https://doi.org/10.26219/heal.aueb.9558

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

"Interdepartmental MSc Program in Business Mathematics."

Innovative Data Analytics and Machine Learning Approaches for Analyzing European Football Shots.

"Thesis submitted to the Master's Program Business Mathematics for the award of the master's degree."

STUDENT'S FULL NAME: Athanasios Kalogiros

SUPERVISING PROFESSOR: Dr Dimitrios Panagopoulos

Athens, September 2024

https://doi.org/10.26219/heal.aueb.9558

This thesis is submitted by the author as partial fulfillment of the requirements for the Master's Program in Business Mathematics at the Athens University of Economics and Business.

It is hereby responsibly declared that this thesis has been written by the undersigned and has not been submitted or evaluated as part of any other postgraduate or undergraduate degree program, in Greece or abroad.

FULL NAME: Athanasios Kalogiros

SIGNATURE:

https://doi.org/10.26219/heal.aueb.9558

ABSTRACT

Over the past decades, the proliferation of the technology has significantly affected most domains, including the field of data. Nowadays, more and more datasets are being created, and the already existing are expanding. The aim of this thesis is to utilize data collected from datasets related to Football, specifically focusing on European football. First, the data were used to provide insights into football through analytical using plots. Additionally, the data were used to train different models to determine whether machine learning can be a contributing factor in predicting outcomes in football. In the first part, the methodology focused on utilizing analytics to derive the results. The Data were analyzed using the Python programming language, and the results were depicted in plots with the appropriate information. In the second part, machine learning methods, specifically supervised learning, were employed to derive the results. Several methods were tested, but finally, five of them were selected for the classification process due to their practicality. In the first part, the results were satisfactorily for most of the teams and players analyzed.An analysis of the shots taken by Messi and Suarez was conducted as part of the study. The analysis focused on the locations and conditions of the shots. A similar analysis was also performed for their teams, Barcelona and Real Madrid, to understand the tactics applied.

Finally, in the presentation of the results, only the shots from Messi and Suarez were presented. The second part of the study focused on testing machine learning models, which initially showed limited effectiveness. However, altering the weights of each variant their performance improved significantly. An effort made to classify the shots into successful (Goal) and unsuccessful (No Goal) using machine learning models. The classification aimed to identify patterns that could predict the likelihood of a shot to result in a Goal, based on specific factors such as shot velocity, the shot angle and the distance to the goalkeeper..

Keywords: Data Scrapping, Data Engineering, Future Engineering, Data Analytics, Supervising Learning, Classification, Python, Plotting & European Football.

https://doi.org/10.26219/heal.aueb.9558

ACKNOWLEDGMENTS

First of all, I would like to express my gratitude to my supervisor, Professor Dimitrios Panagopoulos, for trusting me to start this taught work, and especially, for the support and guidance throughout the process of completing this thesis.

I would like also thanks my classmates, Aggelos Mokakos and Panagiotis Malainos, for selecting me as their partner for the European Football project in the Machine Learning master's course, which was the inspiration for this thesis and the side rock for the whole research.

Additionally, I am grateful to my friends Evangelos Galanis and Ioannis Dimitriou for their guidance and support in the programming aspects of this research, an essential part of this investigation, and finally, i would like to thanks my family and my friends for their support and encouragement, in order to overcome the various obstacles that I faced to reach in the finish of this thesis

https://doi.org/10.26219/heal.aueb.9558

Table of Contents **page**

1. Introduction……………………………………… 6
1.1 Description…………………………………… 6
1.2 Goals …………………………………………… 7
1.3 Meaning……………………………………… 8
1.4 Structure……………………………………… 8
1.5 Significance of Each Chapter……………… 9
2. Literature Review…………………………… 10
2.1 Introduction…………………………………… 10
2.2 Previous Research…………………………… 11
2.3 Gaps…………………………………………… 13
3. Methodology………………………………… 14
3.1 Data…………………………………………… 14
3.1.1 Data Scraping & Engineering…………… 14
3.1.2 Future Engineering………………………… 15
3.1.3 Data Use & Validation…………………… 18
3.2 Analytics……………………………………… 19
3.2.1 Tools……………………………………… 19
3.2.2 The Process……………………………… 20
3.2.3 Target……………………………………… 20
3.3 Machine Learning…………………………… 20
3.3.1 Random Forest…………………………… 22
3.3.2 Logistic Regression……………………… 22
3.3.3 SVM……………………………………… 23
3.3.4 AdaBoost………………………………… 24
3.3.5 Gradient Boost…………………………… 25
4. Results………………………………………… 27
4.1 Analytics……………………………………… 27
4.2 Machine Learning…………………………… 31
4.2.1 Classification Report…………………… 32
4.2.2 Confusion Matrix………………………… 34
4.2.3 ROC Curve……………………………… 34
4.2.4 Future Importance……………………… 36
5. Discussion…………………………………… 41
5.1 Machine Learning…………………………… 41
5.2 Analytics……………………………………… 42
5.3 Summary……………………………………… 42
6. References…………………………………… 44

https://doi.org/10.26219/heal.aueb.9558

Chapter 1

Introduction.

1.1 Description.

As mentioned in the abstract, the amount of the data has faced an unpreceded growth, affecting most sectors of society. Despite football being one of the most popular sports, and a lucrative indusrty, until recently a few efforts have been made to collect and utilize data. As mentioned in the bibliography, only a few experiments have been conducted using simple past data, such as the total number of wins, draws, losses and goals scored, conceded in home or away. The results focused on basic statistical methods with a descriptive target, providing poor performance. However, over the past decade, with the evolution of websites containing information related to football and the development of data scraping methods by several programming languages, football-related have an unstoppable expansion. Betting companies also contribute to this increase in datasets. Daily, they adopt new methods to collect, analyze and utilize data for their benefit, because it plays a major role in their financial and future. Their main target is to predict odds more accurately than bettors, in order to enhance their possibilities. To achieve this, they employ predictive methods, often applying machine learning techniques both supervised and unsupervised. These methods not only can perform effectively in the prediction but also creating a profile for bettors. Nowadays, companies specializing in this field use advanced tools that allow them to accumulate data in real time. The most common method is using real-time images of games for storing data each second. For example, when a pass occurs, information such as position, length, angle, outcome etc. are stored automatically in a json file. Such data were utilized in this thesis to describe and predict the results. The main goal of this thesis is not only to predict the outcome of events in a football games but also to provide deeper insights into the field of football analysis.

The initial effort aimed to visualize the shots taken by specific teams and players. Specifically, shots taken by both Messi and Luis Suarez were plotted. For each shot, the exact positions of shots start and end were used, afterwards, in the second part, used also the position where the player started to carry the ball until the shot. The second approach focused on predicting and evaluating the likelihood of a shot resulting in a goal. Most analyses and companies evaluating shoot probabilities usually give very low chances for shot to end up as a goal. For example, they might estimate a shot has just 1% chance of success, meaning 100 shots would be needed to score a single goal. However, it might only take 10 shots to score 2 goals. This suggests that current analyses may underestimate the true likelihood of a shot leading to a goal, highlighting the need for improved methods that account for additional factors and real-game conditions . This thesis aimed to approach shot evaluation from a different perspective. In the supervised learning methods, different weights were given to the outcomes, with shots that resulted in goals receiving more weight to improve the predictability for goals. Furthermore, the target was to create a model where the true positive rate for both goals and misses would be around 70%, a satisfactory

https://doi.org/10.26219/heal.aueb.9558

percentage for predicting shots. Seven different features used to achieve valuable results, all selected based on author's experience and the previous research in the literature. For all methods, the percentage by which each feature affects the predictability was measured. Logistic Regression, AdaBoost, and Gradient Boost showed both positive and negative feature importance values. This likely stems from their nature: Logistic Regression shows negative values when the features lower the chance of a positive outcome. In AdaBoost and Gradient Boost, the process that corrects mistakes gives negative values to features that that cause misclassifications. This difference will be discussed extensively on chapter four.

1.2 Goals

Based on the bibliography and in previous surveys, the approaches to football data analysis were somewhat general. For instance, the vast majority of the researches has focused solely on predicting match outcomes win, draw and loss. In recent years, however, some innovators have dared to create descriptive models tailored to persons who have specific roles in a team. This inspiration triggers the goals of this thesis: to give insights with the results in different components. Football, as mentioned above, is a sport that influences many aspects and people. Thus, the findings would be intriguing for individuals with various roles. Consequently, each components of the results could be analyzed by many stakeholders, allowing each of them to see them from their unique perspective. This approach would not only be intriguing but also innovative for the football. From coaches to industry betters, something innovative would be appealing and very beneficial. For achieving such results, the only way is to think out of the box and to see the data in another way.

By considering the methodologies of other researchers and by referring to different authors, this thesis tries to shed light on aspects that haven not been explored before. The core idea and the main goals is to produce results that can open new pathways and help people who haven't the opportunity to benefit from the football data. Moreover, the goal is for the results to be as realistic as possible. In simple terms, detailing the characteristics of each shot and its practical application in real match allows various stakeholder to leverage the findings to their advantage. Furthermore, this thesis attempts to incorporate spatial data into the game of football. Analyzing a shot and providing a visual presentation of this can affect many stakeholders in different ways. For example, a shot of Messi will be analyzed differently by his offensive partner, his coach and his opposing defenders. Additionally, this can also alter a team's playing style. Many teams base their tactics on their players or their opponents, which affects both the team's tactics and its finances, as each player brings different value to the team affecting both strategic decisions and economic considerations.

https://doi.org/10.26219/heal.aueb.9558

1.3 Meaning.

It is necessary to conduct surveys to contribute to the evolution of a domain, which can be useful for further advancement. Like each other domain in society that needs to evolve, so does football. First and foremost, it is commonly accepted that the evolution of a domain is essential. This is because it is innate into human's nature and society to contribute to the evolution of each part, as they all affect one another. This thesis aims to advance the evolution of football. By analyzing and utilizing the latest tools, such as Python and machine learning algorithms, it seeks to affect a little more the way matches are viewed. Although there is a limited amount of data publicly and freely available for football, engaging in a research like this could stimulate others to continue the investigation and to go even further.

Additionally, it is necessary to apply the results of an investigation both theoretically and practically. In the theoretical part, the outcomes can help players and coaches understand the significance of football analysis and its value. Since this domain is in the first stages, most of them don't have the appropriate knowledge for this. Only a few clubs, have incorporated analytics into their team, typically those with big budgets. Practically there are numerous applications of football analytics. Players and coaches have the opportunity to analyze every movement of all players from both sides, improving in this way their performance and strategy as well. Also, the betters can use football analytics to predict various events and the likelihood of these events in a match, enhancing in this way their ability to predict different outcomes and their likelihood.

1.4 Structure.

The structure of this thesis is designed to be as simply as possible to help the readers understand all the aspects better. Since it involves an investigation with a huge amount of data and results, it is organized in a manner simple and clear. The first chapter includes the introduction to the survey and consists of five essential parts. 1) description of the problem, 2) the purpose of this thesis, 3) the significance of the subject, 4) the structure of this thesis and 5) a brief description of the chapters. The second chapter contains the literature review, the references in previous investigations, their results, and identifies the gaps in previous researches that this thesis aims to filled by its methodology. The third chapter contains the methodology used to obtain the data and to come up with the results. It consists of four essential parts 1) detailed explanation of the methodology, 2) the reason for using this methodology, 3) the process of collecting the results and 4) the sample and tools used. The fourth chapter contains the analysis of the results. The most valuable parts of this chapter are 1) the method of data analysis and 2) the results obtained from the analysis. The fifth and final chapter is the discussion of the results. In this chapter there are three basic parts 1) the interpretation of the results, 2) the disagreements with the existing bibliography and 3) the primary findings which relate to the scientific part. As mentioned before, this thesis is structured simply.

Therefore, the most extensive chapters are chapter three and four. This emphasis on methodology and results aims to help readers to better understand how these can be applied in other investigations. The remaining chapters try to remain as simple and clear as possible, ensuring a smooth transition to the main part on methodology and results. Finally, the bibliography is a valuable part, but it will not be extensive. Only the most valuable and related to the subject parts will be used. Also, this includes previous researches with gaps in the subject. The bibliography contains introductory and easily accessible sources , selected to focus on the thesis topic. Detailed analysis of each chapter will be provided in the next paragraph in order to be more clear and understandable.

1.5 Significance of each chapter.

In the first chapter, more weight is given on the detailed description and analysis of the problem. The aim of this effort is to make the problem more understandable, ensuring that readers are well informed about the focus of this thesis. Another essential component for understanding is the motivation for engaging with this subject. These two parts together form a well-rounded understanding of the significance of the topic. Finally, this provides a chance to guide the reader smoothly into the thesis's objectives for better comprehension.

In the second chapter, the use of the existing bibliography in conjunction with the information of the first chapter helps the readers comprehend the inspiration behind the topic. Specifically, this chapter contains the results and references from other investigations that form the foundation of this thesis. Additionally, understanding the results and motivation of previous surveys aims to provide an understand to the philosophy of this thesis. Finally and more importantly, referencing in the limitations of previous investigations is crucial for highlighting how this topic is both different and innovate. Filling these gaps and the approach taken are essential too.

In the third chapter, a detailed analysis is provided in the practical part of this thesis. From better understanding, all the tools and steps taken are explained detailed and in a consecutive row. Additionally, understanding the reason why this methodology is chosen plays an important role in interpreting the results. The way these tolls produce the results is another essential part of this chapter. This part is inseparable connected with the next chapter as it forms its foundation. This connection is beneficial not only for a better explanation of the results but also for understanding the process behind this.

In the fourth chapter, the results of both parts of the investigation are presented in a detailed form. Plots, matrices and textual analysis provide a complete reference in the results. Their rational structure and the way they are created, ensure a better understanding of the findings. Additionally, a reference is provided to explain how the pure data of this thesis leads to the results. The target of this effort is to offer a compete explanation of the entire process again as now the methodology has applied.

 https://doi.org/10.26219/heal.aueb.9558

In the fifth and final chapter, a summary of the conclusions is presented, based on the author's knowledge and experience. Additionally, the limitations of the study are discussed, along with anticipated outcomes in relation to existing literature. Furthermore, potential gaps left by the current work, and the need for further research in this field. . All the chapters aim to provide a well understanding of the topic and underscore the significance of a subject with much to offer in the field of the data analysis.

Chapter 2

Literature Review.

2.1 Introduction.

By searching and analyzing the literature on this topic from previous years, it becomes evident that little job has been done. Most of these surveys covered by scientific articles, indicate that the research has not been conducted in depth. The main reason for this is the scarcity of data available in conjunction with the luck of powerful tools for data analysis. Their analysis was simple and focused only on predictive results related to the outcome of a match. Little effort has been made to describe and predict specific events within the game. On the other hand, theses analysis served as the starting point for a very interesting and promising field in both descriptive and predictive analysis. Also, the use of the pure statistical tools aids in better understanding the results generated by the machine learning algorithms.

Nowadays, two main factors contribute to the expansion of the literature related to this topic. Firstly, technology facilitates the scraping and storing of huge amounts of data. Programming languages offer extremely powerful tools to scrape data from the internet every minute. Additionally, many firms use virtual games to simulate football matches, in order to extract data related to specific teams or players Teng Jia, Jirarat Sitthiworachart , John Morris.(2024) . Secondly, the substantial profits that come from this expansion played a critical role. Both teams and betting companies invest more and more money to better analyze and predict the game. Since machine learning is a predictive tool, it performs perfectly with data related to football. Furthermore, the descriptive aspect is essential for managers and players to understand the match and characterize players.

In this chapter, the results from previous investigation and the gaps that exitst in them will be presented. The vast majority of these studies have focused on predicting match outcomes, as previously mentioned, most of them derived from scientific articles, which are limited in scope and have not allowed for deeper analysis. Moreover, understanding the limitation of these surveys was crucial for this thesis.

 https://doi.org/10.26219/heal.aueb.9558

This motivated the formation a thesis centered on both predictive and descriptive results out of the scope of previous research.

2.2 Previous research.

As is mentioned above, little research has done been in the domain of football analytics. Still, despite the limitations of the available data and tools, the researchers have done exceptional work.

Kerr (2025) mentioned that, while football being the most popular sport worldwide, its analytical research has not yet reached the level seen other sectors have. Over the past 17 years, there has been great effort from both professional and academic circles to advance sports analytics. Despite its popularity, football analytics have focused on creating new summary statistics to support traditional game insights, lacking the significant data analysis seen in other sports.

Aladesote, Agbelusi, and Ganiyu (2020) further discussed the huge impact of big data revolution in football, emphasizing its role in improving the viewing experience of the sport. Their research reveals that football is played and bet by over 150 million people in more 200 countries, making it the world's most popular team sport and favorite for betting. Additionally, they estimated the value of sports betting industry, including both legal and illegal markets, at $700 billion and $1 trillion annually, respectively. Finally, in their survey, they stated, ´´the key reason for football's global appeal is its inherent unpredictability . ´´

In Muhammad Yonus Saiedy, Muhammad Aslam Hemmat Qachmas , Dr. Amanullah Faqiri (2020) made a noteworthy effort to predict the outcome of matches in English Premier League EPL with the aim to predict the total grading of the championship. They use predictive tools, specifically supervised learning and classification methods with Random Forest and SVM classifiers. They achieved an accuracy of 49,8% for Random Forest and 54,3% for SVM. Though the accuracy seems low for both methods, the total grading showed little variation in the first 5 and the last 5 teams, which are the areas that attracts the most interest. Finally, their success in the classification process for English Premier League demonstrates the effectiveness of machine learning to predict outcomes on football. They also mentioned that, more light can be shed on this topic.

A similar try was done by Yash Ajgaonkar Kunal Bhoyar, Prof. Anagha Patil andJenil Shah in (2021). Based on the idea that sport prediction is one of the most expanding areas in predictive accuracy as it involves huge amount of money in betting and its popularity, they used supervised learning in the English Premier League matches. Improving with Muhammad Yonus Saiedy, Muhammad Aslam HemmatQachmas , and Dr. Amanullah Faqiri they achieved an accuracy of 67% using Support Vector Machine, 60% using Random Forest and 56% using Naïve Bayes classifier.

 https://doi.org/10.26219/heal.aueb.9558

Victor Chazan, Pantzalis, Christos, Tjortjis in (2020) tried something quite different. They made an effort to predict the performance both of the teams through European Leagues and the center defender players. Data collected from the season 2018-2019 season included team actions (transfer, managerial changes). Two approaches were used for teams: first, to classify teams as performing better or worse than the previous season. The accuracy of this classifier was around 70% indicating a satisfactory but not outstanding result. The second approach was to simulate each match of the season and classify results as home win, drawn and away win, accumulating points to form the total table of the championship. The higher accuracy achieved was 57% for the English Premier League with an RMSE of 11%.The lowest RMSE was 9% for the Spanish La Liga, with an accuracy of 51%. Champions were correctly predicted with an accuracy of 64%. For player performance prediction one approach was used: finding match events and actions that influence central defender's rating. Multiple Linear Regression with Backward Elimination was chosen as method. The method achieved an R-squared of 0.907 and an adjusted of 0.88. Thirteen features were statistically important, including classic defensive action (interceptions, clearances) and attributes (jumping reach, strength), as well as some attacking skills. They conclude that classification techniques have the potential to produce notable results in predicting both team and player's performance.

In (2020) Salil Chincholikar, Niharika Sadul, Parth Thakkar, Ashutosh Amrutkar, Mr. Nikhil Dhavase data analytics have become crucial in the football industry today. The rise of big data enables the extraction of valuable insights to improve player performance, prevent injuries and increase commercial efficiency. Based on this concept, they want to carry out more surveys to explore these possibilities in more detail.. Moreover, they used both historical and current data. They believe that by combining these different types of data, they could predict various aspects of the game. These include match scores, player ratings, and goal analytics across different leagues. . They used an 18-dimensional vector as a metric in neural networks for a classification process. The first component represented a rating of the goalkeeper, the next six components were for a ratings of defenders, the next seven for midfielders and the final four for forwards. For the empty positions, the components were set as zero. When they passed to the neural network, this formed a 36-dimensional vector, with the home team occupying the first 18 dimensions and the away team the final 18 dimensions. With this process, they achieved a precision of more than 70%. allowing them to state that machine learning can be effectively applied to different fields, including sports. This involves analyzing player's and team performance to give valuable insights and predict outcomes in a satisfactory degree.

James H. Hewitt and Oktay Karakuş in (2023) dared to try something different from previous researches by making an effort to go beyond the usual data analysis in football. They applied spatial data to depict shoots and passes made during a game. The first attempt was to display the start and end positions of shots on a plot, mapping their exact positions on the field. The same was done for the passes as well, achieving a clear and understandable depiction of the most frequent areas for shoots and passes. For the passes, they used data from Toni Kroos, considering him as the player with the most successful

https://doi.org/10.26219/heal.aueb.9558

passes. For the shots, they used data from Messi, Henry and Sanchez. The results were exceptional, as they achieved an accurate depiction of the passes and shots with their exact coordinates. For the shots, those that resulted in goals were plotted with a different color than those ended out of the target, achieving a clear separation between them in the plot. Additionally, they valuated the importance of each feature to determine the most threatening shots. This allowed them to calculate the probability of each shot to end as goal. They applied supervised learning to classify the shots, calculating the probabilities afterwards. Another idea they applied was to compare their results with the expected goals provided by StatsBomb, considering these probabilities as a reliable metric. They implemented Logistic Regression and Decision Tree as classification methods, with the results from Decision Tree fitting very well with those of StatsBomb. Furthermore, they measured the expected goal for different players and compared these with their actual goals. The data utilized in the paper showed that most elite finishers produce more Goals than their xG value. Lionel Messi's sample of 415 Goals from 342 xG is a %21.34 overperformance, while Luis Suarez has 125 goals from 107 xG, a %16.82 higher output. Messi scored 0.82 xG, and his overperformance was significantly greater than that of any other player in the sample. Finally, they accomplished a detailed shoot analysis from the 2018 Champions League Final between Real Madrid and Liverpool.

2.3 Gaps.

Football analytics is a fast evolving and promising filed with potential for future search. As in any interesting area, each study leaves its gaps for future exploration. This thesis aims not only to fill the gaps left in previous research, but also to motivate others to go further this investigation. Summarizing the literature, it is clearly understandable that most research in the domain of football analytics has focused only on descriptive mode. Moreover, the descriptive surveys were focused on results relating to the outcome of a match. Conducting a survey that includes analytics to give insights for future application and machine learning methods to predict specific events during a football match is very appealing. The motivation for creating this thesis was based on this idea.

In previous surveys, only James H. Hewitt and Oktay Karakuş in (2023) dared to go beyond the usual investigations and try something different. They combined analytics with the prediction of the likelihood that a shot will result in a goal, introducing new features that enhanced Expected goal (xG) models. Their study incorporated factors such as goalkeeper positioning, player pressure, and the number of defenders between the shoot and the goal. Their results confirms that forwards are the most efficient shooters, followed by midfielders and then defenders, with Messi outperforms even the most forwards in terms of shooting efficiency. Though their survey fills many gaps, it does not explore football analytics in detail. Predicting a player's shot tactics was an area that had not been searched before. Additionally, while some studies have used spatial data to calculate the possibility of a shot resulting in a goal, this aspect has not been analyzed in depth in previous researches. Concerning the futures used for football

https://doi.org/10.26219/heal.aueb.9558

analytics, the movements of a player before a shot have been not represented in a plot, specifically the carry of the ball and its directions a few seconds before the shot. These gaps are both appealing and interesting for investigation by teams and players. Moreover, betting sites and bettors see this an opportunity to enhance their performance in predicting results.

Taking all these into consideration, the limitations in existing literature are an especially appealing topic for exploration. Therefore, the main target of this thesis is to shed light on aspects that have not been investigated before in football analytics. These aspects include, first, the creation of a plot to depict the player's carry of the ball before the shot, and second, the evaluation of the probability of a shot resulting in a goal. This study goes a step further than previous studies by incorporating player movements before each shot, using more spatial data to describe a shot. Also, other studies often gave very low probabilities for shots, like 0,1% meaning 100 shots would be needed for a goal, which is not realistic. In real matches, about 20 shots usually lead to 2 goals. In order to improve this, binary classification used, focusing only whether a shot ended up as goal or not. This helped to remove unnecessary outcomes (e.g., saved, post, out) and gave more realistic probabilities for each shot. Finally, the expansion of the existing research not only fills the gaps but creates potential evidence for further investigation as well. This can motivate future passionate researchers to expand the field and create new paths for improving the domain.

Chapter 3

Methodology

3.1 Data

3.1.1 Data Scraping & Engineering.

The first step attempt was to collect the appropriate data based on previous research. In the most cases, authors used data downloaded by sites that provide them in csv format in a structured form. These data contain specific events such as corners, attacks, shots as a summary. Thought they were very useful and ready to use, they weren't very practical for applying them in a specialized investigation. The data needed to be more detailed and include events related to spatial information. Data of this type can be found on Statsbomb website, a sport site which provides unstructured data in json files. The files were posted in GitHub, and there were two ways to download them: the first was scraping them from the API using a programming language, and the second was downloading them manually. The second way was chosen because it was more straightforward and helped avoid the time required to learn a new method. A

https://doi.org/10.26219/heal.aueb.9558

particular characteristic of these dataset is its reference of each event that occurred during a football match.

Each file, was parsed with the aid of Python to retrieve critical information, which contained one or more matches and all the events occurred every minute and then, stored it automatically in a Python dictionary. The events of the match are stored per minute, along with the player in possession with the ball, the player's team, and his opponents that are near the ball. Additionally, if an event occurred, extra information is added. For shots, the dataset includes the position of the shooter, the angle of shot, the end position of the shoot, the end time of the shoot, and the outcome of the shoot. For passes, the position of the passer, the angle of the pass, the end position of the pass, and the outcome of the pass are available. These real-time data, provided instantly as events occurred, made them highly unique and practical for application, as well as ideal for an innovative study.

The ability to process and transform these data into a format that fits your model was the most special and attractive feature of them. Having a large source of data gives the chance to go beyond the usual analysis, finding new paths that never existsed before, and to solve problems. With the aid of the Python programming language, the data was structured into a Data Frame using pandas and, in this way, the raw unstructured data was transformed into useful features, resulting in a different way of analyzing the match. Using data engineering and data mining methods, spatial data was altered in form without losing its value, on the contrast, it gained more value. As a result, new features were created, inspired by the combination of those used in previous researches, and the objectives of this thesis.

3.1.2 Feature engineering.

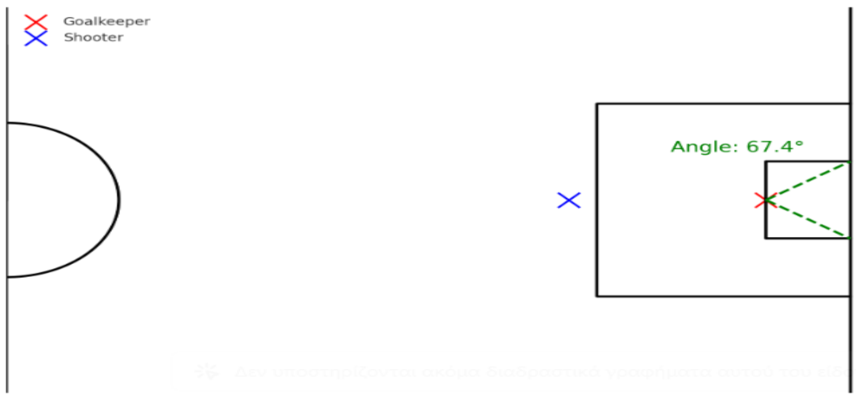
In his section, feature engineering was really important for getting the right results , as it contributed to shape the data in the right way . Based on the experience gained from the literature, the results from analytics and machine learning methods in football analysis are very sensitive to spatial data. Therefore, the process of transforming and creating new features was very careful. The creation of the new variables was based on the characteristics that affect the outcome of a shot, and the innovative ideas inspired by previous studies.

For the analytics part, the results derived from unstructured data. To plot a shot, only the player's position, the end position of the shot and the outcome of the shot (Goal or No Goal) were used. In the second part of the analysis , additional data points were derived considering the player's movements with the ball before the shot. Only the coordinates of the ball's acquisition by the player and the end of his curry, which was the same as the start position of the shot, were used. The big challenge in transforming data and creating new variables was for the machine learning methods. Many attempts were made to identify the most important and effective features for our models, all of which were based specifically on spatial data. Finally, seven of them were used, which are:

https://doi.org/10.26219/heal.aueb.9558

1. **Shot angle**: using the exact coordinate where the shot occurred based on a 2-D coordinate system, the angle formed between the position of the shot and the two-goal posts was calculated. All calculations performed in a Cartesian coordinate system, where the x-axis was set to 120 and the y-axis to 60, representing the dimensions of the field. The center of the goal was positioned at y=30, and the coordinates of the two goalposts were determined by adding and subtracting half of the goal's width (7.32 meters) from the center. This setup ensured accurate representation of the goalposts in the plot..
2. **Velocity of the shot**: this variable didn't appear in the previous surveys, was completely inspired by the combination of the available data and the author's experience as a football player. Its creation was straightforward, using the math library incorporated in Python, the Euclidean Norm (norm 2) between the position where the shot occurred and the position of it ended was calculated. The norm 2 was used because the coordinates were in a 2-D dimension. The calculated norm 2, which represented the distance the ball traveled, was then divided by the duration of the shot to calculate the velocity in meters per second.
3. **Angle of Goalkeeper**: this feature represents the angle formed between goalkeeper's position and the two goalposts at the time of the shot. The angle was calculated using the same methods and mathematical operations as the shot angle variable. This idea behind this feature is that a smaller angle indicates that is in the side that shot occurring, making it more difficult for the shooter to score. Based on the analytics of this thesis, most shots occured to the two sides rather than directly at the center of goalposts. Therefore, if goalkeeper's angle is small, this indicates that the goalkeeper is closer to the side of the shoot, increasing the difficulty for the shooter to score. Figure 3.1 describes this above.

Figure 3.1



4. **Distance to Goal**: the variable was created based on the rationale that the closer a shot occurs to the Goalpost, the more likely it is to result in a goal. Crating this variable was an easy task, using the same methodology as for the Distance to Goalkeeper. This variable refers to the distance of the shot position and the center of the goalposts. The center of the goalpost is

 https://doi.org/10.26219/heal.aueb.9558

represented by the coordinates (120,30). Applying the Distance to goalkeeper with this feature, forms a highly promising combination for predicting valuable results.

5. **Opponents less 5 meters**: this feature formed from the position of opponents close to the shooter. Initially, the Euclidean Norm used to calculate the distance of each opponent from the shooter, and all the elements were stored in a list. Then, a loop was used to parse the list and find all the distances that were less than 5 meters using an If statement, while a variable employed to count the number of opponents in the range when the statement was true. The main idea behind creating this variable was based in the assumption that the more opponents are in the range of 5-meters, the more challenging is for the shooter to score. This a variable refers to the defensive aspect, in contrast to the previous features.
6. **Opponents in the angle**: the inspiration to form this variable came after creating the previous variable, ΄ Opponents less 5 meters'. The name does not fully respond to its application but rather its operation. The core idea behind its creation was to count the opponents that positioned between the shooter and the goalpost. Ideally, the target was to count the opponents inside the angle formed by the shooter and the two posts. However, using a specific space in a plot was very challenging. Therefore, a function was used to count the opponents that were in front of the shooter towards the goalpost. Specifically, this function used as inputs the positions of the shooter, positions of the opponents and the center of goalposts. Then, for each opponent, the function first calculated the clockwise angle between the shooter and the center of the goalpost, and then the clockwise angle between the defender and the shooter. Finally, an if statement was used to determine whether the defender was positioned towards to goalpost.

    - For shots on the right side (shooter΄s angle positive): if the defender΄s clockwise angle relative to shooter was greater than 180°, the defender was considered to be positioned between the shooter and the goalposts. Figure 3.2 & 3.3 describe this with the first for a defender to be between the shooter and the goalpost and the second not.

Figure 3.2

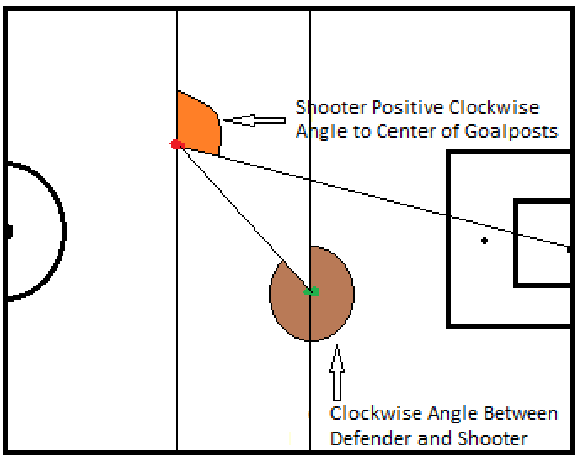
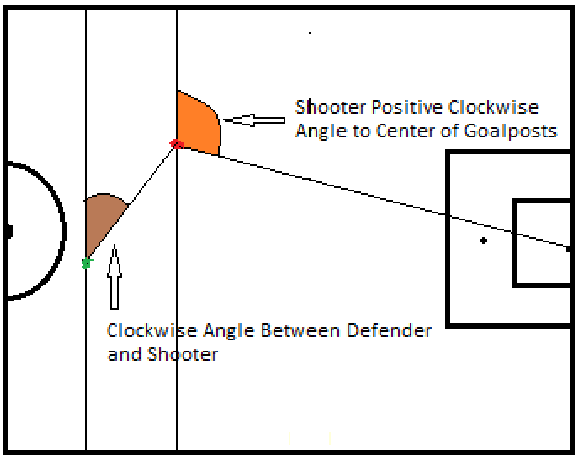


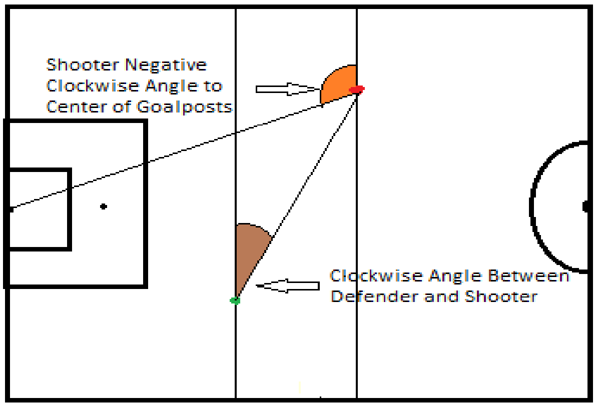
Figure 3.3



 https://doi.org/10.26219/heal.aueb.9558

- For shots on the left side (shooter´s angle negative): if the defender´s clockwise angle relative to shooter was less than 180°, the defender was also considered positioned towards the goalpost. Figure 3.4 describes above the case when the shooter's angle is negative (left side) and the defender is between the shooter and the goalposts.

Figure 3.4



To further refine the blocking condition, only defenders within 5 meters of the shooter were counted as potential blockers. This ensures that only close opponents significantly obstructing the shot are considered. This approach, similar to the 'Opponents less than 5 meters' variable, captures the defensive aspect, contributing to assess how defensive positioning influences the likelihood of a successful shot.

Future engineering was one of the most critical parts of this thesis. If the methods did not perform well, it would have been a very negative sign for the prediction process. The survey demanded features that would be innovative and effective at the same time. Using unstructured data to provide insights in a demanding field like football was both risky and challenging. However, by forming valuable variables with future engineering process, many new paths could be opened. So, the main two reasons for using feature engineering were: to create something new and to inspire other authors to expand the research in this field.

3.1.3 Data use and validation.

At the first attempt, all the downloaded json files were used with the algorithms for creating the analytics plots and executing the machine learning methods. A process like this demands a lot of both time and memory. Afterwards, the files were split in folds, each containing 200 files. The results derived from each fold were quite the same, with some exceptions that will be discussed in Chapter Four. This method aided the process of extracting the results and helped free up memory to conduct other tests. While conducting further tests and observing the data sets, it was noticed that some data were derived from women's championships and the rest from men's championships. Then, the data were divided into

 https://doi.org/10.26219/heal.aueb.9558

men's and women's data, with both parts containing the same features mentioned before. In this attempt, difference was revealed in the importance of feature affecting the results. 'Distance to Goalkeeper ', 'Velocity' and 'Distance to Goal 'were the variables in which the Random Forest method were affected by this change, but the percentage difference was little, so this data separation was not utilized. Similarly, AdaBoost showed some variation in the same variables, but again the difference was little and not considered significant. 'Opponents in Angle' also showed a significant difference in the AdaBoost model, as one was positive and the other negative. However, despite this contrast, both values were too small to have a meaningful impact on the results..

Finally, the initial 200 files were used for both analytics and machine learning methods. Since the methods performed as expected, these files were utilized to avoid issues of underfitting and overfitting, as well as to eliminate the high bias problems, which are associated with underfitting and occur when a model is too simplistic to capture patterns in data. Though the files seem small in number, the created data was over two thousand rows, indicating a valuable sample for analysis. Therefore, the survey was conducted using these valuable and validated data.

3.2 Analytics.

3.2.1 Tools.

Data analytics requires powerful, flexible and practical tools for analyzing and visualizing the results. From data scraping to visualization, every process must be done accurately to avoid missing important information that will impact the final results. These detailed-oriented processes are time-consuming and affect the whole process. For these reasons, Python programming was applied in the Jupyter environment to produce the final results in the analytics section. An analytical review of these processes is provided in the next two paragraphs.

The environment used for writing the Python's code was Jyputer notebook. There are two main reasons for choosing Jupyter notebook: first, Python is known for its gentle learning curve and ease of use, making it an ideal choice for data analysis. Additionally, the author has already gained experience with programming in Jupyter Notebook, which made the process easier and helped run the analysis efficiently. Second, Jupyter notebook gives the opportunity to execute the code cell by cell. This feature enabled the author to break the code in pieces and clarify each operation. As a result, any issues in the operations could be isolated without affecting the rest of the investigation. This separation has led to the creation of a code very clear and understandable in each of its operations, making it more feasible for use by other researchers.

From data scrapping to plotting the results, everything was done by using the Python. Python is one of the most renowned programming languages for Data Science, Data Engineering and Data Analytics with a vast number of users worldwide. Its variety of libraries enabled programmers to perform many tasks within the same environment, such as loading, visualization, statistics, natural language processing, image processing, and more *(Andreas Müller & Sarah Guido,2016)*. Therefore, for a survey like this thesis, Python was the ideal tool. Matplotlib, math, and scikit-learn are the most powerful libraries used in this thesis, with scikit-learn being extensively used. Therefore, Python was first used to

https://doi.org/10.26219/heal.aueb.9558

parse the files and store the appropriate data in a dictionary for analysis. This data scraping process operates well with Python, making it the most practical tool for this task. A factor that improved the data scraping process was the ability to create functions in Python to automate the data storage. Moreover, Python's contribution in the data engineering process was considerable. Operations such as changing data types, cleaning and transforming data were crucial for converting unstructured data into valuable ones. Then, data was stored into a Data Frame using the Pandas library, enabling the transformation into the features referred above. Finally, Python enabled us to execute different algorithms at the same time and provided a variety of results. By using only, the player's name and the outcome of the shot as inputs, the function could produce a plot with all his shots. At the same time, Python could produce a ROC curve, Features Importance plot, Cross-Validation results, a Classification Report and the Confusion Matrix only using a function. All these were factors led the author to use exclusively the Python.

3.2.2 The process

The first step was to create a function in matplotlib library to plot a pitch based to the coordinates used to store the data for depicting the results. The pitch had coordinates labeled with x = 120 and y = 90 to fit the spatial data exactly. Although the entire pitch was plotted, only the shoots on the right side depicted, as they were stored in this way. Afterward, a function was called to plot the information regarding a player's shot. Essentially, this function served as a base for plotting the shots. Additionally, it allowed to plot for each player the successful shots with different color than those that were not successful. The starting position of the shots was plotted with a circle, the ending with an arrow to show its direction and the direction of the shot with a straight line. In the second part of analytics, the player's movement before the shot was also introduced, with the successful shots being recorded first, followed by the unsuccessful ones before their use.

3.2.3 Target.

Depicting data such as these achieved two goals: firstly, it constituted a very simple and not confusing depiction of the most important events that can happen in a football match, and secondly, it is very understandable and easy to apply. Using only three variables in the first part and four in the second as inputs, a variety of insights can be found. Based on real data, these plots witness all the movements of a player and the tactics used to conduct a shot. These insights could be valuable for players to improve their performance and for opponent's manager to understand their tactics. It should be emphasized; the most important aspect is that these analytics provide realistic and applicable results.

3.3 Machine learning.

Over the past ten years with the big data revolution the machine learning method has become essential in every sector that uses data. Machine learning is the use of computer systems that are able to learn and adapt without following explicit instructions, by using algorithms and statistical models to

https://doi.org/10.26219/heal.aueb.9558

analyse and draw inferences from patterns in data *(www.Google.com)*. In simple words, in machine learning a model which receive data is developing to train a model and to create a model which will be used for predictions.

A depiction of this process is provided in fig 3.5.

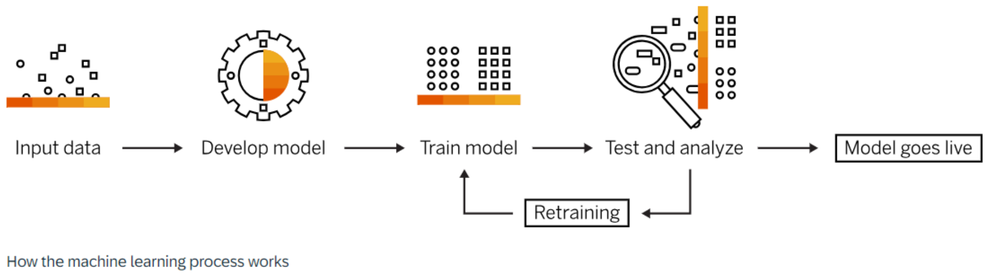


Figure 3.5 the process of Machine Learning, *source www.SAP.com*

They are four types of machine learning methods: Supervised Learning, Semi-Supervised Learning, Unsupervised Learning and Reinforcement Learning. In this thesis, only supervised learning was used, specifically the classification process, to extract the outcomes. Supervised learning can be divided into two types: Classification and regression.

In supervised learning algorithms, the machine is taught by example. Supervised learning consists of input and output data pairs, where the output is labeled with the desired value. By way of an algorithm, the system complies all of the training data over time and begins to determine correlative similarities, differences, and other points of logic until it can predict the answers *(www.SAP.com)*. Supervised learning has two types: Classification and regression. In the classification process, the algorithm uses past data to create classes that share similar traits. Once the model is successfully trained and tested, every new input is analyzed and categorized into the class that is more appropriate. It may not be visible, but the classification technique is used every second in daily life. It is most commonly used in the banking sector, medicine, mailboxes and meteorology. In banking sector, it helps recognize suspicious transactions in accounts. In medicine, it predicts diseases such as if a tumor is malicious or not. In mailboxes, it filters and store mails as spam. As the technology advances, machine learning will continue to gain ground various sectors.

In this thesis, five powerful binary classification models were used.

- Random Forest.
- Logistic Regression.
- Support Vector Machine (SVM).
- AdaBoost (Adaptive Boosting).
- Gradient Boosting.

The concept behind using those models was to create patterns for predicting as accurately as possible the shoots that ended in goals. Those five methods are presented analytically bellow:

3.3.1 Random Forest

Random Forest is essentially a collection of decision trees, where each tree is slightly different from the others. The idea behind random forest is that each tree might is a relatively good job of predicting but will likely overfit on part of the data. If we build many trees, all of which work and overfit in different ways, we can reduce the amount of overfitting by averaging their results. This reduction in overfitting, while retaining the predictive power of trees, can be shown using rigorous mathematics. To implement this strategy, we need to build many decision trees. Each tree should do an acceptable job of predicting the target and should also be different from the other trees. Random forests get their name from injecting randomness into the tree building to ensure each tree different. There are two ways in which the trees in Random Forest are randomized: by selecting the data points used to build a tree and by selecting the features in each split test. *Andreas Müller, Sarah Guido (2016). page 83.* A typical example how Random Forest Algorithm works is presented below in figure 3.6

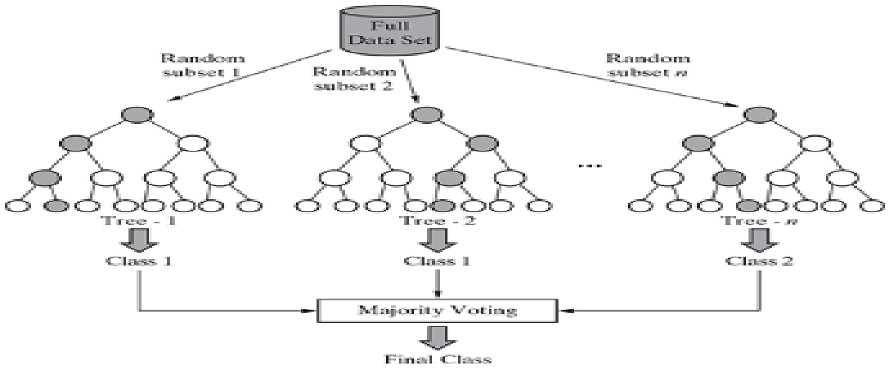


Figure 3.6 Random Forest model, source machine learning, Pearson e-book

3.3.2 Logistic Regression

Logistic Regression is both classification and regression technique depending on the scenario used. Logistic Regression is a type of regression analysis used for predicting the outcome of a categorial dependent variable similar toOrdinary Least Squares (OLS) regression.. In Logistic Regression, dependent variable is binary (0,1) and independent are continuous in nature. The probabilities describing the possible outcomes of a single trial are modelled as a logistic function of the predictor variables. In Logistic Regression a chi-square test is used to gauge how well the Logistic Regression fits the data. The goal of Logistic Regression is to predict the likelihood that the dependent variable is equal to one given certain values of independent. *(Saikat Dutt & Subramanian Chandramouli & Amit Kumar Das ,2019). Machine Learning. Pearson India. page 409, 1st paragraph*. Figure 3.7 show a typical flow of Logistics' Regression model.

https://doi.org/10.26219/heal.aueb.9558

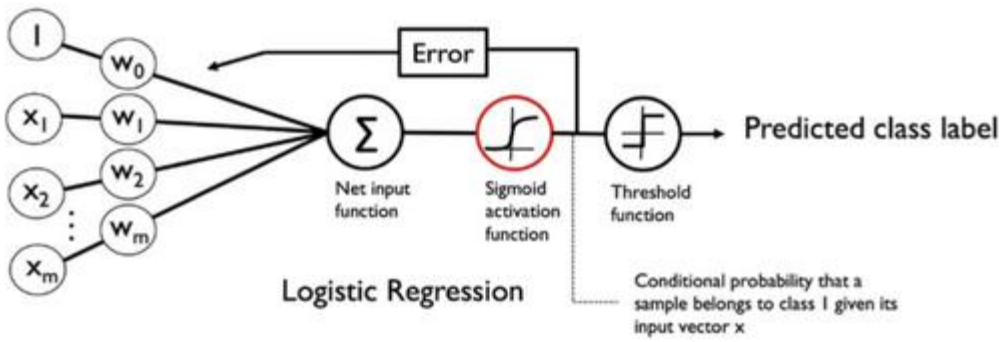


Figure 3.7 the Logistics Regression flow, *source www.Google.com*

3.3.3 SVM

SVM is a model, which can do linear classification as well as regression. SVM is based on the concept of a surface, called a hyperplane, which draws a boundary between data instances plotted in the multi-dimensional feature space. The output prediction of an SVM is one of two conceivable classes which are already defined in the training data. In summary, the SVM algorithm builds an N-dimensional hyperplane model that assigns future instances into one of the two possible output *classes (Saikat Dutt & Subramanian Chandramouli & Amit Kumar Das ,2019). Machine Learning. Pearson India. page 356, 4th paragraph.* Figure 3.8 (a) depict how SVM model categorize data in two-dimensional feature space and (b) multi-dimensional feature space

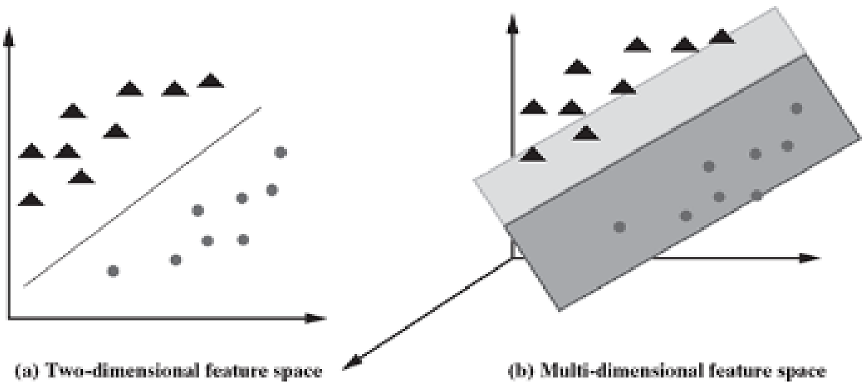


Figure3.8 Linearly separable data instances, *source Machine learning [e-book]*

 https://doi.org/10.26219/heal.aueb.9558

Figure 3.9 below gives a better understanding of SVM model on a 3D dimension dataset

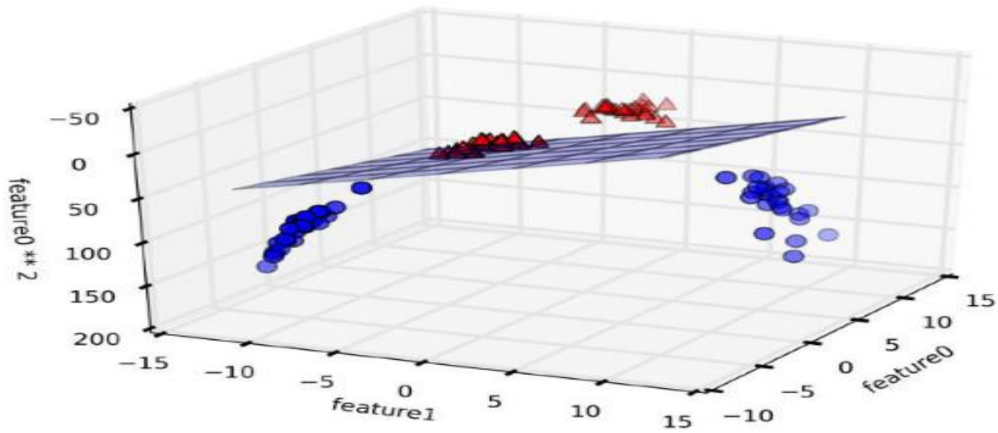


Figure 3.9 SVM model on 3D dimensional dataset, *source introduction to Machine Learning with Python [e-book]*

3.3.4 AdaBoost

AdaBoost ensemble learning technique. It operates iteratively, capitalizing on the errors of weak classifiers – those that achieve just a tad better than mere random predictions – and amplifies their capabilities to mold them into potent classifiers. The essence of AdaBoost lies in its ability to amalgamated several weak classifiers, enhancing the ones that showcase superior performance through higher weight assignments. Although AdaBoost was not crafted explicitly for skewed datasets, its knack for allocating increased weights to underrepresented classes allows it to show promise in addressing CI. This ensures the algorithm zeroes in on accurately categorizing these lesser-represented samples. However, highly skewed class distributions can still present hurdles for the method *(Andrea de Giorgio , Gabriele Cola , Lihui Wang December 2023). page 16, 2nd paragraph.* Figure 3.10 presents a typical structure of AdaBoost model.

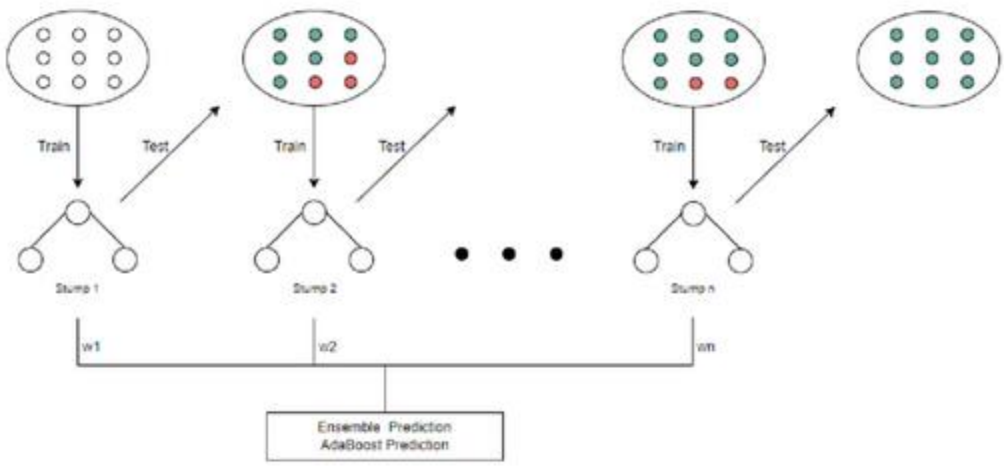


Figure 3.10 AdaBoost classifier, *source* www.Google.com

 https://doi.org/10.26219/heal.aueb.9558

3.3.5 GradientBoost

The Gradient Boost Machine is a generalized version of AdaBoost that is easily adaptable to classification and regression problems. Gradient Boost creates a series of models in the form of a single predictive model. A model in the series is created by building on the previous model's prediction errors in the series. By developing multiple models sequentially, prediction accuracy is increased. Previous models are used in the retrofitting process. Each additional basic model aims to correct the mistakes made. The biggest advantage of boosting methods is to conduct the capability of resampling of instances to give superior results. *(Anil K. Philip, Aliasgar Shahiwala, Mamoon Rashid, Md Faiyazuddin).(2023) page 35-54. Figure 3.11 shows a typical architecture of Gradient Boost model.*

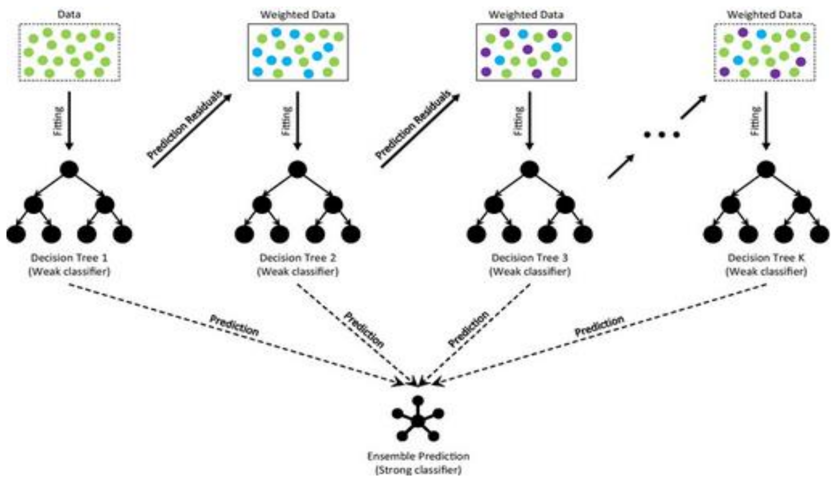


Figure 3.11 Gradient Boost model, source www.Google.com

The target of the applied methods was to achieve the highest true positives scores related to the scored shots. This approach gives more value to shots, consider them as potential goals, and enhances the calculated probabilities of each shot. To achieve this, it was necessary to balance the results because the percentage of the successful shots was comparatively much smaller than that of unsuccessful ones. The first attempt was to use the SMOTE function from imbalanced-learn library. The main function of SMOTE is create new samples very similar to the minority class by using a distance measure. However, using SMOTE the results were not very satisfactory, as the true positive scores were enhanced the true negatives increased significantly as well. Finally, the results were improved by assigning different weights to the outcome classes in each method. The weights assigned to each method were as follows:

- **Random Forest:** 1:20.5
- **SVM:** 1:6.75
- **Logistic Regression:** 1:6.6
- **AdaBoost:** 1:9
- **Gradient Boosting:** 1:9

https://doi.org/10.26219/heal.aueb.9558

These weights were determined after multiple tests to achieve a balance between correctly predicting both successful (Goals) and unsuccessful (No Goals) shots. The weight ratio (e.g. 1 : 6.75 for SVM) represents the balance assigned between the two outcome classes "Goal" (successful shots) and "No Goal" (unsuccessful shots). Specifically, the first number (1) corresponds to the weigh given to the majority class (No Goals), while the second number (e.g. 6.75) represents the weight assigned to the minority class (Goals).

https://doi.org/10.26219/heal.aueb.9558

Chapter 4

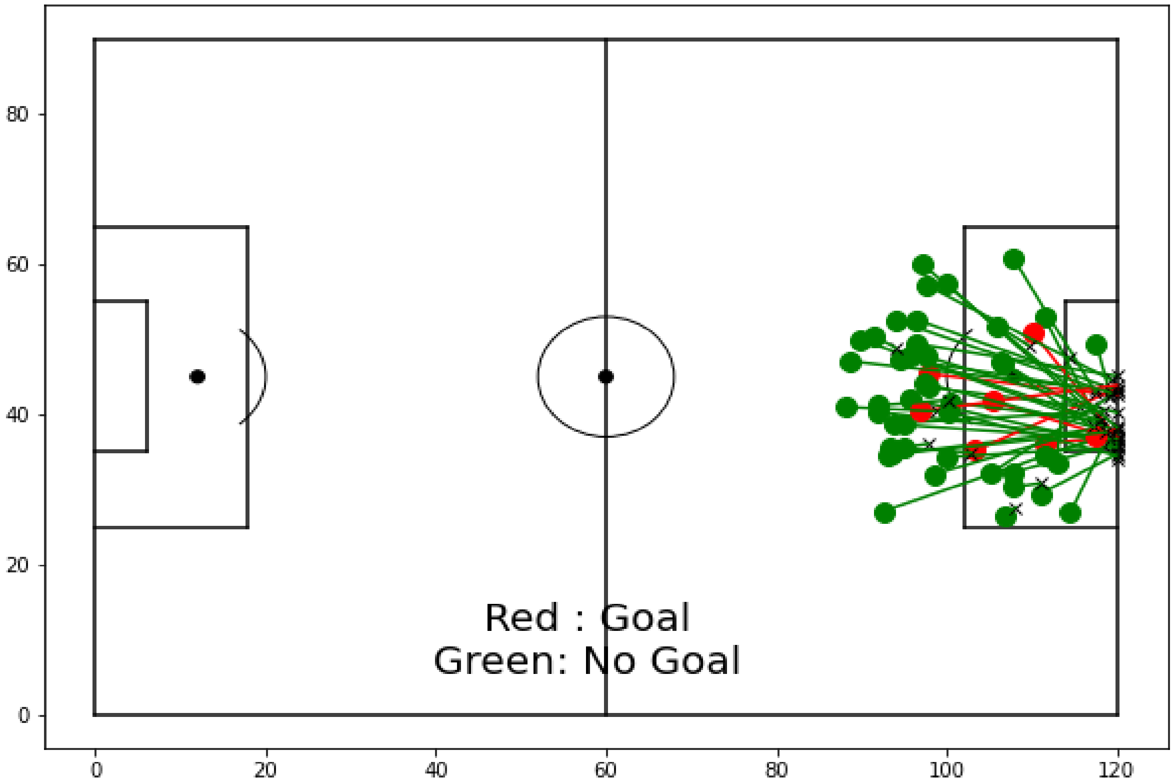
Results.

4.1 Analytics

The first attempt to extract results using analytics methods was conducted similarly to the literature. First, the shots by Messi and Suarez were depicted in plots using only the start and end positions of each shot. Figures 4.1 and 4.2 below present these plots, with figure 4.1 displaying Messi's shots and figure 4.2 Suarez's shots. The successful shots are depicted in Red, while the unsuccessful are shown in Green. Additionally, a straight line connects the start and the end position of each shot to indicate their direction.

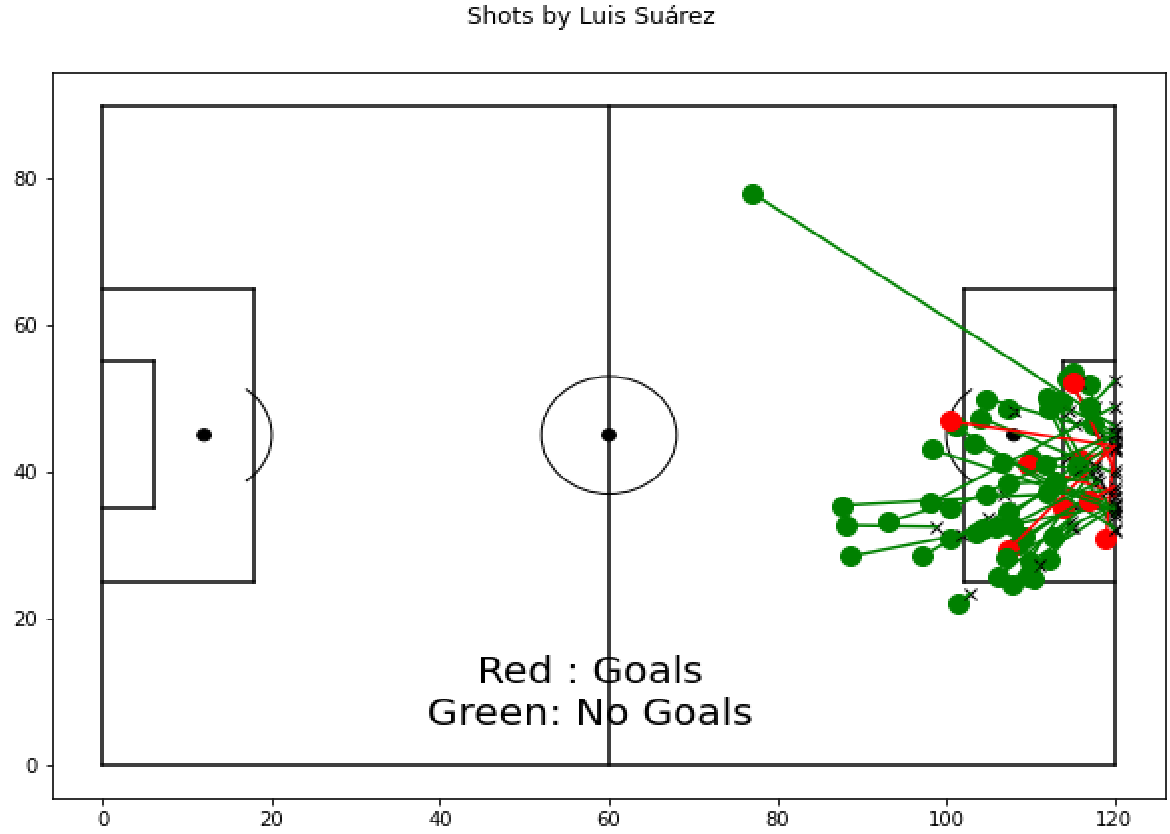
Figure 4.1

Shots by Lionel Messi



https://doi.org/10.26219/heal.aueb.9558

Figure 4.2



As mentioned above, these results are quite similar to the existing literature and were the basis for further research. Adding Velocity and the height of each shot at their final positions could provide additional results, it was not the initial target of this thesis. Moreover, including such details would require fewer shot to have a clear view of each shot. This can be an interesting area for future research.

For going further, the research of this thesis needed to use something both innovating and interesting. The second part of the analysis includes four plots. Figures 4.3 and 4.4 present the successful shots by Messi and Suarez, respectively, this showing their movements with the ball before conducting the shot.

https://doi.org/10.26219/heal.aueb.9558

Figure 4.3

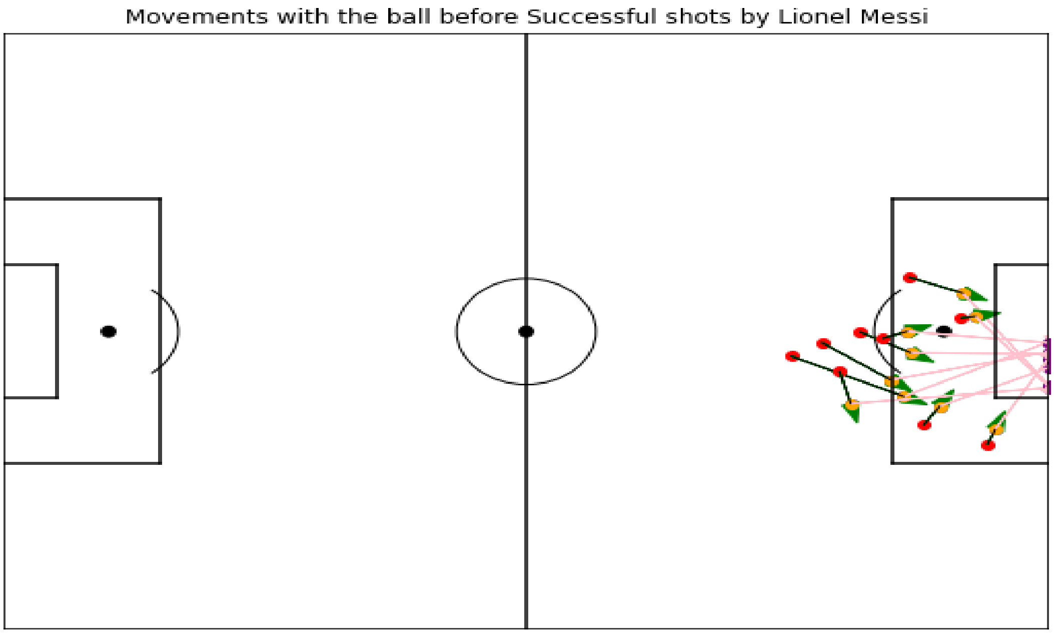
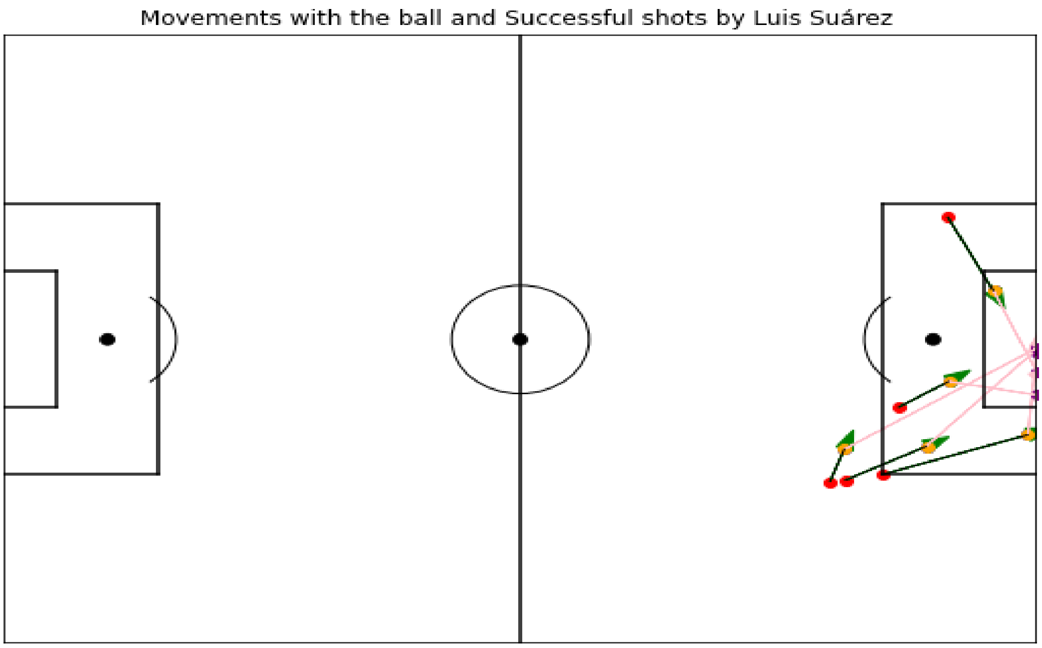


Figure 4.4



In this attempt, fewer shots are depicted for two main reasons: first, there were fewer available data for player's movements and successful shots, secondly, fewer shots were required to provide a clear plot for extracting conclusions. In contrast, figures 4.5 and 4.6 below provide more insights because the available data were much more.

https://doi.org/10.26219/heal.aueb.9558

Figure 4.5

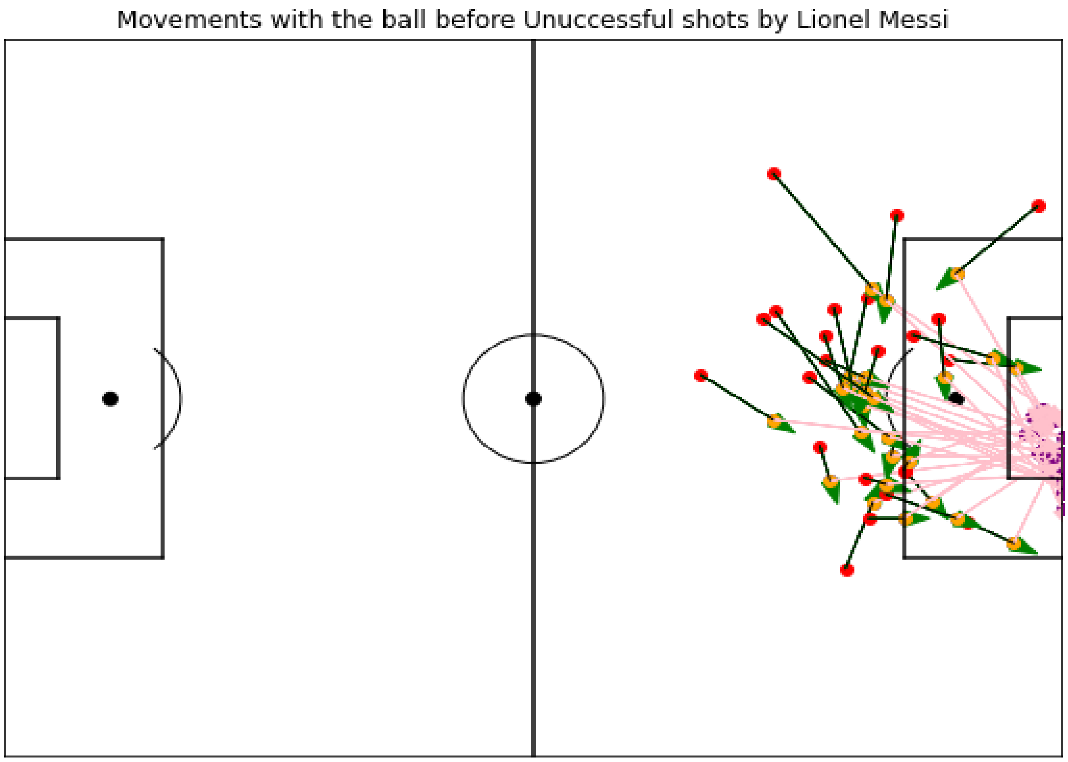
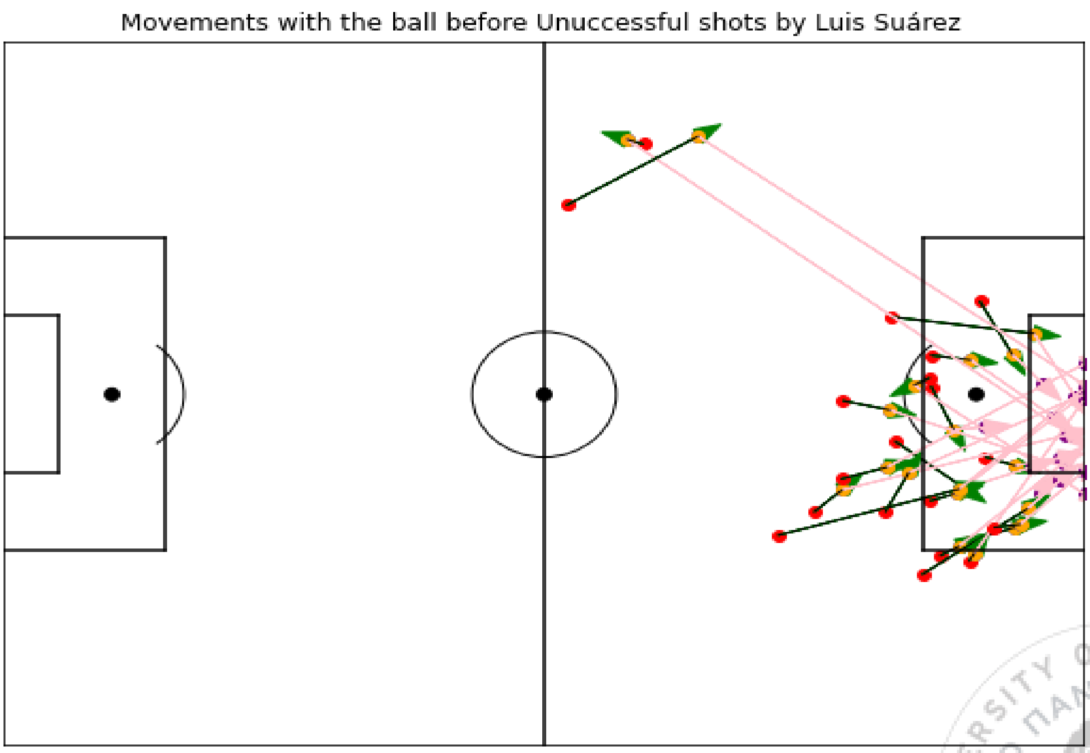


Figure 4.6



https://doi.org/10.26219/heal.aueb.9558

Figures 4.5 and 4.6 above present the unsuccessful shots by Messi and Suarez, respectively, and their movements with the ball before taking the shot. These two pictures provide a clearer view of the shots because the unsuccessful shots were more than the successful ones. The innovative part of this section is incorporating the player's movements before taking the shots. Additionally, the method of plotting is new, using arrows to show the ball's direction and marking its exact end position has not appeared in the existing literature. Messi tend to shoot from outside the penalty area, preferring long-range attempts, with almost all his shots ending on the left side of the goalkeeper, indicating a strong preference for that direction. In contrast, Suarez takes most of his shots from the right side, with a tendency for his shots to end at the center of the goal or the left side of the goalkeeper, suggesting a different approach in finishing. These patterns highlight their shooting styles and positional tendencies, which could be of interest for tactical analysis. Like the analytics in the previous section, these plots have gaps that future researchers could explore. Variables like Velocity and the height of the ball and the ending position of each shot could be added at these plots. However, the limitation of plotting too much information would be still present. Finally, incorporating player's movements with the ball before each shot adds more value to the application of spatial data in football, and creates new areas for investigation.

4.2 Machine Learning

The core of the results from the Machine Learning algorithms were the statistical reports. As mentioned in the third chapter, the main target was to give different weight to the outcomes, treating each goal as a possible Goal rather than calculating the general possibility. The first statistical report is presented in the 4.1 table below.

Model	Accuracy	Cross Validation Score
Random Forest	0.891	0.897
SVM	0.756	0.732
Logistic Regression	0.761	0.760
Ada Boost	0.720	0.896
Gradient Boost	0.751	0.893

Table 4.1 Statistical Report

First of all, the accuracy was calculated to evaluate the overall performance of each algorithm. Although this was not the main focus of investigation, the results were satisfactorily compared with the existing literature. Previous surveys succeeded higher accuracies, but they used simpler variables than those applied in this thesis. The higher accuracy was achieved by the Random Forest Method at 89.1 %, which was not so satisfactory, and this is going to be discussed in the next section. The other three algorithms achieved similar accuracies: SVM at 75.6%, Logistic Regression at 76.1%, Ada Boost at 72% and Gradient Boost at 75.1%.

The next mandatory step was to perform the cross-validation method to evaluate the predictive value of the data. In the existing literature, the cross-validation rarely appears despite it's significant value in investigations. In this survey, the highest scored achieved by the Random Forest method at 89.7%, with small difference from Ada Boost at 89.6% and Gradient Boost at 89.3%. The other two, SVM and Logistic Regression, achieved 73.2% and 76% respectively. Compared to Accuracy, the Cross-Validation score showed more variance, with Ada Boost and Gradient Boost having a difference of more than 10% from their accuracy score.

4.2.1 Classification report

The next statistical report is the classification report. This report was the side rock of the entire machine learning research. The target was to achieve as higher as is possible the recall score for the Successful shots with no negatively affecting the other results in the survey. Table 4.2 presents the classification report.

Clasification Report					
Result	**Precision**	**Recall**	**F1-score**	**Support**	**Method**
No Goal	0.90	0.99	0.94	1051	Random Forest
Goal	0.56	0.11	0.19	131	
No Goal	0.95	0.77	0.85	709	SVM
Goal	0.23	0.63	0.34	79	
No Goal	0.96	0.77	0.85	709	Logistic Regression
Goal	0.26	0.72	0.38	79	
No Goal	0.96	0.72	0.82	709	Ada Boost
Goal	0.22	0.70	0.33	79	
No Goal	0.96	0.76	0.85	709	Gradient Boost
Goal	0.24	0.70	0.36	79	

Table 4.2 Classification Report

As shown in table 4.2 above the classification reports include Precision, Recall, F1-score and Support for shots that resulted in Goals and No Goals. The aim of this thesis was to achieve the highest precision and recall scores for both Goals and No Goals. The precision score for ‘’No Goal’’ and ‘’Goals’ are approximately the same for all the methods, with each algorithm achieving over 90% for ‘’No Goals’’. There is some variation in the precision of ‘’Goal’’, Random Forest method achieved a score of 56% and the rest four less than 26%, with their scores to be approximately the same.

Additionally, the Recall score for both ‘’Goal’’ and ‘’No Goal’’ showed little variation to all algorithms except for Random Forest. Random Forest method achieved a recall of 99% for ‘’No Goal’’ and only 11% for ‘’Goal’’, which explains the high score of accuracy which referred before. This is an unusual compared to the rest of results and to the target of this investigation. Although it was a good predictor for unsuccessful shots, it missed the most successful shots. The other algorithms performed well, achieving scores higher than 70% except for SVM, which reached 63% for ‘’Goal’’.

 https://doi.org/10.26219/heal.aueb.9558

Finally,previous research did not focus much on the recall score as this survey did, making this a valuable result for analyzing the Confusion Matrix.

Figure 4.7

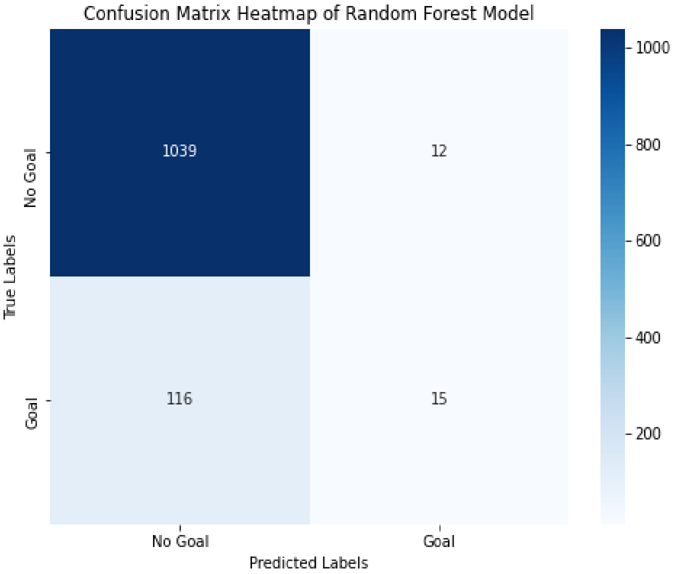


Figure 4.8

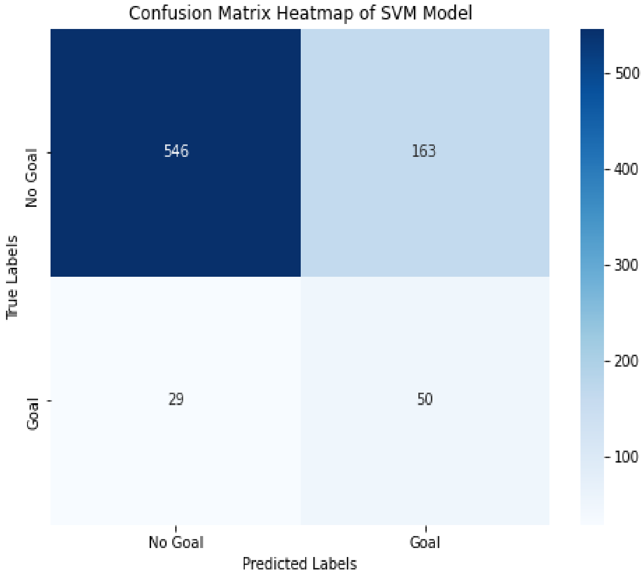


Figure 4.9

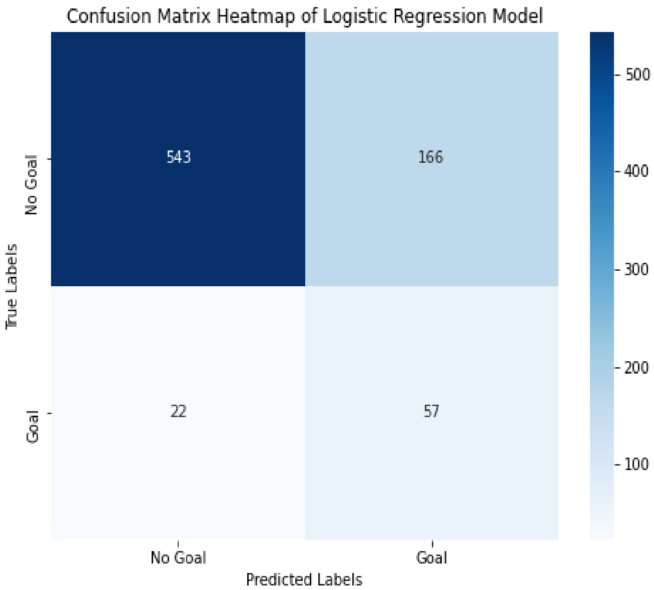


Figure 4.10

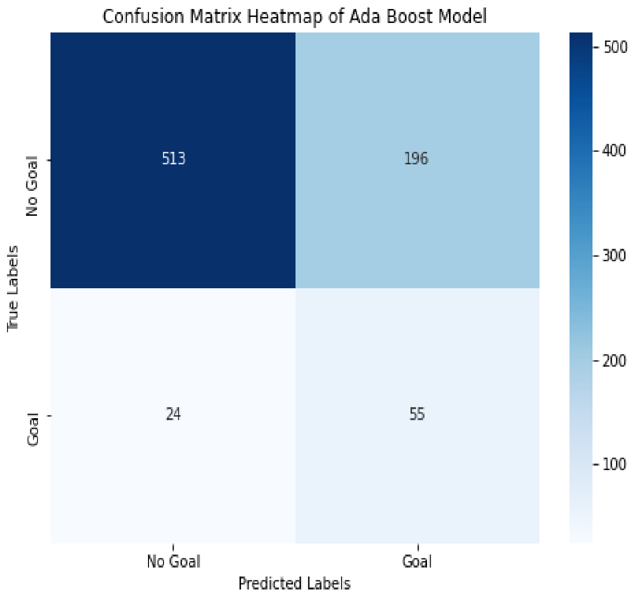
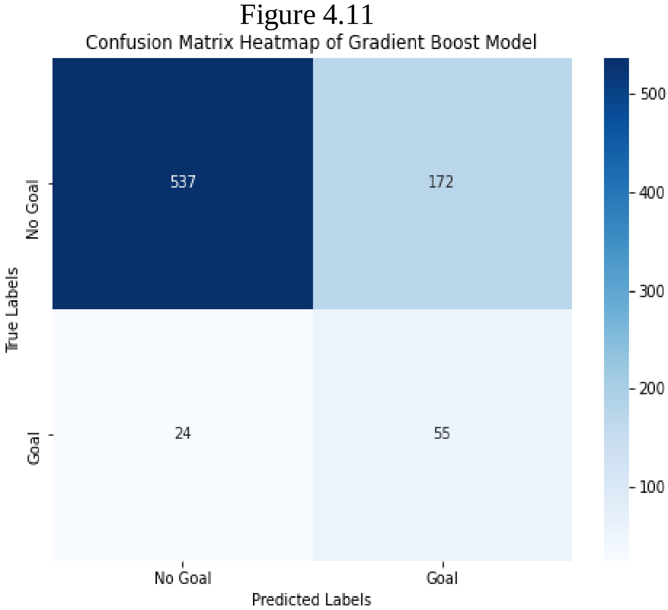


Figure 4.11



https://doi.org/10.26219/heal.aueb.9558

4.2.2 Confusion Matrix

Confusion Matrix is the most well-known report for the classification process of a Machine Learning algorithm. The simple and easy calculation of True/False Positives and True/False Negatives makes it very useful for evaluating an algorithm's performance. Plotting the Confusion Matrix for each algorithm helps better understand the results. The figures above depict the Confusion Matrix for each algorithm. As expected, Random forest showed a poor performance in predicting 'Goal' compared to the other algorithms. The main target was to achieve the highest possible True Positive and True Negative scores for all the algorithms. ''Goal'' considered Positive and ''No Goal'' as negative.

The outcomes were as follows:

- Random Forest to achieve a score of 11% for True Positive and 98% for True Negative.
- SVM 63% for True Positive and 77% for True Negative.
- Logistic Regression 72% for True Positive and 76% for True Negative.
- Ada Boost 70% for True Positive and 72% for True Negative and
- Gradient Boost 69% for True Positive and 75% for True Negative.

The results were satisfactory and presented extensively in table 4.3 for better understanding. All the algorithms performed well based on their Confusion Matrix for both True Positives and True Negative, except Random Forest, which achieved a very high score for True Negatives and a very low score for True Positives. Additionally, the Random Forest showed another anomaly, it revealed more ''Goal'' and ''No goal'' labels compared to the other algorithms, which showed exactly the same True for both ''Goal'' and ''No Goal'' labels. For example, Random Forest revealed 131 True ''Goal'' labels, while the other four algorithms showed only 79.

Confusion Matrix Report				
Model	**T/P**	**F/P**	**T/N**	**F/N**
Random Forest	0,12	0,88	0,98	0,02
SVM	0,64	0,46	0,77	0,23
Logistic Regression	0,72	0,28	0,76	0,24
AdaBoost	0,7	0,3	0,72	0,28
GradientBoost	0,7	0,3	0,75	0,25

Table 4.3 Confusion Matrix Report

4.2.3 ROC Curve.

The ROC curve considers all possible thresholds for a given classifier, but instead of reporting precision and recall, it shows the false positive rate (FPR) against the true positive rate (TPR). Recall that the true positive rate is simply another name for recall, while the false positive rate is the fraction of false positives out of all negative samples. For the ROC curve, the ideal curve is close to the top left: it is preferable a classifier that produces a high recall while keeping a low false positive rate. It means

https://doi.org/10.26219/heal.aueb.9558

that, the True Positive Rate (TPR) is very high while the False Positive Rate very low. *(Andreas Müller & Sarah Guido,2016).*

The ROC curve, like Confusion Matrix, is an indivisible part of evaluating the performance of Machine Learning algorithms. For this investigation, the ROC curve was also evaluated and plotted using Python for its purpose.

Figure 4.12

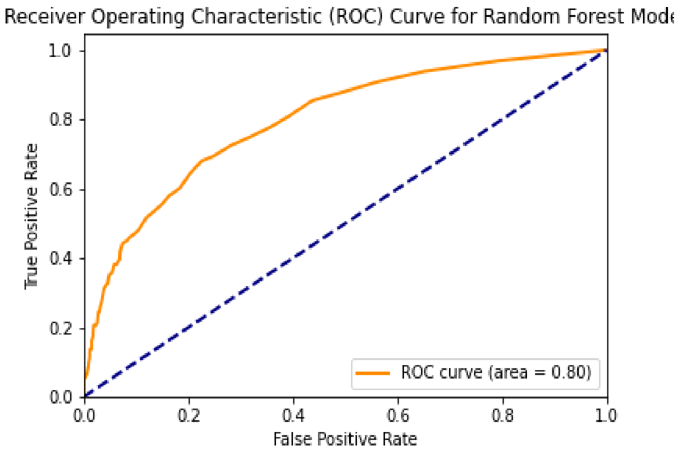


Figure 4.13

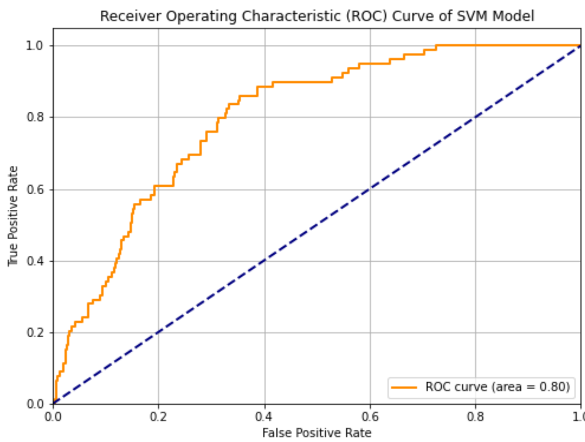


Figure 4.14

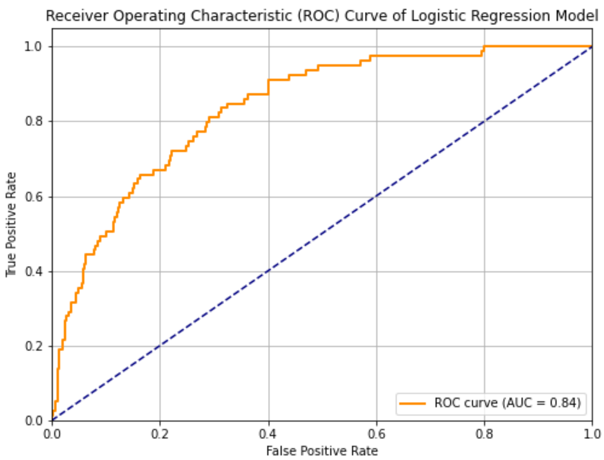


Figure 4.15

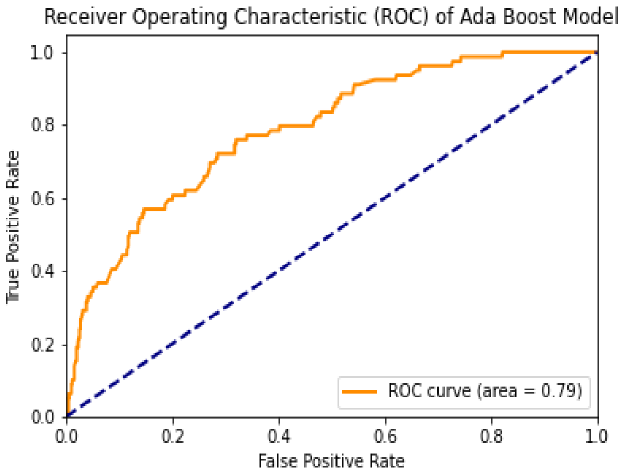
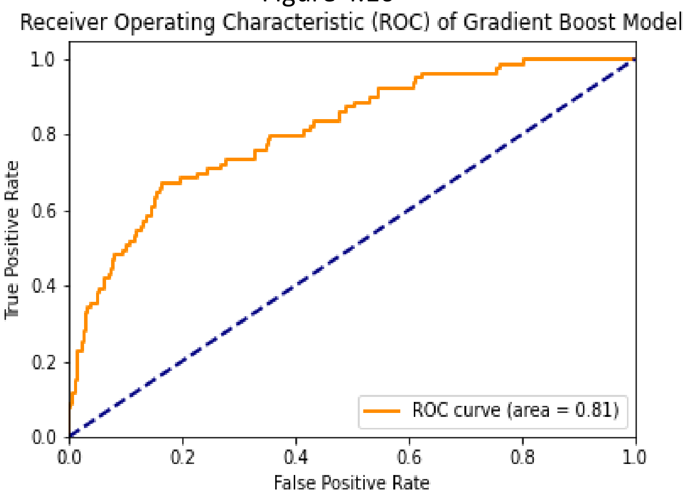


Figure 4.16



The results were quite a bit similar to those of the confusion matrix, with all the algorithms to perform as expected. The outcomes ae presented below. Figure 4.12 presents the ROC curve of Random Forest, figure 4.13 the ROC curve of SVM model, figure 4.14 the ROC curve of Logistic Regression model, figure 4.15 the ROC curve of AdaBoost model and figure 4.16 the ROC curve of Gradient Boost model.

The area under the Roc curve (AUC) was more than 0.8 for all the algorithms except Ada Boost classifier, which achieved a score of 0.79, very close to 0.8, an area which generally considered good for ROC curve value. The specific outcomes were as follows:

- Random Forest achieved a score of 0.8, which was not expected to be so high.
- SVM: 0.8.
- Logistic Regression: 0.84, the highest score, as expected.
- Ada Boost: 0.79 and
- Gradient Boost: 0.81.

The initial goal was to achieve a score higher than 0.85 for each algorithm, but it was not efficient due to the unbalanced sample between ‘’Goal’’ and ‘’No Goal’’. However, the results were considered normal concerning that the True Positive Rate (TPR) was difficult to achieve a score higher than 75%. It should be noted that, in the existing literature, the ROC curve was commonly presented and considered one of the most valuable classification reports for evaluating the performance of each method.

4.2.4 Features Importance.

The final part of the results consists of the plots depicting the degree to which each Feature affected the performance of a specific algorithm in the classification process. For the Logistic Regression, AdaBoost and Gradient Boost algorithms, some variable showed negative importance. A negative importance score indicates that when the model uses this feature, it did not contribute effectively to predict shot outcomes, and usually lead to misclassifications. In contrast Random Forest and SVM algorithms, assigned only positive importance values, because they measure features contributions differently, prioritizing patterns that improve prediction accuracy.

In Tree-based models, features show in most cases a positive importance which reflects their impacts to the predictive power. On the other hand, Linear Models, can have both positive and negative future importance, representing the relationship between the variable and the positive outcome. This investigation includes Tree-based models: Random Forest, Ada Boost and Gradient boost, as well as Linear models: The Logistic Regression and SVM. Though, AdaBoost and Gradient Boost are Tree-based algorithms, they have shown negative score for ‘’Distance to goal’’ feature. This can occur in specific cases, which will be discussed in the fifth chapter. The plots below depict the Importance of each variable for all the algorithms:

https://doi.org/10.26219/heal.aueb.9558

Figure 4.17

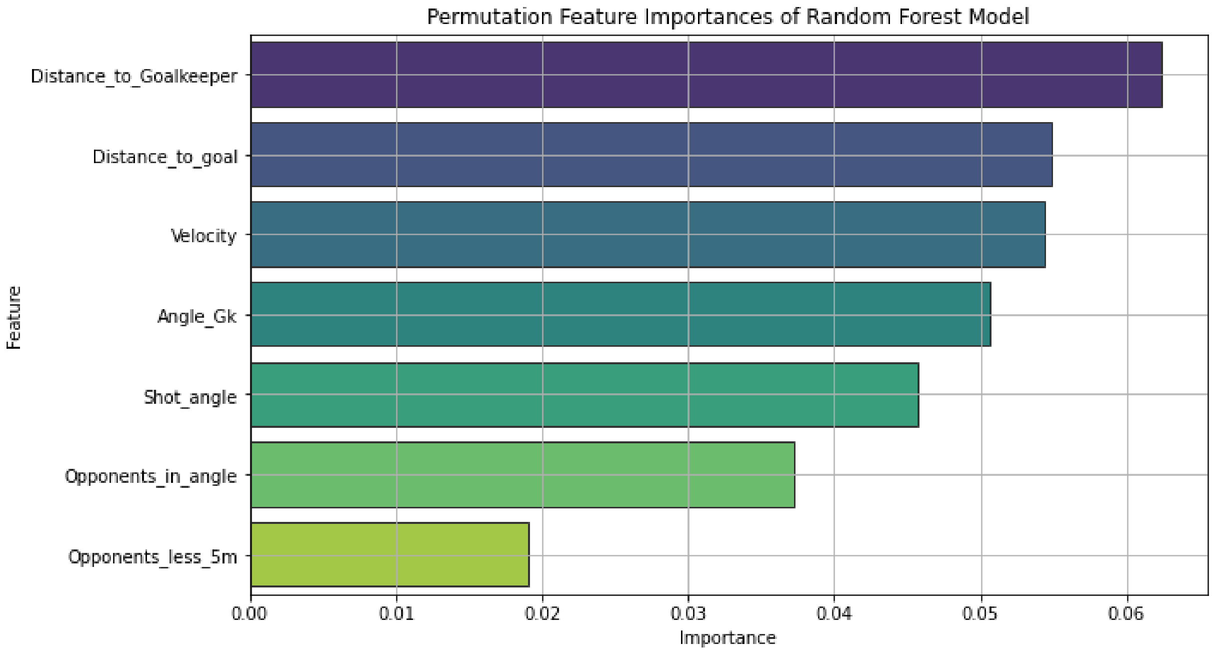


Figure 4.18

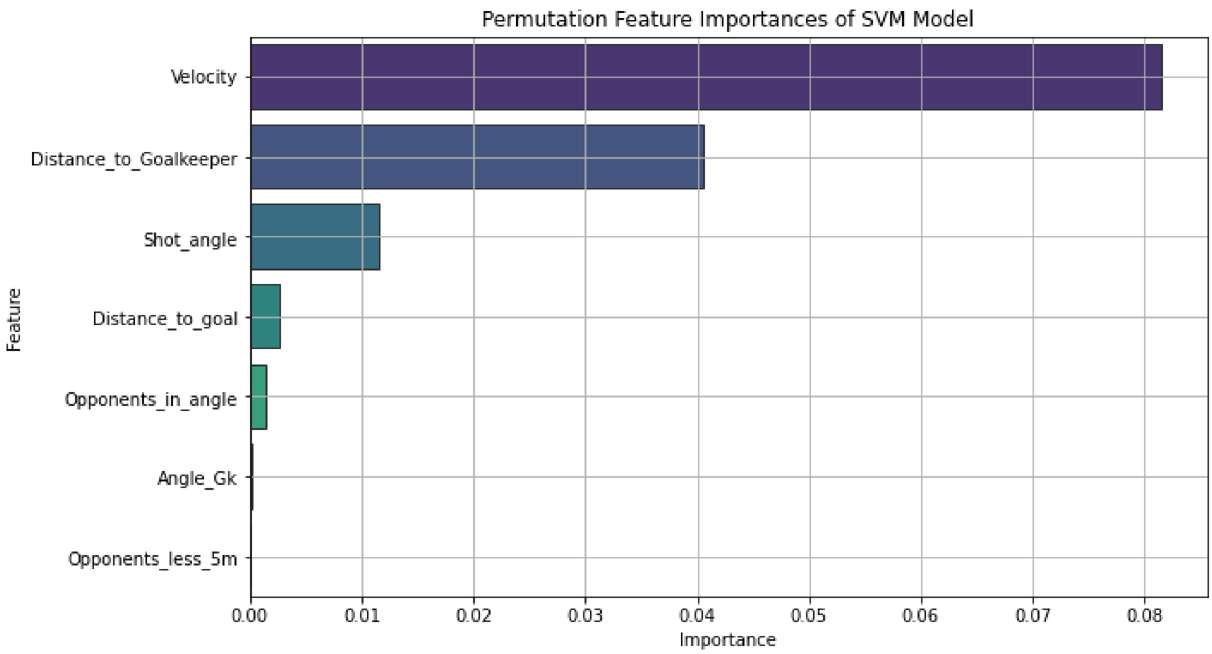


Figure 4.19

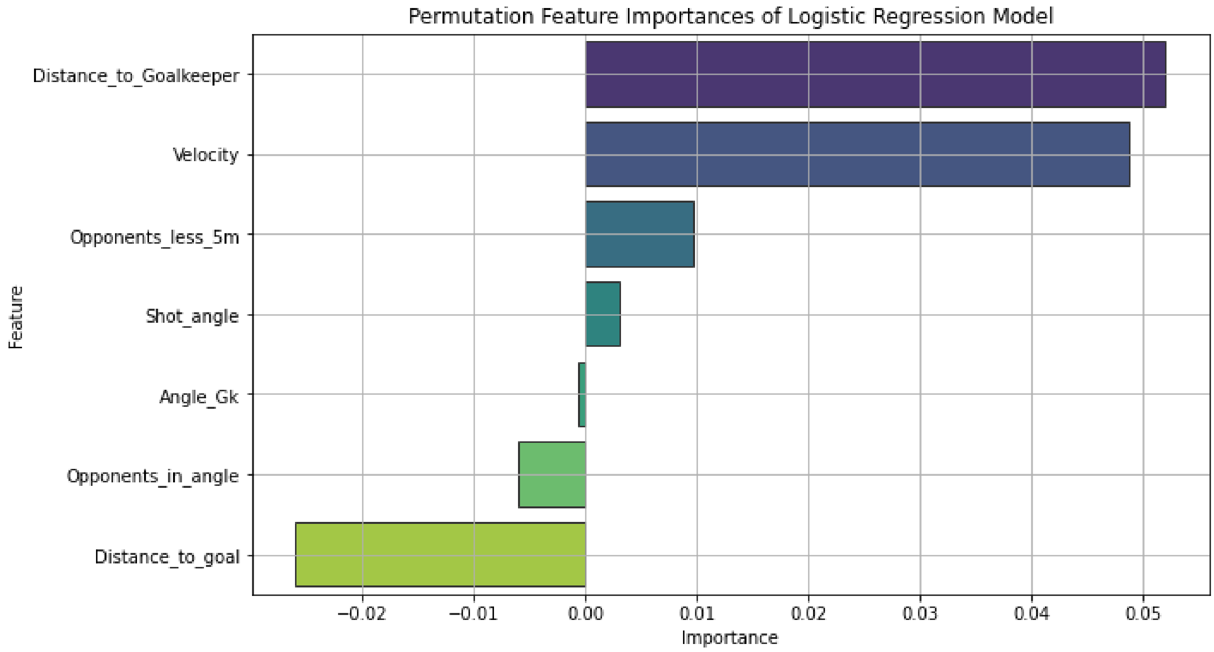
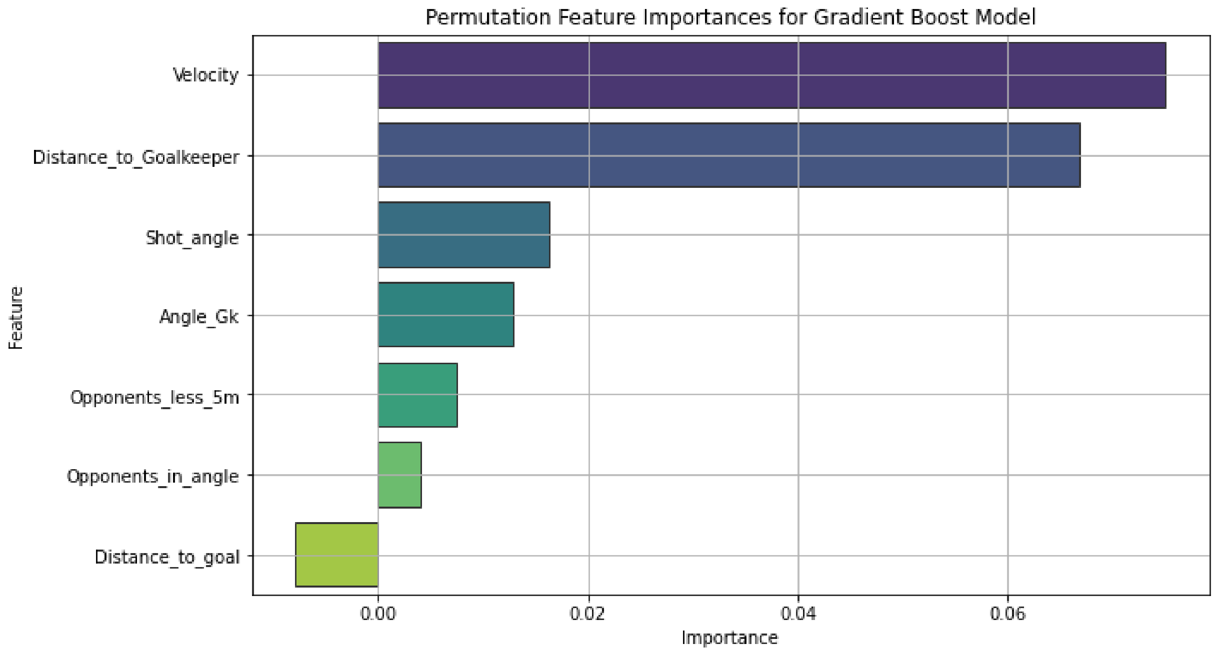
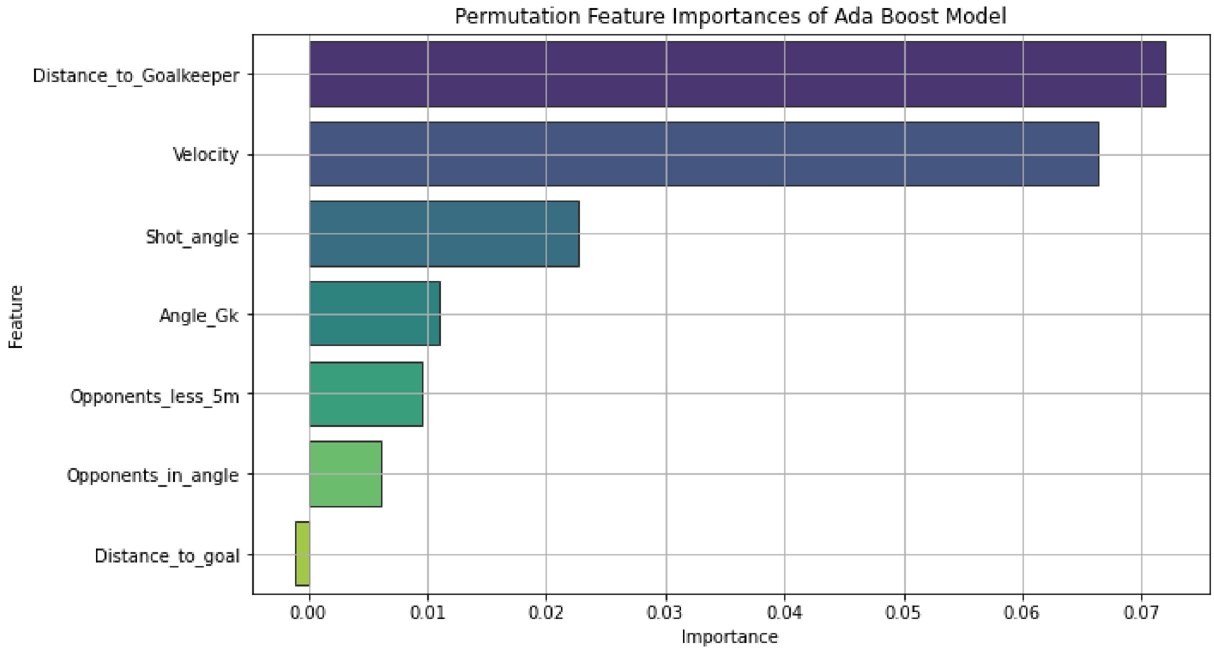


Figure 4.20



 https://doi.org/10.26219/heal.aueb.9558

Figure 4.21



Additionally, table 4.4 below presents thoroughly all the future importance scores for each algorithm:

Future Importance Report					
Future	**Random Forest**	**SVM**	**Logistic Regression**	**Ada Boost**	**Gradient Boost**
Distance to Goalkeeper	0,061	0,041	0,052	0.072	0,064
Velocity	0,05	0,081	0,049	0.067	0,076
Shot Angle	0,046	0,012	0,003	0.03	0,017
Angle GK	0,051	0,001	-0,001	0.011	0,013
Opponents in Angle	0,038	0,003	-0,006	0.006	0,04
Opponents less 5m	0,019	0,000	0,009	0,009	0,006
Distance to Goal	0,054	0,004	-0,026	-0,002	-0,008

Table 4.4. Future Importance Report

The evaluation of each Future importance is the final and most important part of this survey, for two main reasons: first, it assesses the performance of each variable, which is fundamental and impacts significantly the results, second, this gave the opportunity to be presented the importance that can have the spatial data in football. The conclusions regarding future importance are presented below:

- **Distance to Goalkeeper** and **Velocity** are the most influential feature across all models. This indicates that the proximity of the shooter to the goalkeeper and the speed of the shots significantly impact the likelihood of scoring a goal.
- **Shot angle** also has notable impact, particularly in Random Forest and AdaBoost, showing its role in shot success.

 https://doi.org/10.26219/heal.aueb.9558

- The importance of **velocity** is highest in SVM (0.081) and Gradient Boost (0.076), suggesting these models rely more on the ball's speed for prediction.
- **Distance to Goalkeeper** has the highest importance in AdaBoost (0.072), meaning this model emphasizes the shooter's positioning relative to goalkeeper.
- **Distance to Goal** has negative importance in Logistic Regression (-0.026),AdaBoost (-0.002) and Gradient Boost (-0.008), indicating these models do not consider it a key predictor.
- **Angle GK** and **Opponents in Angle** show very low or even negative importance in certain models(e.g. Logistic Regression). This suggests that these factors may not contribute or could even introduce noise in predicting shot success.
- **Opponents less 5m** has close to zero influence across models, indicating defender proximity does not affect the prediction.
- Random Forest and Gradient Boost give similar importance to all features. SVM focuses mostly on Velocity, while Logistic Regression show negative importance to some features because the do not have an impact on scoring outcomes.

Considering the two reasons above, the new variables can open new areas for more investigation in the future and enhance the value of football analytics. Finally, most authors in the previous literature evaluated the Future Importance to highlight the degree that each variable affected their methods. However, they assessed commonly used variable, whereas this thesis applied innovative variable.

 https://doi.org/10.26219/heal.aueb.9558

Chapter 5

Discussion

5.1 Machine Learning.

Based on the previous literature, the results presented in Chapter four were as expected. The initial goal was to achieve a recall score of approximately to 70%, in each method, particularly for the succeeded goals. Based on the existing literature, accuracy scores in football shot prediction range between 65% and 80%. In this study the results fell within this range, with models achieving accuracy between 72% and 76% except Random Forest which achieved 11% in predicting Goal. This issue likely stems from Random Forest ´s tendency to overfit on this type of data. However, its results were included to provide additional insights and comparisons within the study.

The accuracy for all the algorithms was approximately the same and over 70%, except for Random Forest. This indicates that the data used were valid and the algorithms perform well with spatial data. However, in the cross-validation test, AdaBoost and Gradient Boost algorithms showed a difference of about 14%, which is considered significant and reflects an improvement in the performance of the algorithm. The different data samples used in the cross-validation process showed varying results, suggesting that certain patterns or distributions within the data had a stronger influence on the models in some folds than in others, leading to fluctuations in predictive accuracy. In the ROC curve, the results were clearer, with all the algorithms achieving an Area Under the Curve (AUC) of approximately 0.8. While it is usual for the ROC curve to reach a score more than 0.9, based on the complexity and the unbalanced samples, a score of 0.8 is considered as satisfactory. Finally, the results of feature Importance were the most interesting, and a bit unpredictable for two main reasons: first, most variables were newly created, and second, they were selected based on the author's experience. ''Velocity'' and ''Distance to Goalkeeper'' were the two most influential variables across all algorithms. Additionally, the ''Shot Angle'' variable proved to be influential as well, but not in the same degree as the two before. On the other hand, ''GK Angle'', ''Opponents Less 5m'' and ''Opponents in Angle'' din not considerably influence the algorithms. The most interesting variable was ''Distance to Goal''. In Random Forest algorithm, it was the second most influential variable, while in the others algorithms was approximately zero and negative in two of them. If all the reposts could provide the exact impact, positive or negative, the conclusions would be more specific, probably to predict specific tactics used by teams. As expected, ''Velocity'' and the ''Distance to Goalkeeper'' played a crucial role in shaping the results and in allowing this thesis to reach its goal.

This part has two significant limitations, the complexity of the data and the unbalanced sample. These issues could be avoided if the outcomes became more specific and the data were derived from individual teams, players or championships, as each team and player have distinct tactics and playing

https://doi.org/10.26219/heal.aueb.9558

styles. This would help reduce complexity and balance the dataset by considering different playing styles and tactics. This is a valuable opportunity for future researchers and data Engineers to develop new methods and techniques to simplify the process.

5.2 Analytics.

Although these results were first presented in Chapter four, their discussion will be presented last due to their interest. The method applied was very simple as well as practical, making it directly applicable. Despite its simplicity, the insights gained could be highly valuable. Both managers and players from any club can use a similar method to analyze and extract conclusion from the outcomes. in previous research, it was very common to plot shots and passes with their exact coordinates, but this thesis took a different approach by using players' ball-carrying before each shot.

In the initial results, it is evident that both Messi and Suarez employed specific tactics in order to send the ball toward the right Post and at the center of the Goalposts. Messi tends to shoot from various positions in contrast to Suarez who shoots in most cases from the right side of the of the penalty area. This makes Messi an unpredictable shooter and Suarez very predictable. In the second analysis, incorporating their movements with the ball before shooting, their actions became more predictable. Both of them tend to follow a specific pattern, Messi often moves from left to right, while Suarez from right to left. This predictability makes them more vulnerable to opposing defenders. The final shot positions were similar to the first analysis, with both players aiming the right Post and the center of the goal.

This part had two main limitations: first, incorporating of the distance each player covered before each shot, along with their exact movement, could have provided more valuable results in the survey. Additionally, this would have offered more insights and made the findings easier to understand. Second, the tools din not allow to incorporate the height at the end point of each shot. It demands more advanced tools to plot this in 3D and separately to avoid confusion from the accumulated information. Summarizing, these limitations can be a significant opportunity for future researchers to further the investigation.

5.3 Summary

Both methods provided satisfying results while achieving their target. The outcomes derived from the analytics were simple and easy to understand, making them practical and applicable for everyone involved in football. On the other hand, the results from the machine learning methods were more specialized. Although they require a deeper understanding and prior knowledge, they provided valuable insights for analyzing shots in football. Contrary to previous research, the methods used in this thesis add

https://doi.org/10.26219/heal.aueb.9558

more value to shots by considering them as potential goals rather than evaluating their overall possibility of scoring.

In line with the aim of this thesis, the results are considered successful. The two main targets were reached: first, it introduced an innovative approach to evaluate shots from both an Analytics and Machine Learning perspective, and second, this innovation opened new areas in two aspects: the application of new tools and further football analysis. Considering all aspects of this thesis, previous research contributed to its inspiration, while their gaps provided a chance to advance football analysis.

https://doi.org/10.26219/heal.aueb.9558

References

Books

[1] Saikat Dutt & Subramanian Chandramoul & Amit Kumar Das, 2019. Machine Learning [eBook version]. Pearson India.

[2] Andreas Müller & Sarah Guido,2016. Introduction to Machine Learning with Python: A Guide for Data Scientists [eBook version]. O'Reilly Media, Incorporated, 2018

[3] Tom M. Mitchell,1997. Machine Learning [eBook version]. Publisher: McGraw-Hill Science/Engineering/Mat. ISBN: 0070428077

[4] Alex Smola and S.V.N. Vishwanathan,2008. Introduction to Machine Learning [eBook version]. Self-published by Alex.Smola.org

[5] Anil K. Philip, Aliasgar Shahiwala, Mamoon Rashid, Md Faiyazuddin,2023. A Handbook of Artificial Intelligence in Drug Delivery. Paperback ISBN: 9780323899253. eBook ISBN: 9780323903738

[6] Shai Shalev-Shwartz and Shai Ben-David,2014. Understanding Machine Learning: From Theory to Algorithms [eBook version]. Published by Cambridge University Press.

[7] Marc Peter Deisenroth ,A. Aldo Faisal, Cheng Soon Ong.2020. Mathematics For Machine Learning. published by Cambridge University Press

[8] John D. Kelleher & Brendan Tierney. (2018). Data Science [eBook version]. ISBN: 9780262535434. Publisher: The MIT Press

Thesis

[9] Omar El Yousfi,2021.Design and Implementation of a Football Data Analysis Application. University Mohammed First, Multidisciplinary Faculty of Nador, Computer Science department, Nador – Morocco

[10] Brad Kyoshi Donohoo, 2012.Machine Learning Techniques for Energy Optimization In Mobile Embedded Systems. Colorado State University. Department of Electrical and Computer Engineering

[11] Francisco Henriques,2018. How Can Big Data Help Football Clubs Achieve Competitive Advantage. Católica Lisbon School of Business & Economics

Articles

[12] Salil Chincholikar, Niharika Sadul, Parth Thakkar, Ashutosh Amrutkar, Mr. Nikhil Dhavase. (2020). PREDICTIVE ANALYTICS OF FOOTBALL. International Research Journal of Engineering and Technology (IRJET). e-ISSN: 2395-0056. p-ISSN: 2395-0072

[13] O. I. Aladesote, O. Agbelusi & M. Ganiyu. (2020). Predicting Football Match Results with Data Mining Techniques. Journal of Computer Science IJCSIS.

[14**]** Kerr, Matthew G. S. (Matthew George Soeryadjaya) .(2015). Applying machine learning to event data in soccer. Massachusetts Institute of Technology. URI http://hdl.handle.net/1721.1/100607

 https://doi.org/10.26219/heal.aueb.9558

[15] Wisdom, Conor and Javed, Aqib,(2023) Machine Learning for Data Analytics in Football: Quantifying Performance and Enhancing Strategic Decision-Making. Available at SSRN: https://ssrn.com/abstract=4558733 or http://dx.doi.org/10.2139/ssrn.455873 (Football Hackers: The Science and Art of a Data Revolution) Christoph Biermann 2019

[16] Andrea de Giorgio, Gabriele Cola, Lihui Wang. (2023). Systematic review of class imbalance problems in manufacturing. Journal of Manufacturing Systems 71:620-644 71:620 644 DOI:10.1016/j.jmsy.2023.10.014

[17] James Mead , Anthony O'Hare , Paul McMenemy.(2023). Expected goals in football: Improving model performance and demonstrating value . PMID: 37018167: PMC10075453 DOI:10.1371/journal.pone.0282295

[18] Rahul Baboota , Harleen Kaur.(2018). Predictive analysis and modelling football results using machine learning approach for English Premier League. International Journal of Forecasting. DOI:10.1016/j.ijforecast.2018.01.003

[19] Muhammad Yonus Saiedy, Muhammad Aslam HemmatQachmas , Dr. Amanullah Faqiri (2020).Predicting EPL Football Matches Results Using Machine Learning algorithms. International Journal of Engineering Applied Sciences and Technology. ISSN No. 2455-2143, Pages 83-91. Published Online July 2020 in IJEAST (http://www.ijeast.com)

[20] Victor Chazan Pantzalis, Christos Tjortjis (2020). Sports Analytics for Football League Table and Player Performance Prediction. Published in International Conference on 15 July 2020. DOI:10.1109/IISA50023.2020.9284352

[21] Yash Ajgaonkar Kunal Bhoyar Prof. Anagha Patil Jenil Shah .(2021).Prediction of Winning Team using Machine Learning. ISSN: 2278-0181.Published by www.ijert.org NTASU - 2020 Conference Proceeding

[22] James H. Hewitt, Oktay Karakuş.(2023) A machine learning approach for player and position adjusted expected goals in football (soccer). https://doi.org/10.1016/j.fraope.2023.100034

[23] Teng Jia, Jirarat Sitthiworachart , John Morris.(2024) Application of Simulation Technology in Football Training: A Systematic Review of Empirical Studies.

DOI: 10.2174/011875399X277947231228071109

On Line Sources

[24] StatsBomb’s Post for Open Play Built-up in English Premier League: www.linkedin.com
https://www.linkedin.com/posts/statsbomb_with-the-202425-premier-league-season-around-activity-7230147538245357568-yTAm?utm_source=share&utm_medium=member_android

[25] Understanding xG — Part 2: Medium official website. www.Medium.com

https://medium.com/@ishdeepsinghchadha/understanding-xg-part-2-d74d17f62fef

[26] Understanding Expected Goals — xG: Medium official website. www.Medium.com

https://doi.org/10.26219/heal.aueb.9558

https://medium.com/@ishdeepsinghchadha/understanding-expected-goals-xg-4da7bba61e4c

[27] Understanding Position Based Expected Threat(xT): Medium official website. www.Medium.com

https://medium.com/@ishdeepsinghchadha/understanding-expected-threat-xt-part-1-ac3c6198512e

[28] Google official website: www.Google.com

[29] ChatGPT, 2024: www.chatgpt.com