \arg\max_{P_i \in TeamMates(P)} \left\{ \frac{10 + Gain\left(pass\left(P, P_i\right)\right)}{10 + Overall\ risk\left(pass\left(P, P_i\right)\right)} \right\}.$$

Moreover, goal chance describes the probability that a player P will score a goal if he shoots to the goal from his current position. The formula which represents goal chance is the following:

$$Goal\ chance(P) = \frac{goal\ width}{d} * \frac{\min(\alpha, penalty\ angle)}{penalty\ angle} * \frac{1}{1 + Overall\ risk(pass(P, GK))}$$

where:

- *goal width* is the width of the goal area.
- d is the distance of the shooting player P from the goal.
- α is the angle between the lines drawn from P's location to the two endpoints of the goal area.
- *penalty angle* is the angle between the lines drawn from the penalty point to the two endpoints of the goal area.
- GK is the opponent team's goalkeeper.

Finally, decision time represents the length of duration that a pass receiving player has to decide what to do with the ball before the closest opponent player challenges him. It is formulated as:

$$Decision\ time(P) = \frac{d}{max\ speed(closest\ player\ to\ P)}.$$

Given all these, the pass effectiveness score for a sequence of passes (P_1 passes to P_2 who passes to P_3) is described by the following formula:

$$\begin{aligned} Effectiveness(pass(P_1, P_2))_{Next\ Pass:\ pass(P_2, P_3)} = {} & w_1 \times Gain(pass(P_1, P_2)) + \\ & w_2 \times Pass\ advantage(P_2) + \\ & w_3 \times Goal\ chance(P_2) + \\ & w_4 \times Decision\ time(P_2) + \\ & w_5 \times Effectiveness(pass(P_2, P_3)) \end{aligned}$$

where w_i are weights which may be tuned based on insights from domain experts.

The effectiveness of this model was also evaluated by experts and it was proved that it coincides with expert's evaluations on a number of scenarios in 94% of the examples. Additionally, they implemented and incorporated the model in an open–source analytics tool. Drawbacks of the research were the small training dataset, which may lead to overfitting or divergence of the model accuracy, the small number of parameters used and the fact that the model does not take game status into consideration [77].

Horton et al. also exploited the opinion of domain experts, as they used them to rate the features of a pass between two players. Then they performed supervised learning to classify the

passes as Good, OK or Bad. Experimental results produced accuracy of 85.8% on classifying passes [78].

Cintia et al. split the pitch into 100 zones and introduced an innovative passing metric, using the harmonic mean H, of five measurable indicators: total passing volume w, mean players' passing volume μ_p, variance of players' passing volume σ_p, mean zones' passing volume μ_z and variance of zones' passing volume σ_z. So, passing behavior H of a team is formulated as follows:

$$H = \frac{5}{\frac{1}{w} + \frac{1}{\mu_p} + \frac{1}{\sigma_p} + \frac{1}{\mu_z} + \frac{1}{\sigma_z}}.$$

H was proven to be better correlated with success than every other of the five indicators and 2017 Champions' League winners, Real Madrid, was observed to have the best H rating amongst the 32 teams participating in the tournament. They successfully used H in combination with other metrics to predict the final league table of certain championships [74].

Relationship between passes and shots was investigated by another paper: Brooks et al. used data from 2012–2013 season of Spanish La Liga (i.e. the first division) and searched the relationship between the location of a pass and shot opportunities. They regarded possession of ball as a sequence of passes, converted each pass into a feature vector and assigned labels with the value 1 if the possessions ended in a shot or –1 if they did not. They also split the pitch in 18 importance–varying zones, used heat maps and showed that a team can be identified by their passing styles and by where on the pitch they attempt passes. Finally, they ranked players according to their tendency to be involved in leading to shots pass sequences using a novel ranking, the Average Pass Shot Value (APSV) [79].

Passing networks is also a very intriguing subject; players are represented as nodes of a network, while passes between two players are represented as edges between the players–nodes. The edges are weighted based on the amount of passes that are being exchanged between the players–nodes, as seen in Figure 11.

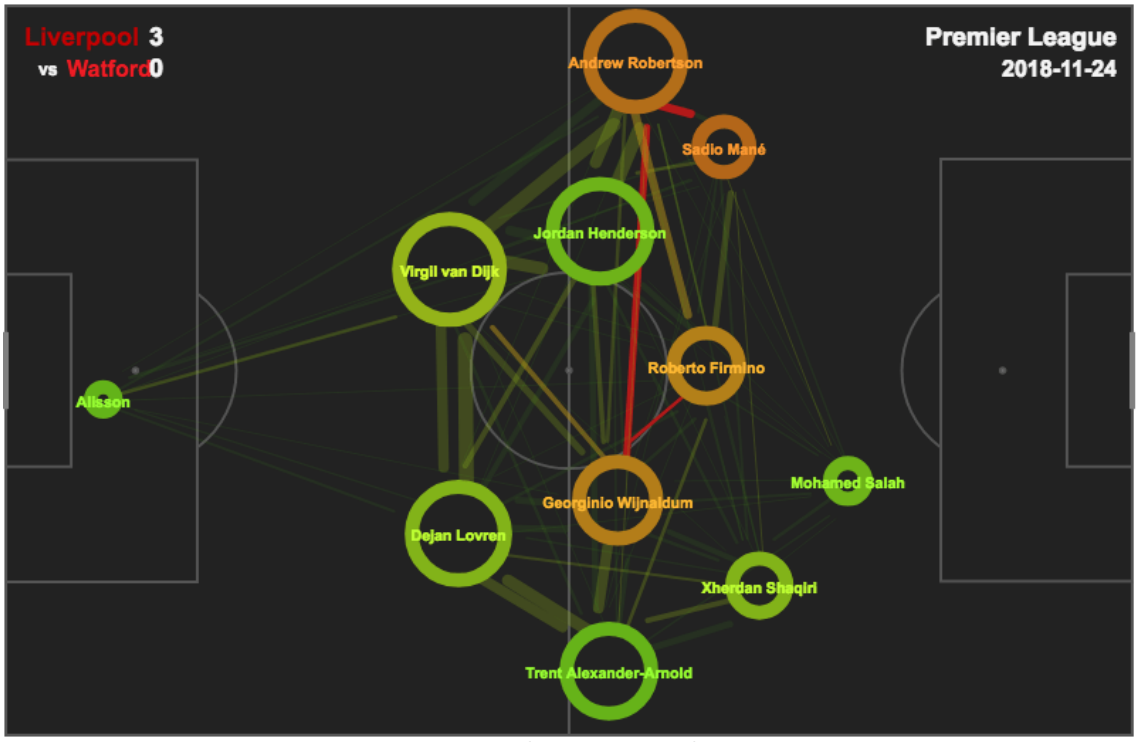


Figure 11: Pass Network of a single game of Liverpool FC **[80]**.

Grund analyzed a dataset of 283,259 passes and applied mixed–effects modeling to 76 repeated observations of the interaction networks and performance of 23 soccer teams. He proved that teams that performed better than others were characterized by networks with high intensity and low centralization [81].

Clemente et al. employed network metrics to improve the offensive processes analysis of teams. The dataset was five matches of the same team. They were possible to assess player connections and measure the strength of the connections between them. Their results indicated high level of importance towards the identification and characterization of the team collective process. They claimed that a network approach tool could be useful for coaches to delve into their team passing skills [82].

Cintia et al. leverage passing networks in several papers, in order to predict football matches outcome, but also leagues final table. They finally conclude that networks are more efficient for long–term predictions of whole competitions [83].

Spatiotemporal data consist of both time and space information. They are very substantial in sports analytics, as by definition, sports involve game modifications over time, while space allocation and exploitation are prerequisite. Furthermore, the advances in image processing made the analysis of positional data a lot easier. Several papers, already presented in the Dissertation, use spatiotemporal data, such as [67], [75], [76] and [78], but it's worth mentioning some more notable works:

Borrie et al. suggest that temporal pattern analysis will lead to a major contribution in deeper understanding of sport performance. The authors detected temporal patterns to find similar pass sequences within games [84].

Gudmundsson and Wolle leveraged positional data only to extract a list of events happening during a match and analyzed player trajectories (i.e. the sequence of all positions during a game) to cluster players by their movement patterns, which they named subtrajectories. The disadvantage of this method is that it proved to be computationally expensive [85].

Tamura and Masuda used data from the Japanese and German leagues and searched for correlation between temporal patterns of formation changes during games and the outcome of a match. The authors proved that managers keep the formation untouched when the team is winning and tend to change formation when the team is losing. So, they generally follow a WSLS strategy (i.e. win–stay, lose–shift). They noted that, surprisingly, change of formation usually does not have a significant impact to the losing team [86].

A computational geometry's tool, called "*Voronoi diagram*" (VD) has been discovered in 1908 by Georgy F. Voronoy, a Russian mathematician. Nevertheless it has been vastly used during the last 30 years in different science fields, such as biology, astrophysics, chemistry, fluid dynamics, engineering, architecture, urban planning, but also ML, computer graphics, robot navigation and more. It is a rather simply–defined and easy–visualized construct; given some points in a plane, their VD divides the plane according to the nearest-neighbor rule: Each point is associated with the region of the plane closest to it. The distance can be measured with various metrics (i.e. Euclidean distance, Manhattan distance etc). Despite their simplicity, VD have surprising mathematical properties and have been proven to be an effective tool in solving seemingly unrelated computational problems [87].

In the same manner, VD have also been used in sports analytics. The plane is the pitch and each point is a player. So, the power of a VD in football is that is shows region dominance of players and consequently of teams, as seen in Figure 12. VD are also dynamic, which means that a VD is changing by the movement of players. So, their analysis during a match, reveals the relevant patterns.

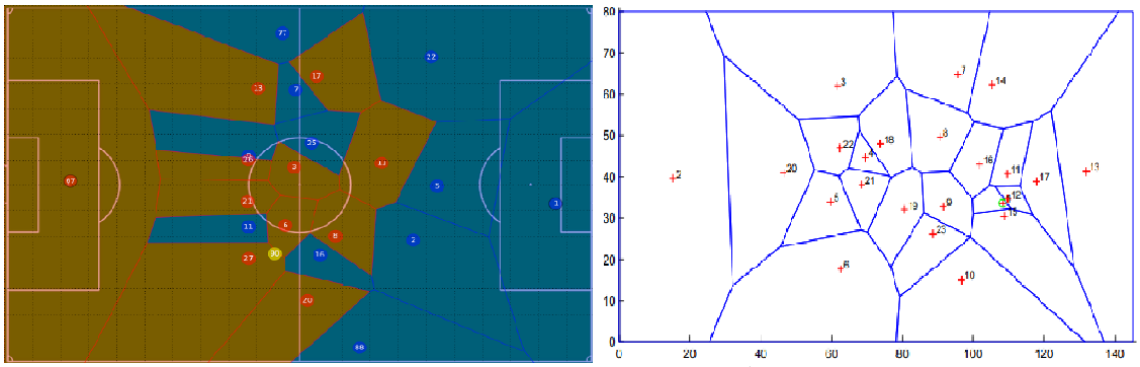


Figure 12: Typical Voronoi Diagrams, showing region dominance for teams and players **[88] [89]**.

The concept of dominant region was firstly introduced by Taki and Hasegawa in 2000. They developed a motion analysis system of team ball games in which the dominant region was used for quantitative evaluation of basic teamwork actions [90]. Dominant region was further developed by Fujimura and Sugihara. The authors proposed a quantitative method for evaluating sport teamwork based on VD and dominant regions. They also constructed and evaluated an efficient and realistic player motion model [91].

Kim also presented a Voronoi analysis of football games. According to his findings, a team having larger area ratio in VD might not win the game, but has most chances to score a goal, thus to win. He also indicated how certain defensive strategies affect the VD and highlighted that Voronoi analysis could be used to evaluate players as well as the homogeneity of a team [89].

Fonseca et al. used a database from futsal games (i.e. a variant of football, being played indoors and on hard court), performed spatial analysis with VD and suggested that it is possible to identify a number of characteristics that can be used to describe players' and teams' spatial behavior. They focused on defensive methods too [92].

2.2.8 Cameras and Wearable Devices

Structured data about football are easy to find. There are a number of web sites that provide accessible databases, tables, rankings, statistics, past results etc. These data can be used in analysis and prediction, but are not always adequate, as they do not usually contain spatiotemporal information about game actions. The easiest and not very costly way to acquire that kind of data is through cameras and video analysis. Some indicative works are presented in this section of the thesis.

In 2010, Poppe et al. proposed a multi–camera video analysis system for soccer sequences. Each camera detected and classified objects (i.e. players, referees and ball), as seen in Figure 13. Then the different projections were merged to record the trajectories of players and the ball. In that way, field actions are described by certain movements (e.g. when some players run towards the opponent team's goalpost, that may imply an attack). Some issues regarding video resolution are the difficulty in detecting the ball when it is close to a player and the discrimination of two players that also are close to each other. The system was evaluated against a public dataset and presented possible improvements for sequences [93].

Figure 13: Object detection on single camera view **[93]**.

In a similar work, Theagarajan et al. designed a capable of generating automated visual analytics and player statistics from recorded on camera football games. The authors evaluated the dataset with team dependent and team independent settings and noticed how these settings affect the performance of the networks. They proved that differences in the generalization ability of the network occur when training the networks on RGB or gray scale images and indicated that the performance of the system are influenced by different game scenarios. They concluded that similar future models should consider augmenting the quantity of images, game events, and data used [94].

Niu et al. proposed a framework to systematically analyze soccer tactics. They detected field lines in order to extract trajectory methods and recognized six typical soccer attack patterns for tactic analysis. Ball state was clearly defined and the identified trajectory was claimed to enable analyzing and improving soccer tactics in terms of the conciseness, clarity, and usability [95].

Kazemi et al. engaged with pose estimation of players, with snapshots from multiple calibrated cameras. They used a random forest classifier to capture the variation in appearance of body parts in 2D images. Then, they combined the 2D part detectors in order to create 3D pose estimations, with the help of 3D appearance likelihoods computed from part detectors that ignore the left and right label of the parts. Problems with mirror symmetric body parts were solved with a latent variable formulation [96].

Even if cameras and computer digitization have been two of the main sources for data mining and analysis at sports during the last decade, they display some drawbacks; data need to be collected manually, which is a time–consuming and tiring procedure. It is also difficult to

measure certain quantities, such as the velocity of a player. Finally, data that are collected via cameras is questionable in terms of validity and quality [97].

Wearables are electronic devices incorporated into items that can be comfortably worn on a body or on clothes and are used for tracking information on real time. They have sensors concerning motion or medical purposes that take a snapshot of your activity and send their data to other devices, as seen in Figure 14. It is a big technological innovation, but even before the wearable technology entered the consumer market, these devices were used in the field of military technology. They were an integral part of the medical and healthcare sector in the army. Nowadays, smart watches, fitness trackers, smart clothing and other types of wearables are ready to take the market by storm [98] but in football the usage of wearable devices was forbidden, so they are not yet fully used by every club. Electronic Performance and Tracking Systems (EPTS) were only recently allowed during matches by FIFA [23].

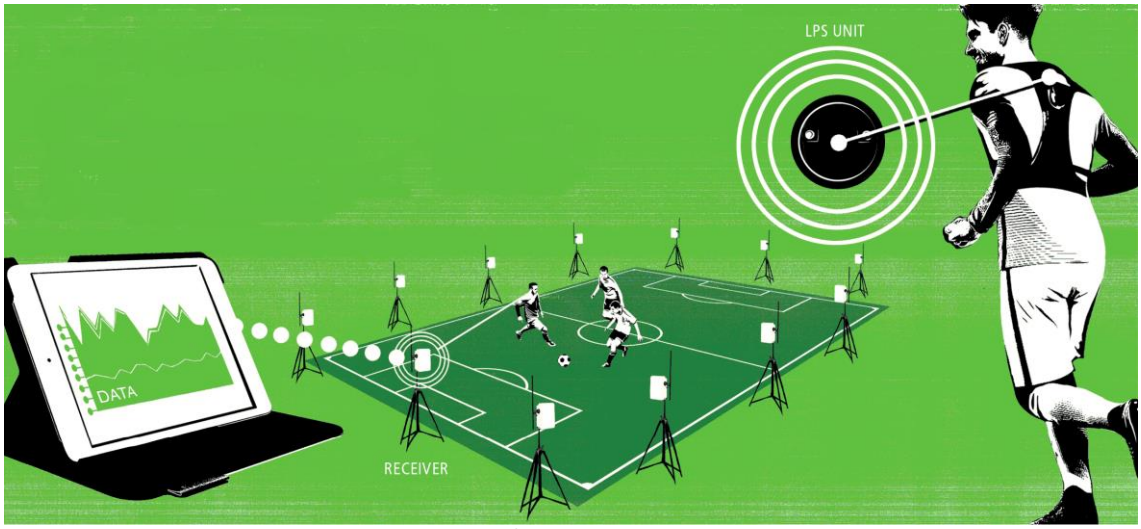


Figure 14: Typical structure of an EPTS device and its data receiver **[99]**.

Nevertheless, there are various reasons why clubs should include wearables in their equipment. One obvious reason is because data acquisition is being conducted faster and easier. Data accessed by EPTS devices is accurate and provide several measurements. In general, wearables are able to measure what cameras cannot. Clubs can use them while training and during the matches. Since wearables are connected to other devices, providing real–time data visualization to coaching staff is feasible. Coaches must be aided in order to take decisions concerning tactics and strategy. An appropriate athletic performance monitoring system should be intuitive, provide useful analytics, feedback and reliable predictions to coaches and athletes in real–time [100].

Moatamed et al. in their work proposed such a tool and used it to measure the readiness score. According to the authors, readiness score is a metric which indicates whether an athlete is ready to perform an exercise or take part in a game. They stressed that sometimes the results of

those metrics are misleading, so an inclusive analysis is required to achieve substantial evaluation [100].

Fernandez et al. proved that, by exploiting non–linear models and by performing fine–grained removal of highly correlated features, it is possible to predict physical variables based on training and match information from EPTS devices. The authors also claimed that previous match data alone is not reliable for predictions. They also proposed that internal metrics, like heart rate exertion should also be used in future works as complementary to tactical data [101].

What the authors of the previous research implied, leads to the next major topic of wearables, which is sports medical science. The medical staff of a team (i.e. physiotherapists, doctors, fitness coaches etc.) can leverage the data coming of the EPTS devices in various ways. They could be able to prevent an injury or detect a fatal emergency (i.e. a heart attack) moments before it happens. In the time of big data, the medical staff should possess the full medical history of players and exploit it with perfect responsibility. One player maybe should be treated in a different way than others during the rehabilitation following an injury. Wearables can collect absolutely vital medical data, which is not approachable even maybe with health examinations. And also, save time, money and lives.

Wilkerson et al. analyzed data extracted from wearable devices to search for relations between average inertial load measured during training and occurrence of musculoskeletal sprains and strains. Coefficient of variation for average inertial load and frequent exposure to game conditions were found to be strongly associated with injury occurrence [102].

Nawrocka and Lukowski experimented with 50 athletes from various sports before and after exercise and extracted 10 variables for each athlete. They detected for differences in brain wave frequencies before and after exercise. The authors proved that the connection between mind and body after and before exercise can specify for frequencies of Delta and Alpha brain waves [103].

However, there are some ethical concerns in big data in sports, derived by wearables. A faulty device, signal interruptions between the device and its interacting environment or the body of the user might produce unreliable or unstable data. This may lead to wrong tactics decision or may harm the athlete's health. Clubs should hire data scientists, specialized in tracking noise in data and unusual patterns, to try and minimize the errors of wearables and to deal with the inevitable algorithm bias [104].

Additionally, data security and protection issues exist. Personal information is not fully protected from hackers or malwares. Sensitive data corresponding to famous athletes worth a vast amount of money and may be intercepted. Even worse, wearables and GPS devices are not used only by professional athletes, but from simple amateurs as well. People's data are not safe and anonymity is not guaranteed too. Therefore, the problem reaches the fringes of violating citizens' personal information and is transformed into a political issue [104].

Finally, there is the ethical dilemma, as stretched by Vermeulen and Sarma: "*advances in research or protection of human subjects?*" By implementing the new technologies described, athletes tend to rely more and more on their biometric data. They learn to trust numbers more than their intuition and visualizations of statistical analyses more than their experience. Data scientists managing the tracking data from wearables are responsible for protecting athletes from potential damage and also for maintaining the integrity and quality of their researches [104].

Bottom–line, wearable devices generate large datasets which can be exploited by clubs to extract hidden knowledge, but can also be used for upgrading the physical activities of amateurs. Data scientists should employ techniques to organize and transform the raw data into mathematical and statistical models which should be used for innovative approaches to training and tactics. Nevertheless, there are considerable risks by the usage of wearables, with security issues and governance of performance by electronic devices being the most important [104].

2.2.9 Player Performance Prediction

It is reasonable to say that one of the most intriguing things to predict in football is the performance of individual players regarding to next season. Clubs are paying millions to buy players. Paris Saint Germain FC bought Neymar Jr. from Barcelona FC, paying an outrageous 222 million euros. His transfer in summer of 2017 caused a transfer domino effect; Barcelona FC bought Coutinho from Liverpool FC for 145 million euros, then Liverpool FC spent 85 million euros to buy van Dijk and 62 million euros to buy Alisson, the most expensive transfers of a defender and a goalkeeper at the time they were agreed. What is known as "Neymar effect" has boosted the money spent in transfers in almost prohibitive levels. So, clubs that are heavily investing in footballers need to have an indication of how their new transfer is going to perform.

Additionally, there is a huge raise of gamblers who bet on fantasy sport games. Those are games in which players build in imaginary team picking real players from different teams. The bettors may spend an imaginary but limited budget in order to build their team. They can also trade or drop players in their own judgment. Virtual teams are competing to each other in terms of statistical performance of the chosen players in real games. Fantasy games vary in terms of complexity, regarding to the point system they use, but are offered by almost every betting company. Surprisingly, the first fantasy game was played in Harvard University, in 1960. By 2016, it was estimated that almost 60 million people were taking part in fantasy sport games just in the USA and Canada and that fantasy was a 7.22 billion USD industry. Under that point of view, it is understandable that betting companies and bettors are really interested to be able to predict player performance [105].

However, the issue of predicting player performance still remains rather unexplored. Later on, the most important relevant works are briefly presented.

Szczepanski in his work proposed a framework suitable for the evaluation of player skills in order to establish their value. He posited that the framework should relate player performance with his underlying skill as well as other exogenous factors and also that there should be a link between individual performance and team success. He analyzed players' shooting and passing, as well as the parameters that affect them. Then the author used a Markov chain model to describe game events depending on the skills of the players involved in the specific game. Therefore, players are evaluated according to their actions and factors of performance, apart from player skills, are being recognized. Thus, players are characterized by multiple weighted skills, their relative value is expressed in terms of team performance and the value of any player is proven to vary for different teams. The drawbacks of his research were that he only examined two offensive skills (shooting and passing), ignoring the defensive contribution of a player and that the dataset contained data from only two seasons. There also were some limitations in terms of statistical analysis [106].

Nsolo et al. investigated the attributes which best predict the success of individual players, based on their position and evaluated different ML algorithms regarding prediction performance. The authors focused on top players of the top five European leagues and valuated players based on different attributes for each player's position. They concluded that forwards tend to have higher performance ratings than other players, so maybe more advanced metrics should be applied on defensive players. Some limitations of their method are that rankings used are expert–based and that their approach does not take the quality of the team mates into account. They also indicated that correlation of rankings with player market values should be included in future researches [107].

Sîrb et al. presented a set of 54 performance criteria, over different playing positions in order to evaluate the performance of players, taking into consideration each player's natural position and the tactical formation that the team deployed in a match. They examined the performance of a Romanian team for the first half of a season as a case study and employed fuzzy sets in order to evaluate the satisfaction degree of each tactical compartment regarding each performance criteria. The authors claimed that their mathematical model has been proven to be very efficient, given that the objectivity of the final result achieved has been tested in the case study [108].

Pariath et al. in their work acknowledged the difficulty in identifying performance value of individual players due to a variety of reasons. They use a data–driven approach and implement a player performance prediction model which learns from various player attributes and skills. The authors managed to establish a relationship between player attribute values, player positions, market value and predicted performance value [109].

Mackay used a different approach. Although his work refers to goal probabilities of EPL teams, there is a section about individual player performance. Players were rated in terms of the difference between goal probabilities for every action they made. The dataset consisted of 260 players and the top 15 players who turned out to be more dangerous for the opponents' defense were actually most famous EPL players; apart from the praise the got by press and the public, they turned out to be the most expensive ones. Again, defensive skills are not the issue [110].

He et al. focused on the market value of players and the model used for performance evaluation was applicable only on forwards. The relation between player market values and player performances was studied and it was shown that market value raise according to performance, as expected. Nevertheless, they noted that there is a ceiling for market value. The authors also researched the notion of over–valued and under–valued players, comparing the market value of a player to the money that was spent for the player's transfer [111].

One model which highly assesses some defenders was proposed by Sáez-Castillo et al., back in 2011. But, again, does not rate them in terms of defending actions, but in terms of goal scoring. The model evaluates the forwards more strictly than midfielders and defenders as far as expected number of goals are concerned [112].

In a very stimulating work, Pappalardo et al. analyzed player performances from 18 different competitions for several years and presented PlayeRank, a powerful data–driven framework. The dataset contained 31 million game events and 21 thousands players. PlayeRank was found to outperform competitive predictive algorithms. They also discussed what distinguishes top players from others and discovered patterns for excellent performances. One of the limitations that emerged was that PlayeRank does not take into consideration off–ball actions, like pressing. The authors also emphasized on the fact that an improved version of the framework should be able to leverage data from other sources, like wearables, GPS and video tracking data [113].

While ending this sector of the thesis, it would be appropriate to make a reference to a research referring to fantasy football, which was discussed above. The analysis was presented online and its objective was to predict the best players for the 2017–2018 season in Italy's Serie A. The dataset included results of 4 seasons and 265 players. 58 of them had a constantly increasing average score (mainly forwards) and 46 of them had a constantly decreasing score (mainly defenders or goalkeepers). The author used a simple mathematical formula, no more than a weighted sum, to predict the score of each player, based only on five parameters; score difference between first and last season, career score, presence of a constant positive/negative trend, number of years playing in Serie A and last season's average score. By that, he proposed whom he believed to be the best 35 players for the fantasy game [114].

More researches have been conducted on player performances for other sports, like baseball, American football, Australian football, cricket and so on.

2.2.10 Player Injuries Prediction

Injuries in sports in general and in football in particular, have a major negative effect not only for the injured players, but also for the clubs they belong to. When a player is injured, a team loses part of its strength. It might be even worse if multiple players are injured, if important players are injured and if they are injured for a long term period. Besides, the medical and rehabilitation expenses are extra charge for the clubs [115].

Injuries also affect players. A player that has recently healed from a serious injury might not be able to perform at acceptable level for some time. Fragile players that get injured multiple times in their career may never reach their full potential. Additionally, injuries may have a long time impact on players, even after their retirement. Injuries that are related to sports are the second most common cause of traumatic brain injury after traffic accidents [116].

Consequently, being able to prevent serious injuries is a substantial task of every club's medical staff. ML techniques are being exploited during the last years for that purpose. In that section of the thesis, some notable works over injury forecasting in football are discussed.

Kampakis collaborated with two EPL clubs to examine predictive modeling for injuries in football. He conducted three different researches: The first one was to predict the recovery time of injuries depicted in Figure 15, using ML algorithms. Then he performed correlation–based subset feature selection to improve the model's performance, which optimized with random forest. Secondly, he tried to predict injuries, studying the relationship between hours of training or match and injuries. More specific, the objective was to find how many hours a footballer may exercise before getting hurt. Finally, he analyzed GPS measurements from training sessions. His goal was to discover intrinsic injuries due to burnout or overtraining. The task was applied on Tottenham Hotspur FC players in order to predict if a player is threatened by an injury in a given week. A number of algorithms were applied; supervised principal component analysis achieved the best results, while the final variables used were no more than 3 or 4 [117].

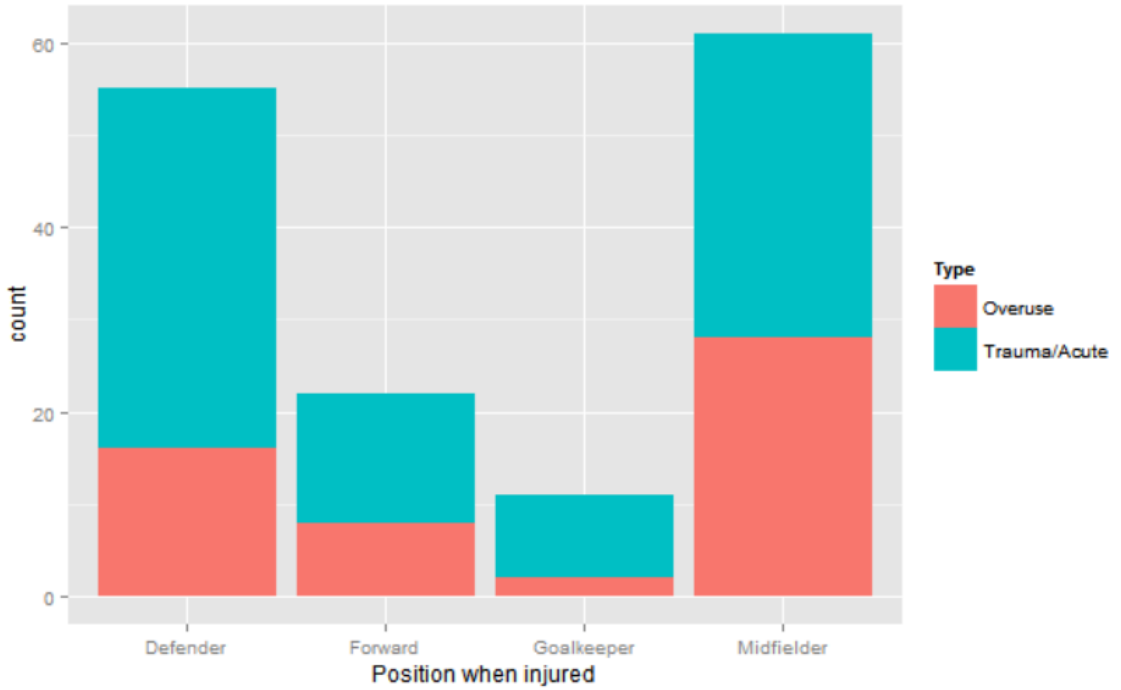


Figure 15: Position of the player injured and severity of the injury **[117]**.

Rossi et al. introduced a multi–dimensional approach to injury prediction in football, based on GPS measurements during training and ML. The injury forecaster they created has a good trade–off between accuracy and interpretability, but also provides the rules to relate an injury with the reason behind it. The authors claim that if being used at the beginning of the season, it can save the club from medical expenses during the season. Future work that authors indicated as possible involves future extraction from official games, with higher physical stress. Additionally, they suggested that it is possible to train personalized predictors, combining GPS data with other types of medical data about each player for several seasons [115].

Martins created a model for injury prediction using EPTS data from training, combined with self–rating subjective data. He used PCA to reduce dimensionality and ROSE sampling technique, along with XGBoost algorithm (XGB) was considered the most accurate. He claimed that his algorithm can be used by coaches to delimit the training load of individual players, as overtraining is highly possible to end to an injury. The author related stress with overtraining and muscular injuries and noted that the most relevant objective features are related to a decrease of sprinting frequency and velocity. Limitations on his research refer to small amount of data as he used data from one team and from one season [118].

Michałowska et al. studied 342 isokinetic records and trained a model about injury risk evaluation for three predefined types of injuries, with ANN. The model returns a value in range (0,1) from various muscle parameters. The accuracy and specificity of the model was good, but sensitivity was not satisfying. The authors suggested that improvement is feasible by increasing

the amount of training data and by combining this data with other features of injury risk assessment [119].

Olmedilla et al. investigated injuries, under a different perspective, as they focused on the players' psychological state and the effect it has on them in terms of injuries. They proved that the key psychological factor related to sport injury occurrence is stress. They presented different models which examine the relationship between stress and injuries and presented stress control strategies. Nevertheless, the aforementioned relationship is still in question, as the authors presented researches where stress control groups of athletes had not significantly less injuries than other athletes. They pointed out that age, gender, competitive level, position and cultural issues must also be taken into consideration in future researches [120].

2.2.11 Uncertainty of Outcome, Competitive Balance and Competitive Intensity

Football, as well as other sports, is not just a sport, but a whole industry. Top leagues and top clubs are not interested only in improving team performances, but also in profiting from the sport. Fans tickets, sponsorships and TV rights are the biggest money sources for leagues and clubs. But in order to maximize their profits, they have to assure that they will upgrade their product, which is football, in various ways.

In 1956, the father of sports economics, Simon Rottenberg, in his article in the Journal of Political Economy, stated the uncertainty of outcome hypothesis (UOH) and claimed that the fans favor the games where there is high competition for the win [121]. His hypothesis is summarized as follows: Less certain winning team leads to higher competition, which leads to higher game attendance. This notion has been verified but also has been challenged by various researchers [122].

Scelles et al. [123] defined the three concepts of the sector title as follows:

- "*Competitive balance* is the domination of a team over one or more seasons"
- "*Competitive intensity* is the importance of a match in the framework of qualification in higher competitions or relegation in lower one"
- "*Uncertainty of outcome* is teams' probabilities to win a competition".

Manasis et al. used mainly economic data from eight leagues of European football from 1959 to 2008. Their goal was to identify the optimal index for the study of competitive balance. According to the authors there was a strong indication that UOH is indeed supported by their model. They also highlight that, regarding attendance, ranking mobility across seasons is more important than the performance of one season [124].

On the other hand, Baidina and Parshakov did not find any evidence that UOH explains the behavioral pattern of the attendees in the RFPL (i.e. the first league of Russian football). They noticed a U–shaped dependence between attendance and uncertainty and they put that down to visiting team effect. The authors claimed that high attendance is more related to seeing a top team as a guest team, rather than on the uncertainty of outcome [122].

Czarnitzky and Stadtmann also agree that reputation of playing teams is more important than uncertainty of outcome in terms of attendance. The authors analyzed uncertainty measures of match outcomes from the German Bundesliga. They also included the uncertainty of champion winner and some other features [125].

Staying in German football, Pawlowski admitted that previous researches could not assess whether football competitions might be at risk of losing fans' interest because of the UOH effect. But he claimed that competitive balance is a very important factor for the vast majority of football fans, while the league is still away from securing that [126].

Scelles et al. emphasized on the importance of competitive intensity and were skeptical against researches that are focused only on competitive balance. Their approached revealed that the impact of uncertainty of outcome for the home team is significantly positive. They propose that future researchers should combine competitive balance with competitive intensity to fully comprehend the relationship between uncertainty of outcome and interest in match attendance [123].

Finally, Buraimo and Simmons focused on television audience preferences and concluded that although in previous years uncertainty of outcome maybe have had slight effect on audiences, during the latest years the hypothesis is not verified, at least as far as television audience is concerned. The authors noted that television viewers rather than championships' competitiveness tend to be more interested in player's talents and advised clubs and leagues to try and develop that [127].

2.2.12 Outstanding Previous Work in other Sports

Completing the literature review of the thesis, it would be appropriate to mention some of the top researches that were conducted and do not have football at the center of their attention. As there are thousands of fine papers, there was a need of reducing the number of candidate researches to cite.

MIT Sloan Sports Analytics Conference is probably the most important sports analytics conference in the world. It is held in Boston, once a year and the inaugural event took place in 2007. Several organizations, like sport clubs, sports data and statistical companies, but also

individuals, such as athletes, GMs and other staff members take part in the conference. Every year, the papers that are considered the most influential are being awarded for their novelty.

Every paper that is presented in this section has already earned a distinction during the last 5 years by MIT. Thus, the previous work that is discussed is not only outstanding, but also recent.

Since Houston Rocket's Daryl Morey is co–chairing the conference, it is not a surprise that a significant proportion of papers refer to basketball. Besides, basketball, as mentioned is one of the most measurable sports. The first four papers to follow engage with player movements and actions on offense.

Wang and Zemel used a variation of neural networks to process data from SportVU and classify team offensive play calls. The problem is rather difficult as there are a number of factors and player interactions that affect classification. They achieved 80% accuracy on unseen data and also pointed out that their model displayed notably transferability across seasons [128].

Similarly, Miller and Bornn implemented a machine learning method for analyzing a database of basketball players' tracks in order to find interpretable patterns. The authors fitted a probabilistic model regarding game actions and used it to describe player behavior. Then they developed a hierarchical model and used it to describe interactions between players. They proved that teams follow repeated offensive structure [129].

Nistala and Guttag used unsupervised machine learning to constructed 3 million trajectory–embeddings from 3 seasons of NBA player tracking data. Trajectory embeddings are 32–dimensional vector representations of player offensive movements. They considered Euclidean distance between trajectory–embeddings an excellent indicator of the visual similarity of the movements they encode [130].

Mortensen and Bornn leveraged a mix of Markov chain represented as a Poisson process and spatial statistic tools to analyze complex movements and attacking patterns. They claimed that their model has more clear interpretation compared to similar researches. They managed to model two dimensional data and proposed the engagement with three dimensional data as future work [131].

Kuehn created a framework for evaluating NBA players based on how their skills are complementary to their teammates. He developed a probabilistic model for ball possessions which takes into account on–play events, opponent teams' lineups and the complementarities between the skills of teammates. Thus, he identified the substitutability between player actions, good and bad teammates and the lineup–specific value a player brings to a team. He concluded that players are mostly payed according to their individual statistics, while they should be assessed based on the under–evaluated complementarities [132].

Felsen and Lucey performed an analysis on 1500 three point shots and captured the high level body movements of a player during a shot. They proved the value of this pose

representation as they quantified attribute differences for made and missed shots and evaluated attribute importance. Finally, they performed a case study on Steph's Curry shooting style [133].

Another research on shot distribution in respect of lineup formation was conducted by Sandholtz et al. They used public data to evaluate the efficiency of shot allocation, but in novelty spatial manner. A method for assessing shot distribution of the lineup was implemented and then exploiting those measurements, they quantified how many points were lost due to inefficient shot allocation [134].

Kaplan et al. studied the effect of the absence of an NBA superstar player from a game for various reasons. They concluded that their absence comes with a significantly negative economic impact for franchises. The authors highlighted some differences between players (i.e. some players' absence is economically harmful for the home team, others' for the away team) and organizations (i.e. games played in larger markets appear to reduce the importance of the absence effect of superstars) [135].

Talukder et al. used ML techniques to predict the probability of injury for NBA players and presented a model for injury prevention during games. They achieved strong accuracy in short–term injury prediction and proposed that players at risk of injury should be strategically rested during the season. Finally, they ranked player injuries based on the economical expenses and financial cost of missed games for the injured player [136].

Baseball is also a lucrative source of SA and several papers based on baseball have been published. Plenty of them try to evaluate and predict pitchers' performances. Pitchers are thought by most to be the most important on field participants of baseball games.

Salmon and Harrison deployed various metrics in order to explore pitchers' behavior and identify their performance trends. They claimed that their research can help in managers' decision making regarding pitchers [137]. On another study, the authors presented a technique for the classification of pitcher–batcher matchups and then simulate potential matchups. The goal was to identify optimal matchup strategies [138].

To assess pitchers' true pitching ability, Shu presented "Arsenal/Zone" rating, a speed and trajectory of pitched balls system. The final model was considered to be better than every mainstream projection system in terms of distinguishing breakout and breakdown pitchers [139].

Martin analyzed 2.5 million pitches from MLB thrown by more than 400 pitchers during a time period of 5 years. He focused on velocity, movement and release points and examined differences between pitches. He concluded that maximum velocity, strike rate and vertical movement are the most important features for assessing strikeout percentage [140].

Glynn and Tokdar detected about abnormalities in MLB performances and ability of players to hit home runs. Trajectories and performance levels of players were dynamically

modeled in regard to players' age. The authors developed a Markov chain Monte Carlo algorithm for Bayesian parameter estimation. The model was found to be comparable with similar methods and the accuracy was proved to increase with age [141].

Paulsen implemented a different approach evaluating a player performance, as he connected it with the years remaining on his contract and examined several alternatives of this relation. Through his study, he concluded that players on the last year of their contract tend to perform on a higher level that their usual standards and also presented a list of evidence on the conclusion he came to [142].

Papers have also been published about American Football. Hochstedler used geospatial data from NFL. His goal was to use quantitative methods in the pursuit of open receiver but also to evaluate elusiveness of players and quarterbacks' decision making, which all are important factors for team performance [143].

Another research that focused on quarterbacks' decision making was conducted by Burke. The author proposed a novel neural network approach to player tracking and passing data and claimed that the results were outstanding. Target selection, expected yardage and pass outcome proved to be substantial for the model. The possibility to adjust the model to individual quarterbacks using transfer learning was also explored [144].

Bornn et al. used training load data from European as well as from American football and exploited acute:chronic workload ratio (ACWR), a metric that takes into consideration the cumulative player workloads in the last 7 and 28 days. By the use of Monte Carlo methods they measured the effect that yearly training calendar has on the ACWR–injury. The authors advocated that trainers may use their model to predict injuries [145].

Kurt et al. offered another point of view on NFL championship through their work. They noticed that a team's average win percentage reduces by almost 4% when facing a team that had extra rest during the previous week. They also indicated the imbalance in the total number of such games between different teams. To overcome the negative effect of competitive imbalance, they implemented a two–phase heuristic approach and created a multitude of possible and balanced schedules that could potentially be applied for an NFL season [146].

Another sport to review is ice hockey. There are not so many statistical tools as in football, baseball or basketball, but since it is a famous sport in North America, scientists have been trying to fill in the gap during the recent years.

Pettigrew introduced a win probability metric for the National Hockey League (NHL) and through this he developed Added Goal Value (AVG), an evaluation metric for player productivity, also suitable for player comparison and for performance prediction. AVG is applicable on goalkeepers too, in order to measure their contribution to the game's outcome [147].

Schulte et al. developed a system that supports decision making in terms of drafting, trading and coaching NHL players. They used a mixed set of game events and location data to cluster players according to their style and roles with ML techniques. The general idea is that clustered players are more easily ranked and are not compared with unequal criteria. Finally, the authors deployed a high resolution Markov game model to analyze every game event and quantify its impact on scoring the next goal [148].

Schuckers presented a statistical generalized additive model which uses historical, demographical and performance data along with traditional scouting in order to predict potential performance of players about to be drafted for the NHL. The author claims that his model has predicted players' careers better than most organizations' GMs [149].

Bornn and Javan conducted a study on ice hockey's face–off (i.e. the method used to restart the game after a goal, with two opponent players trying to gain control of the puck using their sticks). The novelty in their research was that they did not focus on the traditionally metric of win percentage. Instead they focused on other features, as the zone and the side the face–off took place, the areas of the ice the puck was directed to after the face–off and game events following the face–off [150].

The same authors published a research about pace in ice hockey and the effect that player or team pace has on the game. They used spatiotemporal data and concluded that pace is highly correlated with shot quality and danger zone entries but not with successful passes, as with high pace more unsuccessful pass receptions occur [151].

Bagley and Ware presented a method to rate volleyball players and teams on six substantial skills, namely serve, reception, set, attack, block and dig. They noted that is important that the overall rating of players should emerge by weighting these skills, according to each player's position. The results can be used for prediction, as well as for coaching and training [152].

Concerning motorsports, Kataoka and Junkins studied how the muscle use in extreme racing conditions affects IndyCar drivers. The researchers used data from wearable devices and from other sources, dealt with noise, used unsupervised learning (i.e. clustering), as well as visualization techniques and discovered valuable insights. An interesting finding was the potential relaxation points for driver to save muscle fatigue in forearms [153].

Wei et al. conducted a research on tennis to discover patterns of player movement and ball striking. They used data extracted from HawkEye, a computer system which visually tracks the ball's trajectory. They managed to build representation of playing styles and personalize interactions between players according to match events. Based on these, they produced spatiotemporal dominance criteria to predict point distribution [154].

Finally, a paper that concerns various sports was presented by Berry and Fowler. The authors focused on the effect that coaches have on team performance. They introduced Randomization Inference for Leader Effects (RIFLE), a new method for estimating the coaching

effect. The results showed that coaches are responsible for 20–30% of team success, depending on sport examined, which is a significant percentage. Finally, they evaluated individual coaches in terms of their own actual performance against the expected. This could be used by teams as a tool to assess coaching candidates [155].

Chapter 3

3.1 General Terms

In this section of the dissertation, the meaning of the following terms is briefly explained:

- Machine Learning
- Data Mining
- Sports Analytics
- Machine Learning Algorithm
- Data Analysis
- Performance Prediction in Sports

All of them have already been mentioned in Chapter 2, but since they are further mentioned in the dissertation, it would be appropriate to define them before moving on.

3.1.1 Machine Learning

Machine Learning is a science field that gives computers the capability of improving without the need of explicit programming, but rather through past experience, statistical models and algorithms [156] [157]. The most widely quoted definition of ML was given in 1997 by Mitchell in his book "Machine Learning":

"*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.*" [158].

ML is an exciting technology that grows rapidly and is applicable in other science fields, such as finances, health care and more. ML is part of *Artificial Intelligence* (AI) and, as seen in Figure 16, lies at the intersection of *Computer Science* and *Statistics* [156].

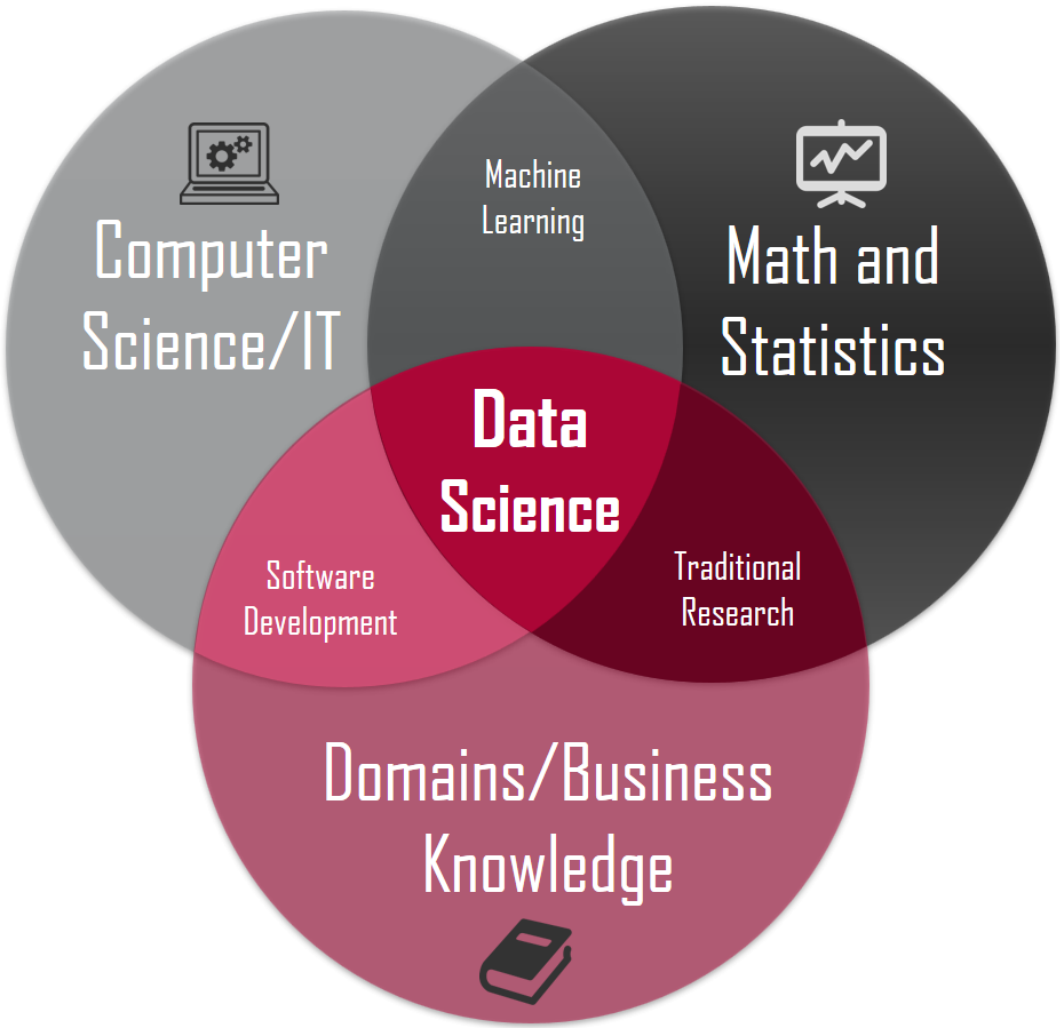


Figure 16: The Venn diagram of a Data Scientist and the position of Machine Learning **[159]**.

ML is split into various branches. There are three main categories and some other sub–categories which are mentioned in this section and are considered part of ML. Specifically, the three main branches are:

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

Other types of ML are *Future Learning*, *Anomaly Detection*, *Association Rules* and more [160]. The structure of ML can be seen in Figure 17:

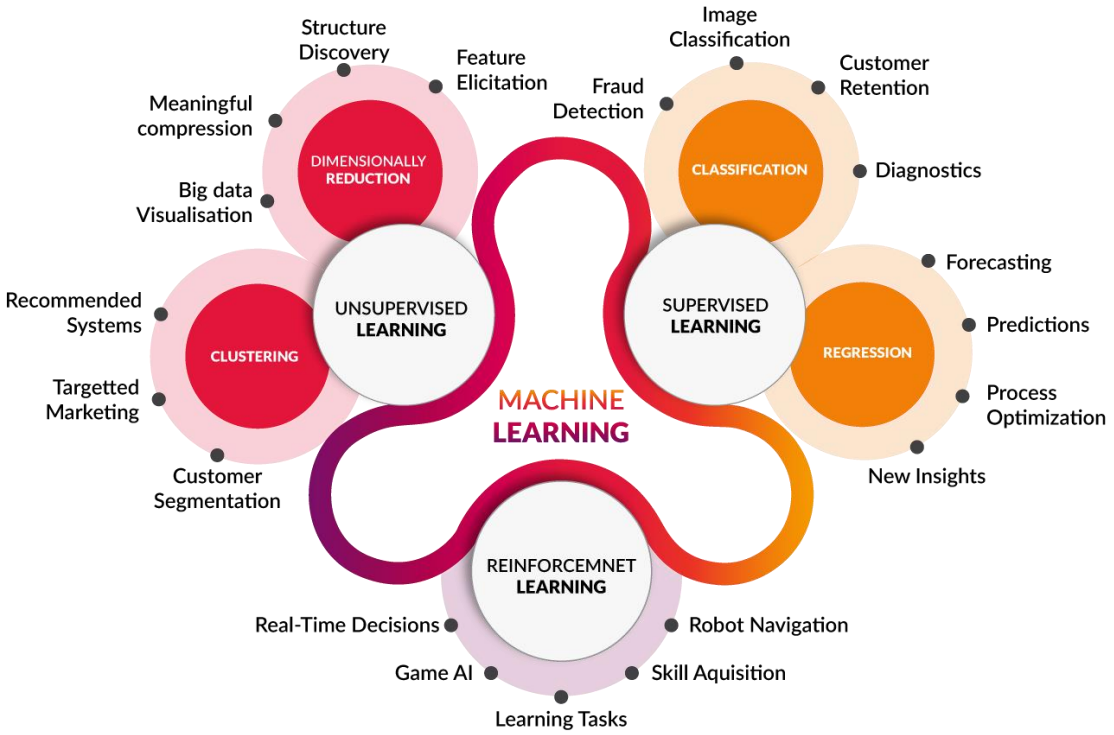


Figure 17: A typical Machine Learning structure with the most important ML tasks **[161]**.

Supervised Learning is the type of ML where data need to consist of input–output pairs. Outputs are assigned by a human supervisor (hence *supervised* learning) and are called *labels*. Therefore, labels are known during training. Samples are being repeatedly fed to the model and for every sample the model returns a prediction, which is compared to the label. Then, after evaluating the model's performance, the model is being fed with another sample and parameters are updated to improve the performance metric and so forth. This procedure is being conducted with historical data. After is being completed, the model is used for making predictions in unknown data. Supervised learning solves two major ML problems: *Classification* and *Regression* [160].

Classification is the process of assigning a category to input data [160]. Several metrics are used to measure success in classification. The most well–known are *Accuracy*, *Precision*, *Recall* (or *Sensitivity*), *Specificity* and *F–score*.

Assuming that there is a prediction on binary (0 or 1) input data, some of the instances are predicted correctly and some are not. The correctly predicted instances are split into two categories: True positives (*TP*) are the 1's that were predicted as 1's and True Negatives (*TN*) are the 0's that were predicted as 0's. Accordingly, there are two categories for the misclassified instances; False positives (*FP*) are the 0's that were predicted as 1's and False negatives (*FN*) are the 1's that were predicted as 0's [162]. Under that assumptions, the following hold:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$F\text{-}score = \frac{2 \times Recall \times Precision}{Recall + Precision}$$

There is no restriction as to which of the aforementioned metrics should be used in a classification problem. This is a decision which should be made upon the nature of the problem. Accuracy, for example, is usually a fine metric, but can easily been misleading in a medical case problem.

Regression is the process of predicting a continuous numerical value for input value [160]. The difference between the predicted number and actual number is called *error*; it indicates how wrong the model is. The goal is to decrease the error. There are different types of errors: The *Mean Square Error* (*MSE*) or the *Root Mean Square Error* (*RMSE*), the *Mean Absolute Error* (*MAE*) or the *Mean Absolute Percentage Error* (*MAPE*) and the *R2–score* [163].

Assuming there is a line that fits n given points. We denote e_t the vertical distance of point t from the line. Then the following hold:

$$MSE = \frac{1}{n}\sum_{t=1}^{n} e_t^{\,2}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^{\,2}}$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n} |e_t|$$

$$MAPE = \frac{100\%}{n}\sum_{t=1}^{n} \left|\frac{e_t}{y_t}\right|$$, where y_t is the actual value of point t

$$R2\text{-}score = 1 - \frac{Error\ \ from\ Linear\ Regression\ Model}{Simple\ Average\ Model}$$

In Figure 18, there are some simple examples for Classification and Regression.

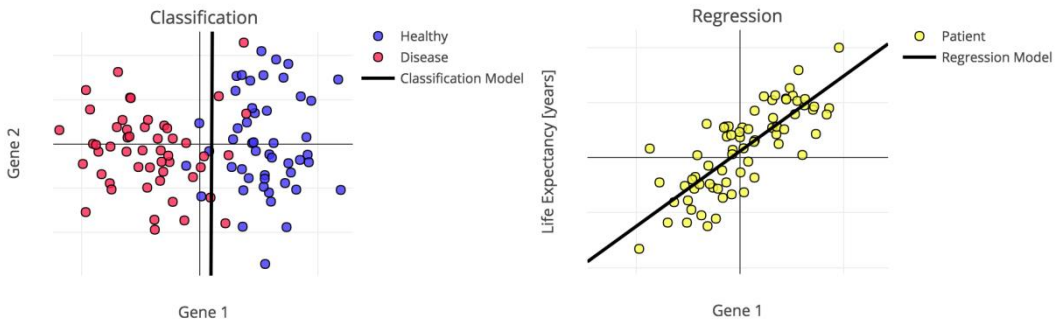


Figure 18: Examples of Classification and Regression models **[160]**.

Unsupervised Learning is the type of ML where the input data are not labeled. Models based on unsupervised learning are used for uncovering previously unknown patterns in the data as depicted in Figure 19. While there are no right or wrong answers, some models perform better than others. Unsupervised Learning is mainly coping with *pattern recognition*, *clustering* and *data dimensionality reduction* [160].

Pattern recognition is the process of automatically recognizing patterns and regularities in data [164].

Clustering is the process of grouping data samples according to a set of features. Groups are called *clusters* [160].

Dimensionality reduction is the process of minimizing the number of features needed to describe data. During the process, some data are lost, but the benefits are more significant [160].

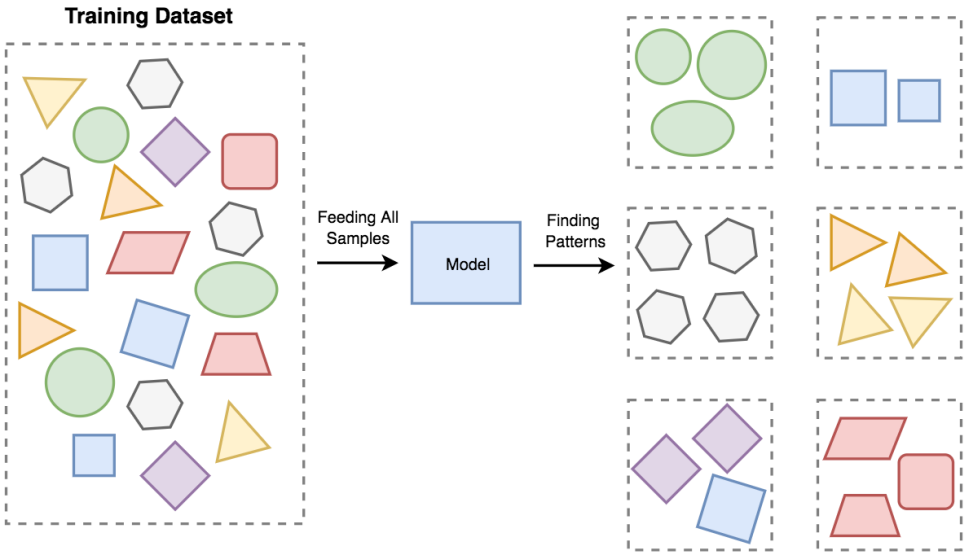


Figure 19: Unsupervised Learning example: pattern recognition and clustering **[160]**.

Reinforcement Learning is the type of ML where a system called *agent* reacts to the information it receives from the *environment* (usually a Markov decision process) by taking an *action*. The information is fed to the agent in form of numerical data, called *state*. After the action, the agent gets a *reward* in form of feedback which is either positive or negative. By

performing trial and error, the agent learns from past decisions and updates its parameters, until the procedure is optimized [160].

3.1.2 Machine Learning Algorithm

A *Machine Learning Algorithm* is a process or a sequence of processes that can adjust itself to learn from data and improve from experience and from exposure to more data, without human intervention. It specifies the way the data are transformed from input to output and how the model learns the appropriate corresponding mapping [165].

The main difference from a traditional algorithm is that traditional algorithms have a fixed set of steps and rules to follow. In contrast, ML algorithms do not have a fixed structure, as they are trained on large datasets, with an intention to learn based on data, contextually.

ML algorithms mixed with datasets enable AI to learn. Algorithms provide a model that AI can use to solve a problem [166]. There are several types of problems and there are countless devices, applications and machines that ground their functionality into ML algorithms.

Some of the most widely used ML algorithms are the following:

- Linear Regression
- Logistic Regression
- Decision Tree
- Random Forest
- Support Vector Machine (SVM)
- Naïve Bayes
- k–Nearest Neighbor (kNN)
- Neural networks
- K–Means
- Apriori
- Principal Component Analysis (PCA)
- AdaBoost
- Gradient boosting algorithms, such as GBM, XGBoost, LightGBM etc.

The majority of them are thoroughly presented in section 3.2 of the thesis.

3.1.3 Data Mining

Data mining (or knowledge discovery in data) is the process of sorting large datasets to identify hidden *patterns* and *trends* and establish *relationships* to solve problems through data

analysis [167]. Data mining combines statistics and artificial intelligence (i.e. neural networks or machine learning) with database management [168].

Algorithms that are mainly used for data mining are classification and regression algorithms along with association rules, which are used to identify relationships among data elements [169]. Applications which are fitting to data mining are those referring to pattern mining, anomaly detection, fraud detection, trend analysis, predictive modeling, descriptive modeling and more [168]. As seen in figure 20, data mining is widely used in business sector.

Figure 20: Data Mining in Business **[170]**.

Companies often use data mining to extract knowledge from raw data. They can learn useful information about their customers in order to plan effective marketing strategies, increase sales and decrease costs (169). Data mining process is split in the following steps, which also can be seen in Figure 21:

- Data *extraction*, *collection*, *transformation* and load in data warehouses.
- *Storage* and *management* into databases on in-house servers or the cloud.
- Data *analysis* from business analysts and information technology professionals with the use of application software.
- *Evaluation* of the importance and innovation for the extracted knowledge.
- *Presentation* of the data analyzed in understandably formats, such as graphs or tables [169] [171] .

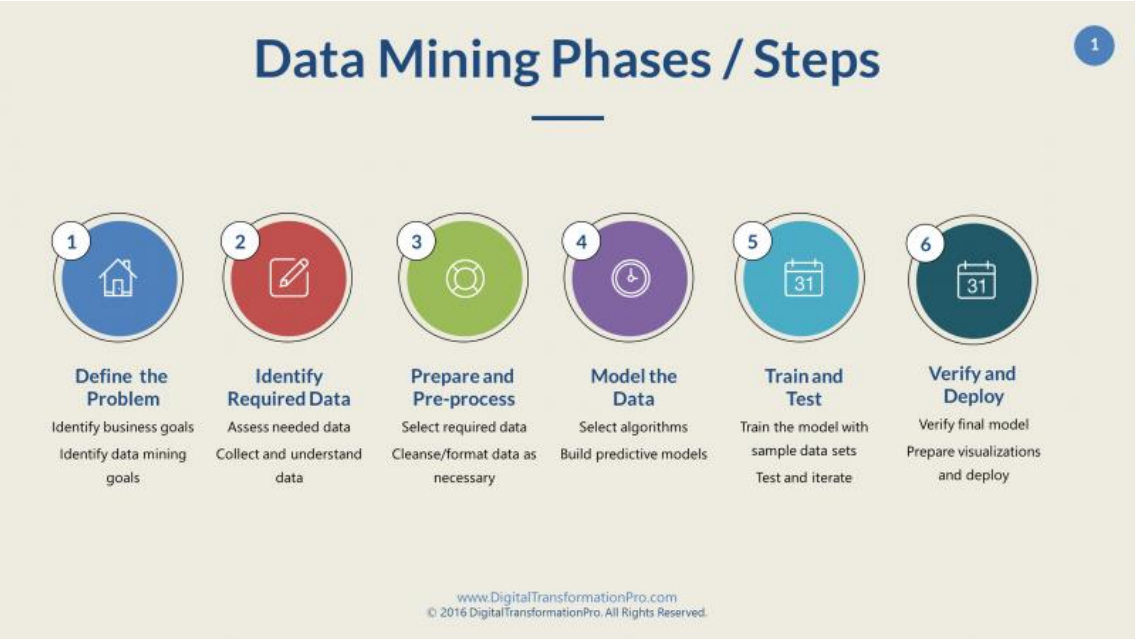


Figure 21: Typical Data Mining procedure **[172]**.

3.1.4 Data Analysis

Data analysis is the process of cleaning, transforming, evaluating and modeling data using statistical tools and logical reasoning to discover useful information and support the business in decision making. There are various data analysis techniques, such as data mining, business intelligence, text analytics and data visualization [173]. Data analysis in statistics is divided into *descriptive statistics, exploratory data analysis (EDA)* and *confirmatory data analysis (CDA)* [174].

Data analysis is incorporated in various aspects of business, but also in everyday life. It is increasingly spread in diverse sectors of society, as seen in Figure 22. As cases become more complex, so does the sophistication of data analysis [175].

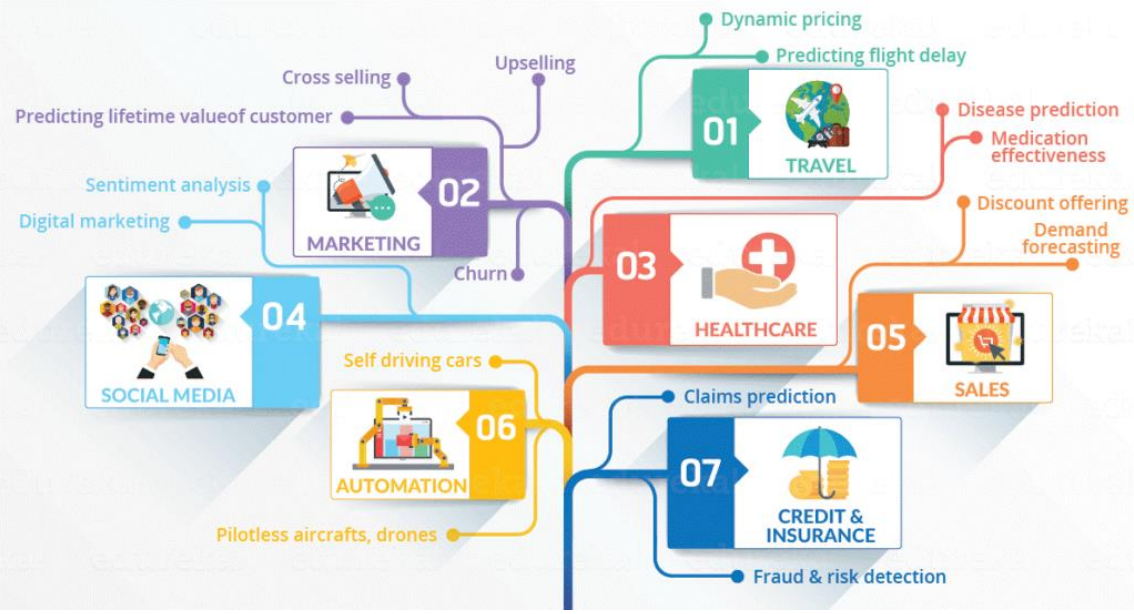


Figure 22: Applications of Data Analysis in society **[175]**.

Data analysis is vital for organizations. Through data analysis, enterprises understand their customers' needs, so they serve them better and increase their revenue. The amount of available data today is too huge to be handled by traditional architectures and infrastructures. Fortunately, there are a number of data management solutions and customer experience management solutions, so enterprises can get effective insights, make decisions on a more solid ground and improve business practices [176].

There is a vast amount of tools, software and platforms suitable for data analysis and visualization of data analyzed; Sisense (Figure 23), Tableau, KNIME, Periscope Data, but also Matlab, Google Analytics, Apache Spark and R programming are only some examples. Plenty of them are also free.

Figure 23: Typical Sisense dashboard **[177]**.

3.1.5 Sports Analytics

Sports analytics is the use of historical data and advanced statistics to measure performance, make decisions but also make predictions regarding to performance and outcomes, in other to gain a competitive sports advantage. SA when deployed by a team or an individual can yield competitive advantages to them against competition [178].

Essentially, sports analysts do not differ from any other data analyst in terms of approach and methodologies. Data acquisition, cleaning, curation and optimization are basic priorities here too. Predictions are daily generated and accuracy is expected.

There are two main categories in SA: the *on–field analytics* and the *off–field analytics*. In on–field approach, analysts gather data from the game and the players (i.e. events, fitness data, spatiotemporal data, biometrics, strategies etc.) to leverage them for improving performance aspects. The off–field approach has been developed recently. It considers parameters associated with business performance of clubs (i.e. finances, trading deals, salaries, stadium attendance, fan support through social media etc.) and has turned out to be as effective as on–field analytics [179].

There has been noticed a boost in the interest for SA during the last 15 years. One thing that contributed to that was the publication of Lewis' book "*Moneyball*" in 2003, but that coincided with the explosion of data available for analysis; innovation of sport science, the

progress in data collection, increased computing power, reduced storage costs and the development of companies such as STATS and Opta, all contributed to that result [180]. As seen in Figure 24, not every club in every sport appreciates the value of SA. There are organizations that spend millions on collecting, managing and analyzing data and others that focus on a different approach.

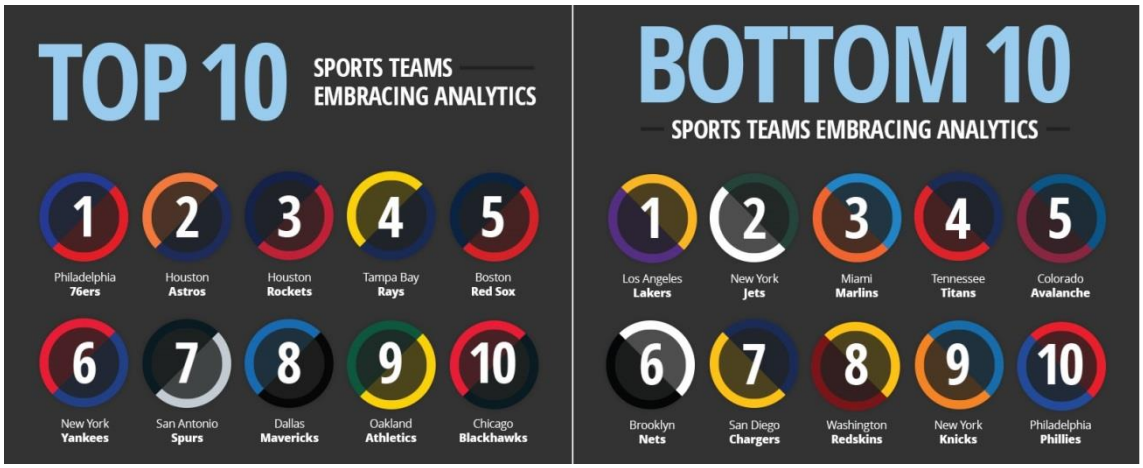


Figure 24: The Top10 and Bottom10 teams in embracing analytics (According to the ESPN Great Analytics Rankings) **[181]**.

To summarize, SA is essential for teams and athletes, as it helps improving their on–field performance, it enhances organization's business performance and contributes in prediction regarding player injuries risk [178].

3.1.6 Performance Prediction in Sports

Performance Prediction is the most common procedure in SA. Sport analysts process data from players and teams under an intended target; the prediction of game results or tournament winner or efficiency for teams and individual players. Forecasts may be destined for short–term or long–term events. For that reason, diverse methods and their corresponding algorithms are deployed.

Organizations' decision makers are very much assisted by predictive models which turn huge sets of raw data into meaningful insights. In that way, performance prediction is a key component of every effective SA process, but is also complex and variegated depending on the club's needs.

3.2 Algorithms and Tools

The most important machine learning algorithms and mainly those used in the experiments are presented in this section. Additionally, the main evaluation techniques and the tools that were used during the whole process are presented, too.

3.2.1 Decision Trees

Decision Tree is a technique used for classification and prediction purposes. As depicted in Figure 25, it is represented by a flowchart–like structure in which, starting from one initial node called *root*, every internal node represents a test on an attribute. Then, a *branch* notes the outcome of this test and finally the produced *leaves* represent the class labels. A category of Tree models is that where the target variable takes a finite set of values; those are called *Classification Trees* [182].

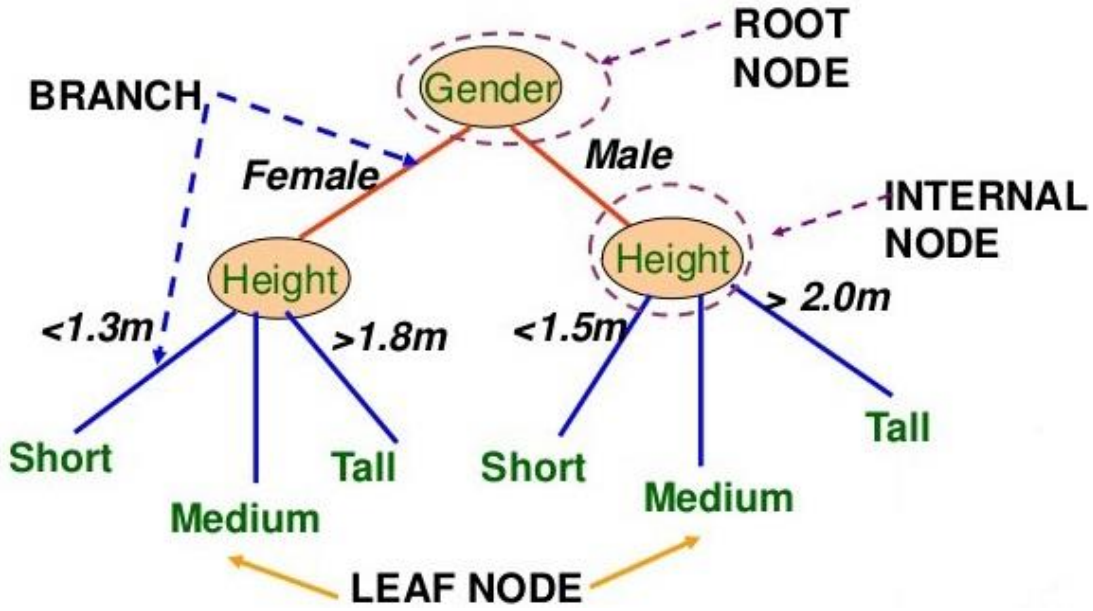


Figure 25: Typical structure of a Decision Tree **[183]**.

Decision Trees are free of ambiguity and robust even in datasets with missing values. Missing values can potentially be a problem, but there are ways to handle that; treating them as null values (if this is meaningful), assign to them the most common value of the attribute, imputation (i.e. trying to predict them using the existing values), surrogate splits (i.e. substitute the main splitter when a record of missing data is encountered) are some ways of handling the missing value problem.

The predictive performance is bound to vary as some trees demonstrate higher generalization accuracy than others. Searching the whole space for the optimal tree is

computational infeasible (due to its exponential size), however, a number of efficient algorithms, like *ID3* or *C4.5*, have been developed and they suggest specific steps to growing a tree. Those algorithms use specific formulas to compute the *Information Gain*, the *Gain Ratio* or the *Gini Index* of each attribute respectively. Those metrics are responsible for the evaluation of the splitting criteria and the selection of the optimized ones.

When a Decision Tree is built, it often contains unnecessary structure and many of its branches reflect anomalies such as repetitions and replications, due to noise or outliers and overfitting of the training set. It is generally recommended to simplify the trees before they are deployed by removing sections of the tree that provide little power to classify instances. This process is called *pruning* and can be seen in Figure 26. There are two main methods of pruning a tree: the *pre–pruning*, which is implemented during the early stages of tree building process and the *post–pruning*, where pruning is implemented after the full growth of the tree.

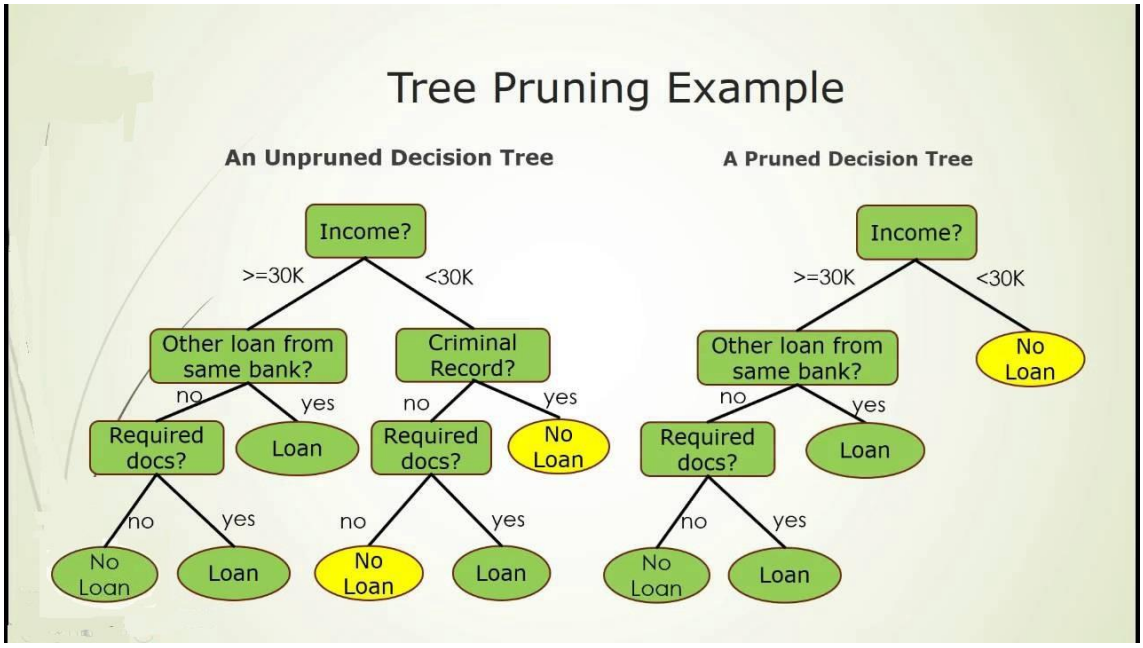


Figure 26: Example of unpruned versus pruned Decision Tree **[184]**.

Decision Trees have some benefits. They are often a preferable method as they are constructed easily and fast compared to other classification techniques. They are computationally inexpensive to build, they are not complex and perform well for small datasets. Handling both numerical and categorical attributes is efficient and despite being simplistic, they achieve accuracy which is comparable to other more expensive and complex models.

Drawbacks on the other hand also exist. The most important is Decision Trees' tendency to overfit the training set. A tree tends to perform poorly if the training set differs from the test set. Algorithms which are responsible for splitting decisions are greedy, so achieving the lowest impossible impurity is nearly impossible.

3.2.2 Random Forests

Random Forests are considered to be the expansion of the Decision Trees. They are an ensemble learning method as they use multiple algorithms to obtain the optimal result and they can be used both for *classification* and *regression* tasks. The method operates by constructing multiple Decision Trees to classify a new object, as seen in Figure 27. Every tree gives a classification result and the forest chooses to classify the object considering the most trees votes in classification and the average in regression. So the uncorrelated decision trees operate as a committee.

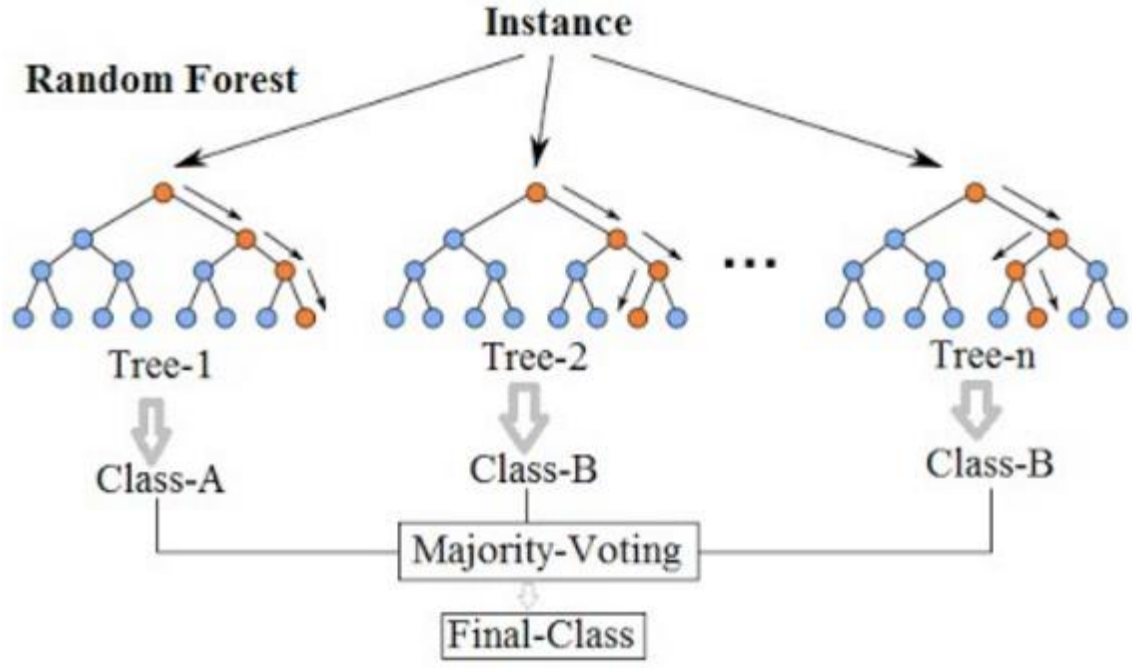


Figure 27: Typical Architecture of a Random Forest model **[185]**.

A Random Forest classifier is used to be preferred compared to a Decision Tree as it is almost certain to produce significantly better results, especially for problems with larger datasets. As many uncorrelated trees exist in the forest, the higher is the probability that accuracy will rise. To ensure that trees are uncorrelated to each other, Random Forest use *bagging* (or bootstrap aggregation); that is allowing each tree to randomly sample from the dataset with replacement, in order to result in different trees. Another method used is the *feature randomness*; each tree in the forest may pick only from a random features' subset. Thus more variation amongst trees exists in the model which results in lower correlation across trees. Finally, trees are both trained on different data thanks to bagging and use different features thanks to feature randomness [186].

Random Forests can be used to rank the importance of variables in a regression or classification problem. Also, despite being a *supervised* learning algorithm, Random Forests can be used in *unsupervised* learning too, by defining a Random Forest dissimilarity measure

between unlabeled data and constructing a random forest predictor that distinguishes the labeled data from generated synthetic data [187]. Predictor variables in Random Forests can unrestrictedly be of any type. Random Forests are not affected in terms of performance from skewed distributions, outliers and missing values [188]. They operate well with large datasets, can handle large numbers of features and are sufficiently robust to noise. On the other hand, the interpretability of Random Forests is limited [189]. Additionally, the complexity and the long training period are also two of the few disadvantages of Random Forests.

3.2.3 Support Vector Machines (SVM)

Support Vector Machines is a supervised ML algorithm which is mainly used for *classification* problems, but also for *regression*. Each data item is depicted as a point (or a *vector*) in n–dimensional space, where n equals the number of features that are used. Essentially, the values of each feature generate the coordinate of each point. The challenge is to find the *hyperplane* that optimizes the differentiation between several classes. That is achieved by maximizing the distance between the hyperplane and the two nearest data points (each one from different class). That distance is called *margin* and the data points which contribute to the discovery of optimal solution are called *support vectors* [190].

In Figure 28 a simple case scenario is depicted, where two classes are linearly separated in a dataset with only two features, thus two dimensions. In this occasion the hyperplane is just a line. But this is not always the case; sometimes, the classification procedure is *non–linear*.

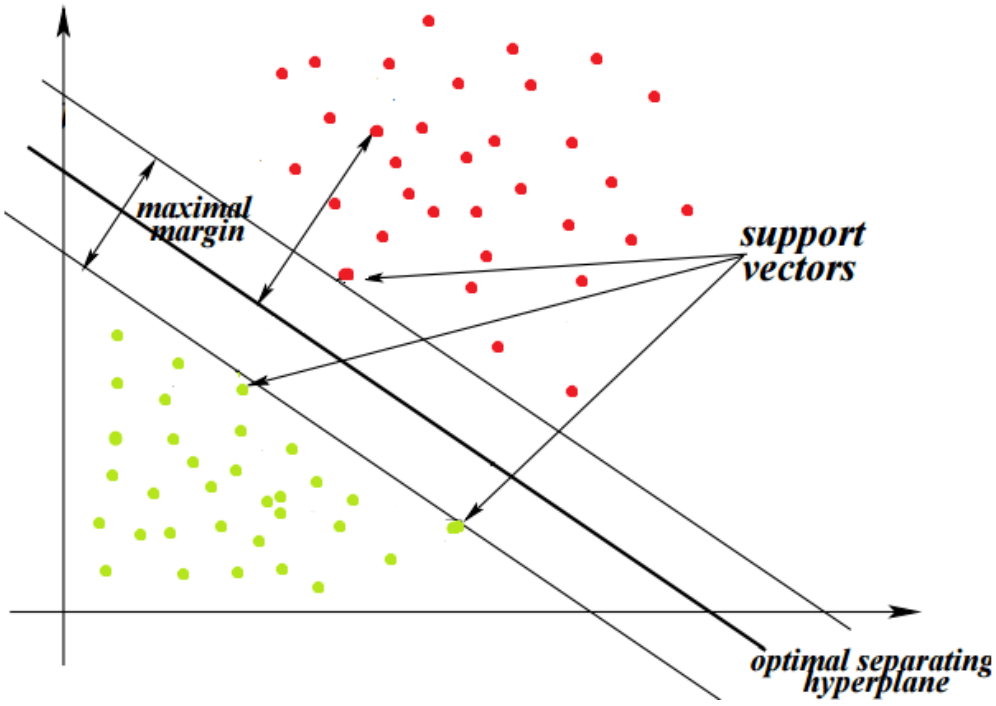


Figure 28: Linear classification with SVM on two-dimensional space for two classes **[191]**.

SVM can easily solve the problem. It incorporates an additional feature in order to make the points be linearly separable, just adding another dimension in the space. This technique is not even required to be performed manually, as SVM have integrated functions, called *kernels*, which make the job done easily. There are various kernels, like *linear*, *polynomial* and *radial basis function* (*rbf*) [190].

In Figure 29, a useful example is given; red squares cannot be linearly separated by green dots in two dimensional space. But with the implementation of a kernel, finally the two classes are segregated in the three dimensional space.

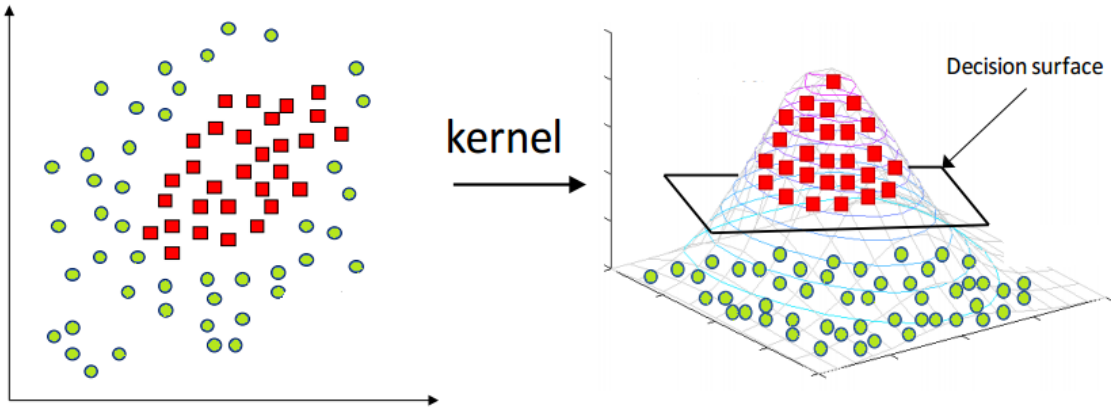


Figure 29: SVM classification with the use of a kernel **[191]**.

Along with the use of kernels that make SVM really efficient, even for nonlinearly separable data, the algorithm has more advantages. Complexity of the training dataset in SVM is characterized by the number of support vectors rather than the dimensionality. So SVM perform well even on high dimensional spaces. Finally, their solution is a global and not a local minimum, so optimization is guaranteed [191].

Nevertheless, SVM have some drawbacks; they are not suitable for large datasets, because the training may be slow and computationally intensive. Also, they are not so effective on handling noise in the dataset, with the major issue to be overlapping classes [191].

3.2.4 Linear Regression

Linear Regression is a simple, basic and widely used type of predictive analysis. The main idea is to find a linear relationship between some independent predicting variables and a dependent variable. In order to do that, the model examines which variables are most significant for predicting the value of the dependent variable and in what way they affect it [192].

The formula that describes a Linear Regression model is the following:

$$y = b + w_1 x_1 + w_2 x_2 + \ldots + w_n x_n$$

where:

- y is the value of the dependent variable.
- $x_1, x_2, \ldots, x_n$ are the values of the independent variables.
- $w1, w2, \ldots, wn$ are the weights corresponding to each independent variable, which are called regression coefficients.
- b is a constant.

In the aforementioned formula, we have many independent variables, thus the model is called *Multiple Linear Regression*. There are many types of Linear Regression. In its simplest form, called simple Linear Regression there is only one independent variable, so the formula is converted as follows:

$$y = b + wx.$$

In Figure 30, a simple Linear Regression example is depicted. Other types of Linear Regression are the *Ordinal Regression*, the *Multinominal Regression*, the *Discriminant Analysis* and the *Logistic Regression*, which –despite its name– is not a regression, but a linear classification algorithm.

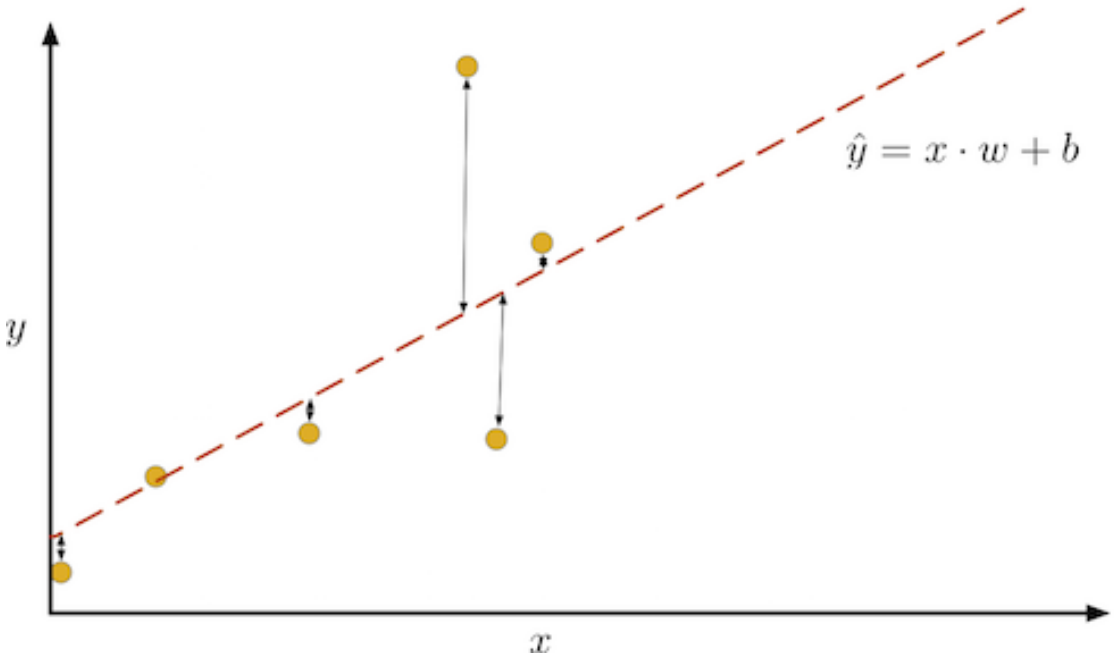


Figure 30: Simple Linear Regression example **[193]**.

At the beginning, there is only a set of data points. Then the algorithm fits a line that best represent the data. But in what way can the best line be drawn? Simple enough, just measuring and aggregating the vertical distance of all the data points from each proposed line and then picking the line that has the minimum aggregated distance. The margin between the observed value of dependent variable y and the expected value of dependent variable $\hat{y}$ is called *error*. There is also an estimation of the unobservable error, which is called *residual*. Errors are

independent and their sum is almost surely non zero, while the sum of the residuals must be equal to 0 and the residuals are not independent. Errors, residuals and some other factors affect the optimization of the line.

Linear Regression is mainly used for determining the strength that predictors have on the dependent variable, for forecasting how much the dependent variable changes with a change in one or more independent variables and for trend forecasting. Adding independent variables to a linear regression model increases the explained variance of the model but by adding too many variables may cause *overfitting*, which reduces model generalizability. Overfitting is addressed with reducing the number of features or with reducing the values of regression coefficients [192].

3.2.5 Neural Networks

An *Artificial Neural Network* (*ANN*) learning algorithm or Neural Network is a computational learning system that uses a set of functions to apprehend and translate a data input of one form into a desired output, usually in another form. It is a ML approach which allows the computer to learn by incorporating new data. The concept was inspired by human neurons and imitates the way a child's brain learns when provided with new information [194]. ANN are mostly used for *classification, clustering, regression, reinforcement learning* and *pattern recognition*.

ANN are usually formed in three layers; *input layer*, *hidden layer* and *output layer*, although the hidden layer may splits into multiple hidden layers. Each layer consists of one or more *nodes*. There is a flow of information from one node to the next. In most types of ANN, the information flows only from the input to the output. However, some types of ANN have more intricate connections, such as feedback paths. The nodes of the input layer do not modify the data; they receive a single value on their input, and pass it through to the nodes of hidden layer, which –like the nodes of the output layer– are modifying the data they receive. Each value from the input layer is duplicated and sent to all of the hidden nodes. Then the values entering a hidden node are multiplied by *weights*, which are predetermined numbers stored in the program. Eventually, the weighted inputs are added to produce a single number which is passed through a nonlinear mathematical function called *activation function* (i.e. sigmoid, tanh, ReLu etc.). Similarly, each of these values is duplicated and applied to the output layer. The active nodes of the final layer combine and modify the data to produce the output values [195]. In Figure 31, a simplistic form of ANN with one hidden layer is depicted.

Neural Networks' main purpose is to adjust the weights according to the error they cause. As weights are slowly adjusted, the Neural Network learns. There are many optimization

functions that are applied on Neural Networks for that purpose, but the most important is *Gradient Descent* [196].

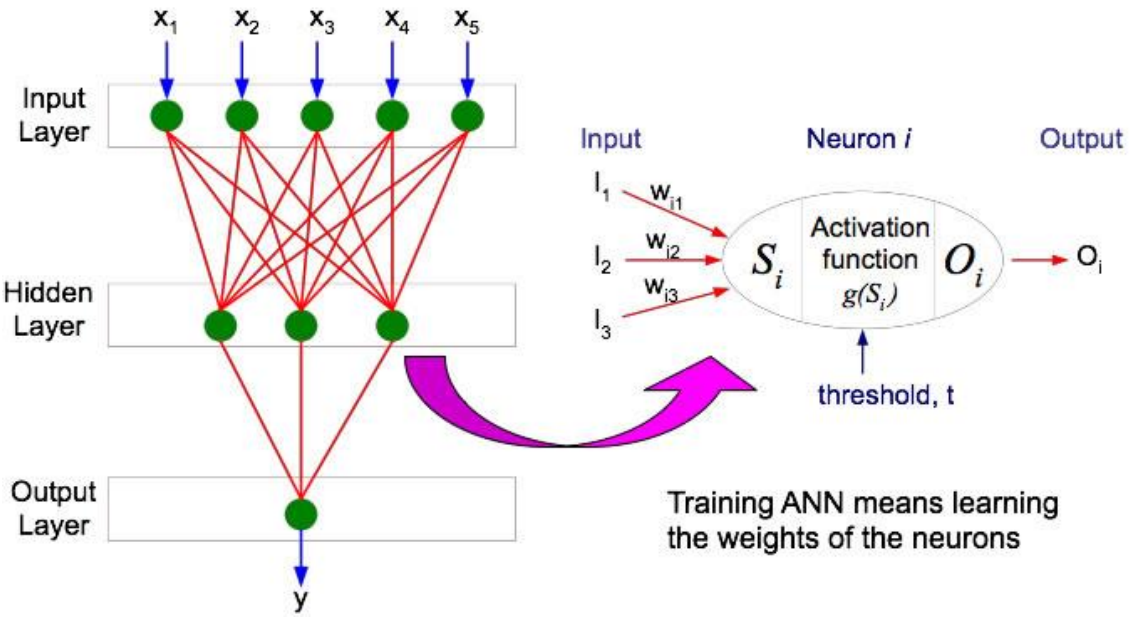


Figure 31: Typical architecture of an ANN with one hidden layer **[197]**.

Earlier versions of Neural Networks, such as perceptron were *shallow* and had one hidden layer at most. On the other hand, *Deep Neural Networks* (*DNN*) are distinguished from common ANN by their depth. In DNN, each layer of nodes trains on a distinct set of features based on the previous layer's output. The features recognized by the nodes are becoming more complex as the network becomes deeper. This property is called *feature hierarchy*. DNN are ideal for *unsupervised learning* (i.e. detection of image similarities, text classification, anomaly detection etc.) as working in a predictive manner, the network assigns probabilities to particular outcomes (i.e. the input image is 97% likely to represent a cat) [196].

Neural Networks can assess many different types of input, like images, videos, files and more so they can be applied to a broad range of fields; image recognition, pattern recognition, self–driving vehicle trajectory prediction, facial recognition, data mining, email spam filtering, medical diagnosis and cancer research are only some of the applications that leverage the beneficial properties of ANN [194].

To conclude, Neural Networks have a number of advantages; they perform better than other learning algorithms when the amount of data is huge, so they can exploit the computational power available nowadays and the breakthroughs in the development of algorithms. Unavoidably, some disadvantages also exist. The main drawback of ANN is their "*black box*" nature; as seen in Figure 32, the user cannot know how or why the Network came up with a certain output as the features the ANN use are not interpreted by humans. Interpretability is vital in some cases, so ANN are not always the ideal solution. Moreover, there is the problem of the development of the Network, especially when control over the algorithm is

needed. *Tensorflow* provides some possibilities, but it is complex and fairly slow in terms of development. Finally, ANN are considered to be computationally expensive compared to simpler solutions [198].

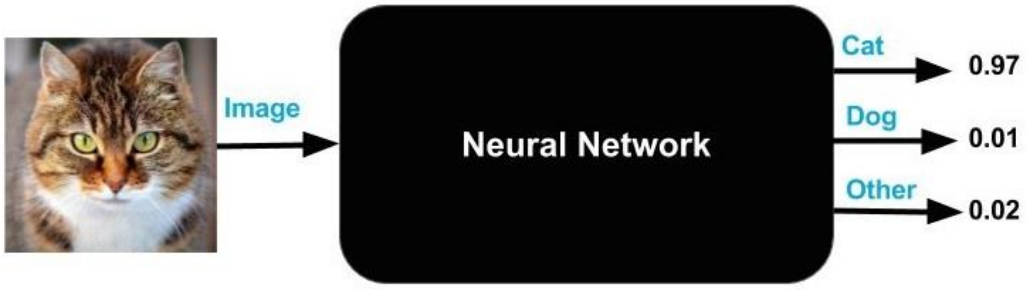


Figure 32: Image recognition. Neural Network has the disadvantage to act as a black box **[199]**.

3.2.6 Jupyter Notebook

Moving on with tools, a very efficient application that was used during the dissertation is presented: Jupyter Notebook is a web–based open source interactive environment used for the creation and share of documents containing code, equations, visualizations and text. It is suitable for data cleansing, data transformation, data visualization, machine learning, big data integration and more.[200].

Jupyter Notebook has some benefits: it allows connection to many kernels and supports over forty programming languages, including Python, R, Julia, Scala and more. Additionally, a notebook can be downloaded, converted in various formats and be shared with others via mail, cloud or GitHub [200].

3.2.7 Weka

Weka (Waikato Environment for Knowledge Analysis) is an open source tool which uses a collection of machine learning algorithms for data mining tasks. It was developed by Computer Science Department of University of Waikato, New Zealand [201].

Apart from the existing algorithms for classification, regression, clustering and association rules, there are various filters for data pre–processing, a number of data visualization tools and also attributes' evaluators for features' selection. Weka supports a number of file formats (e.g. arff, csv etc.) and can also load data from URL's or from SQL databases [201].

Optimizing classifiers' parameters in Weka is easy. Besides, as seen in Figure 33, simplicity is the largest advantage of Weka, as it provides machine learning and data mining techniques without the obligatory knowledge of programming languages.

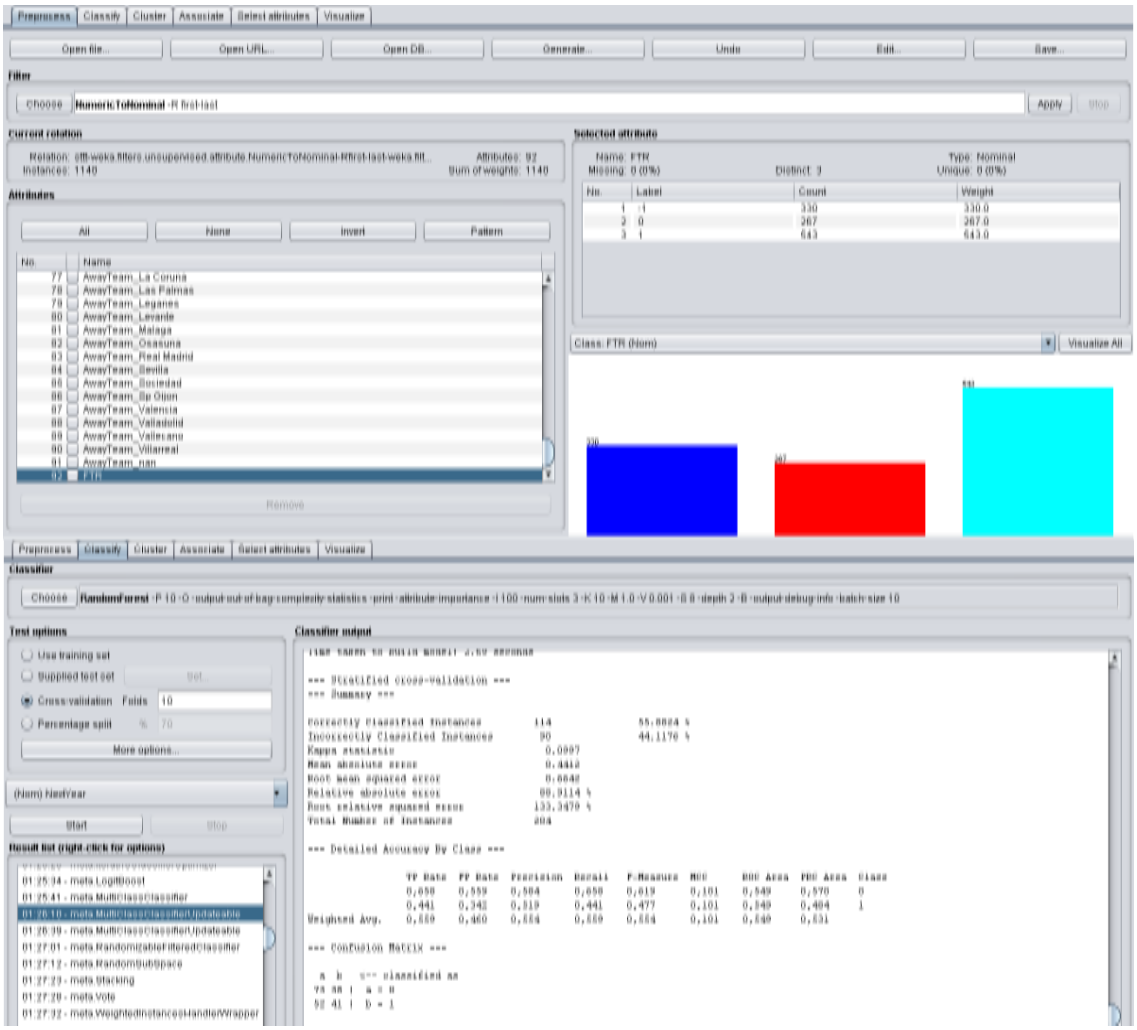


Figure 33: Weka Interfaces: Above, data preprocessing. Below, data classification.

Chapter 4

4.1 Problem Definition

Long–term performance prediction for teams or for individual players is a field that yet needs exploring. Not only coaches, but also sports agents and bookmakers are interested on how a team or a player is going to perform during a whole season compared to previous years. What is discussed in this section is the context of this problem. The objectives of the research are set.

It would be very beneficial for coaches to be able to foresee if the team they are building during the summer preparation has the potential to improve. It would be very enlightening for general managers to know where the team stands in view of the new season in order to take corrective actions before the official matches begin. It would also allow them to develop a crystal clear long–term vision for their team.

As already discussed, football is a huge market that is not only about clubs, managers and footballers, but also about betting companies and gamblers. Special bets like "Which team will win the championship?", "Which teams are going to be relegated?" or "Which coach will be first fired?" are very popular among bettors. But companies need to compute the betting odds with great accuracy, so they have to be able to predict team performance before the season starts.

Predicting team performance before the start of the season is a challenging task, as there are yet no recorded official games to evaluate teams. Moreover, the unique components of football matches make long–term predictions extremely difficult; only few goals are scored per game. Also, there is no clear changeover between the instantaneous change of possession and transition between offense and defense. Moreover, the players' positions and the tactics followed are not fixed and finally, the game has a continuous flow, which complicates the recording of game events [69].

Research may focus on statistics from the previous season and on historical data. Also, some financial data (i.e. net transfer spending, team salaries) and some data concerning player injuries may be exploited in order to contribute in team evaluation process.

The novelty of this research is that some advanced metrics have been used, such as expected goals and expected points. Also, another significant advantage of this research is that, in contrast to other similar researches, it makes performance prediction *before* the season begins and not after some games have already been played and recorded.

Finally, it was only recently that Liverpool's defender Virgil Van Dijk won the "UEFA Player of the Year" award, ahead of exceptional attackers, like Leonel Messi and Cristiano

Ronaldo. But, generally, this is not the case; usually attackers are graded higher than defenders, even if they are not always more influential in team strategy. So, regarding to player evaluation, this research attempts to locate the player skills that make a distinction between defensive and attacking players to fight the bias that is generally regarded. It also clarifies which features makes good defenders really stand out compared to others.

4.2 Approach Followed

The subject of this section is the flow of the events that took place before being able to get meaningful results from our experiments. The way the data were acquired and the preprocessing that was made in order the data to be functional.

The block diagram which summarizes the process followed is depicted in Figure 34:

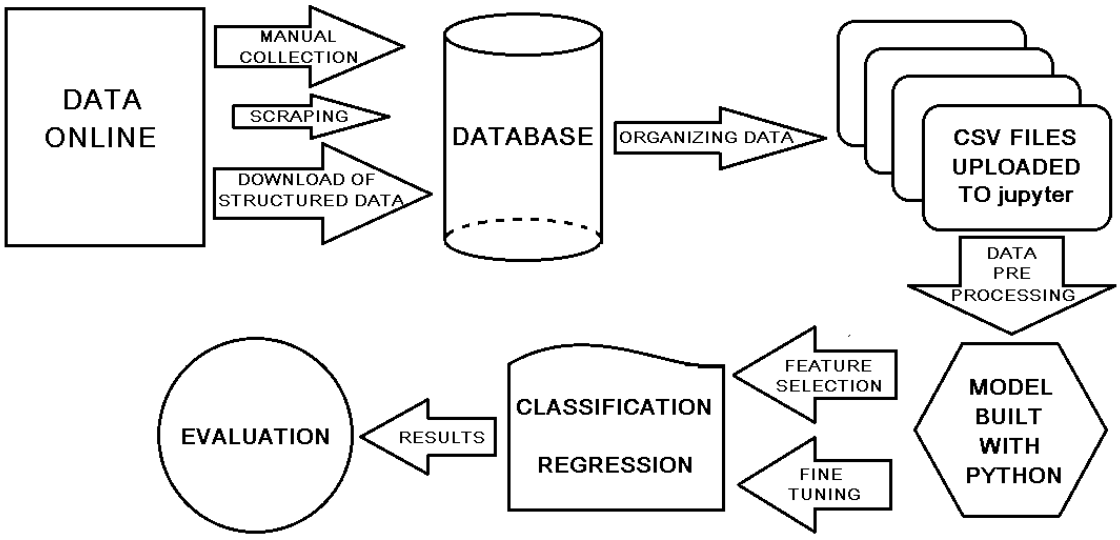


Figure 34: Block diagram of the process followed for the experiments of the thesis.

At first, the appropriate data had to be found. There are a lot of web pages that contain information and statistics regarding to football games and events. The data needed refer to teams and players as well. For the completion of this dissertation the following web pages were used:

- understat.com
- whoscored.com
- transfermarkt.com
- capology.com
- football-data.co.uk
- datahub.io

Some of the data were accessed and collected manually, especially when that was an easy process. However, some of them were scraped from the internet using various scraping tools. Finally, a free database from an expired kaggle competition had been downloaded and used for

the experiments. The database contains data from thousands of players and is extracted from a famous manager simulation game. It demonstrates the ratings of each player to a number of football skills. Players are rated by domain experts.

After the process of data acquisition, there was a large database which needed to be organized. The database was split into different csv files, according to what data were essential for each experiment. Then the csv files were uploaded to jupyter, which was the software that was used for data processing.

Naturally, the data firstly needed to be preprocessed. They were checked for null values, duplicates, noise etc. Python language was used to clean the data and build the models. Then, data transformation and data reduction took place in order to keep only the appropriate features for each classification or regression technique.

Finally, the results that were obtained were evaluated in terms of accuracy, error rates and bias involved. They were also being compared to results of other similar researches in order to estimate the value produced by them.

More information and detailed description about the process followed in each experiment, are given in the next chapter.

Chapter 5

5.1 Experiments

In this section, two experiments are being extensively presented. The experiments are relevant to performance prediction, regarding to team and player performance.

5.1.1 1st Experiment: Team Performance Prediction

The first experiment is divided in two parts: The procedure of the first part can be described as follows: Having a dataset with every team from four important European football national leagues, with more than 40 features for every team for each of the last four years (2015-2018), predict whether a certain team is going to have a better or worse season than the previous year in terms of points. Every previous season is used as training set and the final season (i.e. 2017–2018) is used as test set. It is handled as a binary classification problem (better season / worse season) and the evaluation is conducted by measuring *AccuracyP* as follows:

$$AccuracyP = \frac{Number\ of\ teams\ \ for\ which\ the\ perfromance\ \ prediction\ is\ correct}{Number\ of\ \ total\ teams}.$$

Then, for the second part of the experiment, another method is presented; using almost the same features as in the first part of the experiment, a model is built, in order to simulate every match of the 2018–19 season for the same championships (i.e. 380 matches per championship). Then, the virtual points that are collected by teams are accumulated in order to predict the final league standings. The predicted league table is compared to the actual league table and the evaluation is conducted by calculating the *RMSE* for the championship:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2}$$

where:

- n is the number of teams participating in the championship.
- $\hat{y}_i$ is the predicted points for the i-th team.
- y_i is the actual points for the i-th team.

Also, every model is evaluated for its ability to predict the final outcome of matches played. The evaluation metric is *AccuracyM*, defined as follows:

$$AccuracyM = \frac{Number\ of\ \ games\ with\ correctly\ predicted\ \ outcome}{Number\ of\ \ total\ \ games}.$$

Firstly, it is essential to give a thorough description of how the data were acquired. For the first part of the experiment, the ground on which the dataset was built was scraped from understat.com. It is a web page that contains multiple statistical categories for the teams of the six most important leagues in Europe, but specializes on the expected goals model and its outcomes (i.e. expected points). Then, some other features were added manually. Those were derived by other web pages; whoscored.com was used to fill in some missing statistics, transfermarkt.com and capology.com were used for transfer, managerial and financial related features (i.e. market value, salaries etc.), while some data were manually collected from wikipedia.org and from google.

On this point, it must be noted that certain information extracted by web pages are suffering from subjectivity. Player market values are a fine example of bias. Moreover, some financial data are not always accurate. As an example, despite player salaries data can be found easily, clubs usually do not divulge full details on salaries, so any information concerning finances must be viewed with skepticism.

Predicting the season performance of a team during the summer break is a challenging task. There is a flow of data concerning each team, but actually few of them can have any impact on team performance. For example, it is generally accepted that successful teams usually are more supported than others. However, there is no indication that a team with higher average home attendance has a better chance of winning the title from a least popular team. Consequently, despite being easy to acquire it, an "average home attendance" attribute might not be valuable. Therefore, as Constantinou claims, finding data easily available might be tempting, but this data are not always useful [65]. Therefore, the features that were used for the experiment are the ones that were considered more relative to team performance. Those features could be divided in three main categories:

1. The first category involves past data, which were generated during the last five years. This mainly refers to performance indicators from previous seasons (e.g. team average points).
2. The second category, which contains more attributes than the other two, includes every interesting team statistical feature only from the season that has just ended (i.e. wins, goals, XGoals, shots per game, shots per game against, possession percentage, pass percentage, Pezzali score and more).
3. The third category contains data that are not measurable by team performance. One example is the difference between team transfer expenditure and income or the team percentage of the total transfer fees payed by all teams. This category's attributes are

generated during the summer break, so most of them are independent from the previous season, but they are very likely to have an impact on the new season's performance.

The list of features and a small description for each feature are presented below:

First Category

- **Prev5**: average team position during the last 5 championships.
- **DfP5**: difference between team position in the previous championship and the average team position during the last 5 championships.
- **PosRow**: indicates how many years in a row the team has improved or worsened its position in the final table.
- **PTSPrev5**: average team points during the last 5 championships.
- **DfPTSP5**: difference between team points in the previous championship and the average team points during the last 5 championships.
- **PTSRow**: indicates how many years in a row the team has improved or worsened its points finally gathered.

Second Category

- **Pos**: team position in the previous championship.
- **M**: matches played by team in the previous championship.
- **W**: matches won by team in the previous championship.
- **D**: matches drawn by team in the previous championship.
- **L**: matches lost by team in the previous championship.
- **G**: goals scored by team in the previous championship.
- **GA**: goals scored against team in the previous championship.
- **PTS**: points won by the team in the previous championship.
- **xG**: team expected goals in the previous championship.
- **DGxG**: difference between goals and expected goals for the team in the previous championship.
- **NPxG**: team non–penalties expected goals in the previous championship.
- **xGA**: team expected goals against in the previous championship.
- **DGAxGA**: difference between goals against and expected goals against the team in the previous championship.
- **NPxGA**: non–penalties expected goals against team in the previous championship.
- **NPxGD**: Difference between for and against non–penalties goals for the team in the previous championship.

- **PPDA**: passes allowed per defensive action in the opposition half in the previous championship.
- **OPPDA**: opponent passes allowed per defensive action in the opposition half in the previous championship.
- **DC**: passes completed within an estimated 20 yards of goal (crosses excluded) in the previous championship.
- **ODC**: opponent passes completed within an estimated 20 yards of goal (crosses excluded) in the previous championship.
- **xPTS**: team expected points in the previous championship.
- **DPTSxPTS**: difference between team points and expected points in the previous championship.
- **RatG_xG**: ratio between goals and expected goals in the previous championship.
- **RatGA_xGA**: ratio between goals against and expected goals against in the previous championship.
- **Pezz**: Pezzali score (fully explained in section 2.2.6) in the previous championship
- **RatPTS_xPTS**: ratio between points and expected points in the previous championship.
- **Spg**: shots per game in the previous championship.
- **Poss**: overall percentage of team ball possession in the previous championship.
- **PasPct**: overall percentage of team completed passes in the previous championship.
- **Rat**: previous championship's overall team rating, using the evaluation made by whoscored.com.
- **SCpg**: shots against per game in the previous championship.
- **S6YBpg**: shots made from the 6 yard box per game in the previous championship.
- **SPApg**: shots made from the penalty area per game in the previous championship.
- **Keypg**: key passes given per game in the previous championship.
- **Asspg**: assists given per game in the previous championship.
- **POT**: overall team ball possession in the opposition's third in the previous championship.

Third Category

- **Net**: difference between transfer expenditure and income.
- **NetPct**: team percentage of the total transfer fees payed.
- **ChMan**: (binary) indicates whether a team has changed manager during the summer break or not.

- **Prom**: (binary) indicated whether a team is newly promoted or not.
- **Eur**: indicates whether a team is participating in a European cup or not. If yes, it differentiates between Champions' League and Europa League teams.
- **EuGam**: shows how many games the team has played for the European cup it participates to.
- **ValPct**: the team percentage of market value for the upcoming season.
- **SalPct**: the team percentage of salaries payed to players for the upcoming season.

Finally, in the dataset, there is the "**NextYear**" column, which is the target attribute. It is binary and corresponds to whether the team is going to have a better or worse season than the previous one in terms of points won.

During the preprocessing phase, the data were checked for null values and for duplicates (there were not any). Then, some attributes from the original datasets were removed as they were considered irrelevant with the research or were considered as noisy data that would not add any value to the outcome. Those were team statistics, like yellow and red cards, tackles, interceptions, offsides, fouls, aerials won and more. Thus, the attributes that were described above were the ones to stay in the dataset.

The first problem to handle was that not every team of the previous championship takes part in the new one; there are teams that are relegated and teams that are promoted. How the relegated teams perform in the next season is out of scope of this research, as they will play in a lower division. In addition it is meaningless to have historical data about newly promoted teams, because the data would refer to a different league than the one that is studied. Therefore, the bottom three teams of every league were dropped from the dataset, as they would not participate in the next championship after being relegated. As for the newly promoted teams, some adjustments had to be made; calculating the average team points of the last five seasons, if a team was playing in a lower division during that time, the points of the bottom league team were assigned to it. Calculating the average team position of the last five seasons, if a team was playing at a lower division during that time, it was assigned with a position number just after the bottom league position (i.e. if the league has 20 teams, the team was assigned with position number 21).

As it is later explained in the dissertation, those adjustments caused certain problems. Newly promoted teams do not necessarily have the same strength as teams that have just got relegated. Thus, the way they are described by features and attributes assigned to them might not be representative of their actual status. Furthermore, if there are two or three newly promoted teams, they are all assigned with the same values for the corresponding variables, which is not efficient. Therefore, the validity of this method is questionable.

After that, the discrete datasets were concatenated to each other and a bigger dataframe was created in python, using pandas. Teams were encoded using dummy variables and the final data used is depicted in Figure 35. It consisted of 272 rows and almost 50 attributes (team dummy variables are not included)

In [5]: df

Out[5]:

	Pos	Prev5	DfP5	PosRow	PTSPrev5	DfPTSP5	PTSRow	M	W	D	...	Team_Tottenham	Team_Udinese	Team_Valencia	Team_Vallecano	Team_Veron
0	1	3.0	-2.0	3	75.6	11.4	4	38	26	9	...	0	0	0	0	
1	2	2.4	-0.4	-1	78.2	0.8	-1	38	24	7	...	0	0	0	0	
2	3	3.2	-0.2	2	73.0	2.0	-1	38	22	9	...	0	0	0	0	
3	4	2.6	1.4	1	81.4	-11.4	1	38	20	10	...	0	0	0	0	
4	5	4.8	0.2	1	68.4	-4.4	-2	38	19	7	...	1	0	0	0	
5	6	6.0	0.0	-1	63.6	-1.6	-1	38	18	8	...	0	0	0	0	
6	7	17.0	-10.0	4	34.8	25.2	4	38	18	6	...	0	0	0	0	
7	8	14.8	-6.8	1	37.4	18.6	1	38	16	8	...	0	0	0	0	
8	9	12.0	-3.0	3	46.0	8.0	2	38	15	9	...	0	0	0	0	
9	10	19.0	-9.0	3	29.4	18.6	3	38	13	9	...	0	0	0	0	
10	11	6.6	4.4	-1	61.2	-14.2	-1	38	12	11	...	0	0	0	0	
11	12	16.2	-4.2	1	35.8	11.2	1	38	12	11	...	0	0	0	0	
12	13	13.4	-0.4	1	39.6	4.4	1	38	11	11	...	0	0	0	0	
13	14	21.0	-7.0	2	26.4	14.6	2	38	11	8	...	0	0	0	0	
14	15	12.8	2.2	-1	44.0	-5.0	-1	38	10	9	...	0	0	0	0	
15	16	13.4	2.6	-1	42.6	-4.6	-4	38	7	17	...	0	0	0	0	
16	17	12.2	4.8	-2	45.8	-7.8	-2	38	10	8	...	0	0	0	0	

Figure 35: representation of the data used for the first part of the first experiment.

The data were split into train and test set. The size of the test set was almost 30%. Train and test set were standardized using python's StandardScaler. Then, multiple classifiers were used to classify the test set teams into teams that are going to have a better season next year and in teams that are going to have a worse season next year compared to the previous championship. Python's scikit–learn package provides two valuable methods: Grid search was used for model tuning and Cross Validation with 10 iterations was used for testing the effectiveness of the model used. An additional way to locate the most important features is to draw the feature importance graph, as depicted in Figure 36, using the advantages that are offered by matplotlib, a python plotting library.

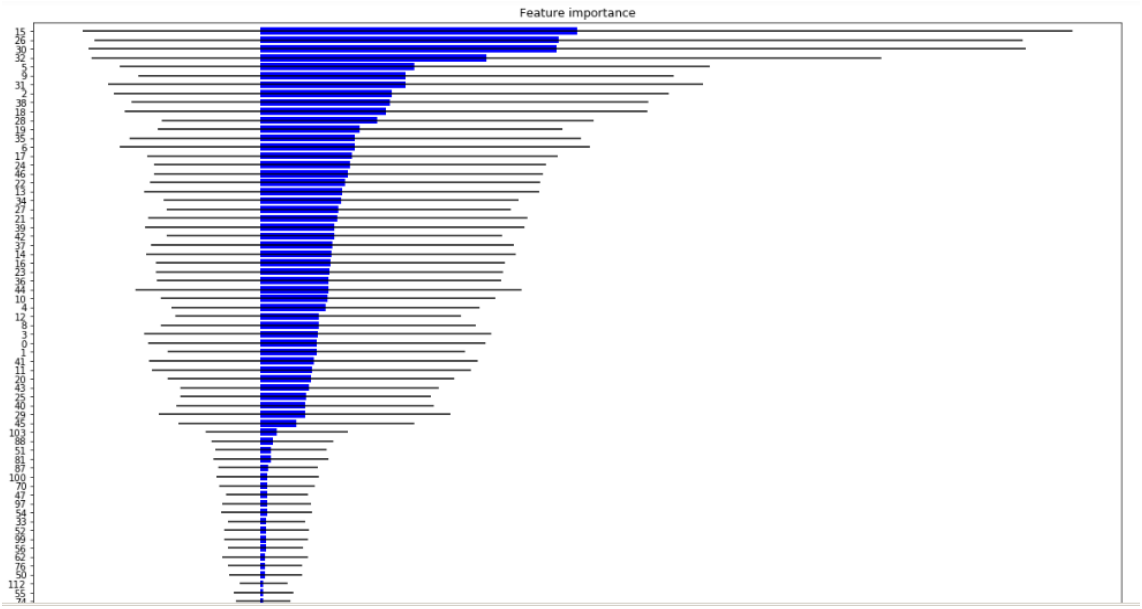


Figure 36: Feature Importance to track the most important features for the training model.

According to the graph, the best attribute is the difference between goals and expected goals for the team in the previous championship, followed by difference between points and expected points for the team in the previous championship. Also, the difference between goals against and expected goals against is high in the list. Naturally, the corresponding ratios are also very important for the model. This observation highlights how important those advanced statistics can be in football. Other important attributes are the difference between team points in the previous championship and the average team points during the last 5 championships, the number of draws, the difference between team position in the previous championship and the average team position during the last 5 championships, the overall percentage of team completed passes in the previous championship and the team percentage of the total transfer fees payed during the summer break.

On the other hand, the attributes that do not contribute as expected are the promotion indicator, the overall team ball possession in the opposition's third in the previous championship and –quite surprisingly– the indicator of managerial change to the club. It looks like that changing manager is not an important factor regarding team performance. However, it must be noted that the attribute used in this research only takes into consideration the fact that a new manager has been appointed to the club or not. But does not factor what caused the managerial change in the club; if the team was underperforming with the old manager, then a new one would probably make the team perform better. If the team was overperforming and the old manager was appointed to a bigger club due to his success, then maybe the team would perform worse under the new manager. Similar researches in the future may deal with this issue.

On the model build on Jupyter, Random Forest was the classifier that achieved the highest accuracy, with more than 70% accuracy and with standard deviation less than 10%. The results are depicted in Figure 37:

```
In [33]: X = df.iloc[:,:-1].values
         y = df['NextYear'].values
         X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=5/17)
         X_train = sc.fit_transform(X_train)
         X_test = sc.transform(X_test)
         classifier = RandomForestClassifier(max_features='sqrt',min_samples_split=13,n_estimators=262,min_samples_leaf=1)
         classifier.fit(X_train,y_train)
         predictions = classifier.predict(X_test)
         print(classification_report(y_test,predictions))
         print(confusion_matrix(y_test,predictions))
         print("Accuracy:",metrics.accuracy_score(y_test, predictions))
         acc = cross_val_score(estimator = classifier, X = X_train, y = y_train, cv = 10)
         print("Mean is" , acc.mean())
         print("Standard deviation is" , acc.std())

                       precision    recall  f1-score   support

                    0       0.63      0.78      0.70        41
                    1       0.69      0.51      0.59        39

            micro avg       0.65      0.65      0.65        80
            macro avg       0.66      0.65      0.64        80
         weighted avg       0.66      0.65      0.64        80

         [[32  9]
          [19 20]]
         Accuracy: 0.65
         Mean is 0.7141025641025641
         Standard deviation is 0.08247448713915890
```

Figure 37: Code for training the model and estimating mean accuracy and standard deviation of the test set, using a Random Forest Classifier with 10–fold cross validation. The classifier's parameters were defined after grid search.

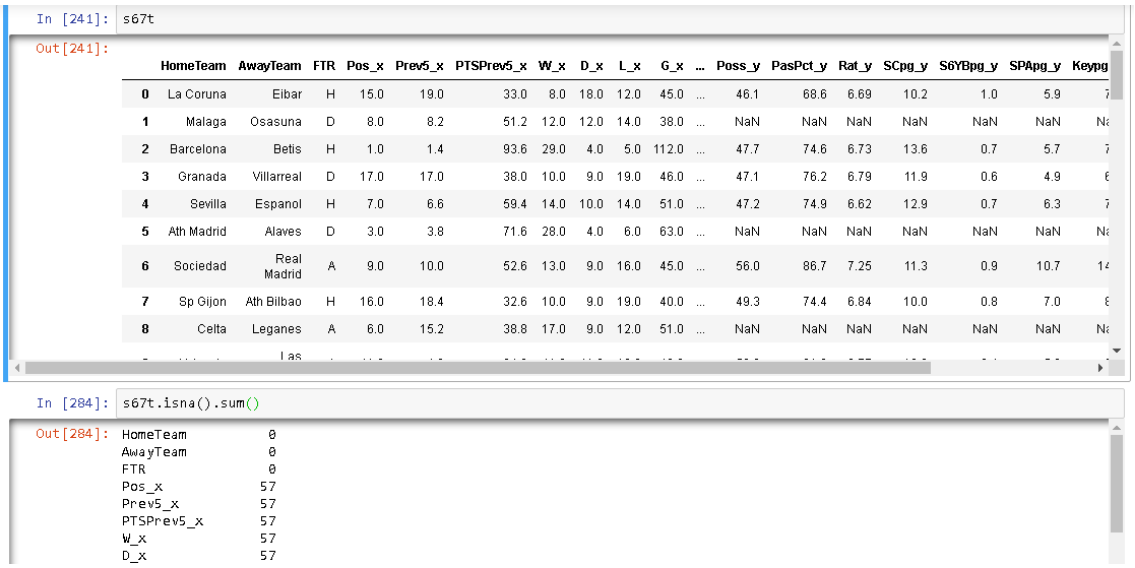
Experimenting with Weka, similar results were achieved. This time the best of the classifiers used was a Decision Tree, J48 in particular, which is an implementation of the C4.5 algorithm, deployed for classification purposes. J48 also had 71% accuracy.

Finally, when the final season was handled as the test set, the results were approximately the same: 69% accuracy and 9% standard deviation with Random Forest.

For the second part of the experiment, more databases were used. They were downloaded from football-data.co.uk and from datahub.io. That data pertained to the results of every game of the four championships presented in the first part of the experiment. The original databases that were downloaded included many attributes, except from the match's results. Most commonly observed attributes were game statistics and betting odds about a number of categories from various betting companies.

Every unnecessary attribute was removed and the datasets was merged with the datasets of the first part of the experiment. This process resulted into a new dataset, which contained every match from a football season with its full time result and with statistical, financial and historical data about the two teams involved in each game.

Naturally, a problem came up: some of the teams participating in championship games lack any data, because they are newly promoted. Thus, as depicted in Figure 38 there were some missing values in the dataset to be handled properly.



```
In [241]: s67t
```

Out[241]:

	HomeTeam	AwayTeam	FTR	Pos_x	Prev5_x	PTSPrev5_x	W_x	D_x	L_x	G_x	...	Poss_y	PasPct_y	Rat_y	SCpg_y	S6YBpg_y	SPApg_y	Keypg
0	La Coruna	Eibar	H	15.0	19.0	33.0	8.0	18.0	12.0	45.0	...	46.1	68.6	6.69	10.2	1.0	5.9	7
1	Malaga	Osasuna	D	8.0	8.2	51.2	12.0	12.0	14.0	38.0	...	NaN	NaN	NaN	NaN	NaN	NaN	Na
2	Barcelona	Betis	H	1.0	1.4	93.6	29.0	4.0	5.0	112.0	...	47.7	74.6	6.73	13.6	0.7	5.7	7
3	Granada	Villarreal	D	17.0	17.0	38.0	10.0	9.0	19.0	46.0	...	47.1	76.2	6.79	11.9	0.6	4.9	8
4	Sevilla	Espanol	H	7.0	6.6	59.4	14.0	10.0	14.0	51.0	...	47.2	74.9	6.62	12.9	0.7	6.3	7
5	Ath Madrid	Alaves	D	3.0	3.8	71.6	28.0	4.0	6.0	63.0	...	NaN	NaN	NaN	NaN	NaN	NaN	Na
6	Sociedad	Real Madrid	A	9.0	10.0	52.6	13.0	9.0	16.0	45.0	...	56.0	86.7	7.25	11.3	0.9	10.7	14
7	Sp Gijon	Ath Bilbao	H	16.0	18.4	32.6	10.0	9.0	19.0	40.0	...	49.3	74.4	6.84	10.0	0.8	7.0	8
8	Celta	Leganes	A	6.0	15.2	38.8	17.0	9.0	12.0	51.0	...	NaN	NaN	NaN	NaN	NaN	NaN	Na

```
In [284]: s67t.isna().sum()
```

Out[284]:

```
HomeTeam      0
AwayTeam      0
FTR           0
Pos_x        57
Prev5_x      57
PTSPrev5_x   57
W_x          57
D_x          57
```

Figure 38: The problem with missing values in the dataset: Osasuna, Alaves and Leganes were newly promoted in the 2016–2017 La Liga championship. As a result, no data were available for those clubs.

The method chosen to fill the missing values was the following: those teams were considered to be the weakest teams in the league and were assigned with the maximum, the minimum or the mean value of the corresponding attribute, according to the nature of the attribute (i.e. if the attribute was the number of expected goals against, those teams were assigned with the maximum value. If the attribute was the points collected, those teams were assigned with the minimum value. If the attribute was the draws achieved, those teams were assigned with the mean value etc.)

The next step was to combine the attributes of the home and the away team by subtracting the corresponding pairs (e.g. if the home team had 18 wins in the previous season and the away team had 24 wins in the previous season, a new attribute would be created and its value for the aforementioned example would be –6). Finally, home and away teams were encoded using dummy variables and the first three seasons of each national championship were concatenated.

The dataset that was created was split to a training set and a validation set. The last season was kept separately from other seasons, the target (i.e. the full time result) was hidden and was used as a test set. The training/validation set consisted of 1140 rows (i.e. equal to the number of games during a period of three years) and almost 40 attributes (team dummy variables are not included). Some attributes from the first part of the problem were excluded from the second one as the subtraction of the attributes' values was not meaningful. The test set consisted of 380 rows and the same attributes as in the training set. An example of the final datasets is depicted in Figure 39:

In [14]: sttt

Out[14]:

	Pos	Prev5	PTSPrev5	W	D	L	G	GA	PTS	xG	...	AwayTeam_Real Madrid	AwayTeam_Sevilla	AwayTeam_Sociedad	AwayTeam_Sp Gijon
0	4.0	3.4	-8.2	-9.0	1.000000	8.0	-29.0	3.0	-26.0	-23.31	...	0	1	0	0
1	-15.0	-16.0	35.4	16.0	-0.117647	-15.0	38.0	-39.0	43.0	27.49	...	0	0	0	0
2	-4.0	0.8	-2.6	3.0	3.000000	-6.0	14.0	-13.0	12.0	10.01	...	0	0	0	0
3	5.0	6.0	-14.8	-4.0	1.000000	3.0	-9.0	9.0	-11.0	4.38	...	0	0	1	0
4	7.0	11.0	-22.8	-7.0	-7.000000	14.0	-24.0	36.0	-28.0	-7.27	...	0	0	0	0
5	6.0	6.6	-39.4	-15.0	6.000000	9.0	-68.0	20.0	-39.0	-57.44	...	0	0	0	0
6	12.0	9.8	-20.4	-9.0	-2.882353	11.0	-19.0	31.0	-25.0	-27.21	...	0	0	0	0
7	7.0	-5.4	10.2	-4.0	-2.000000	6.0	-13.0	23.0	-14.0	-23.59	...	0	0	0	0
8	16.0	19.0	-62.0	-23.0	7.117647	15.0	-89.0	30.0	-57.0	-66.21	...	1	0	0	0
9	2.0	-3.2	7.8	-2.0	6.000000	-4.0	-5.0	9.0	0.0	7.83	...	0	0	0	0
10	-4.0	-0.8	5.4	3.0	2.000000	-5.0	1.0	-14.0	11.0	12.79	...	0	0	0	0
11	-8.0	-8.4	46.0	16.0	-4.000000	-12.0	68.0	-27.0	44.0	56.76	...	0	0	0	0
12	-3.0	2.4	-5.2	-2.0	8.000000	-6.0	1.0	-24.0	2.0	11.10	...	0	0	0	0
13	-16.0	-19.0	62.0	23.0	-7.117647	-15.0	89.0	-30.0	57.0	66.21	...	0	0	0	0
14	-6.0	-9.2	20.2	4.0	3.882353	-7.0	15.0	-17.0	11.0	3.93	...	0	0	0	1
15	9.0	13.0	-25.2	-6.0	-2.000000	8.0	-8.0	14.0	-20.0	-15.98	...	0	0	0	0

Figure 39: Final dataset from Spanish La Liga, seasons 2015–2017, used as a training and validation set.

Standardization of the data, parameter tuning and cross validation techniques were used again, exactly as described for the first part of the experiment. Multiple classifiers were deployed to predict the outcome of the games and therefore the championships' final standings.

Team percentage of market value for the upcoming season, expected points and non–penalties expected goals turned out to be the most important features, but not by a big margin from all other attributes.

During this process two problems that are extensively presented in Chapter 5.2 came up; the first problem was that almost every classifier used had the tendency to favor big teams over smaller ones. The other problem was that most of the models built faced difficulties in predicting draws.

Despite the drawbacks, the results achieved could be considered promising. They are comparable to results from similar researches, while the advantage of this research is that the experiments can be concluded at the beginning of the season, with no official games played and recorded. The best accuracy for the outcome of the games was 57% for the English Premier Division and the smallest RMSE for team points was 9, achieved for Spanish La Liga. French Ligue 1 produced the worst results, both in terms of accuracy and RMSE. It was, actually, a very strange season, as teams that usually topped the table had a bad or even an awful season. AS Monaco FC finished the championship in 16th place with only 36 points, while the previous year had won the second place with 80 points and two years earlier had won the champion scoring 95 points. On the contrary, there were teams that performed surprisingly well (i.e. Lille, Reims, Nimes). Thus, prediction about Ligue 1 was not easy. The results from each league are presented in the following Tables 1 to 4. The best result for each league is noted with green color and the worst result with red.

Table 1: Results from English Premier League

CLASSIFIER	AccuracyM	RMSE
Naïve Bayes	55	17
Decision Tree	45	12.9
Random Forest	56	14.3
KNN	48	15.3
SVM (rbf kernel)	54	18.2
SVM (polynomial kernel)	**57**	**11**
XGBoost	52	17.3

Table 2: Results from Spanish La Liga

CLASSIFIER	AccuracyM	RMSE
Naïve Bayes	47	23.7
Decision Tree	39	14.9
Random Forest	48	17.7
KNN	46	13.3
SVM (rbf kernel)	**51**	13.8
SVM (polynomial kernel)	47	**9**
XGBoost	45	17.4

Table 3: Results from Italian Serie A

CLASSIFIER	AccuracyM	RMSE
Naïve Bayes	**53**	19.7
Decision Tree	41	**11.3**
Random Forest	40	14.4
KNN	47	14.7
SVM (rbf kernel)	52	19
SVM (polynomial kernel)	50	12.2
XGBoost	42	14.5

Table 4: Results from French Legue 1

CLASSIFIER	AccuracyM	RMSE
Naïve Bayes	42	28.2
Decision Tree	39	17.6
Random Forest	**45**	24.8
KNN	39	20.9
SVM (rbf kernel)	43	22.2
SVM (polynomial kernel)	43	**17.3**
XGBoost	44	21.8

It is obvious that SVM classifier with polynomial kernel gives the best results. Individually, there are classifiers that compete or outnumber its AccuracyM and RMSE, but SVM with polynomial kernel is observed to steadily achieve good results in every league studied, so it is regarded as a benchmark for this research from now on.

Overall, the best result in terms of RMSE was achieved from Spanish La Liga, where the classifier predicted the final league table with surprising –given the few attributes used– accuracy. In Figure 40, the real vs the predicted league tables are shown and compared.

ACTUAL TABLE		PREDICTED TABLE	
1. Barcelona	87	1. Barcelona	83
2. Atletico Madrid	76	2. Atletico Madrid	75
3. Real Madrid	68	3. Real Madrid	64
4. Valencia	61	4. Valencia	58
5. Sevilla	59	5. Sevilla	57
6. Getafe	59	6. Getafe	57
7. Espanyol	53	7. Real Betis	57
8. Athletic Bilbao	53	8. Eibar	57
9. Real Sociedad	50	9. Celta Vigo	57
10. Real Betis	50	10. Villarreal	55
11. Alaves	50	11. Athletic Bilbao	54
12. Eibar	47	12. Real Sociedad	54
13. Leganes	45	13. Leganes	54
14. Villarreal	44	14. Espanyol	51
15. Levante	44	15. Alaves	51
16. Celta Vigo	41	16. Levante	51
17. Valladolid	41	17. Valladolid	51
18. Girona	37	18. Girona	51
19. Huesca	33	19. Vallecano	50
20. Vallecano	32	20. Huesca	49

Figure 40: Spanish La Liga 2018–2019 actual vs predicted table.

In the green color are the teams that won European qualification through Champions League, in the blue color those teams that won European qualification through Europa League and in red color the teams that were relegated after the end of the season. The classifier has done a terrific job in predicting those teams, as its only mistake was that Espanyol instead of Real Betis was the last team to win a European qualification. The classifier correctly predicted the champion, but also the ranking of the first six teams in the league.

In the aforementioned example, SVM with polynomial kernel succeeded not to overestimate the top teams, a problem which was often observed throughout most of the classifiers, but on the other hand overestimated the bottom teams instead. One other problem was its inefficiency in predicting draws, as very few games' outcomes were predicted as "draw".

Despite their divergence and despite how small or big the AccuracyM and the RMSE were in every case, most of the classifiers correctly predicted the league champion. Specifically, the champion was correctly predicted in 64% of the cases. The equivalent accuracy was very good, regarding the teams that won European qualification and mainly those amongst them that qualified for Champions' League, as shown in Table 5 below. Results for the relegated teams were also acceptable. Europa League teams were the exception, as the prediction accuracy was

poor. This happened because those teams may have been either overestimated, hence predicted for Champions League qualification, or underestimated, so predicted for no European qualification.

Table 5: Accuracy in predicting champion, teams that won European qualification and teams that were relegated

	Premier League	La Liga	Serie A	Ligue 1	Overall
Champion	71% 5/7	71% 5/7	57% 4/7	57% 4/7	64% 18/28
Europe	86% 42/49	76% 37/49	82% 40/49	46% 13/28	75% 132/175
Champions League	79% 22/28	86% 24/28	71% 20/28	57% 12/21	74% 78/105
Europa League	38% 8/21	29% 6/21	29% 6/21	0% 0/7	29% 20/70
Relegation	52% 11/21	48% 10/21	57% 12/21	10% 2/21	42% 35/84

It would also be interesting to search for factors that increase the RMSE and try to eliminate their negative contribution if positive. A new attribute, named "diff" was created. The attribute refers to the difference between team predicted minus actual points. The goal is to measure and analyze its correlation with other attributes. For that cause, a correlation heatmap from a python visualization library, named seaborn was created for every league. A part of the heatmap for Spanish La Liga can be seen in Figure 41:

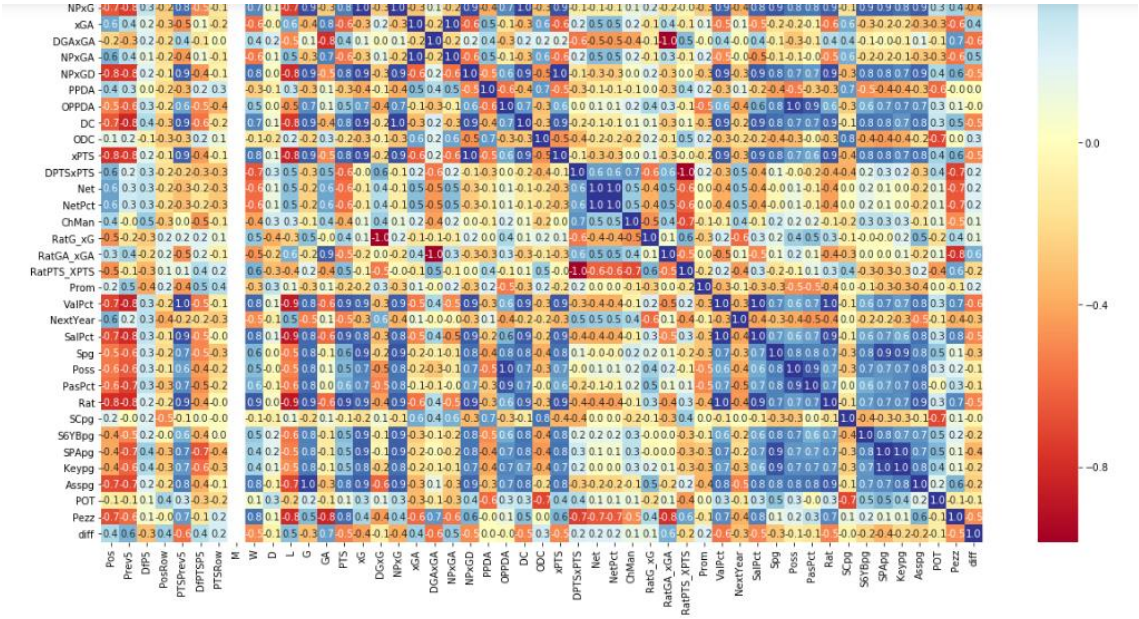


Figure 41: Python's seaborn correlation heatmap for Spanish La Liga.

Observing the correlation between attribute "diff" and all other attributes, next thing that can be done is to draw some jointplots (also from python's seaborn). Some examples are depicted in Figures 42 and 43.

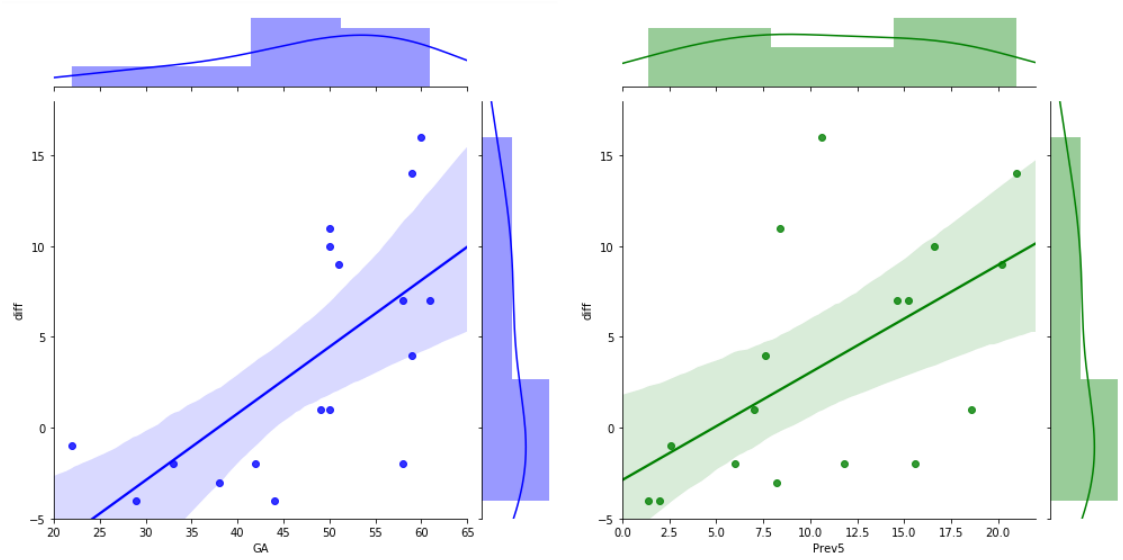


Figure 42: On the left, the goals against vs predicted/actual points difference and on the right overall teams' position during the last five years vs prdeicted/actual points difference for Spanish La Liga.

On the left graph, the importance of goals against with regard to an accurate prediction must be noted. Teams that have conceded less than 50 goals seem to have a small error in respect to points, while teams with more than 50 goals conceded tend to have bigger error. This is confirmed by the graph on the right; it seems that prediction is better for teams that generally finish in the first 8 league positions. Unfortunately, this assumption does not hold for every league.

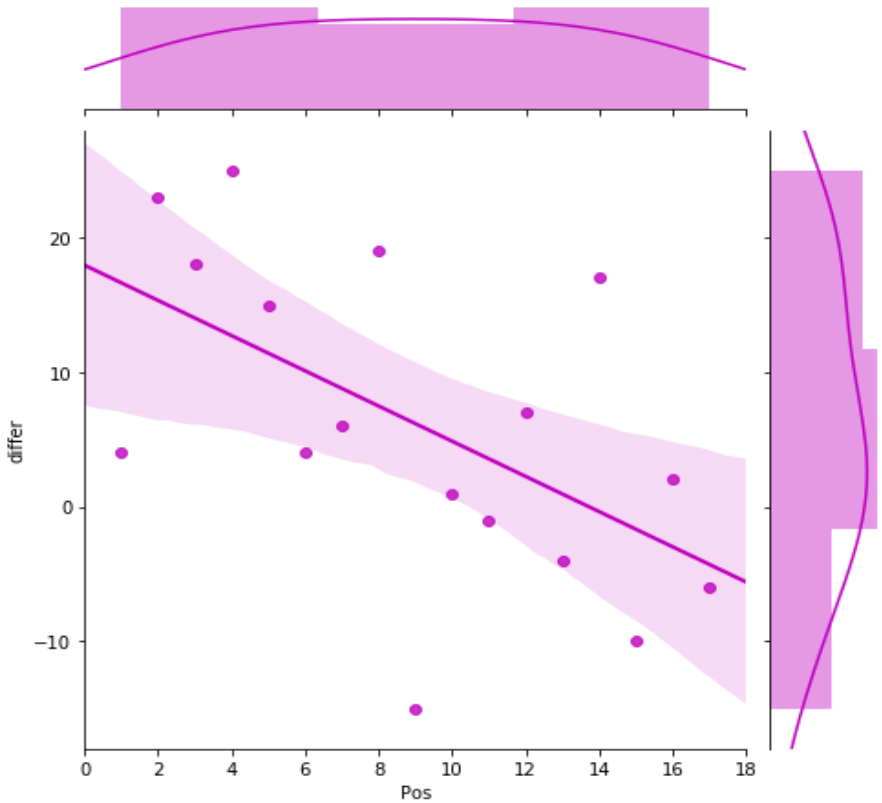


Figure 43: Position of previous championship vs the difference between predicted and actual points for English Premier League.

In the graph above, it is easy to see that the error is bigger for English teams that have finished the previous championship in top places. The reason behind this is the classifier's bias towards big clubs, which leads to them being overrated. Additionally, there has been no significantly strong correlation between "diff" and any other attribute for the Italian Serie A, which is discouraging for that part of the research. As a conclusion, it is difficult to infer any rules about the reasons behind the increase of RMSE.

Another method that could be implemented in order to decrease the error is to find the regression line between Predicted PTS and Actual PTS and then try to adjust the predicted points. Points prediction for every team of the four championships and for any classifier utilized was used to draw the plot and to calculate the regression formula. The regression line is shown in Figure 44 and the simple regression formula turned out to be the following:

$$Y = 25.625 + 0.4875x$$

where:

- Y is the adjusted team predicted points and
- x is the initially predicted points.

Naturally, Y needs to be rounded to be an integer.

Figure 44: Regression line between Predicted PTS and Actual PTS.

This adjustment turned out to be very efficient in cases where the RMSE was high. The largest RMSE of the experiment, as seen in Table 4, was 28.2. Specifically, this result emerges by implementing a Naïve Bayes classifier to 2019 season of French Ligue 1. In this occasion, as depicted in Figure 45, RMSE was vastly improved to 16.1, while the difference between

predicted and actual points was also narrowed, especially for certain teams (i.e. Lyon, Marseille, Reims, Nimes, Toulouse, Monaco) for which the difference was initially large.

		INITIAL PREDICTION						ADJUSTED PREDICTION		
Team	**Act PTS**	**W**	**D**	**L**	**Pred PTS**	**y-Y**	**(y-Y)^2**	**Pred PTS**	**y-Y**	**(y-Y)^2**
Paris SG	91	37	1	0	112	21	441	80	-11	121
Lille	75	12	9	17	45	-30	900	48	-27	729
Lyon	72	32	3	3	99	27	729	74	2	4
St Etienne	66	18	10	10	64	-2	4	57	-9	81
Marseille	61	31	4	3	97	36	1296	73	12	144
Montpellier	59	10	18	10	48	-11	121	49	-10	100
Nice	56	24	7	7	79	23	529	64	8	64
Reims	55	0	5	33	5	-50	2500	28	-27	729
Nimes	53	1	6	31	9	-44	1936	30	-23	529
Rennes	52	19	10	9	67	15	225	58	6	36
Strasbourg	49	6	10	22	28	-21	441	39	-10	100
Nantes	48	14	11	13	53	5	25	51	3	9
Angers	46	14	9	15	51	5	25	50	4	16
Bordeaux	41	18	11	9	65	24	576	57	16	256
Amiens	38	7	2	29	23	-15	225	37	-1	1
Toulouse	38	0	7	31	7	-31	961	29	-9	81
Monaco	36	34	3	1	105	69	4761	77	41	1681
Dijon	34	7	14	17	35	1	1	43	9	81
Caen	33	11	0	27	33	0	0	42	9	81
Guingamp	27	11	8	19	41	14	196	46	19	361
		306	148	306			794,6			260,2
Naïve Bayes							**28,1887**			**16,1307**
Acc: 42%										

Figure 45: Initial predictions and adjusted predictions for French Ligue 1.

Additionally, this adjustment dampens the negative effect of models' inefficiency in predicting draws; league's cumulative predicted points, after the adjustment, are less and closer to actual cumulative league points.

Unfortunately, this method does not improve models that had already produced good results. In the case of Spanish La Liga and the SVM classifier with polynomial kernel which has the lowest RMSE, adjustment of points lead to the increase of RMSE from 9 to 10.2. In another case from England's Premier League, RMSE increased from 11 to 12.6.

Finally, as far as AccuracyM is concerned, another aspect of the experiment is the following: Instead of using the previous three seasons as a train set and the last season as an evaluation set, use the first 10 game days of 2018–2019 season as a training set and the rest 28 game days as a test set. Accuracy was better than those presented in Tables 1–4. In Spanish La Liga, AccuracyM raised to 70%, as seen in Figure 46. Therefore, it is proven that present season's data can boost the accuracy of the model in a lucrative manner.

```
                precision    recall  f1-score   support

            -1       0.72      0.38      0.50        76
             0       0.59      0.78      0.67        79
             1       0.83      0.90      0.86       125

   micro avg       0.72      0.72      0.73       280
   macro avg       0.72      0.69      0.68       280
weighted avg       0.73      0.72      0.71       280

[[ 29  35  12]
 [  6  62  11]
 [  5   8 112]]
Accuracy: 0.725
Mean is 0.7043939393939395
Standard deviation is 0.11155148041316686
```

Figure 46: Accuracy in predicting game results after 10 game days from Spanish La Liga have been analyzed.

5.1.2 2nd Experiment: Player Performance Prediction

This experiment was focused on individual players and more specifically on central defenders. The reasons behind that decision are the following:

In rating systems, goals, assists and other attacking actions are being subsidized more than interceptions, clearances and other defensive actions. Thus, there is a bias toward forwards and attacking midfielders. Goals are actually considered the most important element of football, so sometimes defenders' contribution to a team is underestimated. Consequently, there are almost no researches focused on central defenders.

Additionally, while it is easy enough to rate attacking players, according to the goals they score, the key passes and the assists they give, in the case of central defenders it is not straightforward what makes a good player.

So, the purpose of this research is to examine the characteristics and the statistics from central defenders in comparison with their season rating and decide which of them contribute more to distinguish a central defender as a top class player.

The data used for this experiment were downloaded from kaggle.com. It is a database derived from a famous game (Football Manager 2017) which contains data from almost 150,000 players. The data refer to player attributes, playing positions and some demographic features. The database was narrowed down to 59 players, as only central defenders, playing in English Premier League and having participated in at least 10 league games for the 2016–2017 season were selected.

The next step was to collect statistics for those 59 players for the 2016–2017 season from whoscored.com. Data from multiple statistical categories were selected, but the main focus was on statistics regarding defensive actions. Also, some team statistics were collected; despite

demanding to build a model based on player performance, it must be acknowledged that the team the footballer plays for has an impact both on his statistics but also on his overall rating.

Once this procedure was done, the two datasets were merged, duplicates were removed (some cases of namesakes) and the dataset was ready for use, as depicted in Figure 47:

In [5]: d

Out[5]:

	Rank	Player	Team	Age	IntCaps	IntGoals	U21Caps	U21Goals	Height	Weight	...	TeamTotalAerias	TeamWonAerials	TeamLostAerials	Tea
0	1	Nicolas Otamendi	Manchester City	28	38	2	0	0	183	81	...	29.5	15.4	14.1	
1	2	Virgil van Dijk	Southampton	24	7	0	3	0	193	91	...	35.5	17.1	18.5	
2	3	Aleksandar Kolarov	Manchester City	30	64	8	11	1	187	83	...	29.5	15.4	14.1	
3	4	Shkodran Mustafi	Arsenal	24	12	1	7	0	184	82	...	32.4	16.7	15.7	
4	5	Marcos Rojo	Manchester United	26	49	2	0	0	187	80	...	29.9	16.4	13.6	
5	6	Eric Bailly	Manchester United	22	14	0	0	0	187	77	...	29.9	16.4	13.6	
6	7	Cesar Azpilicueta	Chelsea	26	16	0	22	1	178	78	...	28.8	14.2	14.6	
7	8	Jan Vertonghen	Tottenham	29	82	6	0	0	189	79	...	28.4	15.0	13.3	
8	9	Laurent Koscielny	Arsenal	30	36	1	0	0	186	75	...	32.4	16.7	15.7	
9	10	Phil Jones	Manchester United	24	20	0	9	0	185	72	...	29.9	16.4	13.6	

Figure 47: Top–10 English Premier League central defenders, based on whoscored.com rating system. Additional attributes from Football Manager 2017 and downloaded from kaggle.com.

Player attributes are in range 1 to 20, with 1 considered bad and 20 considered excellent. Statistics for players and teams are calculated per 90 minutes (i.e. a full match time). As it is easily understandable, domain experts' opinion was vastly used in the experiment, as in the game of Football Manager, attributes are given to players by football scouters who collaborate with the game's publishing company. Also, whoscored.com rating is a very complex system that takes into consideration every game event before assigning a player with a rating.

The first approach to the problem was to normalize every numeric value of the dataset, so every attribute's range was transformed to range 0 to 1. Then a multiple regression model was built with every possible feature. Despite the simplicity of this approach, some useful early conclusions were drawn regarding to which features contribute more to central defenders' competency. It seems that for the examined dataset, interceptions are the most important characteristic, followed by team overall rating, as expected. Other important statistics were the total attempted tackles and the total succeeded tackles, the aerials won and the team clearances made. Players' best attributes turned out to be their jumping reach, their versatility, their acceleration and their first touch on the ball. Players' age also turned out to be an important feature, but contrary to the belief that experienced central defenders play better than younger ones, it turned out that –for the model built for the examined dataset– younger central defenders perform better than older central defenders. This is also highlighted by Figure 47; none of the top–10 central defenders is older than 30 years old.

Another approach that was followed was to split the dataset's features into three categories: the first category contained player characteristics and attributes. The second category contained player statistics and the third category contained team statistics. Again, the target was to build three multiple linear regression models (i.e. player attributes based and statistics based), but with fewer independent variables than in the first approach.

A detailed recording of features is presented below:

First Category

- **FirstTouch**: How well a player is able to control a ball when he/she receives it and then set it up for his/her next action.
- **Heading**: How well a player can head the ball and how well he/she can head the ball into his/her intended area.
- **Marking**: How well a player is able to cover an opponent to make him/her a less viable option to pass the ball to.
- **Passing**: How accurately a player can pass the ball.
- **Tackling**: How well a player can win the ball from an opponent without committing a foul.
- **Technique**: How good a player's basic technique is.
- **Aggression**: Determines how likely a player will choose to get involved in a physical situation.
- **Anticipation**: How well a player can predict movements of his/her teammates and his/her opponents.
- **Bravery**: Determines if a player is willing to perform an action that risks pain or even an injury.
- **Composure**: The extent to which a player is not affected by mental pressure when he/she has to make a decision or make an action.
- **Concentration**: How well a player is able to keep his/her focus during a game or training.
- **Vision**: How well a player is able to see available options to him/her when he/she has the ball.
- **Decisions**: How well and how quickly a player can evaluate the available options and then choose which action he/she will perform.
- **Determination**: How well and how good a player will try to succeed in his/her actions during a game and training in mentally exhausting situations.

- **Positioning**: How well a player is able to position himself/herself in defensive situations if his/her opponent has the ball.
- **Strength**: Determines how well a player is able to exert physical force on an opposition player.
- **Jumping**: How high a player's head can reach while jumping.
- **Acceleration**: How quickly a player can reach his/her maximum speed while starting to run.
- **Pace**: How fast a player can run when he/she is at his/her maximum speed.
- **Stamina**: How well a player can retain his/her fitness while exerting during a match or team training.
- **Consistency**: Determines if a player is able to perform on a stable level instead of experiencing unexpected drops of his/her form.
- **Dirtiness**: The level of 'dirtiness' a player has during a game.
- **ImportantMatches**: How well a player can manage the big and high pressured matches during a season.
- **InjuryProness**: Determines how likely a player could suffer an injury.
- **Versatility**: Determines how well a player is able to play out of his favourite position and role.
- **Adaptability**: How well a player can adapt to new situations, for example, moving to a new club in a new country.
- **Ambition**: How much ambition player has to play at the highest level he/she can.
- **Loyalty**: How loyal player is in general.
- **Pressure**: How well player is able to handle pressure on and off the field.
- **Professional**: How professional player's conduct is on and off the field.
- **Sportsmanship**: How sporting player's conduct is on and off the field.
- **Temperament**: How well player keeps his/her temper on and off the field.
- **Controversy**: How controversial player is off the field. This is a negative attribute.
- **Rating**: Player's rating (the dependent variable of the dataset).

<u>Second Category</u>

- **Age**: Player's age.
- **IntCaps**: Player's games with his/her country's national team.
- **IntGoals**: Player's goals with his/her country's national team.
- **U21Caps**: Player's games with his/her country's U21 national team.
- **U21Goals**: Player's goals with his/her country's U21 national team.

- **Height**: Player's height.
- **Weight**: Player's weight.
- **Apps**: Player's appearances during the season.
- **Mins**: Player's total minutes of football played during the season.
- **TotalSuccTackles**: Successful tackles made by player per 90min.
- **DribledPast**: Unsuccessful tackles made by player per 90min.
- **TotalAttemTackles**: Total tackles attempted by player per 90min.
- **Interceptions**: Player's interceptions per 90min.
- **Clear**: Player's clearances per 90min.
- **ShotsBlocked**: Shots blocked by player per 90min.
- **CrossBlocked**: Crosses blocked by player per 90min.
- **PassBlocked**: Passes blocked by player per 90min.
- **Total Aerials**: Total aerial combats the player has participated in per 90min.
- **Won Aerials**: Aerial combats won by player per 90min.
- **Lost Aerials**: Aerial combats lost by player per 90min.
- **Fouled**: Fouls won by player per 90min.
- **Fouls**: Fouls made by player per 90min.
- **Yellow**: Yellow cards collected by player per 90min.
- **Red**: Red cards collected by player per 90min.
- **Total Passes**: Passes made by player per 90min.
- **AccLB**: Accurate long balls made by player per 90min.
- **InAccLB**: Inaccurate long balls made by player per 90min.
- **AccSP**: Accurate short passes made by player per 90min.
- **InAccLB**: Inaccurate short passes made by player per 90min.
- **Total Key**: Total key passes made by player per 90min.
- **Long Key**: Long key passes made by player per 90min.
- **Short Key**: Short key passes made by player per 90min.
- **Total Shots**: Total shots made by player per 90min.
- **OutOfBox Shots**: Out of box shots made by player per 90min.
- **SixYardBox Shots**: Six yard box shots made by player per 90min.
- **PenaltyArea Shots**: Penalty area shots made by player per 90min.
- **Total Goals**: Total goals scored by player per 90min.
- **OutOfBox Golas**: Out of box goals scored by player per 90min.
- **SixYardBox Goals**: Six yard box goals scored by player per 90min.
- **PenaltyArea Goals**: Penalty area goals scored by player per 90min.

- **Unsuccessful Dribbles**: Player's unsuccessful dribble attempts per 90min.
- **Successful Dribbles**: Player's successful dribble attempts per 90min.
- **Total Dribbles**: Player's total dribble attempts per 90min.
- **UnsuccessfulTouches**: Player's unsuccessful attempts to receive the ball per 90min.
- **Dispossesed**: Number of times that player lost the ball per 90min.
- **Rating**: Player's rating (the dependent variable of the dataset).

Third Category

- **TeamTotalSuccTackles**: Successful tackles made by team per 90min.
- **TeamDribbledPast**: Unsuccessful tackles made by team per 90min.
- **TeamTotalAttemTackles**: Total tackles attempted by team per 90min.
- **TeamInterception**: Team interceptions per 90min.
- **TeamFouled**: Fouls won by team per 90min.
- **TeamFouls**: Fouls made by team per 90min.
- **TeamYellow**: Yellow cards collected by team per 90min.
- **TeamRed**: Red cards collected by team per 90min.
- **TeamClearances**: Team clearances per 90min.
- **TeamShotsBlocked**: Shots blocked by team per 90min.
- **TeamCrossesBlocked**: Crosses blocked by team per 90min.
- **TeamPassesBlocked**: Passes blocked by team per 90min.
- **TeamUnsuccessfulTouches**: Team unsuccessful attempts to receive the ball per 90min.
- **TeamDispossessed**: Number of times that team lost the ball per 90min.
- **TeamTotalAerials**: Total aerial combats the team has participated in per 90min.
- **TeamWonAerials**: Aerial combats won by team per 90min.
- **TeamLostAerials**: Aerial combats lost by team per 90min.
- **TeamTotalPasses**: Passes made by team per 90min.
- **TeamAccLB**: Accurate long balls made by team per 90min.
- **TeamInAccLB**: Inaccurate long balls made by team per 90min.
- **TeamAccSP**: Accurate short passes made by team per 90min.
- **TeamInAccSP**: Inaccurate short passes made by team per 90min.
- **TeamRating**: Overall team rating.
- **Rating**: Player's rating (the dependent variable of the dataset).

The method used for the implementation of this part of the experiment was *Backward Elimination*. In this method, there is a primary model with every feature selected. Then the feature with the largest P–value is excluded from the set of features. Naturally, when a feature is excluded, a new model must be built, so the remaining features are assigned with a new P–value. Once again, feature with the largest P–value is excluded; a new model is built and so on. The whole procedure stops when the P–values of the remaining features are below a threshold (usually 0.1 or 0.05). The final model is built with those remaining features, which are the most important.

For the first category of features (i.e. player characteristics and attributes), the final model was built, as seen in Figure 48, with seven features, which seem to be the most influential for a central defender; "FirstTouch", "Passing", "Vision", "Determination", "Strength", "ImportantMatches" and "Professional".

```
In [80]: X_opt = X[:, [0, 8, 11, 19, 21, 23, 30, 37]]
         reg_OLS = sm.OLS(endog = y, exog = X_opt).fit()
         reg_OLS.summary()
```

Out[80]:

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.478
Model:	OLS	**Adj. R-squared:**	0.406
Method:	Least Squares	**F-statistic:**	6.674
Date:	Mon, 04 Nov 2019	**Prob (F-statistic):**	1.29e-05
Time:	04:46:28	**Log-Likelihood:**	24.173
No. Observations:	59	**AIC:**	-32.35
Df Residuals:	51	**BIC:**	-15.73
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.0251	0.409	14.732	0.000	5.204	6.846
x1	0.0686	0.015	4.571	0.000	0.038	0.099
x2	0.0281	0.017	1.664	0.102	-0.006	0.062
x3	-0.0255	0.013	-1.998	0.051	-0.051	0.000
x4	0.0395	0.014	2.843	0.006	0.012	0.067
x5	-0.0354	0.017	-2.109	0.040	-0.069	-0.002
x6	0.0344	0.013	2.693	0.010	0.009	0.060
x7	-0.0274	0.012	-2.245	0.029	-0.052	-0.003

Omnibus:	1.036	**Durbin-Watson:**	0.888
Prob(Omnibus):	0.596	**Jarque-Bera (JB):**	0.852
Skew:	0.291	**Prob(JB):**	0.653
Kurtosis:	2.913	**Cond. No.**	629.

Figure 48: Multiple Linear Regression model with 7 independent variables

The five assumptions of linear regression were also verified for this model; firstly, there is an indication of linearity in the model. Also, the expectation (mean) of residuals is almost zero ($-5.9 \cdot 10^{-15}$) and it appears that there is no (perfect) multicollinearity between the features.

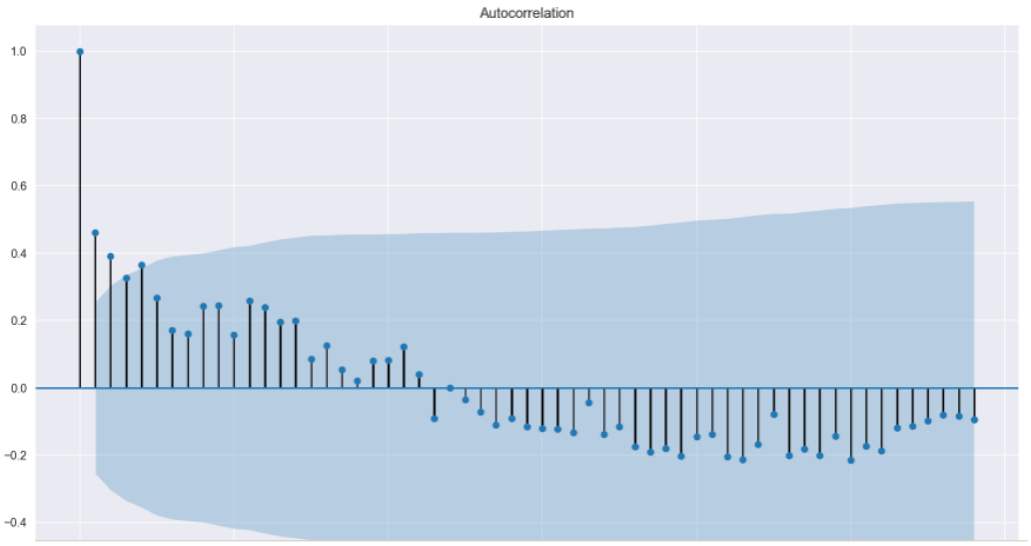
Additionally, as seen in Figure 49, by performing a Breusch–Pagan test it is proven that there is no heteroscedasticity in the model:

```
In [128]: from statsmodels.stats.diagnostic import het_breuschpagan
          from statsmodels.stats.diagnostic import het_white
          import pandas as pd
          import statsmodels.api as sm
          from statsmodels.formula.api import ols

          bp_test = het_breuschpagan(reg_OLS.resid, reg_OLS.model.exog)

In [133]: labels = ['LM Statistic', 'LM-Test p-value', 'F-Statistic', 'F-Test p-value']
          print(dict(zip(labels, bp_test)))

          {'LM Statistic': 10.514948836216204, 'LM-Test p-value': 0.1612206126199397, 'F-Statistic': 1.5800522245668718, 'F-Test p-v
          alue': 0.1626549190037327}

          Since p-value is >0.05, there is no heteroskedasticity
```

Figure 49: Tests prove the absence of heteroscedasticity in the model.

Nevertheless, the final assumption is not met; The Durbin–Watson test, as seen in Figure 50, gives a value much lower than 2, which implies that there is a positive autocorrelation between features. Also, the R–squared and the adjusted R–squared were relatively low, as seen in Figure 48 (under 0.5).

```
In [141]: import statsmodels.tsa.api as smt

          acf = smt.graphics.plot_acf(reg_OLS.resid, lags=58 , alpha=0.05)
          acf.show()
```




```
In [142]: from statsmodels.stats.stattools import durbin_watson
          dw = durbin_watson(reg_OLS.resid)

In [143]: dw

Out[143]: 0.887571097751408
```

Figure 50: Positive autocorrelation with Durbin-Watson test for the first set of features (i.e. player attributes)

Considering that the dependent and the independent variables emerged from two different sources (Football Manager 2017 and whoscored.com), the results could be described as encouraging. Thus, it was expected that data from one source only, would produce better results. Indeed, results were much better for the second model built. The features of the second category (all derived from whoscored.com) were the independent variables, while player rating (also derived from whoscored.com) was, again, the dependent variable.

This time the final model, after backward elimination, consisted of 12 features, with very low P–values, while, as seen in Figure 51, R–squared was 0.867 and adjusted R–squared was 0.833, much better from the first model. The 12 features were: "Age", "IntCaps", "U21Caps", "Mins", "TotalSuccTackles", "Interceptions", "Won Aerials", "Fouls", "InAccSP", "Total Key", "Total Goals", "OutOfBox Goals". Naturally, the fouls made by the player and the inaccurate short passes have a negative coefficient, while the same happens with age, as also observed at the beginning of the experiment.

Out[191]:

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.867
Model:	OLS	**Adj. R-squared:**	0.833
Method:	Least Squares	**F-statistic:**	25.03
Date:	Mon, 11 Nov 2019	**Prob (F-statistic):**	3.39e-16
Time:	02:24:20	**Log-Likelihood:**	64.550
No. Observations:	59	**AIC:**	-103.1
Df Residuals:	46	**BIC:**	-76.09
Df Model:	12		
Covariance Type:	nonrobust		

Omnibus:	1.400	**Durbin-Watson:**	1.912
Prob(Omnibus):	0.496	**Jarque-Bera (JB):**	1.161
Skew:	-0.136	**Prob(JB):**	0.560
Kurtosis:	2.369	**Cond. No.**	1.41e+05

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.9602	0.142	49.122	0.000	6.675	7.245
x1	-0.0327	0.004	-7.776	0.000	-0.041	-0.024
x2	0.0032	0.001	5.650	0.000	0.002	0.004
x3	-0.0113	0.002	-4.635	0.000	-0.016	-0.006
x4	6.782e-05	2.04e-05	3.317	0.002	2.67e-05	0.000
x5	0.2326	0.039	6.017	0.000	0.155	0.310
x6	0.0972	0.028	3.525	0.001	0.042	0.153
x7	0.1261	0.021	6.127	0.000	0.085	0.168
x8	-0.1559	0.042	-3.749	0.000	-0.240	-0.072
x9	-0.0481	0.019	-2.588	0.013	-0.086	-0.011
x10	0.7259	0.124	5.862	0.000	0.477	0.975
x11	0.9718	0.226	4.292	0.000	0.516	1.428
x12	2.1514	0.743	2.896	0.006	0.656	3.647

Figure 51: Final model built with player statistics as independent variables.

Additionally, the five assumptions of linear regression are met; Linearity of the model is obvious, as seen in Figure 52. The expectation (mean) of residuals is almost zero ($1.1 \cdot 10^{-13}$) and it there is no (perfect) multicollinearity between the features, as seen in Figure 53. The Breusch–Pagan test gives a p–value of 0.44, so there is no heteroscedasticity and the Durbin–Watson test gives a value of 1.91, so there is almost no autocorrelation between the features.

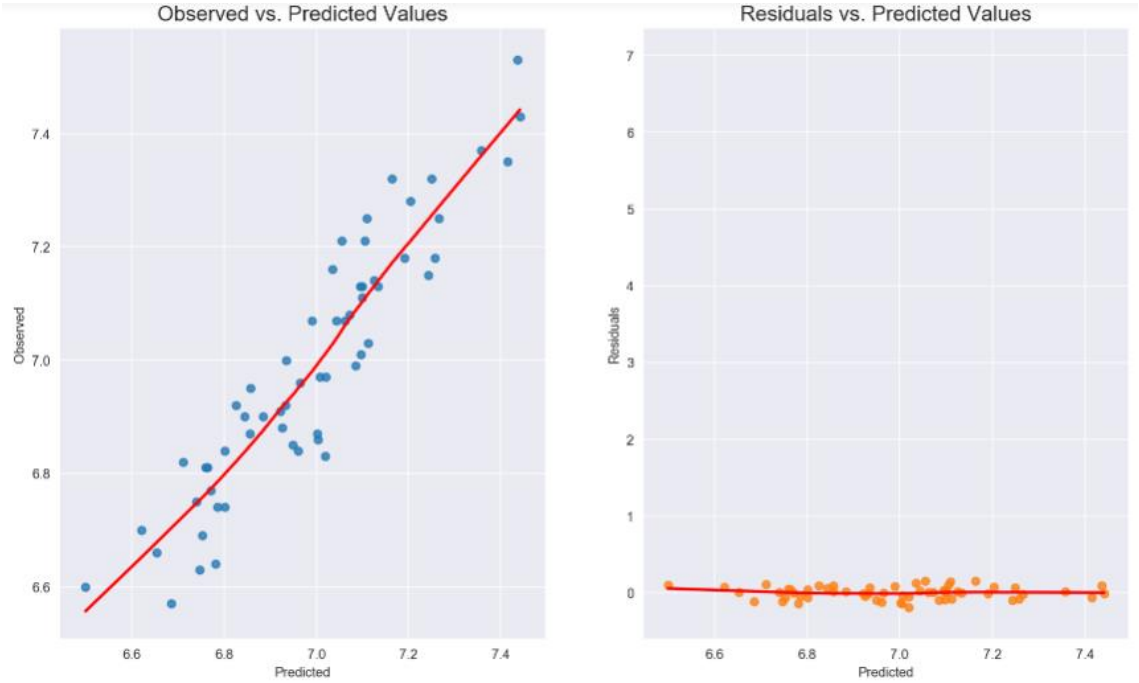


Figure 52: Linearity of the second model.

```
In [194]: #Checking if there is no (perfect) multicollinearity
          from statsmodels.stats.outliers_influence import variance_inflation_factor
          vif = [variance_inflation_factor(X_opt, i) for i in range(X_opt.shape[1])]
          pd.DataFrame({'vif': vif[1:]}, index=X).T

In [195]: vif[1:]

Out[195]: [1.5670672787924178,
           1.676320241272661,
           1.4996186134964986,
           1.270625839510386,
           1.651518009800938,
           1.6944430486712 78,
           1.7161171743474606,
           1.298339729286175,
           2.0420188718104866,
           1.4521319328890463,
           1.3785258443082928,
           1.2660055158329069]
```

Figure 53: Variance inflation factor for the 12 features of the model. All values are close to 1, thus the features are almost uncorrelated to each other.

The third set of features (i.e. team statistics) does not help to build a satisfactory model. There is an indication that the only of those variables worth noting is "TeamRating". It was decided to incorporate "TeamRating" in the second model, in order to exploit that feature. Indeed, by updating the model, adding "TeamRating" as its 13th feature, there has been a slight improvement to the model; R–squared rose to 0.907 and adjusted R–squared rose to 0.88.

In conclusion, the most critical attributes and game actions for predicting the performance of a central defender can be summarized in the following list. It must be highlighted that attacking skills are not absent from the list, following the way modern defenders are expected to play:

- Interceptions
- Clearances
- Aerials Won
- Tackles
- Jumping reach
- Versatility
- Acceleration
- First touch on ball

- Age
- Passing
- Vision
- Determination
- Strength
- Professionalism and ability to perform well in important games
- International Caps
- Minutes Played
- Fouls
- Inaccurate short passes
- Key passes
- Goals
- Team's rating

5.2 Evaluation of Results

In this section a self–assessment is conducted. Problems that came up during the process and the solutions given are discussed. Results of the experiments are evaluated and threats to validity are mentioned, too.

The first problem encountered was the abundance of data regarding team and player statistics, financials, attributes etc. It was practically impossible to use every free online data acquired, so the selection of appropriate data was a challenging task. Luckily, the models produced from the datasets were not computationally intensive, so the approach followed was to include as many attributes relative to the research as possible, in order not to miss out important information. Later, during the future selection phase, some less important attributes were removed. Indeed, certain team statistics (e.g. number of cards, number of fouls, games' attendance) or player statistics (e.g. dribbles, assists and saves) for central defenders are irrelevant of their success and were directly removed from the database. Besides, feature importance graphs at early stages of pre–processing phase indicated some very weak features, which were also removed.

Conversely, the acquisition of certain data considered substantial for research on football analytics was very difficult. Data regarding player injuries and data from wearable devices are mostly defined as private personal details. Thus, there are no such free online data to be used for the experiments.

Another problem was the handling of newly promoted teams, as statistics from the previous season were generated for a lower division, so they could not be used. Concerning those teams, values were automatically assigned to some variables. This action was a necessity, but in certain circumstances the predicted team performance did not meet some teams' real potential. As an example, English club Wolverhampton Wanderers finished the champion in

seventh place with 57 points winning a European qualification, but most models predicted that the team would get relegated.

Another issue was that most models were biased in favor of big clubs. There was a case that a classifier considered that French club Paris Saint Germain was so much better than their opponents that predicted 38 wins in 38 games. An attribute that could be used as a penalization factor in cases of overestimation would balance out the aforementioned bias.

A similar problem was encountered because of the models' difficulty in predicting draws. The solution behind that problem usually lies on the proper usage of cost sensitive classifiers or by tuning the weights of classes.

In both cases, the proposed solutions were tested. Despite the fact that they resolved the issues that were deployed for, they both failed to extract better results than the ones already achieved. Hence, they were not included in the models. A more persistent customization of those methods in the requirements of the datasets would possibly solve the problems faced and give slightly better results. However, such an extensive procedure for a –probably– not remarkable improvement exceeds the scope of this thesis.

As mentioned in the previous section, the prediction for French Ligue 1 was proven to be a difficult task. Furthermore, every championship had its own particularities, so rules extracted from one league do not necessarily apply to others. Therefore, despite building a universal model is feasible, as proven by experimental results, the exploration of the differences between leagues would probably provide better research opportunities.

This issue also reflects to the second experiment. The database used consisted of only 59 defenders based only in England, because it would be inefficient to include players from different leagues. A defender with a low rating playing in English Premier League, which is the most competitive league, would probably got rated higher if he was playing at French Ligue 1, which is a champion with less requirements than English Premier League.

Finally, there was the issue with using domain expert opinion in rating player attributes. Despite generally describing well enough player abilities and potential, scout reports have often been misleading; players have been overestimated or underestimated in the past, while intentional tampering with ratings and attributes should not be ruled out. Additionally, financial–based data (i.e. market value, salaries) usually suffer from inaccuracies and cannot be fully trusted.

Unfortunately, football games are not affected only by team ability and player skills. There are some external factors that cannot be predicted; when a team has to play an important European match during the week, maybe the manager will rest some of his best players for the league game and his team will be lined up with a weaker roster than expected. Luck is also an imponderable factor. Long term injuries of important players are also part of the game. Strange results in games where one or both teams are not in the need of victory (usually at the end of the

season) are often observed. Finally, betting odds inevitably have an influence in games' outcome in one way or another. All those drawbacks, which can be viewed as threats to validity, prove what is mentioned from the beginning of the dissertation; long–term sports prediction is very demanding and may not always provide meaningful results.

Nevertheless, the results of the experiments conducted in the dissertation can be described as good or even impressive in certain occasions, like the case of Spanish league simulation depicted in Figure 40.

Accuracy level for the first part of the first experiment can be described as satisfactory, given the fact that it is a long term prediction with no official games' data and statistics available. An expert (i.e. a coach or a general manager or a bookmaker) could exploit the results of the experiment, along with his own intuition and make certain decisions relevant to his employment.

The main achievement of the dissertation is the second part of the experiment, where the models used predicted the most famous champions' final table with great accuracy. Also, as seen in Figure 46, classifiers are able to predict almost 2 out of 3 games' outcomes when the model is applied in the midst of the season. Consequently, this implementation can be vastly used for betting purposes under certain circumstances. Provably, planning a profitable betting strategy based on experimental results and –apparently– in some human expertise, is not impossible.

Finally, the second experiment succeeds into locating a set of attributes and skills that a central defender must improve in order to be considered a top class player. Of course every player is different and has his own playing style, but it would be very useful for coaches to have a specific targeting when training a player. Long–term player prediction performance could also be a huge contribution to fantasy sports games. The experiment resulted in a variety of features. Some of them –unsurprisingly– were the main defending actions and attributes, but in an interesting manner, some were also found to be attacking actions or attacking attributes.

Chapter 6

6.1 Conclusions

In this research, two fundamental cases of sports analytics were studied; team performance prediction and player performance prediction.

For the first experiment, the goal was to predict how each team of four important European leagues (English Premier League, Spanish La Liga, Italian Serie A and French Ligue 1) will perform during 2018–2019 season. The data available were only historical data (from 2015 onwards) and information about team actions (transfers, managerial changes e.t.c.) during the summer of 2018, just before the beginning of the season. No data from official matches were included in the data used. Two approaches were followed to address this issue.

For the first approach, the target was to classify teams in those that would perform better than last season and those that would perform worse than last season in terms of points collected. Results could be described as satisfactory, but not impressive, as the accuracy of the classifiers deployed reached the level of 71%. In this approach, no distinction between the examined championships was made, so the model used could be described as universal.

Another approach for team performance prediction achieved remarkable results; the idea was to simulate every match of the season and classify their results as home win, draw or away win. At the end of this procedure, each team's points were accumulated and a predicted league final table was extracted. The effectiveness of the model was measured with two metrics: accuracy of the predicted match outcomes and root mean square error of predicted vs actual team points in the league table. Highest accuracy achieved was 57% for English Premier League and lowest RMSE was 9 for Spanish La Liga. Additionally, the champion was correctly predicted in 64% of the times and the teams that won European qualification were correctly predicted in 75% of the times. These results are very satisfactory and comparable to results of similar researches, but have the advantage of being obtained without any official game data available. Also, this time, the four championships were separately studied and differences between them were evident.

Final steps for the first experiment were to try and adjust point prediction by using a simple regression formula between predicted vs actual points. Additionally, applying prediction after using the first ten match days of the season as a training set was suggested as an alternative. In that case, accuracy of predicted match outcomes was impressively raised to 70%.

The second experiment was about defining which attributes and game actions are mainly influencing a central defender's game rating. Dataset consisted of 59 central defenders having

played at least 10 games for English Premier League 2016–2017 season. The method used was multiple linear regression with backward elimination and the evaluation metrics were R–squared and adjusted R–squared.

Findings were quite interesting, as for a quite satisfying 0.907 R–squared and 0.88 adjusted R–squared, thirteen features were proved to be really statistically significant. Classic defensive actions like interceptions and clearances were amongst them, along with player attributes more suitable for defenders, such as jumping reach and strength. The interesting part was that some attacking skills, such as passing, and some attacking game actions, such as number of key passes and number of goals were also found to have an impact on rating central defenders. This fact stresses the change of playing approach from central defenders nowadays and points out that they are more complete footballers that their predecessors had been twenty or thirty years ago.

A final conclusion statement is that sports analytics are already a hot trend and they will get more important for teams, as the data generated by players during training or matches are collected with greater efficiency. There will be a demand for fast and purposeful analysis, due to the amount of data obtained. Therefore, every club will need to hire sport analysts in order to have a complete training staff and not left behind by their competition. There will also be a need of innovative and efficient software and tools, customized on the needs and the culture of every organization, as they can be a great asset for clubs, managers, coaches and players. A whole industry on sports analytics is already being built and grows fast.

6.2 Future Work

The experiments of this dissertation have proven that it is possible to make long–term predictions about team and player performance, so it is reasonable that researchers will work in the same direction in the future, trying to resolve some issues that were out of the scope of this thesis or trying to improve the experimental results of this thesis.

Data from cameras and wearables would be an invaluable asset to any sports analytics research. Future works on sports analytics should focus their attention on gathering and leveraging data from those devices.

Another idea would be the evaluation of newly promoted teams' abilities and the study of their performance in order to comprehend what are the factors that lead them to be successful or not.

Problem with models' bias in favor of bigger clubs and difficulties in predicting a draw were not fully resolved. Cost sensitive classifiers and tuning of the classes' weights did not improve the experimental results. Hence, to scientists who will make a similar research in the

future, it is suggested to further delve deeper into those methods or implement a different approach to solve the aforementioned problems.

What was generally observed in this research and must preoccupy researchers in the future is the major divergence displayed on results extracted from different leagues. Therefore, fundamental differences between leagues should be specified, otherwise models could only be applicable on individual leagues and not become universal. That was also an issue at the second experiment of the dissertation and resulted in the usage of only 59 instances in the dataset, coming only from one league.

Additionally, it would be very useful if future researchers took into consideration some aspects that were not examined in this dissertation; player fatigue because of consecutive matches, starting lineup rotation due to European matches and important player long–term injuries are factors that can affect player or team performance, but at the same time can make a model very complex. Nevertheless, if this complexity is confronted, that information could be a great asset for the research.

Finally, a very stimulating process would be to investigate whether a profitable betting strategy could be built upon the experimental results of this dissertation or other similar works.

References

[1] H. Chadwick, "1860 Beadle's Dime Base Ball Player," [Online]. Available: https://vbba.org/rules-and-customs/1860-beadles-full-text/.

[2] Neyer, Rob, "Sabermetrics," [Online]. Available: https://www.britannica.com/sports/sabermetrics.

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

[4] R. Lederer, "Abstracts From The Abstracts," 14 November 2004. [Online]. Available: http://baseballanalysts.com/archives/2004/11/abstracts_from_20.php.

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

[6] L. Steinber, "CHANGING THE GAME: The Rise of Sports Analytics," 18 August 2015. [Online]. Available: https://www.forbes.com/sites/leighsteinberg/2015/08/18/changing-the-game-the-rise-of-sports-analytics/#8a6c59a4c1fd.

[7] D. Leewood, "Moneyball," 2 October 2011. [Online]. Available: https://en.wikipedia.org/wiki/Moneyball#/media/File:MONEYBALLchart.png.

[8] R. Jazayerli, "The Curious Have Won," 3 November 2016. [Online]. Available: https://www.theringer.com/2016/11/3/16038274/2016-world-series-chicago-cubs-theo-epstein-analytics-war-9f1248c44eb7.

[9] P. Ghosh, "Dallas Cowboys And The Indian: How A Computer Statistician From Uttar Pradesh Helped Create 'America's Team'," 25 October 2013. [Online]. Available: https://www.ibtimes.com/dallas-cowboys-indian-how-computer-statistician-uttar-pradesh-helped-create-americas-team-1441358.

[10] K. Lyons, "Lloyd Lowell Messersmith and the Origins of Notational Analysis," Centre for Notational Analysis, Cardiff Institute of Higher Education, Cardiff, 1994.

[11] J. a. P. R. Krause, "TimeOut Feature: The Early Days of Basketball Analytics," 19 July 2016. [Online]. Available: https://www.nabc.com/nabc_releases/timeout_features/2016/timeout-analytics.

[12] "STATS SportVU® football player tracking," [Online]. Available: https://www.stats.com/sportvu-football/.

[13] "Sportvu Review - Pro Sports Analytics & Stats Platform," [Online]. Available: https://www.infinitepowersolutions.com/sportvu/.

[14] "Sports analytics have been around as long as… sports," 19 April 2019. [Online]. Available: https://thrive.dxc.technology/2019/04/19/sports-analytics-have-been-around-as-long-as-sports/.

[15] A. Clemens. [Online]. Available: http://www.austinclemens.com/shotcharts/#teams.

[16] "NBA Team Three Pointers Attempted per Game," June 2019. [Online]. Available: https://www.teamrankings.com/nba/stat/three-pointers-attempted-per-game.

[17] N. Yau, "Goodbye, Mid-Range Shot," [Online]. Available: https://flowingdata.com/2019/01/15/goodbye-mid-range-shot/.

[18] B. a. S. S. Pande, "Formula 1: From the real world to tracks," 25 October 2011. [Online]. Available: https://economictimes.indiatimes.com/PDAET/articleshow/msid-10482296,curpg-2.cms?from=mdr.

[19] A. Wooden, "The Secret to Formula 1's Success? Big Talent Meets Big Data," [Online]. Available: https://www.intel.co.uk/content/www/uk/en/it-management/cloud-analytic-hub/big-data-powers-f1.html.

[20] A. Woodie, "Go Fast and Win: The Big Data Analytics of F1 Racing," 19 April 2018. [Online]. Available: https://www.datanami.com/2018/04/19/go-fast-and-win-the-big-data-analytics-of-f1-racing/.

[21] C. McDonald, "Fast Cars, Big Data - How Streaming Data Can Help Formula 1," [Online]. Available: https://mapr.com/blog/fast-cars-fast-data-formula1/.

[22] A. Sundar, "Here is everything you want to know about sports analytics," 24 August 2018. [Online]. Available: https://dataconomy.com/2018/08/here-is-everything-you-want-to-know-about-sports-analytics/.

[23] K. a. T. C. Apostolou, "Sports Analytics algorithms for performance prediction," International Hellenic University, School of Science and Technology, Thessaloniki, Greece, 2018.

[24] J. Wilson, "Inverting the Pyramid: The History of Football Tactics," Orion Books, 2009, pp. 138–144, 288–295, 301–303.

[25] A. Rathke, "An examination of expected goals and shot efficiency in soccer," no. 12(2proc), pp. S514 - S529, 2017.

[26] "Understat," [Online]. Available: understat.com.

[27] "FiveThirtyEight," [Online]. Available: https://projects.fivethirtyeight.com/soccer-predictions/.

[28] "WhoScored," [Online]. Available: whoscored.com.

[29] B. J. Coleman, "Identifying the 'Players' in Sports Analytics Research," *INFORMS Journal on Applied Analytics*, vol. 42, no. 2, pp. 109-118, 1 April 2012.

[30] E. -. A. O. H. -. L. R. Morgulev, "Sports analytics and the big-data era," *International Journal of Data Science and Analytics*, vol. 5, no. 4, pp. 213-222, June 2018.

[31] L. a. H. J. G. Passfield, "A Mine of Information: Can Sports Analytics Provide Wisdom From Your Data?," *International Journal of Sports Physiology and Performance*, vol. 12, no. 7, pp. 1-17, December 2016.

[32] N. Silver, The Signal and the Noise: Why So Many Predictions Fail - but Some Don't, New York: The Penguin Press, 2012, pp. 81-86.

[33] M. -. R. Babbar, "A systematic review of sports analytics," Lal Bahadur Shastri Institute of Management, Delhi, Delhi, India, 2019.

[34] M. J. a. C. S. G. Dixon, "Modelling Association Football Scores and Inefficiencies in the Football Betting Market," Lancaster University, UK, 1997.

[35] H. a. S. O. Rue, "Predicting and Retrospective Analysis of Soccer Matchs in a League," Department of Mathematical Science, NTNU, Norway, 2000.

[36] M. -. D. M. -. L. A. -. R. M. Crowder, "Dynamic Modelling and Prediction of English Football League Matches for Betting," *Journal of the Royal Statistical Society*, vol. 51, no. 2, pp. 157-168, 2002.

[37] R. a. R. C. Pollard, "Measuring the effectiveness of playing strategies at soccer," *Journal of the Royal Statistical Society*, vol. 46, no. 4, pp. 541-550, 1997.

[38] J. Goddard, "Regression models for forecasting goals and match results in association football," *International Journal of Forecasting*, vol. 21, no. 2, pp. 331-340, 2005.

[39] C. -. L.-B. J. -. R. E. Lago-Peñas, "Differences in performance indicators between winning and losing teams in the UEFA Champions League," *Journal of Human Kinetics*, vol. 27, no. 1, pp. 135-146, 2011.

[40] K. a. N. A. Harrop, "Performance indicators that predict success in an English professional League One soccer team," *International Journal of Performance Analysis in Sport*, vol. 14, no. 3, pp. 907-920, 2014.

[41] L. -. P. Z. -. L. H. -. G. M.-A. Mao, "Identifying keys to win in the Chinese professional soccer league," vol. 16, no. 3, pp. 935-947, 2016.

[42] M. -. Z. H. -. B. U. Tavakol, "Feature Extraction and Aggregation for Predicting the EURO 2016," in *ECML/PKDD 2016, Machine Learning and Data Mining for Sports Analytics Workshop*, Riva del Garda, Italy, 2016.

[43] N. a. J. Y. Tax, "Predicting The Dutch Football Competition Using Public Data: A Machine Learning Approach," *Transactions of knowledge and data engineering,* vol. 10, no. 10, 2015.

[44] M. A. -. G.-L. M. -. L. C. -. S. J. Gomez, "Effects of game location and final outcome on game-related statistics in each zone of the pitch in professional football," *European Journal of Sport Science,* vol. 12, no. 5, pp. 393-398, 2011.

[45] A. -. F. N. E. -. N. M. Joseph, "Predicting football results using Bayesian nets and other machine learning techniques," *Knowledge-Based Systems,* vol. 19, no. 7, pp. 544-553, 2006.

[46] A. a. T. J. McCabe, "Artificial Intelligence in Sports Prediction," 2008.

[47] J. a. R. A. Hucaljuk, "Predicting football scores using machine learning techniques," in *MIPRO, 2011 Proceedings of the 34th International Convention*, Opatija, Croatia, 2011.

[48] S. a. A. A. Kampakis, "Using Twitter to predict football outcomes," 2014.

[49] H. -. W. H. -. W. A. Lepschy, "How to be Successful in Football: A Systematic Review," *The Open Sports Sciences Journal,* vol. 11, no. 1, pp. 3-23, 2018.

[50] H. -. M. J. -. S. F. a. Z. Y. Broich, "Statistical Analysis for the First Bundesliga in the Current Soccer Season," *Progress in Applied Mathematics,* vol. 7, no. 2, pp. 1-8, 2014.

[51] A. -. M. E. -. B. J. -. B. E. Kapidžić, "Differences in some parameters of situation efficiency between winning and defeated teams at two levels of competition," *Sport SPA,* vol. 7, no. 2, pp. 21-28, 2010.

[52] C. -. F. L. -. F.-G. A. -. Z. A. Ruiz-Ruiz, "Analysis of entries into the penalty area as a performance indicator in soccer," *European Journal of Sport Science,* vol. 13, no. 3, pp. 241-248, 2013.

[53] A. -. L. B. -. P. M. -. J. V. Janković, "Influence of certain tactical attacking patterns on the result achieved by the teams participants of the 2010 FIFA world cup in south africa," *Physical Culture,* vol. 65, no. 1, 2011.

[54] V. -. Y. A. -. P. S. -. S. D. Armatas, "Evaluation of goals scored in top ranking soccer matches: Greek "Superleague" 2006-2007," *Serbian Journal of Sports Sciences,* vol. 3, no. 1, 2009.

[55] V. -. Y. A.-. Z. G. -. S. D. -. P. S. -. F. N. Armatas, "Differences In Offensive Actions Between Top And Last Teams In Greek First Soccer Division. A Retrospective Study 1998-2008," *Journal of Physical Education and Sport,* vol. 23, no. 2, 2009.

[56] F. M. Clemente, "Study of successful soccer teams on FIFA World Cup 2010 through notational analysis," *Pamukkale Journal of Sport Sciences,* vol. 3, no. 3, pp. 90-103, 2012.

[57] P. -. B. A. -. H. M. -. V. T. Luhtanen, "A comparative tournament analysis between the EURO 1996 and 2000 in soccer," *International Journal of Performance Analysis in Sport,* vol. 1, no. 1, 2001.

[58] E. -. G. A. -. G. I. -. K. S. -. A. F. Bekris, "Winners and losers in top level soccer. How do they differ?," *Journal of Physical Education and Sport,* vol. 14, no. 3, pp. 398-405, 2014.

[59] M. -. F. I. Hughes, "Analysis of passing sequences, shots and goals in soccer," *Journal of Sports Sciences,* vol. 23, no. 5, pp. 509-514, 2005.

[60] J. a. D. J. Van Haaren, "Predicting the Final League Tables of Domestic Football Leagues," in *5th international conference on mathematics in sport,* Leuven, Belgium, 2015.

[61] C. Herbinet, "Predicting Football Results Using Machine Learning Techniques," Department of Computing, Imperial College of Science, Technology and Medicine, London, 2018.

[62] L. Knorr-Held, "Dynamic Rating of Sports Teams," Institute for Statistics, LMU, Munchen, Munchen, Germany, 1997.

[63] A. a. F. N. Constantinou, "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,* vol. 9, no. 1, pp. 37-50, 2013.

[64] L. M. a. A. H. Hvattuma, "Using ELO ratings for match result prediction in association football," *International Journal of Forecasting,* vol. 26, pp. 460-470, 2010.

[65] A. a. F. N. Constantinou, "Towards smart-data: Improving predictive accuracy in long-term football team performance," *Knowledge-Based Systems,* vol. 124, pp. 93-104, 2017.

[66] C. -. A. S. -. P. R. -. J. B. -. S. L. Wright, "Factors associated with goals and goal scoring opportunities in professional soccer," *International Journal of Performance Analysis in Sport,* vol. 11, no. 3, pp. 438-449, 2011.

[67] P. -. B. A. -. M. M.-. C. P. -. M. I. Lucey, "“Quality vs Quantity”: Improved Shot Prediction in Soccer using Strategic Features from Spatiotemporal Data," in *MIT Sloan Sports Analytics Conference 2015,* Pittsburgh, USA, 2015.

[68] H. -. v. E. R. -. P. M. Eggels, "Explaining soccer match outcomes with goal scoring opportunities predictive analytics," Eindhoven University of Technology, Eindhoven, The Netherlands, 2016.

[69] J. Oberstone, "Differentiating the Top English Premier League Football Clubs from the Rest of the Pack: Identifying the Keys to Success," *Journal of Quantitative Analysis in Sports,* vol. 5, no. 3, 2009.

[70] S. G. -. S. S. -. Z. A. Hall, "Testing Causality Between Team Performance and Payroll - The Cases of Major League Baseball and English Soccer," *Journal of Sports Economics,* vol. 3, no. 2, pp. 149-168, 2002.

[71] M. a. O. T.-E. Kringstad, "Can sporting success in Norwegian football be predicted from budgeted revenues?".

[72] D. -. F. B. -. J. T. Coates, "Superstar Salaries and Soccer Success: The Impact of Designated Players in Major League Soccer," *Journal of Sports Economics,* vol. 17, no. 7, pp. 716-735, 2014.

[73] J. a. M. M. Gerhards, "Who wins the championship? Market value and team composition as predictors of success in the top European football leagues," *European Societies,* vol. 19, no. 3, pp. 223-242, 2017.

[74] P. -. P. L. -. P. D. -. G. F. -. M. M. Cintia, "The harsh rule of the goals: Data-driven performance indicators for football teams," in *IEEE International Conference on Data Science and Advanced Analytics, Paris, France*, 2015.

[75] L. -. K. H. -. R. P. Gyarmati, "Searching for a Unique Style in Soccer," 2014.

[76] A. -. L. P. -. C. P. -. Y. Y. -. S. S. -. M. I. Bialkowski, "Identifying Team Style in Soccer Using Formations Learned from Spatiotemporal Tracking Data," in *2014 IEEE International Conference on Data Mining Workshop, Shenzhen, China*, 2014.

[77] A. -. U. A. -. D. E. Cakmak, "Computational Modeling of Pass Effectiveness in Soccer," *Advances in Complex Systems,* vol. 21, no. 3-4, 2018.

[78] M. -. G. J. -. C. S. -. E. J. Horton, "Classification of Passes in Football Matches using Spatiotemporal Data," arXiv:1407.5093, 2014.

[79] J. -. K. M. -. G. J. Brooks, "Using machine learning to draw inferences from pass location data in soccer," *Statistical Analysis and Data Mining: The ASA Data Science Journal,* vol. 9, no. 5, pp. 338-349, 2016.

[80] Robertson, Grace, "STATSBOMB," 12 December 2018. [Online]. Available: https://statsbomb.com/2018/12/has-jurgen-klopp-changed-liverpools-approach-for-good/.

[81] T. U. Grund, "Network structure and team performance: The case of English Premier League soccer teams," vol. 34, no. 4, pp. 682-690, 2012.

[82] F. M. -. C. M. S. -. M. F. M. L. -. M. R. S. Clemente, "Using Network Metrics in Soccer: A Macro-Analysis," *Journal of Human Kinetics,* vol. 45, no. 1, p. 123–134, 2015.

[83] P. -. R. S. -. P. L. Cintia, "A network-based approach to evaluate the performance of

football teams," in *Machine Learning and Data Mining for Sports Analytics workshop (MLSA'15), ECML/PKDD conference 2015*, 2015.

[84] A. -. J. G. K. -. M. M. Borrie, "Temporal pattern analysis and its applicability in sport: an explanation and exemplar data," *Journal of Sports Sciences,* vol. 20, no. 10, pp. 845-852, 2002.

[85] J. a. W. T. Gudmundsson, "Towards Automated Football Analysis: Algorithms and Data Structures," in *10th Australasian Conference on Mathematics and Computers in Sport,* 2010.

[86] K. a. M. N. Tamura, "Win-stay lose-shift strategy in formation changes in football," *EPJ Data Science,* vol. 4, no. 9, 2015.

[87] F. Aurenhammer, "Voronoi diagrams—a survey of a fundamental geometric data structure," *ACM Computing Surveys,* vol. 23, no. 3, pp. 345-405, 1991.

[88] E. -. U. A. -. I. M. F. -. G. O. -. C. A. Delibas, "Interactive Exploratory Soccer Data Analytics," *INFOR Information Systems and Operational Research,* vol. 57, no. 2, pp. 141-164, 2019.

[89] S. Kim, "Voronoi Analysis of a Soccer Game," *Nonlinear Analysis: Modelling and Control,* vol. 9, no. 3, p. 233–240, 2004.

[90] T. a. H. J.-i. Taki, "Visualization of dominant region in team games and its application to teamwork analysis," in *Computer Graphics International, 2000*, 2000.

[91] A. -. S. K. Fujimura, "Geometric analysis and quantitative evaluation of sport teamwork," *Systems and Computers in Japan,* vol. 36, no. 6, p. 49–58, 2005.

[92] S. -. M. J. -. T. B. -. A. D. -. L. A. Fonseca1, "Measuring spatial interaction behavior in team sports using superimposed Voronoi diagrams," *International Journal of Performance Analysis in Sport,* vol. 13, no. 1, pp. 179-189, 2013.

[93] C. -. V. S. -. D. B. S. -. V. d. W. R. Poppe, "Multi-Camera Analysis of Soccer Sequences," in *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance, Boston, USA*, Gent, Belgium, 2010.

[94] R. -. P. F. -. Z. X. -. B. B. Theagarajan, "Soccer: Who Has the Ball? Generating Visual Analytics and Player Statistics," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA*, 2018.

[95] Z. -. G. X. -. T. Q. Niu, "Tactic analysis based on real-world ball trajectory in soccer video," *Pattern Recognition,* vol. 45, no. 5, pp. 1937-1947, 2012.

[96] V. -. B. M. -. A. H. -. S. J. Kazemi, "Multi-view Body Part Recognition with," in *British Machine Vision Conference, Bristol, UK*, Stockholm, Sweden, 2013.

[97] D. R. -. D. C. -. C. J. -. R. J. R. -. V. J. E. Seshadri, "Wearable Devices for Sports: New Integrated Technologies Allow Coaches, Physicians, and Trainers to Better Understand the Physical Demands of Athletes in Real time," *IEEE Pulse,* vol. 8, no. 1, pp. 38-43, 2017.

[98] "Happiest Minds," [Online]. Available: https://www.happiestminds.com/Insights/wearable-technology/.

[99] "EPTS ELECTRONIC PERFORMANCE AND TRACKING SYSTEMS," FIFA, [Online]. Available: https://football-technology.fifa.com/en/media-tiles/epts/.

[100] B. -. D. S. -. G. M. -. K. M. -. M. C. -. L. M. -. S. M. Moatamed, "Sport analytics platform for athletic readiness assessment," in *2017 IEEE Healthcare Innovations and Point of Care Technologies, Bethesda, MD, USA*, 2017.

[101] J. -. M. D. -. D. A. G. -. A. M. -. G. R. Fernandez, "From Training to Match Performance: A Predictive and Explanatory Study on Novel Tracking Data," in *IEEE 16th International Conference on Data Mining Workshops*, Barcelona, Spain, 2016.

[102] G. B. -. G. A. -. A. J. R. -. K. C. M. -. C. M. A. Wilkerson, "Utilization of Practice Session Average Inertial Load to Quantify College Football Injury Risk," *Journal of*

Strength and Conditioning Research, vol. 30, no. 9, pp. 2369-2374(6), 2016.

[103] M. a. L. M. Nawrocka, "Biofeedback EEG data integration and visualization analytics for endurance exercise practices: Data integration and visualization analytics of biofeedback EEG," in *2017 IEEE International Conference on Big Data (Big Data), Boston, MA, USA*, 2017.

[104] E. a. S. V. Vermeulen, "Big data in sport analytics: applications and risks," in *International Conference on Industrial Engineering and Operations Management, Pretoria, South Africa*, Pretoria, South Africa, 2018.

[105] "Fantasy Sports & Gaming Association," [Online]. Available: https://thefsga.org/.

[106] Ł. Szczepanski, "Assessing the skill of football players using statistical methods," Salford Business School, University of Salford, Slaford, UK, 2015.

[107] E. -. L. P. -. N. C. Nsolo, "Player Valuation in European Football," in *5th Workshop on Machine Learning and Data Mining for Sports Analytics co-located with 2018 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2018), Dublin, Ireland*, Linkoping, Sweden, 2018.

[108] L. -. M. A. -. N. F. Sîrb, "The Exercise of Prediction Process o fPerformance within Football Sports Management by Using Fuzzy Logic from the Perspective of Value Analysis on Tactical Compartments of Game of the Football Players," *Journal of Knowledge Management, Economics and Information Technology*, vol. 5, no. 2, 2015.

[109] R. -. S. S. -. S. A. -. M. J. Pariath, "Player Performance Prediction in Football Game," in *Second International Conference on Electronics, Communication and Aerospace Technology (ICECA),Coimbatore, Tamil Nadu, India*, 2018.

[110] N. Mackay, "Predicting goal probabilities for possessions in football," Vrije Universiteit Amsterdam, Amsterdam, The Netherlands, 2017.

[111] M. -. C. R. -. K. A. He, "Football Player's Performance and Market Value," in *2nd workshop of sports analytics, European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), Porto, Portugal*, 2015.

[112] A. J. -. R.-A. J. -. P.-S. J. M. Sáez-Castillo, "Expected number of goals depending on intrinsic and extrinsic factors of a football player. An application to professional Spanish football league," *European Journal of Sport Science*, vol. 13, no. 2, pp. 1-12, 2011.

[113] L. -. C. P. -. F. P. -M. E. -. P. D. -. G. F. Pappalardo, "PlayeRank: data-driven performance evaluation and player ranking in soccer via a machine learning approach," 2018.

[114] D. Totaro, "A simple method to predict player performance using Fantasy Football data," 2 August 2017. [Online]. Available: https://towardsdatascience.com/a-simple-method-to-predict-player-performance-using-fantasy-football-data-8b2d3adb3a1a.

[115] A. -P. L. -C. P. -. I. F. M. -. F. J. -. M. D. Rossi, "Effective injury forecasting in soccer with GPS training data and machine learning," *PLoS ONE*, vol. 13, no. 7, 2018.

[116] E. A. -. S. M. V. -. C. W. C. Zillmer, Principles of neuropsychology, Belmont, CA, USA: Wadsworth/Thomson Learning, 2001.

[117] S. Kampakis, "Predictive modelling of football injuries," arXiv:1609.07480, London, UK, 2016.

[118] B. G. P. Martins, "Predicting the risk of injury of professional football players with machine learning," Lisbon, Portugal, 2018.

[119] M. -. W. T. -. G. J. K. -. G. M. Michałowska, "Artificial neural networks in knee injury risk evaluation among professional football players," in *22nd International Conference on Computer Methods in Mechanics, Lublin, Poland*, Poznań, Poland, 2018.

[120] A. Z. -. R. V. J. -. O. E. Olemdilla, "Predicting and preventing sport injuries: the role of stress," in *Sports Injuries: Prevention, Management and Risk Factors*, G. Hopkins, Ed.,

Nova Science Publishers, Inc., 2014.

[121] P. J. Sloane, "Rottenberg and the Economics of Sport after 50 Years: An Evaluation," *IZA Discussion Papers,* vol. 2175, 2006.

[122] K. -. P. P. Baidina, "Uncertainty of Outcome and Attendance: Evidence from Russian Football," Moscow, Russia, 2017.

[123] N. -. D. C. -. B. L. -. G. D. -. A. W. Scelles, "Competitive balance versus competitive intensity before a match: Is one of these concepts more relevant in explaining attendance?The case of the French football Ligue 1 over the period 2008-2011," *Applied Economics,* vol. 45, no. 29, pp. 4184-4192, 2013.

[124] V. -. N. I. -. R. J. Manasis, "Measuring Competitive Balance and Uncertainty of Outcome Hypothesis in European Football," 2015.

[125] D. a. S. G. Czarnitzki, "Uncertainty of outcome versus reputation: Empirical evidence for the First German Football Division," *Empirical Economics,* vol. 27, no. 1, p. 101–112, 2002.

[126] T. Pawlowski, "Testing the Uncertainty of Outcome Hypothesis in European Professional Football: A Stated Preference Approach," *Journal of Sports Economics,* vol. 14, no. 4, pp. 341-367, 2013.

[127] B. a. S. R. Buraimo, "Uncertainty of Outcome or Star Quality? Television Audience Demand for English Premier League Football," *International Journal of the Economics of Business,* vol. 22, no. 3, pp. 449-469, 2015.

[128] K.-C. a. Z. R. Wang, "Classifying NBA Offensive Plays Using Neural Networks," in *MIT Sloan Sports Analytics Conference 2016, Boston, MA, USA*, Toronto, 2016.

[129] A. C. a. B. L. Miller, "Possession Sketches: Mapping NBA Strategies," in *MIT Sloan Sports Analytics Conference 2017, Boston, MA, USA*, 2017.

[130] A. a. G. J. Nistala, "Using Deep Learning to Understand Patterns of Player Movement in the NBA," in *MIT Sloan Sports Analytics Conference 2019, Boston, MA, USA*, 2019.

[131] J. a. B. L. Mortensen, "From Markov models to Poisson point processes: Modeling movement in the NBA," in *MIT Sloan Sports Analytics Conference 2019, Boston, MA, USA*, 2019.

[132] J. Kuehn, "Accounting for Complementary Skill Sets When Evaluating NBA Players' Values to a Specific Team," in *MIT Sloan Sports Analytics Conference 2016, Boston, MA, USA*, 2016.

[133] P. a. L. P. Felsen, "'Body Shots': Analyzing Shooting Styles in the NBA using Body Pose," MIT Sloan Sports Analytics Conference 2017, Boston, MA, USA, 2017.

[134] N. -. M. J. -. B. L. Sandholtz, "Chuckers: Measuring Lineup Shot Distribution Optimality Using Spatial Allocative Efficiency Models," in *MIT Sloan Sports Analytics Conference 2019, Boston, MA, USA*, 2019.

[135] S. -. R. V. -. G. C. -. S. A. -. P. D. -. W. J. -. Z. D. Kaplan, "The Economic Impact of NBA Superstars: Evidence from Missed Games using Ticket Microdata from a Secondary Marketplace," in *MIT Sloan Sports Analytics Conference 2019, Boston, MA, USA*, 2019.

[136] H. -. V. T. -. F. G. -. H. C. -. H. J. -. K. A. -. M. M. -. S. D. -. S. S. Talukder, "Preventing in-game injuries for NBA players," in *MIT Sloan Sports Analytics Conference 2016, Boston, MA, USA*, 2016.

[137] J. L. a. H. W. K. Salmon, "Tracking Pitcher Performance with Instantaneous Component ERA and Moving Averages," in *MIT Sloan Sports Analytics Conference 2016, Boston, MA, USA*, 2016.

[138] W. K. a. S. J. L. Harrison, "Leveraging Pitcher-Batter Matchups for Optimal Game Strategy," in *MIT Sloan Sports Analytics Conference 2019, Boston, MA, USA*, 2019.

[139] P. Z. Shu, "Arsenal/Zone Rating: A PitchF/X based pitcher projection system," in *MIT Sloan Sports Analytics Conference 2016, Boston, MA, USA*, 2016.

[140] E. P. Martin, "Predicting Major League Baseball Strikeout Rates from Differences in Velocity and Movement Among Player Pitch Types," in *MIT Sloan Sports Analytics Conference 2019, Boston, MA, USA*, 2019.

[141] C. a. T. S. T. Glynn, "A switching dynamic generalized linear model to detect abnormal performances in Major League Baseball," in *MIT Sloan Sports Analytics Conference 2017, Boston, MA, USA*, 2017.

[142] R. Paulsen, "New Evidence in the Study of Shirking in Major League Baseball," in *MIT Sloan Sports Analytics Conference 2019, Boston, MA, USA*, 2019.

[143] J. Hochstedler, "Finding the Open Receiver: A Quantitative Geospatial Analysis of Quarterback Decision-Making," in *MIT Sloan Sports Analytics Conference 2016, Boston, MA, USA*, 2016.

[144] B. Burke, "DeepQB: Deep Learning with Player Tracking to Quantify Quarterback Decision-Making & Performance," in *MIT Sloan Sports Analytics Conference 2019, Boston, MA, USA*, 2019.

[145] L. -. W. P. -. N. D. Bornn, "Training Schedule Confounds the Relationship between Acute:Chronic Workload Ratio and Injury," in *MIT Sloan Sports Analytics Conference 2019, Boston, MA, USA*, 2019.

[146] M. -. K. M. -. P. N. K. -. C. K. Kurt, "Alleviating Competitive Imbalance in NFL Schedules: An Integer-Programming Approach," in *MIT Sloan Sports Analytics Conference 2015, Boston, MA, USA*, 2015.

[147] S. Pettigrew, "Assessing the offensive productivity of NHL players using in-game win probabilities," in *MIT Sloan Sports Analytics Conference 2015, Boston, MA, USA*, 2015.

[148] O. -. Z. Z. -. J. M. -. D. P. Schulte, "Apples-to-Apples: Clustering and Ranking NHL Players Using Location Information and Scoring Impact," in *MIT Sloan Sports Analytics Conference 2017, Boston, MA, USA*, 2017.

[149] M. E. Schuckers, "Draft by Numbers: Using Data and Analytics to Improve National Hockey League (NHL) Player Selection," in *MIT Sloan Sports Analytics Conference 2016, Boston, MA, USA*, 2016.

[150] N. -. Y. D. -. B. C. -. B. L. -. J. M. Czuzoj-Shulman, "Winning Isn't Everything A contextual analysis of hockey face-offs," in *MIT Sloan Sports Analytics Conference 2019, Boston, MA, USA*, 2019.

[151] D. -. B. C. -. B. L. -. J. M. Yu, "Playing Fast Not Loose: Evaluating team-level pace of play in ice hockey using spatio-temporal possession data," in *MIT Sloan Sports Analytics Conference 2019, Boston, MA, USA*, 2019.

[152] C. -. W. B. Bagley, "Bump, Set, Spike: Using Analytics to Rate Volleyball Teams and Players," in *MIT Sloan Sports Analytics Conference 2017, Boston, MA, USA*, 2017.

[153] Y. -. J. D. Kataoka, "Mining Muscle Use Data for Fatigue Reduction in IndyCar," in *MIT Sloan Sports Analytics Conference 2017, Boston, MA, USA*, 2017.

[154] X. -. L. P. -. M. S. -. R. M. -. S. S. Wei, "“The Thin Edge of the Wedge”: Accurately Predicting Shot Outcomes in Tennis using Style and Context Priors," in *MIT Sloan Sports Analytics Conference 2016, Boston, MA, USA*, 2016.

[155] C. R. a. F. A. Berry, "How Much Do Coaches Matter?," in *MIT Sloan Sports Analytics Conference 2019, Boston, MA, USA*, 2019.

[156] M. a. M. T. Jordan, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255-260, 17 July 2015.

[157] "Geeks for geeks," [Online]. Available: https://www.geeksforgeeks.org/machine-learning/

[158] T. Mitchell, Machine Learning, McGraw-Hill Science/Engineering/Math, 1997.

[159] M. Barber, "Data science concepts you need to know! Part 1," 14 January 2018. [Online]. Available: https://towardsdatascience.com/introduction-to-statistics-e9d72d818745.

[160] K. Krzyk, "Coding Deep Learning For Beginners: Types of Machine Learning," 25 July

2018. [Online]. Available: https://towardsdatascience.com/coding-deep-learning-for-beginners-types-of-machine-learning-b9e651e1ed9d.

[161] "COGNITIVE COMPUTING AND MACHINE LEARNING," [Online]. Available: http://www.cognub.com/index.php/cognitive-platform/.

[162] S. Ghoneim, "Accuracy, Recall, Precision, F-Score & Specificity, which to optimize on? Based on your project, which performance metric to improve on?," 2 April 2019. [Online]. Available: https://towardsdatascience.com/accuracy-recall-precision-f-score-specificity-which-to-optimize-on-867d3f11124.

[163] V. Karbhari, "How to evaluate regression models? Data Science Interview Questions around model evaluation metrics," 18 December 2018. [Online]. Available: https://medium.com/acing-ai/how-to-evaluate-regression-models-d183b4f5853d.

[164] C. M. Bishop, Pattern Recognition and Machine Learning, Cambridge: Springer, 2006.

[165] C. Nicholson, "A.I. Wiki, A Beginner's Guide to Important Topics in AI, Machine Learning, and Deep Learning," [Online]. Available: https://skymind.ai/wiki/machine-learning-algorithms.

[166] M. Ellis, "What Are Machine Learning Algorithms? Here's How They Work," 7 November 2018. [Online]. Available: https://www.makeuseof.com/tag/machine-learning-algorithms/.

[167] M. Rouse, "How to solve your TMI problem: Data science analytics to the rescue," February 2019. [Online]. Available: https://searchsqlserver.techtarget.com/definition/data-mining.

[168] C. Clifton, "Data mining," [Online]. Available: https://www.britannica.com/technology/data-mining/Pattern-mining.

[169] "Data Mining," [Online]. Available: https://www.techopedia.com/definition/1181/data-mining.

[170] "Advantages of Data Mining," [Online]. Available: https://www.educba.com/advantages-of-data-mining/.

[171] A. Twin, "Data Mining," 18 August 2019. [Online]. Available: https://www.investopedia.com/terms/d/datamining.asp.

[172] R. Dontha, "Data Mining Steps," 19 December 2018. [Online]. Available: https://digitaltransformationpro.com/data-mining-steps/.

[173] J. Sridhar, "What Is Data Analysis and Why Is It Important?," 12 February 2018. [Online]. Available: https://www.makeuseof.com/tag/what-is-data-analysis/.

[174] "What is data analysis?," PAT RESEARCH, [Online]. Available: https://www.predictiveanalyticstoday.com/data-analysis/.

[175] R. van Dijk, "5 Advanced Data Analysis Techniques Applied to People Analytics," [Online]. Available: https://www.analyticsinhr.com/blog/advanced-data-analysis-techniques-applied-to-people-analytics/.

[176] M. Galetto, "What Is Data Analysis?," 20 January 2016. [Online]. Available: https://www.ngdata.com/what-is-data-analysis/.

[177] "What's new in Sisense 5.7.5," [Online]. Available: https://www.sisense.com/releases/sisense-5-7-5/.

[178] V. Holman, "What is Sports Analytics?," 15 November 2018. [Online]. Available: https://www.agilesportsanalytics.com/what-is-sports-analytics/.

[179] "The Evolution and Future of Analytics in Sport," 22 June 2017. [Online]. Available: https://www.proemsports.com/single-post/2017/06/22/The-Evolution-and-Future-of-Analytics-in-Sport.

[180] B. a. M. V. Alamar, "Beyond 'Moneyball': Rapidly evolving world of sports analytics, Part I," 2011. [Online]. Available: http://analytics-magazine.org/beyond-moneyball-the-rapidly-evolving-world-of-sports-analytics-part-i/.

[181] J. McKinney, "The top 10 pro sports teams embracing analytics.," 2 December 2015. [Online]. Available: https://www.domo.com/blog/the-top-10-pro-sports-teams-embracing-analytics/.

[182] H. a. K. S. Sharma, "A Survey on Decision Tree Algorithms of Classification in Data Mining," *International Journal of Science and Research,* vol. 5, no. 4, 2016.

[183] M. N. Abdul Hamid, "Classification Using Decision tree," 15 October 2015. [Online]. Available: https://www.slideshare.net/knottisme/classification-using-decision-tree-53984611.

[184] C. Sehra, "Decision Trees Explained Easily," 19 January 2018. [Online]. Available: https://medium.com/@chiragsehra42/decision-trees-explained-easily-28f23241248.

[185] W. Koehrsen, "Random Forest Simple Explanation," 27 December 2017. [Online]. Available: https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d.

[186] T. Yiu, "Understanding Random Forest. How the Algorithm Works and Why it Is So Effective," 12 June 2019. [Online]. Available: https://towardsdatascience.com/understanding-random-forest-58381e0602d2.

[187] L. Breiman, "Random Forests," *Machine Learning,* vol. 45, no. 1, pp. 5-32, October 2001.

[188] G. -. M. M. -. C. S. Carvajal, "Components of Artificial Intelligence and Data Analytics," in *Intelligent Digital Oil and Gas Fields*, Elsevier, 2018, p. 374.

[189] R. Pal, "Overview of predictive modeling based on genomic characterizations," in *Predictive Modeling of Drug Sensitivity*, Elsevier, 2017, p. 354.

[190] R. Sunil, "Understanding Support Vector Machine algorithm from examples (along with code)," 13 September 2017. [Online]. Available: https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/.

[191] R. Jain, "Simple Tutorial on SVM and Parameter Tuning in Python and R," 21 February 2017. [Online]. Available: https://www.hackerearth.com/blog/developers/simple-tutorial-svm-parameter-tuning-python-r/.

[192] "What is Linear Regression?," 2013. [Online]. Available: https://www.statisticssolutions.com/what-is-linear-regression/.

[193] "Linear regression from scratch," 2017. [Online]. Available: https://gluon.mxnet.io/chapter02_supervised-learning/linear-regression-scratch.html.

[194] "What is a Neural Network?," [Online]. Available: https://deepai.org/machine-learning-glossary-and-terms/neural-network.

[195] S. W. Smith, "Neural Networks (and more!)," in *The Scientist and Engineer's Guide to Digital Signal Processing*, California Technical Publishing, 1997.

[196] C. Nicholson, "A Beginner's Guide to Neural Networks and Deep Learning," [Online]. Available: https://skymind.ai/wiki/neural-network.

[197] L. Rhodes, "Artificial Neural Networks (ANN)," 2018. [Online]. Available: http://jcsites.juniata.edu/faculty/rhodes/ml/ann.htm.

[198] N. Donges, "4 REASONS WHY DEEP LEARNING AND NEURAL NETWORKS AREN'T ALWAYS THE RIGHT CHOICE," 24 July 2019. [Online]. Available: https://builtin.com/data-science/disadvantages-neural-networks.

[199] S. Mallick, "Neural Networks : A 30,000 Feet View for Beginners," 2 May 2017. [Online]. Available: https://www.learnopencv.com/neural-networks-a-30000-feet-view-for-beginners/.

[200] "Project Jupyter," [Online]. Available: https://jupyter.org/.

[201] S. Iqbal, "Weka Presentation," 7 January 2013. [Online]. Available: https://www.slideshare.net/SaeedIqbal1/weka-presentation.