A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF ÇANKIRI KARATEKİN UNIVERSITY

PREDICTING THE VALUE OF FOOTBALL PLAYER WITH THE IMPACT OF COVID-19 ON THE MARKET VALUE OF THE PLAYER

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
ELECTRONICS AND COMPUTER ENGINEERING

BY

HUSAM HASAN ATIYAH ATIYAH

ÇANKIRI

2023

PREDICTING THE VALUE OF FOOTBALL PLAYER WITH THE IMPACT OF COVID-19 ON THE MARKET VALUE OF THE PLAYER

By Husam Hasan Atiyah ATIYAH

July 2023

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science

Advisor : Asst. Prof. Dr. Seda ŞAHİN

Examining Committee Members:

Chairman : Prof. Dr. Mehmet Reşit TOLUN
Software Engineering
Çankaya University

Member : Asst. Prof. Dr. Seda ŞAHİN
Computer Engineering
Çankırı Karatekin University

Member : Asst. Prof. Dr. Mustafa TEKE
Electrical and Electronics Engineering
Çankırı Karatekin University

Approved for the Graduate School of Natural and Applied Sciences

Prof. Dr. Hamit ALYAR
Director of Graduate School

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Husam Hasan Atiyah ATIYAH

ABSTRACT

PREDICTING THE VALUE OF FOOTBALL PLAYER WITH THE IMPACT OF COVID-19 ON THE MARKET VALUE OF THE PLAYER

Husam Hasan Atiyah ATIYAH
Master of Science in Electronics and Computer Engineering
Advisor: Asst. Prof. Dr. Seda ŞAHİN
July 2023

The unprecedented onset of the Covid-19 pandemic has pervaded countless industries worldwide, with football being a major one to bear its brunt. This global health crisis ushered in a new era where normalcy was challenged and socio-cultural constructs were altered, leading to profound implications on football, both as a sport and an industry. Football, a nexus of economic and cultural intersections, witnessed several changes, most notably the effect on player valuations in the market. The pandemic brought life to a standstill, ensnaring societies in its clutches and forcing them into isolation. This sudden cessation had a ripple effect on football players, amplifying feelings of anxiety, tension, and uncertainty, thus altering their psychosocial dynamics. Consequentially, this shift in psychological state, combined with the disruption of regular football activities, contributed to fluctuations in the market prices of these athletes. Traditional approaches and methodologies, once relied upon for gauging player prices, particularly in the five major European football leagues (English, Spanish, Italian, German, and French), began showing signs of obsolescence in the face of these pandemic-induced challenges. So, this study presents the performance assessment of five machine learning algorithms—Linear Regression, Lasso Regression, Random Forest, Gradient Boosting, and K-Nearest Neighbors—on a regression problem. The evaluation metrics utilized are Mean Absolute Error (MAE) and Root Mean Absolute Error (RMAE). Linear Regression falls short in MAE and RMAE. Lasso Regression exhibits the highest prediction errors, hinting at possible overfitting. Random Forest provides a balanced outcome with moderate errors. Gradient Boosting stands out in terms of prediction accuracy, having the smallest MAE and RMAE. Meanwhile, K-Nearest Neighbors parallels Linear Regression's performance.

The study suggests that Gradient Boosting might be the most accurate for predictions based on MAE and RMAE. However, the ideal model selection is contingent upon the specific objectives of the regression problem, which could prioritize minimizing prediction errors.

2023, 67 pages

Keywords: Machine learning methods, Covid-19, Football player value prediction, Gradient boosting

ÖZET

FUTBOLCUNUN DEĞERİNİ COVID-19'UN OYUNCUNUN PİYASA DEĞERİNE ETKİSİ İLE TAHMİN ETME

Husam Hasan Atiyah ATIYAH
Elektronik ve Bilgisayar Mühendisliği, Yüksek Lisans
Tez Danışmanı: Dr. Öğr. Üyesi Seda ŞAHİN
Temmuz 2023

Covid-19 pandemisinin eşi benzeri görülmemiş başlangıcı, dünya genelinde sayısız sektörü etkisi altına aldı, futbol da bu darbenin büyük bir kısmını üzerine aldı. Bu küresel sağlık krizi, normalliğin zorlandığı ve sosyo-kültürel yapıların değiştirildiği yeni bir dönemi başlattı, bu da futbola, hem bir spor hem de bir endüstri olarak derin etkiler yarattı. Ekonomik ve kültürel kesişimlerin bir merkezi olan futbol, özellikle piyasadaki oyuncu değerlemeleri üzerindeki etkisi olmak üzere birkaç değişikliğe tanık oldu. Pandemi, yaşamı durma noktasına getirdi, toplumları pençesine alarak onları izolasyona zorladı. Bu ani durma, futbol oyuncularına sıçrayarak kaygı, gerginlik ve belirsizlik duygularını artırdı ve böylece onların psikososyal dinamiklerini değiştirdi. Sonuç olarak, bu psikolojik durum değişikliği, düzenli futbol faaliyetlerinin kesintisine eklenerek, bu atletlerin piyasa fiyatlarında dalgalanmalara neden oldu. Özellikle beş büyük Avrupa futbol ligi (İngiliz, İspanyol, İtalyan, Alman ve Fransız) için oyuncu fiyatlarını ölçmede bir zamanlar güvendikleri geleneksel yaklaşımlar ve metodolojiler, bu pandemi kaynaklı zorluklar karşısında eski olduğunu göstermeye başladı. Bu nedenle, bu çalışma, bir regresyon problemi üzerinde beş makine öğrenimi algoritmasının -Lineer Regresyon, Lasso Regresyon, Rastgele Orman, Gradyan Arttırma ve K-En Yakın Komşular- performans değerlendirmesini sunmaktadır. Kullanılan değerlendirme metrikleri Ortalama Mutlak Hata (MAE) ve Kök Ortalama Mutlak Hata (RMAE)'dir. Lineer Regresyon, MAE ve RMAE'de yetersiz kalıyor. Lasso Regresyonu, en yüksek tahmin hatalarını göstererek olası aşırı uyuma işaret ediyor. Rastgele Orman, ılımlı hatalarla dengeli bir sonuç sağlıyor. Gradyan Arttırma, en küçük MAE ve RMAE ile tahmin doğruluğunda öne çıkıyor. Bu arada, K-En Yakın Komşular, Lineer Regresyon'un

performansına paralel ilerliyor. Çalışma, Gradyan Arttırma'nın, MAE ve RMAE'ye dayalı tahminler için en doğru olabileceğini öne sürüyor. Ancak, ideal model seçimi, tahmin hatalarını minimize etmeye öncelik verebilecek regresyon probleminin özel amaçlarına bağlıdır.

2023, 67 sayfa

Anahtar Kelimeler: Makine öğrenimi yöntemleri, Covid-19, Futbolcu değer tahmini, Gradyan artırma

PREFACE AND ACKNOWLEDGEMENTS

I would like to thank my thesis advisor, Asst. Prof. Dr. Seda ŞAHİN, for her patience, guidance and understanding.

Husam Hasan Atiyah ATIYAH
Çankırı-2023

CONTENTS

ABSTRACT i
ÖZET iii
PREFACE AND ACKNOWLEDGEMENTS v
CONTENTS vi
LIST OF ABBREVIATIONS viii
LIST OF FIGURES ix
LIST OF TABLES x
1. INTRODUCTION 1
1.1 Problem Statement 2
1.2 Aim of Thesis 3
1.3 Structure of Thesis 3
2. LITERATURE REVIEW 5
3. BACKGROUND AND THEORY 24
3.1 Implementıng ML Models 24
3.1.1 Linear regression 26
3.1.2 Lasso regression 27
3.1.3 Random forest 29
3.1.4 Gradient boosting 30
3.1.5 K-nearest neighbors 31
4. MATERIALS AND METHODS 33
4.1 Dataset 33
4.2 Programming Environment 35
4.3 Model Design 36
4.4 Data Source 37
4.5 Data Pre-processing 38
4.5.1 Cleaning Data 40
4.5.2 Standard Scaler 41
4.5.3 Data Normalization 43
4.6 Feature selection 45
4.7 Data Split 45

4.7.1 Training set **45**
4.7.2 Validation set **46**
4.7.3 Test set **46**
4.7.4 Train/test split **47**
4.7.5 Cross-validation **47**
5. RESULTS AND DISCUSSION **50**
5.1 Experimental Setting **50**
5.2 Regression with Base Parameters (First Form) **50**
6. CONCLUSION AND FUTURE WORK **55**
REFERENCES **56**
APPENDICES **61**
CURRICULUM VITAE **67**

LIST OF ABBREVIATIONS

ANOVA	Analysis of variance
A-W	Attacking work rate
Att	Attendance
BP	Best position
Covid-19	Coronavirus disease 2019
DT	Decision tree
DEF	Defending
D-W	Defending work rate
DRI	Dribbling
Ea	Electronic arts
EPL	English premier league
FK	Free kick
GLS	Generalized least squares
GK	Goal keeping
FIFA	International federation of association football
KNN	K-nearest neighbors
LOOCV	Leave one out cross validation
LR	Linear regression
ML	Machine learning
MAE	Mean absolute error
OLS	Ordinary least squares
PAS	Passing
RF	Random forest
RMSE	Root-mean-square deviation
R2	R-squared
UEFA	Union of european football associations
WHO	World health organization

LIST OF FIGURES

Figure 2.1 The annual value of the Europe team before and at Covid-19 (Franck and Nüesch 2012) 7
Figure 2.2 Aggregated value change for the top player (Andrea 2021)..... 8
Figure 2.3 The annual value of the Europe team before and at COVID-19 (Andrea 2021). 9
Figure 2.4 Professional football club source of outcome and income (Beheshti *et al.* 2022). 12
Figure 2.5 Decision tree flow chart (Beheshti *et al.* 2022). 13
Figure 3.1 Diagram of the base machine learning algorithms 25
Figure 3.2 LR algorithm (Kavita 2021) 26
Figure 3.3 Lasso regression algorithm (Kumar 2021) 28
Figure 3.4 RF algoritham (Kumar 2021) 29
Figure 3.5 K-nearest neighbors algorithm (Saikumar 2023) 31
Figure 4.1 Diagram of the dataset collection 34
Figure 4.2 Diagram of the proposed system 36
Figure 4.3 Screenshot of the raw data 37
Figure 4.4 Screen shot of the data description 38
Figure 4.5 Screenshot of reading the data set from files 38
Figure 4.6 Pre-processing steps..... 39
Figure 4.7 Screen shot of the two files combination..... 39
Figure 4.8 Diagram of train-test-split (Michael 2022)..... 47
Figure 4.9 Diagram of data spilt cross-validation (Prashant 2021) 49
Figure 5.1 The regressor mean absolute error for algorithms..... 52
Figure 5.2 The regressor root mean square error for algorithms 52
Figure 5.3 The regressor score with machine learning algorithm method..... 53

LIST OF TABLES

Table 2.1 Related literature review 23
Table 3.1 Parameters optimization of LR 27
Table 3.2 The parameters optimization of the lasso regressor 29
Table 3.3 Parameters optimization of the RF regressor 30
Table 3.4 Parameters optimization of grediant boosting 31
Table 3.5 Parameters optimization of k-nearest neighbors 32
Table 4.1 Description of original dataset 35
Table 4.2 Parameters optimization of standard scaler 43
Table 4.3 Parameters optimization of data normalization 45
Table 5.1 Libraries used in model implementation 50
Table 5.3 Results based on algorithms 51
Table 5.4 Evaluation of the studies in the literature 54

1. INTRODUCTION

One of the sensitive fields in the world of the economy is football measured by the number of participants and spectators and football is the most popular game in the world. The turnover of European football clubs alone was estimated at 27 billion US dollars in 2017.

It thus makes an important contribution to the global economy. Demand for football stars has skyrocketed over the past few decades, with players being valued at over €100 million. These numbers are well above historical trade numbers for the normal rate of inflation. From a manager's point of view, the most important decision that football clubs have to make choose a player. Player transfers have a significant impact on a club's chances of success. For this reason, researchers from various fields have studied the factors that affect money transfer fees. Recently, researchers have started paying close attention to players market values. A player's market value is the estimated amount for which a team could sell a player's contract to another team. While transfer fees represent prices actually paid in the market, market values are estimated transfer fees and therefore play an important role in transfer negotiations. Professional footballers such as team managers and sports journalists have historically estimated market value, while crowdsourcing sites such as Transfermarkt (www.transfermarkt.com) have proven helpful in estimating market value in recent years. However, data-based methods for estimating the market value of football are not yet widely used (Mustafa *et al.* 2022).

Also, the market transfer of these fields is having a huge impact on the team's future so the prediction of the value for the football player is important and the COVID-19 pandemic change accuracy of prediction that has been done before and in these research, we will restructure. The COVID-19 pandemic is one of the major global issues faced by health organizations. As of November 19, 2020, the total number of people worldwide confirmed to have been infected with SARS-COV-2 is more than 56.4 million, while the total number of fatalities from the coronavirus is more than million, thereby proving that COVID-19 cases are surging worldwide (Bryson *et al.* 2013).

Review the general characteristics of a soccer player's predictions and show the player's characteristics in terms of both physical and statistical characteristics. Age is one of the important things that will be an indicator of exemplary gratitude because it reflects both commitment and achievement (Behravan and Razavi 2021).

Stock price forecasting is a task that counts as a difficult one because it is highly dependent on the demand for the stock and acutely there is not a variable that shows or helps to predict accurately daily demand for stocks. However, the Efficient Markets Hypothesis (EMH) states that stock prices are also highly dependent on the information those new, and many of those sources of information are the opinions of people on a platform such as social media. And those in social media the user opinions about a player in football or a product that a company produces can determine a company's reputation or player's reputation and influence public decisions to make them decide to buy stock in that company or the value of the player in the transfer market which counts as same nowadays. An appropriate analysis is required when using opinions as primary data. One of the best examples that I can say of using opinions as an analysis of data and one of them Sentiment analysis is the process of determining people's opinions about something, in this case, some company's products. There is some research on predicting stock prices with sentiment analysis. Bollen's research says that user opinions on social media, such as Facebook TikTok twitter any kind of platform, can predict his DJIA score of 87 above accuracy it makes this idea obvious that there is a strong link between these two most people use age as a way to measure a show's rating. Keep in mind that player ratings tend to rise until the mid-20s and then begin to decline. Apart from that, it has been found that the size of the player leads to a decisive increase in salary returns. It shows a large header capacity, which makes it more likely to score or anticipate goals. Another characteristic that has been examined in player valuation inquire.

1.1 Problem Statement

The problem arise with prediction of the football player price when the COVID-19 happened which cannot depend on the pervious way of prediction to predict the value of football player and the reason for that is the player price depend on the economy of the

clubs in the league and from 2000 to 2018 have quite constant value so for that reason pervious research didn't count it as feature to use in the prediction but thing changed when COVID-19 happened so for that reason our research focus to predict during COVID-19 and other timeline by bring new feature to the prediction which fan attendance because fan-attendance will decide most percentage of the economy of club. The annual value of the Europe team is very high and my research aim is predicting the value of football player in any type of drop down or constant in annual value if it happen will my prediction make accurate prediction for the football player value and the newest example we have the COVID-19 which lead to drop down the annual value of the European team.

1.2 Aim of Thesis

This thesis proposes applying the cluster model to predict the market value of soccer players to extract information from the data on a SOFA website during a pandemic. Five traditional machine learning algorithm models were used in this thesis (Linear, Lasso, Gradient Boosting, KNN and Random Forest (RF)) In this study two steps were achieved:

- First step, all data were used to predict the market value of players without the option of specific player features
- In the second step, unlike other methods used

Literary, linear and nonlinear approaches to predicting market value were tested.

The dataset of all models is used and compared, and the main purpose of this thesis is to improve prediction of the market value of football players during the Corona epidemic.

1.3 Structure of Thesis

The summary of this report is prediction of Covid-19 impact based on market value of FIFA (The Federation international de football association) players by using machine learning, so the chapters are divided as following:

- Chapter One provided an overview of the proposed project with objectives based on existing problem. Besides, some related works were reviewed.
- Chapter Two previews the related work
- Chapter Three discusses the theoretical idea and the proposed technology with brief explanations.
- Chapter Four shows the results of datasets analysis and then discuss it.
- Chapter Five concludes the research study with some recommendations and future scope.

2. LITERATURE REVIEW

The performance of the parameters has an influence on the value of football players in general but this research that they have done was in German Bundesliga. Hereby, for example, they show that not only do we have to use a normal or a causal way to predict the market value what they mean by that all the research uses the common feature and they mention should be using something like relative operating times or movement of the player on a stadium in the analysis their method they analyzed all players in the season 2015/16 whichever player that at least played once and that season that they bring the data. Because the form of the functionality is not clear in the football market value and other research before not have been cleared this idea accurately that is why the connection of the value of football players in the market to the impacting those variables, they have done the analysis using Boosted Regression Trees to be able to treat or understand the different scale levels and nonlinearities. They found the influence their value is impacted by some factor and those are money that comes from TV with high ranking in the preseason and number of goal and high accurate crossing the analysis that they have done show that the graph shows that impact of a variable on the market value of players is not linear. In their perspective the value of football players the one that they choose is the German league they think that depends significantly less on sports performance than other research showed in this cause they think that it has less influence on their value but the research showed what we mean by that is their research the player that preselected in those club that famous which those are every season have to impact in the league the player on those teams causes a clear difference in the market values of the individual players. Also, some variable linearity is little between variable which mean by that cause their not that much linearity in the variable but this means in their perspective that other research has done a lot of their research based on a linear method to predict their value, and those variable that are linear which is often the year of birth of player in another word his age (Müller *et al.* 2017).

Bhravan and Razavi have created a model of machine learning. Their reasoning Using a hybrid regression approach - a combination of particle crowd optimization (PSO) and support vector regression (SVR). These results show that their strategy has clear

advantages over other strategies for assessing a soccer player's show rating, group factors for soccer player ratings, player position and they analyzed the impact. Consistent with their emergence, relapse studies have found that group level, the month of birth, alliance, playtime, and player age affect a player's promotional value. They also played in midfield attacks and showed that the players born in the first quarter of the year were the most important. In agreement with the author's opinion, the RMSE and MAE of the strategy were 2,819,286 and 711,029.413, respectively, and the displayed results were 5,793,474 and 3,241,733, respectively (Behravan and Razavi 2021).

Müller *et al.* (2017) presented different strategies for assessing gamers' ratings. They used a dataset that included various characteristics such as player characteristics (age, position, and nationality), player performance, ubiquitous, and so on. They analyzed the player's rating and the subsequent impact of various variables. Apart from that, the author clarified the limits of crowdsource evaluation strategy in the article. They examined the impact of various components in transferring player ratings and determine the most important components. In this inference, he used data from 150 well-known attackers captured in Transfemarkt.de and adopted a GLS strategy (least squares generalization) to determine key factors. Based on his findings, the number of goals, assists, group-wide ratings, and FIFA rating priorities influenced the ad ratings of attacking players. Take the player's role as a positive point of his contemplation and focus on it.

Stanojevic and Gyarmati (2016) have shown a method for assessing the advertising value of 12,858 players based on player execution information. We created several models using targeted learning and player execution information collected from the website transfer market. UK and the sports analysis company Instant. These models were built using 45 indicators. The results show that the created show is superior to the widely used transfermarkt.com show scorer in anticipation of group execution.

One of the studies presented a strategy for assessing or assessing player-based transfer performance based on five common variables (age, accuracy, win, confirmation, and adaptability) and It is seem to conflict with the player's feedback rating. Therefore, the

main limitation of this study is its reliance on community reviews. These may be biased or lack comprehensive knowledge. The title was crafted by her hand (Herm *et al.* 2019).

Also, one of the researchers investigated the effect of the talent and popularity of players on its value. They tried to measure the talent of players using 20 criteria. Using the OLS regression model, they conclude that the popularity of players increases their market value (Herm and Callsen 2019).

And as we know the COVID-19 have been impacted the annual, and it was not of their advantage at all the problem with COVID-19 in and it shown in the Figure 2.1 football like we can say for marketing it impacted but in same the trade done online so the flow not stopped but in the football area that problem not able to fix because the football cannot be containing online or anything like that so in beginning of the pandemic the team fan will not be allowed.

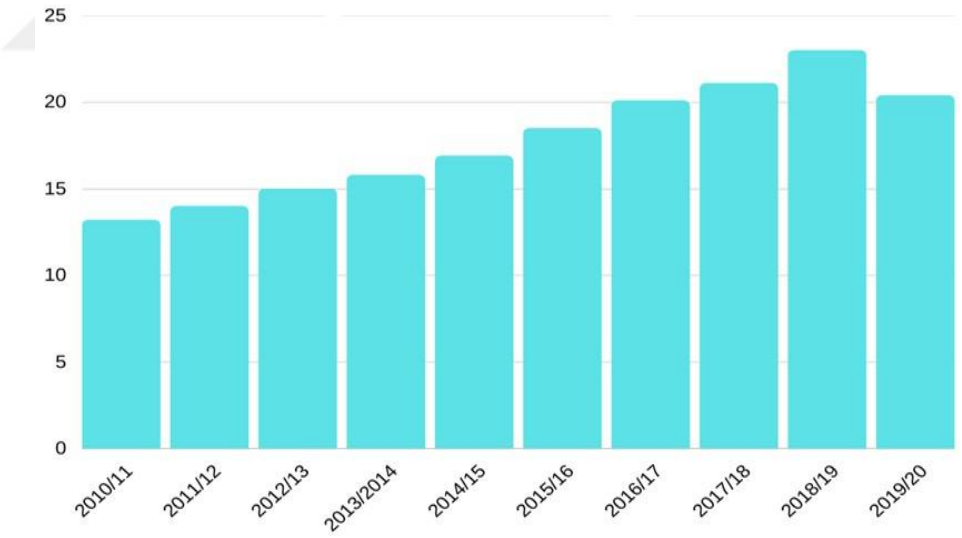

Figure 2.1 The annual value of the Europe team before and at Covid-19 (Franck and Nüesch 2012)

Fans were not able to go to or in another word attend to the game for that manner the football team profit have been decreased and also after pandemics get intense the football have been stop for while so the prediction of the football player was not easy or the

research before that have been done these type of prediction and their prediction will struggle because they are not count these pandemic in another word that we can say is that they did not predict that if fan not attending for the studied a lot other factor, but that will impact the prediction of the market and in the Figure 2.1 shows the annual for COVID-19 and before (Franck and Nüesch 2012). And the other impact of COVID-19 on the football player is that the amount they not able to play their stamina and their physicality level decrease and that leads to more injury when football comes back and there a lot of football players get impacted and the player value change also so this is another factor for changing the value of football players in Figure 2.2, we can see that the average football player value change in 2020.

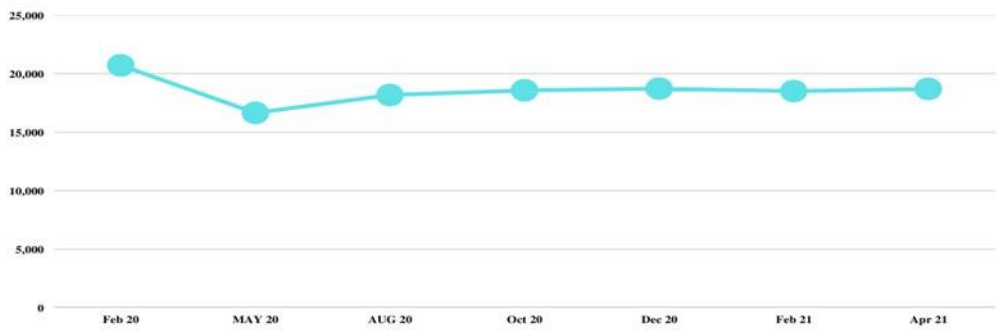

Figure 2.2 Aggregated value change for the top player (Andrea 2021)

So, we aim to build a prediction that will count or care about the fan attending and the football physical trait and also different factor like injury of the player like we said before one of the parts that will impact the value as well and so on the main idea is that we study what we did from the statistic that we got from a different resource which leads for a new factor that will impact the value of the football player so what we want to do is the prediction method will count this new factor or in another word new feature so that the prediction accuracy will increase also during the any pandemic

All European football players have been affected by COVID-19, with UEFA in particular reducing revenues by 21% to € 3 billion, reducing the associated budget burden. The UEFA gave development rates to partial accessions, continued to broadcast to clubs, and

reimbursement to broadcast and commercial accomplices was stunned in the future. seasons and why this research is important or in other words why we should care about this research in the first place the amount of money in the football annual market is huge we can show that in Figure 2.3 for the year 19 to 20 and 18 to 19so that that that year one of the years the annual has been dropped but still the amount is high (Andrea 2021).

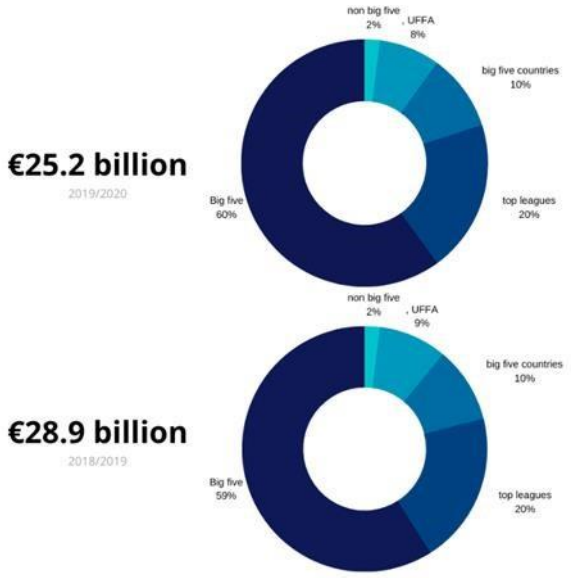

Figure 2.3 The annual value of the Europe team before and at COVID-19 (Andrea 2021).

So, this research among other research shows money that these fields how much important to team player value in Figure 2.3, we can see about 3 billion euros has dropped and the team in mid of COVID-19 does not realize that they should lower the player budget because their market value will drop in those time or another word the team like Barcelona that the player contract had a high budget and that why they impacted when COVID-19 came up to show the profit of the football team for 2010 and that what makes it clear why these type of research is important for that field (Andrea 2021).

The stagflation that happens to the economy impact business in a high way which makes the price of goods in higher price and the amount of output lower in this case the global economy is impacted as well and the service same as goods are impacted negatively and all that lead to the bad economy in the world in general and we can put that in three. Different part that first one supply side which includes the goods and service and the other

one side of the demand the interest rate got lower and that's a downside for the economy, so what we can observe is that football not related directly to the COVID-19 because people was still watching football from tv but the reason the fan was not attending the game which clubs make out most of these money of this and commercial stopped (García *et al.* 2016).

Other predictions that have been done so far did not pay attention to the world economy because it was consistent with the problem when COVID-19 happened it lower the economy has been discussed before and this led to lower player market value and the prediction that before they depend on the factored like the skill of the player age, etc... so the prediction will not be accurate if we only depend on this feature for prediction cause the player value not only related to his skill and the personal attribute it depends on the clubs economy so as well the clubs economy depend on the fan which the main source of the club and statistic show that especially in the deloitte website can see that clearly also football has more fan in the world compare to the basketball and any other sport in football the cup that in Europe, the prize will get higher when the fan is more attendant it to that cup (Thomas *et al.* 2021).

COVID-19 continues to pose a threat to the population due to its ongoing spread. It has warned of a global health crisis since its first instance of human infection. COVID-19 has already led to the death of a lot of humans we can say a million human lives (Who 2020), prompting many countries in the world to take action to contain the spread of COVID-19. The speediest virus government measure is distancing in the social, a way of maintaining distance between people to make less spread of the COVID-19 virus). This idea of people being far away from each other measures have a great impact on people's daily life as they are analogous to quarantines in many areas. For example, most countries in the world, especially in Europe, have closed places such like educational places schools, and sports in general. Some countries have gone into complete lockdowns, so their people to not go outside their homes, and almost all governments have banned events most likely the public ones including football we can say the sport of all kinds and we can observe it was also noted that organizations of all kinds are suffering from the pandemic, this crisis that lead to problem in their economy and in another word sports

clubs have a major impact on the economies of many countries. Football is currently the largest participation, impact and revenue sport in the world, impacting not only the football market and major sports sectors but also the social, and cultural sectors. With revenues collapsing due to the impact of COVID-19, an elite football club is struggling to contain the economic impact of this virus. Also, the football industry is more careful than other industries regarding potential recovery from this virus scenario. Because as we said before football can't be sustainable without fans. The impact of this virus has hit football particularly high rate with clubs being rather pessimistic about the prospects for the future season. the interventions in Public health such as social distancing measures were impactful at the same time do not prevent the virus from re-emerging.

Even if government intervention proves partially effective, it will not return to pre-crisis levels. Given the idiosyncrasies of the organization of football in general, recently responded to need for this type of topic and researched about it about the impact of COVID-19 also football from a variety of perspectives and organizational types. Recent studies have attempted to analyze and predict the economic impact of this crisis in many different contexts to contribute to this discussion. More specifically, this study explores how COVID-19 responses are perceived for managing crises in the PFC which stands for professional football clubs in five countries in Europe, it is meant to be understood in the first place by analyzing how it responds also shown in the Figure 2.4 (Hammerschmidt *et al.* 2021).

Predication of value ether that is football or market industry in general mostly done by using algorithm like linear regression (LR) and the idea or the concept of LR was invented in 1849. LR is a type of test statistic type on the specific type of data set find the relation between the variable statistical tests like, t-test or and (ANOVA)tests, or tests like Fisher's exact, Chi-square, they don't take these covariates/confounders in to equation or accounting them in their method during analyses.

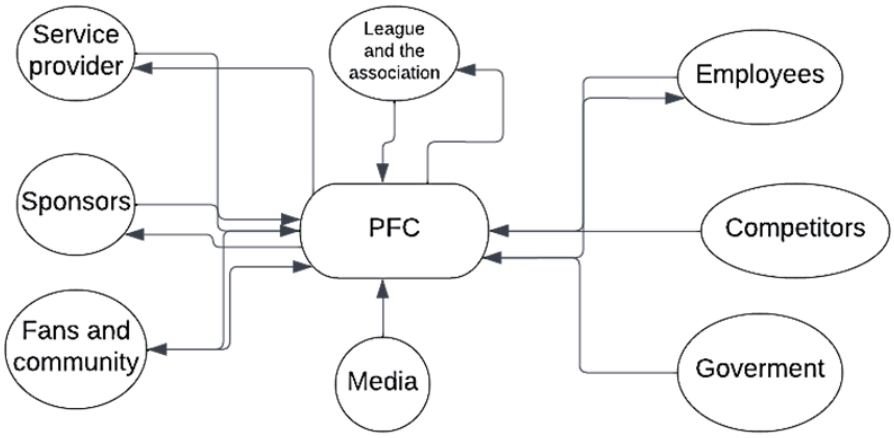

Figure 2.4 Professional football club source of outcome and income (Beheshti *et al.* 2022).

However, regression are the tests the let the user or research or anyone who use this algorithm to control confounders in explaining the relation between those variable while in another hand correlation the measure of strength between two variables with each other, this relationship between variables more than two can be described by regression analysis mathematically. It predicts the value of dependent variable based on independent variable which can more than one so the difference is clear between this two terms. The LR analysis use equation that relate to Equation (2.1).

$$y = mx + c \tag{2.1}$$

where x is the score of the independent variable, m is the regression coefficient, c is the constant.

As we discuss the LR, we can go to another one which is Machine learning models are usually divided into supervised and unsupervised learning algorithms. Defining (labeling) both the dependent and independent parameters create a supervised model. Conversely, if there are no defined (unlabeled) parameters, the unsupervised method will be used. This article focuses on a particular supervised model called RF., it is important to define the decision trees, and the another one called ensemble models, and bootstraps that are

essential to understanding RF models. Decision trees are used for both regression and classification problems. They visually flow like a tree, hence the name. For classification, we start at the root of the tree and follow a binary split based on the variable results until we reach the leaf nodes and get the final binary result. An example of a decision tree is shown below in Figure 2.5.

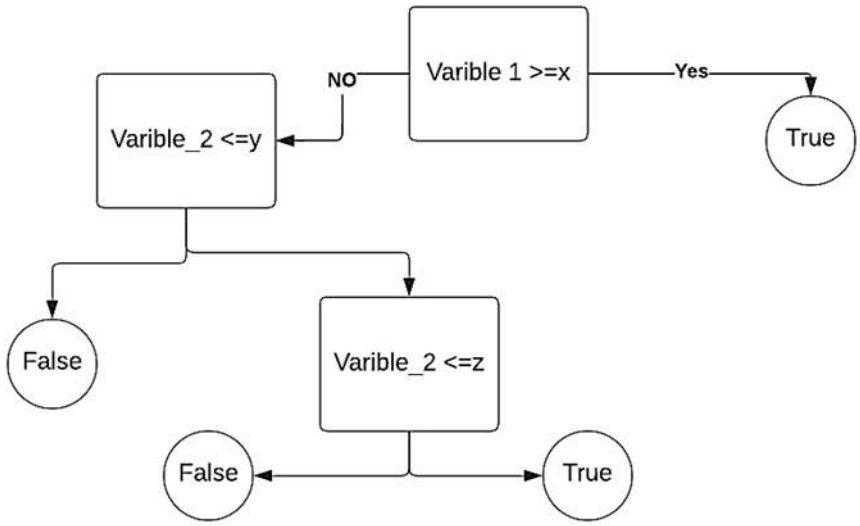

Figure 2.5 Decision tree flow chart (Beheshti *et al.* 2022).

Decision tree starts from Variable 1 and is split according to certain criteria. If yes, the decision tree classifies as true (false or true can be viewed as any binary value such as 1-0, yes-no, etc.). if the answer is no, after that the discussion tree proceeds to the next node and the process repeats so when the decision tree reaches the node that calls the leaf node and determines the result or the outcome. and this is called the ensemble type of learning it is done by using more than one model trained on the same data, and after that result, the average will be taken of each model to ultimately find the stronger prediction/classification result. The RF has to bootstrap which one of their process a subset of a data set over several iterations that have been specified and the number of variables is specified also.

As we said before the result has been averaged together to get a more meaningful result. Application of ensemble models one of their examples is bootstrapping. The ensemble learning method is combined in the RF with frameworks of decision trees to build several drawn which is random in the decision trees from the data and average the results, often for strong prediction/classification. Returns connected results. To keep this contest fair. Instead, we'll cover some general concepts and tips, then focus on the RF model (Beheshti *et al.* 2022).

The point of the research was to predict the stock market of Indonesian using the analysis that I discussed before. Rank tweets using as well this algorithm I discussed before that called (RF) algorithms to calculate sentiment towards companies. Sentiment analysis results are used to predict a company's stock price. Create a predictive model using the LR method. Our experiments show that the model that has been used for a prediction that uses the hybrid functions as predictors gives a determination of 0.9989 and 0.9983 as the best prediction coefficients (Cakra and Distiawan 2015, Frenger *et al.* 2019).

Mustafa A. Al-Asadi and Sakir Tasdemir conduct a research to predict value of football player by using different algorithm they showed in their research that RF was better approach than the decision tree and reason for that is the RF with less input give more accuracy which in another word the RF result showed that the better predicting for the value which they done (Al-Asadi and Tasdemır 2022).

By using 12,000 football player to predict their value the data set for this research was from transfer market website and the other one which is in state a company for sport information they used many different predictors and done by (Stanojevic and Gyarmati 2016).

Lukas barbscak done his research which was quite different with the other research he wanted to fine the reason the make the player of football expensive that was the question that want to find the answer for, by using small data set with two different methodology first one he used all the player with feature like contract remain and which nationality the player belong to etc. and the other one which was only one player Neymar because he

was highest value in the market of football in the world which got coefficient of determination about 0.92 (Barbuscak 2018).

Another research which want to find the relation between two important factors one of them is value of the player in market the other factor was performance they want to find the amount of relation that this two can impact each other. so they bring dataset from transfer market which was contain La Liga player only from Europe the feature was skill of the player like dribbling and goal scoring assist the common attribute for the forward like head accuracy the concluded that the player performance impacts the value of that player (He *et al.* 2019).

We found another research that also predicting the value of the player based on their performance with addition of how many search that done for that player in Wikipedia and by using the LR which the model R squared was about 0.72 and, by using the EPL league and the data come from FPL which is fantasy premier league that can give this info about player in the EPL league (Maurya 2018).

Another research which we can say different than before by the collected data which contain five seasons not like any other research that have been discuses which using only one seasons and by using to type of information for player in another word using different type of feature which can be divided into two group one of them is personal information which include the information of the player race and the age etc. and the other group of the data set that have been used is the skill of the player and use two different method one of them KNN and the other one OLS which they conclude that the information of the player not have impact on their value not like the player skill or player attribute (He 2019).

Predicting the Iranian player value by Rasool Nazari and Saied Azari which they used linear model to predict the value and the find that the age of the player related to the value of the player in high manner also the player position and number of match that played during that season (Nazari and Azari 2021).

This research is different than before that discuses because of using Xgboost method and the data set come from FIFA 18 and the parameter of the Xgboost come from player position that done analysis way to get this parameter what different about this research got good accuracy and they used new method to predicting (Daokang *et al.* 2021).

Another research also discusses professional football clubs investing billions of euros in player recruitment every year. How is the player price detrimental in the transfer market? their research shows an econometric model that explains the main factors that lead to or are involved in finding the amount of fee in the transfer market of the professional soccer player. The model that in this research is used to determine the amount of fee and another purpose is Multiple LR which is used by a lot of other research to predict the value and this research did the same thing with the paid fee of the club count in their research as target variable that the work with.

The data they have used includes more than 2000 trades of a professional soccer player in the world transferred for a specific amount of money from clubs in the top league which top league meant the five most famous leagues in Europe between July 2012 and November 2021. In particular, the paper emphasizes the importance of considering the remaining contract period that ties players to their clubs, and some key factors are ignored in the other research that has been studied by this research and their research shows that the model that they build a statically type of model can lead to explain and show more than 80% of the different transfer fees for a professional football player for a top famous league which is again the famous league mean five leagues that have more trophy and attendance in the champion league and their research shows a different way to develop the football economy in outcome to predict the value of football players which again the player in this research is the same as our research top five leagues.

Transfer negotiations, sale or acquisition of clubs, bank loans, fundraising, financial planning and communications, litigation, and more. As we know they are strong and the model that they build and developed and, the way they made it or in other words the way they create them, have weaknesses in some areas. We can start with the first point, there is not a lot of showing real data or transparency in the soccer area means that not all data

is used or the research point is they can’t get all the data that they want to cause it is not easy to approach by people or researcher, especially regarding the fee that is in the transfer of football club for their player, is authenticated. While sufficient accuracy can be achieved, having access to the data from the main source which we mean FIFA because they have data with all the details, especially besides markups and sales commissions, will be a real help to any research that wants to do this type of thing or want to get data to work with the writer of research said clearly that getting data is not an easy job, is the easier to do this type of research if they have the model with an amount of trusted data. It can no longer be included in the approach. Among these aspects we can mention in particular: the club's particular financial situation (e.g. urgent need for funds), disagreements between the player and the coach or other team members, a team of different footballers plans for employment by Exempt players same position, discipline issues, physical issues affecting the performance in the future, superstar effect in highly popular players, etc. soft skills, etc. Alot of certain factors considered by the player in the market of football during the negotiations done while the player has been transferred may explain some basic relation between the fee that they thought will be and the real one. So their research wants to show that the thing they done will be much better with the high information value that leads to developed econometric models confirm, the development of statistical models for estimating remittance fees and assessing remittance value based on scientific evidence, as described in this article, has broad applications for market players. Below is a no exhaustive list of tasks that the author has already performed (Poli *et al.* 2022).

In this research, we want to predict the outcome of match after and during COVID-19 their target is to find if there is a difference in the way the game ended in a football match during the year that this pandemic showed up they mean COVID-19 pandemic, the high number of football in the professional area in sporting events, not a been done in their time that supposed to be done behind closed doors and another thing that is obvious from the research observation is that home teams can't get the support of their fans and derive less advantage from playing on their home field, thereby reducing their chances of winning. We are seeking to see if this belief is true instead of five leagues they used four leagues in Europe in their research. In their study, they propose a model by name of

bayesian hierarchical poisson to showcase the variable that decides home advantage from the fans on their game which again they used four top leagues which only French league not included in their research. These features that are in their research led to improving the machine learning predictive performance for matches in football for the match that only done in the four leagues in Europe played after the pandemic disruption. Their study statistically analyzes the outcome that has been impacted by the COVID-19 pandemic of course in the football match in terms of expected outcomes which can be shown in goal difference. They also show that the variable that they estimated in their research from the model they work with it shows an advantage of home, which they mean by home in their research are they games that are done in the club's stadium which instead of this term they use home game. Finally, their study says that the inclusion of these variables as some features that can be added to the research will lead to improves performance of the models that predict different soccer matches. The non-disclosure policy introduced by the COVID-19 pandemic has changed. With parameters that reflect the impact of the pandemic, we can predict more accurate outcomes for games played behind closed doors after the COVID-19 hiatus (Lee *et al.* 2022).

Their paper examines the impact of professional football team strategy on player values in value to the transferred player. Their study is using a player in their data set from the 2018/2019 season of the Bundesliga, Series a Premier League, La Liga, and French Ligue. Pricing is shown by a common attribute in this research which a lot of other research used which are those attributes as the following age and how the player performance goal and passing data, and also team-related stats and the writer of this research is particularly interested in the importance of team data and how it relates to game strategy. The results of their work show strong proof of the strategy that players play they mean that every team has different tactic and that will impact player price, enabling optimization of team rosters for maximum value.

Football as we know one of the businesses with a high amount of money conduct gained and spend in this business in other words as the research say it is a multi-billion-dollar type of business and making a profit from the player will not be always the primary target for their life but also it is obvious they want at some point the money for their life and

make a better life, but it has grown in popularity as costs have risen and the introduction of UEFA's Financial Fair Play rules has allowed players to there is a lot of focus on selling smartly. The main purpose of their research can be easily explained as that to investigate are there is a reflation between the main pint that have been discussed at the beginning of the research which is team strategy and player value so they not saying there is a relationship they want to investigate which this research, in my opinion, is one of the good ones. The hypothesis is that, due to the additional cost of attacking players, attacking players are less likely to compete with players in the transfer market despite having the same performance values as players from other clubs could be sold at a higher price. They used the usual method which is LR on the data they got, which the data as we said in beginning, and most research using the same data, not from the same source but top five leagues in soccer for the 2018/2019 season to define the most impacted parameter.

For the variable that is common in a lot of research which is essential to predict the value of football player which are goals scored and accuracy of shooting is age the most important. The primary objective of this research was to explore the relationship between player value and team strategy. The findings revealed that the state of a team significantly influences player value, even when accounting for differences in club quality. This means that the transfer price a club can potentially achieve when selling a player depends not only on the how player plays and change the result for the team what they mean by that is the player may change a lot of games result, but another thing that they found which impact the value is the team achievement. And it is different for each position in their study what they mean by that is the impact are not same for every position striker get more credit for team achievement (Knapp *et al.* 2020).

In the document that other people have done they present as a prerequisite in science to get a master's for management info with a specialization Soccer is Europe's most popular sport and players play a key role. In recent years, we have contributed to player price inflation that can create futile problems for European football teams. It is therefore important to understand some factors that influence the transfer market to predict the value of the player on a football field. Recent research suggests that the value of the transfer market depends on many factors, including performance metrics. These

measurements quantify the skill each player performs on the field. Despite advances in this field, few studies have considered the specifications of each item when selecting performance indicators.

This study, therefore, aims to understand the impact of performance indicators on players in Europe's top six leagues over the past three seasons, considering each position group. A fixed effects regression was performed and the results showed that the market value of goalkeepers was positively correlated with the number of games without a goal. Defenders and his headers per minute are negatively correlated for defenders. Assists have a positive influence on the value of players that play in the position of mid which in football is called midfielder while goals have a significant influence on the player value and attackers. The study that they have done are aims to provide relevant insights for agents playing an active role in the football transfer market (Ferreira *et al.* 2022)

The goal of this research that they do it also another research that wants to find the value of football player price so their goal is to analyze the different impacts of different types of variables (club and the league that the player plays in, the year the player is born(age), the place) and the crowdsourced transfer market value of players in Europe's top leagues which are again it for man football player. All players (n=2259) from the as many other research that we observed it the same here top five leagues is being used to know which top five leagues are as a reader can look at any other research that has been reviewed here from the 17/18 season. Same the data two most famous websites that provide that and in their research obtained from the open source which is the transfer market registered the value for the current economy is named as this (PRESENT) and for the maximum economy (VMAX) of all professional contract players. Regression analysis showed that team size, the data that amount of age and the team that the player play, and of course the league also the player play in had a significant impact ($p<0.05$). The analysis showed that players as attacking midfielders and were born in the first quarter of the year with current values R0.33 of R2 and maximum this is confirmed by both points rated as the most economical in 0.33 of R2. Football that is international associations must therefore consider the finances should be equal between each competition, not just at the national

level, as this will impact the European league directly as their research says (Felipe and FernandezLuna 2020).

Also, another research mention Player transfers make up football club teams and is an essential role in achieving better result. Choosing a good type of player is a prerequisite for high performance in football. Therefore, transfer rumors are a big topic in the soccer world this summer. This research aims to predict using some type of graph which is called graph theory the player transfers in the future and their article examines the networks that initially how the world of football formed and also the ownership of small worlds these networks have. So they have done it which they have created a dating chart for professional soccer players based on whether they were teammates. Create a graph also for a manager in football who is similar to the player. In this case, more than one coach is considered related if the same club are being coached by these two coaches or more they also analyzed the networks that have improved between teams over the past 14 years with links showing player transfers.

They use type metrics which are in a chart with information that provides transfer market to build a model for the data mining that can predict player transfer future. their model can be used for different types of things and one of that is to predict who will move to the league of your choice. The most important traits when buying a player vary from league to league, but in all cases studied, the trait extracted from the graph is one of the most important factors. These features will help their model in a specific way which are improving the performance of predicting the player transfer model and provided a reasonable chance that a transfer would occur. In recent years, the popularity of network science has made it possible to study in their research saying more connections and studying more networks that are complex can yield information that can be hard to get and can lead to a significant impact on player transfer prediction. They show that this model can be used to build powerful functions that improve the performance (Kovacs and Toka 2022).

Also, another article mentions that the worldwide transfer in soccer especially in Europe has less improvement in technology on the football field. Over this time the research says

about twenty years the way of dealing with the economy have changed entirely also the money that has been in this business also change which they mention increased exponentially. Studying them physicality of some type of variable is a kind type of challenge and struggle for the research in itself and the thing related to it or the item that depends on it on many different factors and the same process will be done for a player that has a different type of variable that depends on the data that came from the main source is challenging and hard. So, in their research, they try to do the best way to obtain soccer player data and apply a suitable model to it extract a type of information that can be used fully in this manner and reduce the difference between the price that has been estimated and the final price. So they do that, they crowdsource the data and applying a type of method that much other research used and for their research, they used LR along with the Opt index. they do something else to find results using another method which is not common what I mean by that I didn't see other research done it is that they use neural networks which show the difference between two models that one used LR and other one use neural networks (Patnaik and Praharaj 2019).

Moreover, in another research Football is the sport with the most fans on the planet, so you can expect exorbitant sums of money to be processed. Soccer clubs spend a lot of money, especially when it comes to recruiting playersTo identify the variables that impact the market value of football players, a sample of the top 35 Conmebol team players from the 2017/2018 season was considered. To achieve this objective, the method of multiple linear regression (LR) with two models was applied. Initially, only performance-related variables were considered. In the second model, the first model included variables of extrinsic nature. The results reveal the variables that best describe a player's market value. Various validation tests were applied to the final model to demonstrate its effectiveness. Finally, it was concluded that the level of influence players have is also a determinant of market value (Gallegos Orrala *et al.* 2019).

Another article mention that transfer markets in football player lead to a lot of research cause as we know this type of business with a huge amount of money conducted in it so they mention that in the financial system and control. In this paper, we endorse an excessive stage evaluation technique for classifying participants primarily depending

totally on the performance overall at some point in current seasons. Numerous records evaluation strategies consisting of regression evaluation, characteristic choosing, and cluster evaluation are provided for grouping the football player in a period of their development as player and switch fee.

Specifically, with the aid of using gathering and reading records from whole scored, the most important targeted soccer information website, we've got described gamers into 4 different groups and it contains 1. The player with low state and occasional switch (2) Low overall performance and excessive with this type of fee that count as high (3) excessive type of player state and excessive switch fee and (4) excessive overall performance and occasional high fee for that player. The consequences with inside the segment display that, with the variations of positions, there are unique required abilities that have an effect on the overall performance of gamers (Kim and Bui 2021).

Table 2.1 Related literature review

REFERENCE	**MODEL**
Poli (2022)	Multi LR
Lee and Kim (2022)	Logistic R,MLR ,RFL,SVM
Felipe and Feirnadez-luna (2020)	Regression analysis
Kovacs and Toka (2021)	Clustering
Patnaik, D,Praharaj	Multi LR
Gallegos and Jorrala (2019)	LR
Kim and Bui (2019)	Regression analysis
Knapp (2020)	Regression Model
Ferreira,M.D.S.B.P (2022)	OLS – Regression
Daokany and Caixink (2021)	Multi LR
Muller and Simon (2017)	Multi LR
Aazari *et al.* (2021)	Multi LR
Berkeley (2019)	Regression model
Bhravan *et al.* (2021)	Logistic Regression, SVC
Hammer Schmid (2021)	LR , MLR,RT,Rndam forest
Herm (2019)	LR
Franck and S.Nuesch (2012)	Fixed effects Regression
Cakra and Frenger (2019)	Squares Regression
Andrea (2021)	MLR , OLS Estimation
AL-aasdi and Tasdemir (2022)	Muitple Regression

3. BACKGROUND AND THEORY

In this section, we will go through how we get the data and the method that we used.

3.1 Implementıng ML Models

After performing a series of preparation procedures, the data set is divided into the next phase. The dataset contains 2940 records that are used in this method. Since the separation method is random, the test set (33 percent) and the training set (67 percent). These are made using the snake instructions below training is necessary for ML to produce a good classification model for the selected data set. So, the training data provided in the training phase is an essential function in developing a robust machine learning model. During the training process, the data collection is analyzed to produce a model that attempts to generalize the relationship between attribute values and label values. It's the most a critical step in our strategy. This stage is performed for improve the detection rate and reduce the false alarm rate of the proposed method slope optimization technology many algorithms give an effective classification and then give good results in related business. We chose these algorithms with default values, wherever they are application is submitted to obtain satisfactory results and to comply with the prediction of the soccer players market value. Next, we will combine it with optimization technology. Then select the appropriate options for the modeling system. There several hardware regression techniques are used according to the above information is entered into the system through the label and these technologies, they are as follows: Implementing ML Models:

- LR
- Lasso
- RF
- Gradient boosting
- K- Nearest Neighbors

After preprocessing data, we will try to test all classification algorithms on our own dataset without any optimization techniques, only with base algorithms to show the

effects of each optimization technique that will be used later as shown in Figure 3.1 below:

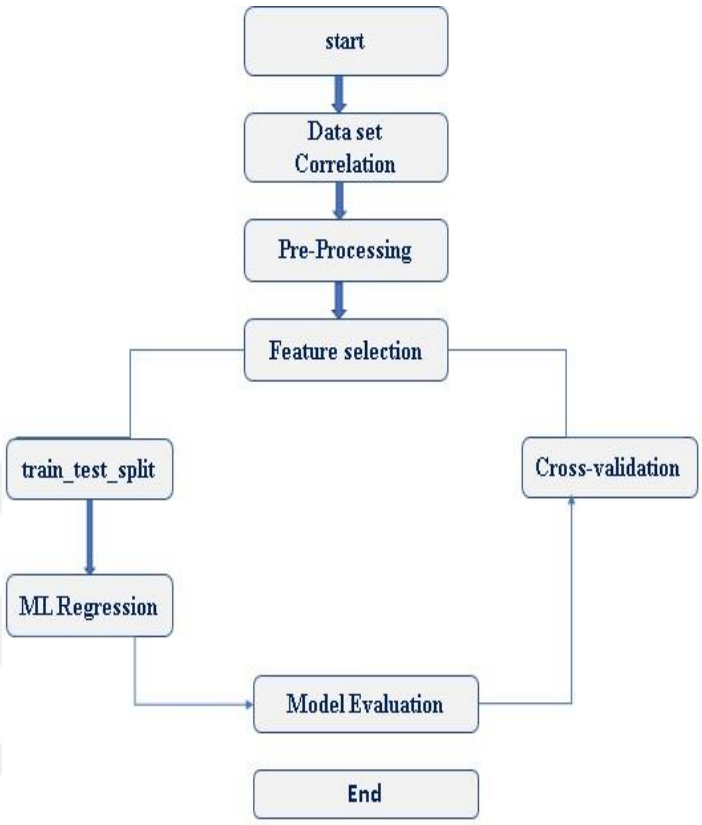

Figure 3.1 Diagram of the base machine learning algorithms

Design options will be provided to select the structure of the model while you are at it work with ML. The best model architecture for a model is not always clear. Thus, it is helpful to be able to try a few different approachesEmbracing the principles of machine learning, we will let the computer undertake the fundamental tasks and autonomously select the best model architecture. Design options will be provided to define the structure of the form while you are at it work with machine learning. The best model architecture for a model is not always clear. Thus, it is helpful to be able to try a few different approaches. In the spirit of machine learning, we will have the computer to do the basic work investigate and choose the best model architecture on its own. The parameters in machine learning are almost infinite, and it is impossible to experiment all in one research regarding methods, numbers and learning mechanisms. Instead, it takes a long time to test

it, so we relied on his experience with methods and numbers made by other researchers and made some by trial and error. We will use the grid search technique to choose the parameters because it is relatively efficient compared to other techniques and more accurate implementation than others. Below are all the algorithms and parameters used in this study.

3.1.1 Linear regression

To predict the market value of soccer players, the LR algorithm has been implemented. LR is one of the most regression systems used in forecasting. LR contains a number of optional parameters that are important to the model careers. The most important parameters have been modified in Equation (3.2).

$$Y_i = \beta_0 + \beta_1 X_i \tag{3.2}$$

Where Y_i is a dependent variable ,β_0 is a constant variable ,β_1 is a slop coefficient X_i s an independent variable.

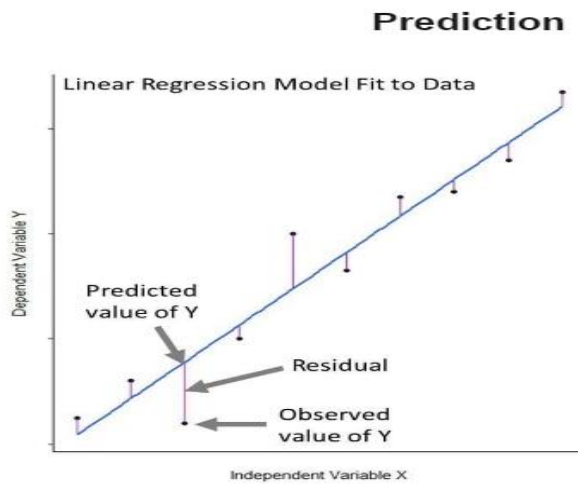

Figure 3.2 LR algorithm (Kavita 2021)

On the first step with everyone 45 Features, LR parameters set as follows: { Copy = 'False', fit_intercept = true, N_jobs = '-1'}. Where punishment is used Sets the penalty

criterion and uses fit_intercept to regulate whether it is static or not in the second step, the RFE was applied to the balanced dataset. Specific hyper parameters: LR is a simple and straightforward algorithm that fits a linear equation to the data. It does not have hyper parameters in the traditional sense. The model's coefficients are determined directly from the data using techniques like ordinary least squares (OLS) or Gradient descent. fit_intercept :Whether to calculate the intercept for this model. copy_X :If True, X will be copied; else, it may be overwritten. n_jobs :The number of jobs to use for the computation. This will only provide speedup in case of sufficiently large problems.

Table 3.1 Parameters optimization of LR

	FIT_INTERCEPT	**COPY_X**	**N_JOBS**
	True	True	0
	False	False	-1
Best Parameters	True	False	-1

3.1.2 Lasso regression

Lasso Regression, short for Least Absolute Shrinkage and Selection Operator Regression, is a LR technique that incorporates regularization to improve the model's performance and handle multicollinearity (high correlation between predictor variables). In traditional LR, the objective is to minimize the sum of squared residuals between the predicted values and the actual target values. However, in Lasso Regression, an additional regularization term is added to the cost function. This regularization term is the sum of the absolute values of the coefficients multiplied by a hyper parameter called alpha (also known as lambda). The addition of the regularization term in Lasso Regression introduces a constraint on the magnitude of the coefficients. As a result, it encourages sparsity in the model by driving the coefficients of less important features to zero. This property makes lasso regression useful for feature selection as it automatically identifies and selects the most relevant features. The cost function of lasso regression can be represented as: Cost = Sum of squared residuals + alpha * Sum of absolute values of coefficients. The alpha parameter controls the strength of the regularization. A higher value of alpha results in more coefficients being pushed towards zero leading to more aggressive feature selection Equation (3.2)

$$\text{Min } \frac{1}{2n \text{ samples}} \parallel Xw - y \parallel \frac{2}{2} + \text{a} \parallel \text{w} \parallel 1 \quad (3.2)$$

Where a is a constant, a ‖ w ‖ 1least-squares penalty and ‖w‖1 is the ℓ1-norm of the coefficient vector.

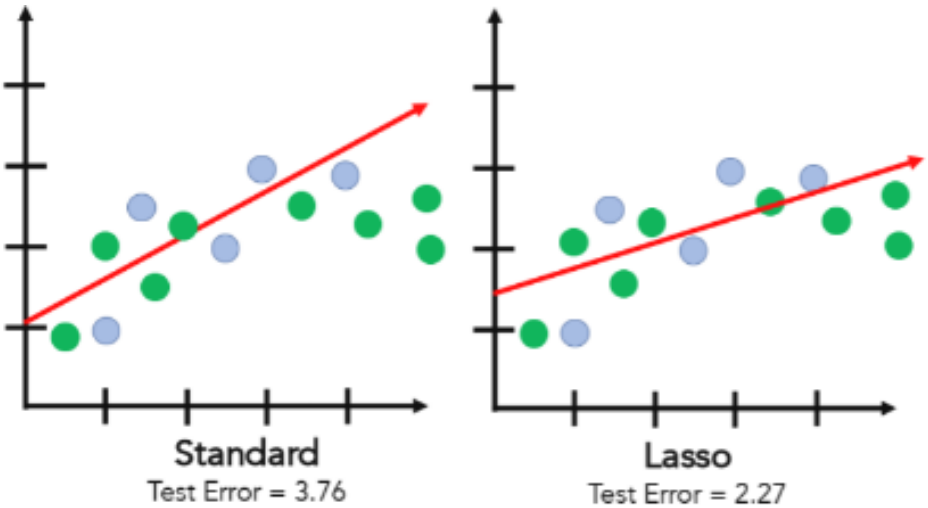

Figure 3.3 Lasso regression algorithm (Kumar 2021)

There are some parameters that will be used in the optimization and they are set as follows:

- (Alpha ="1.0”, Precompute ="n_features" , Fit_intercept="False") as shown in Table 4.4 and they consist of three parameters:
- Alpha: Alpha must be a non-negative float i.e. in [0, inf).
- precompute: Whether to use a precomputed Gram matrix to speed up calculations. fit_intercept: Whether to calculate the intercept for this model.

Table 3.2 shows the parameters optimization of the lasso regressor

Table 3.2 The parameters optimization of the lasso regressor

	ALPHA	PRECOMPUTE	FIT_INTERCEPT
	0	N_features	True
	1.0		False
	0.01		
Best parameters	1.0	N_features	False

3.1.3 Random forest

RF Regressor is an ensemble learning method based on the RF algorithm for regression tasks. It is a powerful and versatile machine learning algorithm that can be used to predict continuous numeric values. RF regressor works by combining multiple decision trees, where each tree is built on a different subset of the data and uses a random subset of features. The final prediction is obtained by averaging or taking the majority vote of the predictions from individual trees as shown in Equation (3.3).

$$MSE = \frac{1}{n}\sum_{i-1}^{n} (Yi - Yi)^2 \quad (3.3)$$

Where $\frac{1}{n}\sum_{i-1}^{n}$ $the\ MSE\ is\ the\ mean$, $(Yi - \acute{Y}i)^2$ the squares of the errors, n predictions is generated from a sample of n data points on all variables, Yi the vector of observed values of the variable being predicted and Ý being the predicted values.

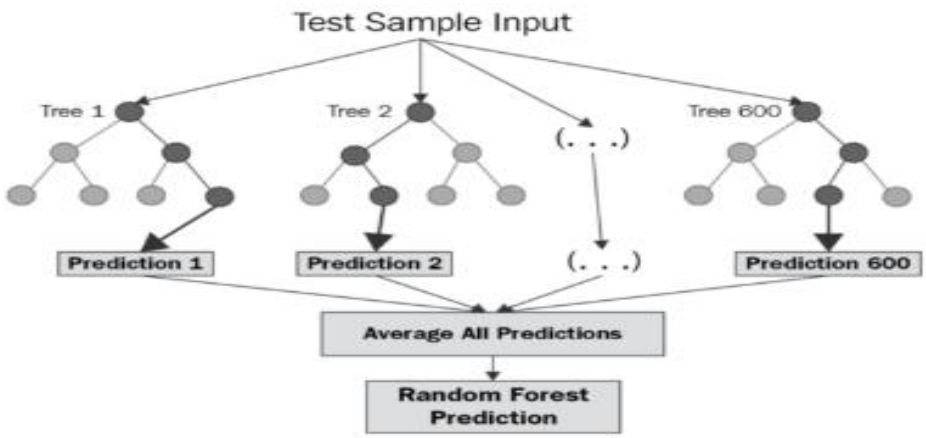

Figure 3.4 RF algoritham (Kumar 2021)

There are some parameters that will be used in this model and they are as follows: (Min_samples_split="5”, N_estimators="100”, Max_depth="5”) as shown in Table 3.3 where: n_estimators: The numeral of trees in the forest. random_state: Controls both the randomness of the bootstrapping of the samples used when building trees (if bootstrap=True) and the sampling of the features to consider when looking for the best split at each node (if max_features < n_features) Max_depth: The maximum depth of the tree.

Table 3.3 Parameters optimization of the RF regressor

	N_ESTIMATORS	RANDOM_STATE	MAX_DEPTH
	2	0	5
	5		15
Best parameters	5	0	5

3.1.4 Gradient boosting

Gradient Boosting Regressor is (ML) algorithm that belongs to the family of gradient boosting methods. It is designed to perform regression tasks and is known for its ability to produce highly accurate predictions. Gradient Boosting Regressor works by building an ensemble of weak regression models, typically decision trees, in a sequential manner. It starts by fitting an initial model to the data and then iteratively builds additional models that focus on the errors or residuals made by the previous models, Equation (3.4).

$$\boldsymbol{L} = \frac{1}{n}\sum_{i} (\acute{Y}\, i - yi)^2 \tag{3.4}$$

Where $\acute{Y}\, i$ the predicted value, yi the observed value and n the number of samples

The predictions from all the models are combined to make the final prediction.There are some parameters that will be used in this form is as follows (Min_samples_split="10" , N_estimators="1900", Max_depth="5") as shown in Table 3.4, where: n_estimators: The numeral of trees in the forest.Max_leaf_nodes: best nodes are defined as the relative minimizing impurity. Max_depth: The maximum depth of the tree.

Table 3.4 Parameters optimization of grediant boosting

	MIN_SAMPLES_SPLIT	N_ESTIMATORS	MAX_DEPTH
	5	100	6
	10	200	5
		1900	7
Best parameters	10	100	5

3.1.5 K-nearest neighbors

The KNN model was used to predict the market value of soccer players. KNN is one of the most basic regression models which are mainly based on KNN regression dataset voting in the beginning Step with all the features, KNN was applied to the balanced dataset as shows in Equation (3.5)

$$pk(y) = \sum_{i=1}^{n} K(\mathrm{y} - \mathrm{Xi}\,;\mathrm{h}) \tag{3.5}$$

h is controlled by the bandwidth parameter, Y the density estimate at a point and Xi ; i =1…n a group of points is given by.

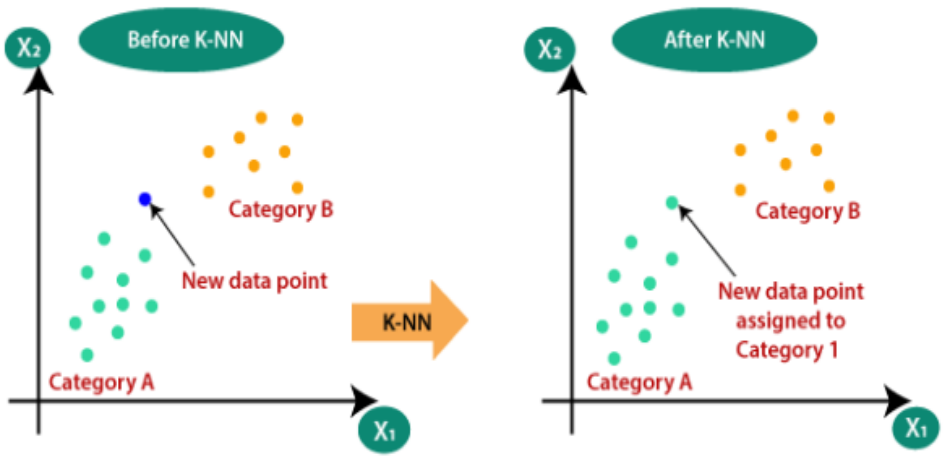

Figure 3.5 K-nearest neighbors algorithm (Saikumar 2023)

KNN Parameters are set to {n_neighbors=1, weights='distance', Algorithm=' auto '}. In the second step, the RFE was applied to the balanced dataset. As shown in Table 3.5.

N_neighbors: Number of neighbors to use by default for kneighbors queries.Weights: Weight function used in prediction. Possible values. N_jobs : The number of parallel jobs to run for neighbors search.

Table 3.5 Parameters optimization of k-nearest neighbors

	N_NEIGHBORS	**WEIGHTS**	**ALGORITHM**
	1	uniform	Auto
	2	Distance	ball_tree
	5	callable	kd_tree
			Brute
Best parameters	2	Distance	Auto

4. MATERIALS AND METHODS

The used approach in this study is based on high-vision ML which means that the algorithm uses samples of data to infer the model and test the model using untested data in building the model. This data allows us to compare the predicted values with the real values. In this thesis, the expected values are the market value of soccer players during the Corona pandemic, where each sample was represented by a set of variables that represent the player's performance his skill and the audience. The proposed approach consists of a set of steps as follows:

4.1 Dataset

In the research, the data includes details for the top 5 leagues in Europe and the transfer market data is the fan attendance collected from (Sofifa .com) consists of 2940 rows and 49 columns that are made by web scraping and the feature is ("Name", "OVA", "POT", "Team and Contract", "BP", "Value", "Wage", "Crossing" "Finishing", "Heading Accuracy", "Short Passing", "Dribbling", "Curve", "FK Accuracy" Long Passing, Ball Control, Acceleration, Running Speed, Agility, Reflexes, Balance, Shooting Power, Jumping, Stamina, Strength Long Shots, Aggression, Interceptions, Positioning, Vision, Penalties, Composure, Pointing, Standing Tackle, Sliding Tackle, Diving Goalkeeper', 'Goalkeeper Handling', 'Goalkeeper Kicking', 'Goalkeeper Position', 'Goalkeeper Reflexes', 'Total', 'SM', 'A-W', 'D-W', 'IR' and ' PAS', 'DRI' and 'DEF') which respectively stands for name stands for player's full name and ova generally means for player which is the average skill the player has and POT stands for player's influence from 0 to 100 also Team and Contract is the name of the team he plays for The player and the term of the contract also BP stands for the best possible position for the player. Its play value is the player's market value. Crossing is the exact amount of crossing from 0 to 100. The opponent player curves the ball and passing is the amount of curve the ball the player can do from 0 to 100 FK. Accuracy is free kick accuracy and long passing like cross and short pass but for greater distance, most other features are obvious only SM which means skill move and A-w means offensive work rate too D-W means defensive work rate, IR means reputation of intent and in this data obtained from the site (sofifa.com) contains file. We

have converted the data file to in CSV format, and then they showed us the specification for that data, which is 44 columns, training file with 2940 records, as shown in Figure 4.1 below.

	OVA	POT	Value	Crossing	Finishing	Heading Accuracy	Short Passing	Dribbling	Curve	FK Accuracy	...	GK Reflexes	Total	SM	A/W	D/W	IR	PAS	DRI	DEF	att
0	92	92	69500000.0	85	91	70	91	96	93	93	...	8	2206	4	0	0	5	91	95	34	41.589
1	92	92	119500000.0	71	95	90	85	85	79	85	...	10	2211	4	2	1	5	79	86	44	40.146
2	91	91	129000000.0	81	93	59	85	92	84	69	...	14	2236	4	2	1	4	82	91	45	53.008
3	91	91	125500000.0	94	83	55	93	88	85	83	...	13	2303	4	2	2	4	93	87	64	52.738
4	91	91	45000000.0	84	94	90	80	86	81	79	...	11	2188	5	2	1	5	80	86	34	73.156
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2935	56	70	325000.0	40	36	61	66	54	45	40	...	13	1570	2	1	1	1	55	56	58	12.738
2936	56	61	170000.0	10	8	11	28	12	11	9	...	56	926	1	1	1	1	52	56	32	31.965
2937	56	72	325000.0	11	7	14	21	13	11	11	...	62	866	1	1	1	1	55	62	26	16.773
2938	56	70	325000.0	53	45	34	56	62	60	45	...	12	1445	2	1	1	1	54	63	32	9.932
2939	56	66	300000.0	51	36	54	53	33	29	26	...	8	1357	2	1	0	1	43	47	54	9.932

2940 rows × 45 columns

Figure 4.1 Diagram of the dataset collection

Application of the proposed model to predict the market value of soccer players described in this chapter. Python is used for modeling and testing. It is implemented in four stages: pre-processing of input data, feature selection or extraction, data augmentation, regression, and testing. Pre-processing of input data involves preparing a dataset of soccer players to be used as input for the identification system. Feature selection and regression are the basic processes in the identification system and model. Similar architecture and features were called in for training and testing. First, we will practice optimization results by hyper parameter tuning. Secondly, the two feature technologies extraction will be used to improve results. Third, the feature selection process it will test three filter, distortion, and inline techniques. Finally, we will practice improve results by over-sampling and then choosing the most effective technique that yields best results with the regression process. The regression process is used five algorithms, KNN, decision tree slope, RF slope, Gradient boost ramp, and XGBoost ramp. The goal is to determine the best identification system model. This chapter also presents the results of implementing the proposed models

for systems for determining the market value of soccer players. Python is used to perform the test enter data and obtain regression results as shows in Table 4.1.

Table 4.1 Description of original dataset

FEATURE	MEAN	STD	MIN	MAX
Value	1.04	1.65	4.5	1.94
Overall rating	72.46	7.14	56	92
Potential	77.71	5.07	60	95
Dribbling	62.47	20.18	5	96
Curve	53.97	20.13	6	93
FK Accuracy	46.91	18.92	4	94
Crossing	55.55	19.86	8	55.55
finishing	50.96	21.41	4	95
Heading accuracy	57.01	19.009	5	93
Short passing	65.98	15.168	12	93
Skill moves	2.636	0.89	1	5
SM	2.636	0.89	1	5
A/W	1.27	0.55	0	2
D/W	0.51	0	1	2
IR	1.43	0.72	1	5
PAS	64.66	10.15	30	93
DRI	69.64	9.64	36	95
DRI	69.647	9.64	36	95
DEF	56.10	18.03	17	91
Att	25.33	15.10	5.90	73

4.2 Programming Environment

We utilized Google Colab, a cloud-based IDE offered by Google, for this project. Colab facilitates the execution of Python code and provides a platform similar to Jupyter Notebook, pre-loaded with data science and machine learning libraries like Tensor Flow and Pandas. It supports GPU and TPU acceleration, making it suitable for deep learning tasks. Being cloud-based, it allows seamless collaboration on notebooks and is accessible from any online device. While it offers free resources, there are limitations which can be overcome with the subscription-based Colab Pro. Colab has emerged as a preferred tool for data tasks due to its comprehensive features and ease of use (Doe 2023).

4.3 Model Design

Model Design as explained in the last chapter, it is difficult to predict the player's market value when the Corona pandemic occurs and to identify weaknesses in predicting the player's value. Machine learning is good for making a prediction model and our system is designed to predict the market value of football players, and this can be accomplished by creating a model. Using data from 2,940 players, the graph illustrates the general layout of the model at a high level.

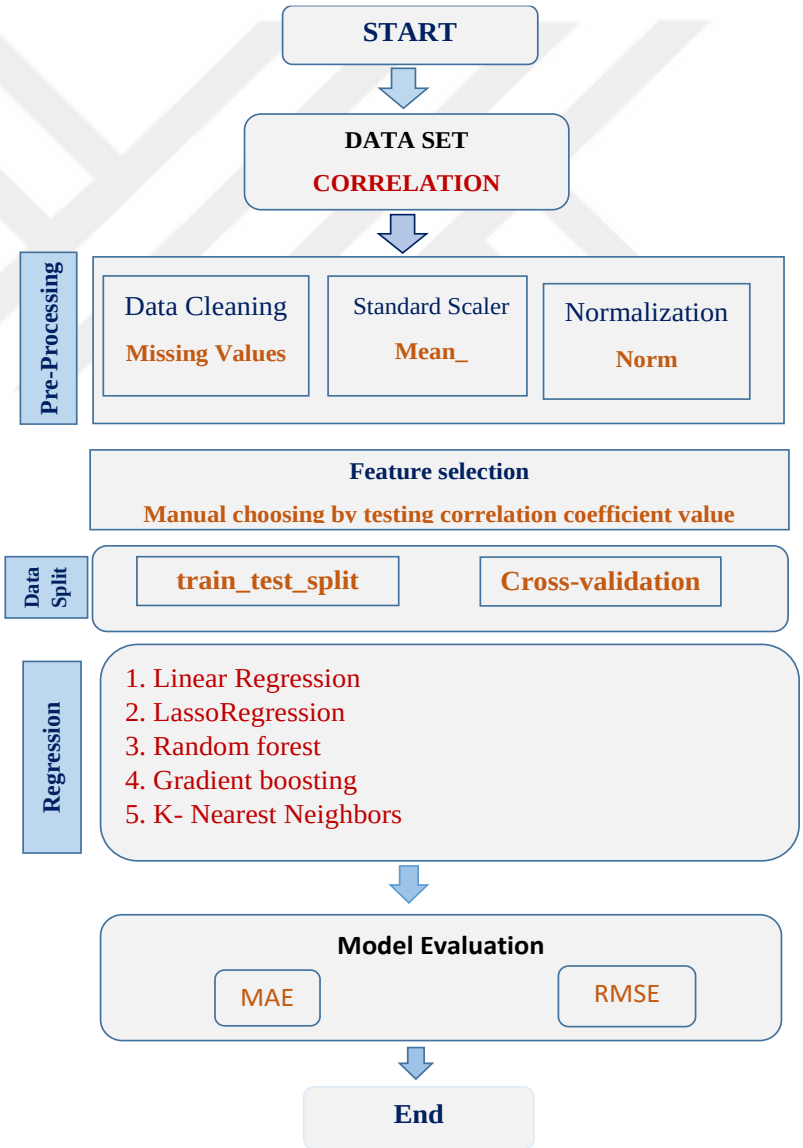

Figure 4.2 Diagram of the proposed system

4.4 Data Source

In this research, we used a set of official data containing 2940 players, and the data set represents the skills and behavior of football players and contains 2940 records and 48 columns, as shown in Figure 4.3.

```
df
```

	Name	OVA	POT	Team & Contract	BP	Value	Crossing	Finishing	Heading Accuracy	Short Passing	...	GK Reflexes	Total	SM	A/W	D/W	IR	PAS	DRI	DEF	att
0	L. Messi	92	92	Paris Saint-Germain	CAM	69500000.0	85	91	70	91	...	8	2206	4	0	0	5	91	95	34	41.589
1	R. Lewandowski	92	92	FC Bayern München	ST	119500000.0	71	95	90	85	...	10	2211	4	2	1	5	79	86	44	40.146
2	M. Salah	91	91	Liverpool	RW	129000000.0	81	93	59	85	...	14	2236	4	2	1	4	82	91	45	53.008
3	K. De Bruyne	91	91	Manchester City	CM	125500000.0	94	83	55	93	...	13	2303	4	2	2	4	93	87	64	52.738
4	Cristiano Ronaldo	91	91	Manchester United	ST	45000000.0	84	94	90	80	...	11	2188	5	2	1	5	80	86	34	73.156
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2935	T. Le Bris	56	70	FC Lorient	CDM	325000.0	40	36	61	66	...	13	1570	2	1	1	1	55	56	58	12.738
2936	M. Köbbing	56	61	. FC Köln	GK	170000.0	10	8	11	28	...	56	926	1	1	1	1	52	56	32	31.965
2937	A. Schulz	56	72	DSC Arminia Bielefeld	GK	325000.0	11	7	14	21	...	62	866	1	1	1	1	55	62	26	16.773
2938	M. Torlak	56	70	SpVgg Greuther Fürth	CAM	325000.0	53	45	34	56	...	12	1445	2	1	1	1	54	63	32	9.932
2939	E. Kratzer	56	66	SpVgg Greuther Fürth	RB	300000.0	51	36	54	53	...	8	1357	2	1	0	1	43	47	54	9.932

2940 rows × 48 columns

```
df.head()
```

	Name	OVA	POT	Team & Contract	BP	Value	Crossing	Finishing	Heading Accuracy	Short Passing	...	GK Reflexes	Total	SM	A/W	D/W	IR	PAS	DRI	DEF	att
0	L. Messi	92	92	Paris Saint-Germain	CAM	69500000.0	85	91	70	91	...	8	2206	4	0	0	5	91	95	34	41.589
1	R. Lewandowski	92	92	FC Bayern München	ST	119500000.0	71	95	90	85	...	10	2211	4	2	1	5	79	86	44	40.146
2	M. Salah	91	91	Liverpool	RW	129000000.0	81	93	59	85	...	14	2236	4	2	1	4	82	91	45	53.008
3	K. De Bruyne	91	91	Manchester City	CM	125500000.0	94	83	55	93	...	13	2303	4	2	2	4	93	87	64	52.738
4	Cristiano Ronaldo	91	91	Manchester United	ST	45000000.0	84	94	90	80	...	11	2188	5	2	1	5	80	86	34	73.156

5 rows × 48 columns

Figure 4.3 Screenshot of the raw data

The data frame now contains numeric data with 2940 records and 48 columns. The file description contains a lot of information for each column:

mean -> mean (average) value.
count -> the number of non-blank values.
STD -> standard deviation Max -> the maximum value min -> minimum value.
75% -> percentage 75%*.
50% -> percentage 50%*.
25% -> 25% percentile*.

As shown in Figure 4.4, the data value is 6.72 megabytes.

	OVA	POT	Value	Crossing	Finishing
count	2940.000000	2940.000000	2.940000e+03	2940.000000	2940.000000
mean	72.461905	77.715986	1.045142e+07	55.557143	50.962925
std	7.143994	5.073259	1.650950e+07	19.865346	21.413578
min	56.000000	60.000000	4.500000e+04	8.000000	4.000000
25%	68.000000	74.000000	1.600000e+06	44.000000	35.000000
50%	74.000000	78.000000	4.100000e+06	61.000000	55.000000
75%	77.000000	81.000000	1.250000e+07	70.000000	69.000000
max	92.000000	95.000000	1.940000e+08	94.000000	95.000000

8 rows × 45 columns

Figure 4.4 Screen shot of the data description

4.5 Data Pre-processing

Preprocessing is an important and critical phase of any strategy to increase model performance. Because of it contains a variety of technologies. In the next step we will describe some techniques designed to improve feature extraction and remove unnecessary information (Figure 4.5)(García *et al.* 2016).

	OVA	POT	Value	Reactions	Shot Power	Vision	Composure	Total	IR	PAS	DRI
0	92	92	69500000.0	93	86	95	96	2206	5	91	95
1	92	92	119500000.0	93	90	81	88	2211	5	79	86
2	91	91	129000000.0	94	83	86	92	2236	4	82	91
3	91	91	125500000.0	91	92	94	89	2303	4	93	87
4	91	91	45000000.0	94	94	76	95	2188	5	80	86
...	...	...	...	...	...	...	...	...	...	...	...
2935	56	70	325000.0	57	58	50	49	1570	1	55	56
2936	56	61	170000.0	49	39	44	37	926	1	52	56
2937	56	72	325000.0	53	41	33	31	866	1	55	62
2938	56	70	325000.0	56	57	56	54	1445	1	54	63
2939	56	66	300000.0	44	44	21	42	1357	1	43	47

2940 rows × 11 columns

Figure 4.5 Screenshot of reading the data set from files

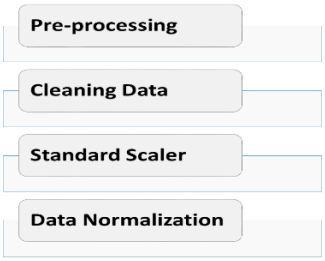

Figure 4.6 Pre-processing steps

There are two processing techniques.

- If there is a difference in the number of unique values in the target column between the train and the test suite, we must remove extraneous classes or unique values to balance the data. However, doing so is risky because it can happen you lose potentially useful data in a particular category.

- We may find a solution to the problem so if our data has columns what are they. Each unique value should be removed because it does not make use of it Regression. After applying this approach, the data set becomes 44 columns or as shown in the following Figure 4.7.

This is the last step in preparing the data before entering into the three main stages of data processing Build the form. The data is now ready for pre-processing and feature selection and testing the efficiency of the proposed algorithms.

	OVA	POT	Value	Crossing	Finishing	Heading Accuracy	Short Passing	Dribbling	Curve	FK Accuracy	...	GK Refleces	Total	SM	A/W	D/W	IR	PAS	DRI	DEF	att
0	92	92	69500000.0	85	91	70	91	96	93	93	...	8	2206	4	0	0	5	91	95	34	41.589
1	92	92	119500000.0	71	95	90	85	85	79	85	...	10	2211	4	2	1	5	79	86	44	40.146
2	91	91	129000000.0	81	93	59	85	92	84	69	...	14	2236	4	2	1	4	82	91	45	53.008
3	91	91	125500000.0	94	83	55	93	88	85	83	...	13	2303	4	2	2	4	93	87	64	52.738
4	91	91	45000000.0	84	94	90	80	86	81	79	...	11	2188	5	2	1	5	80	86	34	73.156
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2935	56	70	325000.0	40	36	61	66	54	45	40	...	13	1570	2	1	1	1	55	56	58	12.738
2936	56	61	170000.0	10	8	11	28	12	11	9	...	56	926	1	1	1	1	52	56	32	31.965
2937	56	72	325000.0	11	7	14	21	13	11	11	...	62	866	1	1	1	1	55	62	26	16.773
2938	56	70	325000.0	53	45	34	56	62	60	45	...	12	1445	2	1	1	1	54	63	32	9.932
2939	56	66	300000.0	51	36	54	53	33	29	26	...	8	1357	2	1	0	1	43	47	54	9.932

2940 rows × 45 columns

Figure 4.7 Screen shot of the two files combination

4.5.1 Cleaning Data

When an observation is missing data or any cell of a recorded row is null, for a particular variable, this is known as 'missing data'. As a common occurrence, missing data can seriously jeopardize conclusions that can be drawn from the data. Although there are many options to deal with the problem of missing data, the data is cleaned by removing spaces, whether they are of a non-zero value. Otherwise, this syntax in Python is "class sklearn impute. Simple Imputer (*, Missing values = Nan, Strategy = 'mean')" Missing values just got a lot easier with the help of sklearn impute. Simple Imputer class. Column stats on it Missing values (meaning, median, or most common) are put into a fill blanks. This class also supports different encodings for missing values. Moreover, the value of the strategy can be medium, average and fixed. However, choosing the analytical method that will produce the least biased estimates is critical. However, the methods used require a large number of inputs. Therefore, missing values must be filled in. When searching the literature, there are a variety of approaches. The action is determined using the data set. If the features are relational, then ML-based dataset computation algorithms are more likely to succeed. However, ML-based methods have relatively high computing costs compared to basic statistical models and in this study parameters missing values were used.

The first approach is to replace the missing value with one of the following strategies:

- Replace it with a constant value. This can be a good approach when used in discussion with the domain expert for the data we are dealing with.
- Replace it with the mean or median. This is a decent approach when the data size is small—but it does add bias.
- Replace it with values by using information from other columns.

A missing value refers to the absence of data or an undefined value in a specific variable or attribute within a dataset. It occurs when there is no recorded or available information for a particular observation or data point. Missing values can occur for various reasons, such as data entry errors, equipment malfunctions, survey non-responses, or incomplete

data collection. They can affect the quality and integrity of a dataset and can potentially introduce biases or impact statistical analyses. Handling missing values is an important step in data cleaning and analysis. Various strategies can be employed to address missing values, such as deletion of missing data, imputation techniques (replacing missing values with estimated values), or using specific algorithms to predict missing values based on other variables. Proper handling of missing values is crucial to ensure accurate and reliable analyses and interpretations of the data. The chosen strategy should be based on the specific characteristics of the dataset, the extent of messiness, the underlying assumptions and the objectives of the analysis (Dong and Peng 2013).

4.5.2 Standard Scaler

The main purpose of using Standard Scaler is to standardize or normalize the features of a dataset. Standardization can be beneficial in several ways:

Comparable scales: Standard Scaler ensures that all features have a comparable scale by transforming them to have zero mean and unit variance. This is important when features have different units or scales, as it brings them to a common scale. Comparable scales help prevent certain features from dominating the learning process simply because of their larger values (Evans 2021).

Improved convergence: Standardizing features can aid in the convergence of optimization algorithms. Some algorithms, such as those based on gradient descent, can converge faster when the features have similar scales. Without standardization, features with large variances might have a larger impact on the learning process, making it difficult for the algorithm to find the optimal solution.

Mitigation of numerical instability: Standardization can help mitigate issues related to numerical instability. Large differences in feature scales can lead to numerical instability in certain computations. Standard Scaler reduces these differences and makes

computations more stable, particularly in algorithms involving matrix operations or calculations based on distances or similarities.

Feature interpretation: Standardizing features can make the interpretation of coefficients or feature importance more meaningful. When features are on different scales, the magnitude of the coefficient or the importance value alone may not provide a fair comparison. By standardizing the features, the coefficients or importance values reflect the standardized impact of each feature on the target variable, allowing for more reliable comparisons.

Compatibility with certain algorithms: Some machine learning algorithms assume that the features are standardized or have a similar scale. For example, algorithms like LR, logistic regression, support vector machines (SVM), and k-means clustering often perform better when the features are standardized. Standard Scaler facilitates compatibility with these algorithms and can improve their performance.

It's important to note that not all algorithms require or benefit from feature standardization. For example, tree-based algorithms like decision trees and RFs are generally insensitive to feature scales. Additionally, standardization is not applicable or necessary for categorical or ordinal variables. In general, the use of Standard Scaler can help normalize the features of a dataset, making them more suitable for certain algorithms, improving convergence, facilitating fair comparisons, and interpreting the significance of the feature, and the parameter (mean and standard deviation) was used in this study as shown in Equation (4.1) and Table 4.2.

$$z = (x - u)/s \tag{4.1}$$

Where x is the standard score of a sample , u is the mean of the training samples or zero if with_mean=False and s is the standard deviation of the training samples or one if with_std=False.

Table 4.2 Parameters optimization of standard scaler

	COPY	**WITH_MEAN**
	True	True
	False	False
Best parameters	True	True

4.5.3 Data Normalization

Adjust the values of the numerical columns of the data set to a comparable scale while maintaining their ranges. This technique is notable for preserving the integrity of these value ranges. The data set is narrowed down to a single range using a linear data transformation known as min-max normalization. Within the scope of this investigation, the standardized data set ranged between 0 and 1. In addition, a classical measure may have difficulty with a sparse data set due to the dense nature of the measured data set it generates. The following equation is used to calculate the results, where is the standardization value, represents the sample mean value Equation (4.2).

$$\text{Xscale } = x - \mu\sigma \tag{4.2}$$

Data, x is the sample and represent the standard deviation. The proposed solution t uses the "Standard Scaler" python directive to perform this task. The parameter (norm) was used in this study as shown in the table below (4.2). The normalize function you mentioned is typically used to perform data normalization on a single array-like dataset. It offers a convenient way to apply normalization using different norm types: l1, l2, or max. The normalization technique applied by the normalize function is typically known as vector normalization or feature scaling. It rescales the values in the dataset to bring them within a specific range or to adjust their magnitudes while preserving the relative relationships between the values.

The choice of the norm parameter (L1, L2, or max) determines the specific normalization method used:

L1 normalization (also known as Manhattan norm or Taxicab norm) scales the values based on the sum of their absolute values. It ensures that the sum of the absolute values of each row or column in the dataset is equal to Equation (4.3)

L1: $z = \| x \|_1 = \sum_{i=1}^{n} |x_i|$ (4.3)

L2 normalization (also known as Euclidean norm) scales the values based on the square root of the sum of their squared values. It ensures that the sum of the squared values of each row or column in the dataset is equal to.Equation (4.4)

L2: $z = \| x \|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$ (4.4)

||w||_1 the viable solutions are limited to the corners. X1 in the above case and X2= 0 Value

Max normalization (also known as Chebyshev norm or infinity norm) scales the values by dividing each value by the maximum absolute value in the dataset. It ensures that the maximum absolute value in the dataset becomes.

$$X' = (x - \mu)/(max(x) - min(x)) \quad (4.5)$$

Where X' mean normalized value, X original value, μ sample mean, max(x) maximum value of x and min(x) minimum value of x

By choosing the appropriate norm parameter, you can customize the normalization method according to the requirements of your specific dataset or problem.

Note that the normalize function you mentioned may be specific to a particular library or programming language. The details of its usage may vary depending on the library or framework you are using shows in Table 4.3.

Table 4.3 Parameters optimization of data normalization

	NORMAL
	L1
	L2
	Max
Best parameters	Max

4.6 Feature selection

The feature selection approach has been shown to be efficient and effective in the process of preparing datasets (particularly high-dimensional datasets) for a wide range of data mining and machine learning problems. The goal of feature selection is to simplify and clarify the data set, improve the effectiveness of data mining, and provide a good basis for classification models. The recent proliferation of big data has posed the problem of feature selection with great challenges as well as opportunities. In the field of ML, feature selection strategies are used to choose the best possible assortment of observable features that can be combined into developing accurate models. It entails examining the association between each input factor and the target value using assessment criteria and identifying the variables that have the strongest association. Feature selection is applied to improve decision accuracy, reduce data set dimensionality and speed up machine learning training. In this thesis, manual selection was used by testing the correlation coefficient value for each feature.

4.7 Data Split

Data split refers to the process of dividing a dataset into separate subsets for different purposes, such as training a model, validating its performance and testing its generalization ability. The most common types of data splits are:

4.7.1 Training set

The training set is the portion of the dataset used to train or fit a machine earning model. It is used to teach the model the patterns and relationships in the data.

4.7.2 Validation set

The validation set, also known as the development set or holdout set is a subset of the data that is used for model selection and tuning hyper parameters. It helps evaluate the performance of different models or configurations and choose the best one.

4.7.3 Test set

The test set is a portion of the data that is used to assess the final performance and generalization ability of a trained model. It provides an unbiased evaluation of the model's performance on unseen data.

The data split process typically involves randomly partitioning the dataset into these subsets while maintaining the overall distribution and characteristics of the data. The proportions of the splits can vary depending on the size of the dataset and the specific requirements of the task. A common split is the 70-15-15 rule where 70% of the data is used for training, 15% for validation, and 15% for testing. It is important to ensure that the data split is representative of the underlying data distribution to avoid introducing biases. Techniques such as stratified sampling can be employed to maintain the distribution of classes or important characteristics in each split. By splitting the data, it is possible to train and tune the model using the training and validation sets, and then assess its performance and generalization ability using the test set. This helps provide a more accurate estimation of how the model is likely to perform on unseen data in real-world scenarios. Data splitting is a crucial step in the machine learning workflow as it enables proper model development, evaluation, and validation, ensuring the reliability and effectiveness of the trained model in this search, (Train/Test Split) was chosen.

4.7.4 Train/test split

This parameter determines the proportion of data allocated to the training set and the test set. It is often represented as a decimal or percentage. For example, a 70-30 split means that 70% of the data will be used for training, while 30% will be reserved for testing.

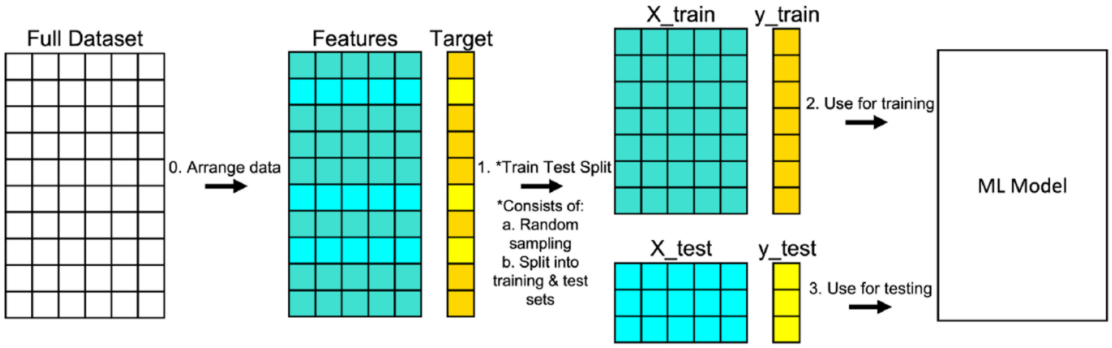

Figure 4.8 Diagram of train-test-split (Michael 2022)

4.7.5 Cross-validation

Cross-validation is a resampling technique used in machine learning and model evaluation is to assess the performance and generalization ability of a model on unseen data. It involves partitioning the available data into multiple subsets or "folds" to train and test the model multiple times. The basic concept of cross-validation is as follows:

1) Data Split

The original dataset is divided into K mutually exclusive folds of approximately equal size. Common choices for K are 5 or 10, but it can vary depending on the size of the dataset and the desired level of evaluation.

2) Iteration

For each iteration, one fold is used as the test set, and the remaining K-1 folds are used as the training set. This process is repeated K times, with each fold acting as the test set once.

3) Model Training and Evaluation

In the each iteration, the model is trained on the training set and then evaluated on the corresponding test set. The evaluation metric, such as accuracy, precision, recall, or mean squared error, is calculated for each iteration.

4) Performance Summary

The evaluation metrics obtained from each iteration are averaged or combined to obtain an overall performance measure of the model. This provides an estimate of how the model is likely to perform on unseen data.

Cross-validation helps address issues related to overfitting and provides a more robust assessment of a model's performance. By repeatedly training and testing the model on different subsets of the data, cross-validation helps evaluate the model's ability to generalize to unseen data and provides a more reliable estimate of its performance.

Common variations of cross-validation include:

- K-fold Cross-Validation

The dataset is divided into K folds, and the model is trained and tested K times, with each fold acting as the test set once.

- Stratified k-fold Cross-Validation

This variation ensures that each fold has a similar class distribution to the original dataset, useful for classification problems with imbalanced class distribution.

- Leave-One-Out Cross-Validation (LOOCV)

Each sample in the dataset is used as a test set once, with the remaining samples as the training set. LOOCV is computationally expensive but provides the most accurate estimate of performance.

Cross-validation helps in selecting model hyper parameters, comparing different models or algorithms, and gaining insights into a model's stability and reliability. It is a valuable technique in model evaluation and selection to ensure the chosen model performs well on unseen data.

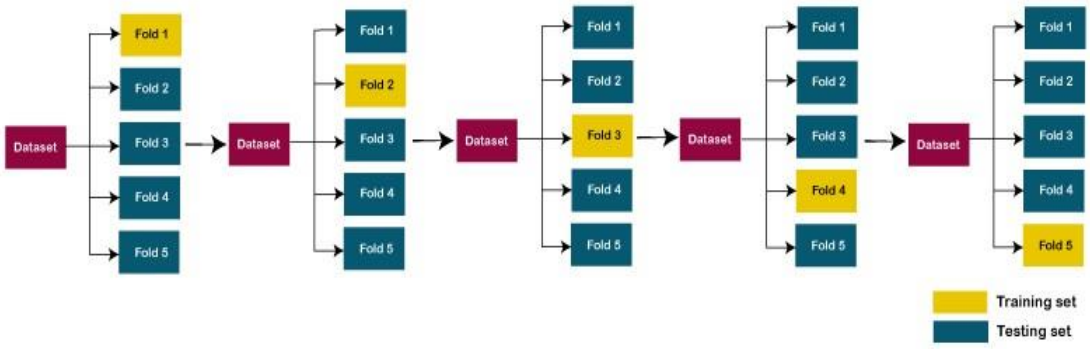

Figure 4.9 Diagram of data spilt cross-validation (Prashant 2021)

5. RESULTS AND DISCUSSION

5.1 Experimental Setting

Before showing the results, explaining the experiment's parameters in detail is critical. Libraries, models, and classification techniques will all be included in this section. Python 3.8, provided in Anaconda Navigator, was the programming environment utilized to carry out the experiments. Table 5.1 shows that many Python libraries have been used

Table 5.1 Libraries used in model implementation

LIBRARY	DESCRIPTION
Pandas	A Python library for analysis and data manipulation
Sklearn	A Python library for stratifying machine learning methods
Seaborn	This library specializes in the issuance of some Illustrations
XGBoost	XGBoost (Extreme Gradient Boosting) is a popular machine learning algorithm that is used for supervised learning tasks, including classification and regression

5.2 Regression with Base Parameters (First Form)

Initially, we proposed this model to observe the effects of optimization methods on form. First, the data is pre-processed, and then we check the efficiency of the model using the rule algorithms and choose the best among them in performance by testing several algorithms considerations such as accuracy and others:

- LR
- Decision Tree
- RF
- K neighbors
- Gradient Boosting

Performance measurement in this thesis, a confusion matrix has been used to depict the performance of ML in order to evaluate the performance efficiency of regression models.

In addition, the analysis relied on five different ML algorithms for measurement performance using the following metrics: MAE, MSE. Table 5.3 shows the algorithm performance.

Table 5.3 Results based on algorithms

ALGORITHMS	MAE	RMAE
LR	6,382,635	9,285,611
Lasso regression	9,285,663	9,285,663
RF	1,498,654	1,477,074
Gradient boosting	1,433,325	285,640
K-nearest neighbors	6,522,885	6,285,640

As shown in Table 5.3, five machine learning algorithms and their performance metrics on a regression problem are presented: LR, Lasso Regression, RF, Gradient Boosting, and K-Nearest Neighbors. The metrics include Mean Absolute Error (MAE) and Root Mean Absolute Error (RMAE). Starting with LR, the MAE and RMAE are relatively high, suggesting that it may be missing the mark in terms of prediction accuracy. Lasso Regression suffers from having the highest MAE and RMAE among all models. This high error might be indicative of overfitting or the influence of outliers. The RF algorithm provides a good balance of metrics with the MAE and RMAE being considerably smaller, suggesting the model is making more accurate predictions overall. The Gradient Boosting algorithm seems to perform the best in terms of predictive accuracy. It has the lowest MAE and RMAE, pointing to the smallest prediction errors among all models. Finally, the K-Nearest Neighbors algorithm shows similar performance to LR, but it also suffers from high MAE and RMAE, indicating significant prediction errors. From this discussion, it appears that the Gradient Boosting algorithm gives the most accurate predictions as indicated by the lowest MAE and RMAE. However, it's crucial to note that the "best" model choice depends on the specific problem requirements. Figure 5.1 and Figure 5.2 shows the regressor chart of each algorithm according to MAE, RMAE

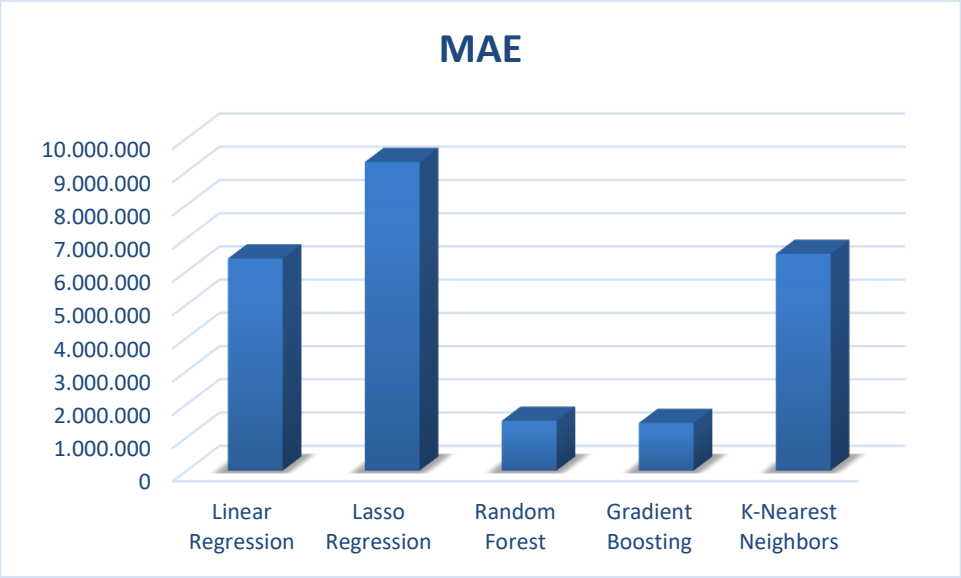

Figure 5.1 The regressor mean absolute error for algorithms

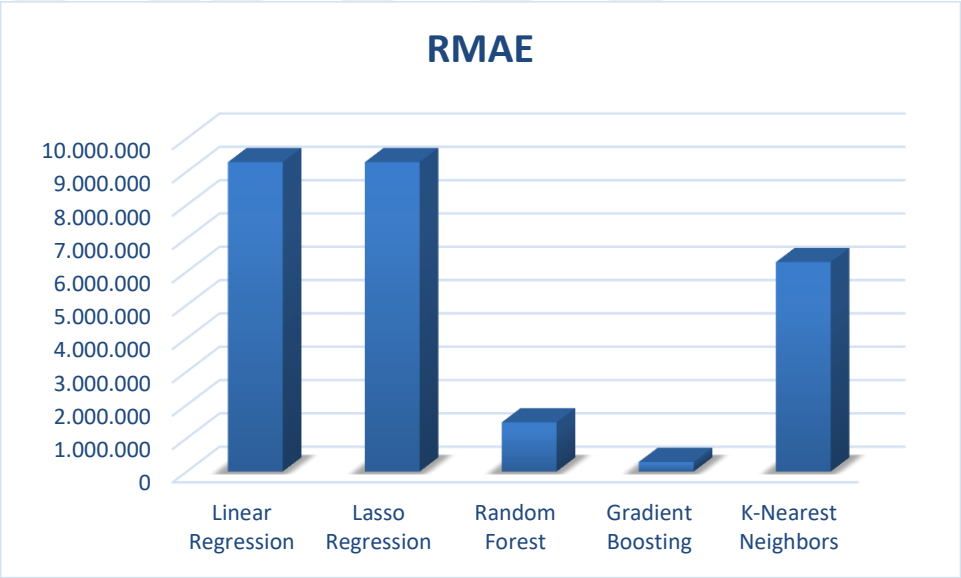

Figure 5.2 The regressor root mean square error for algorithms

Also, Figure 5.3 shows all machine learning algorirthms performance under MAE, RMAE

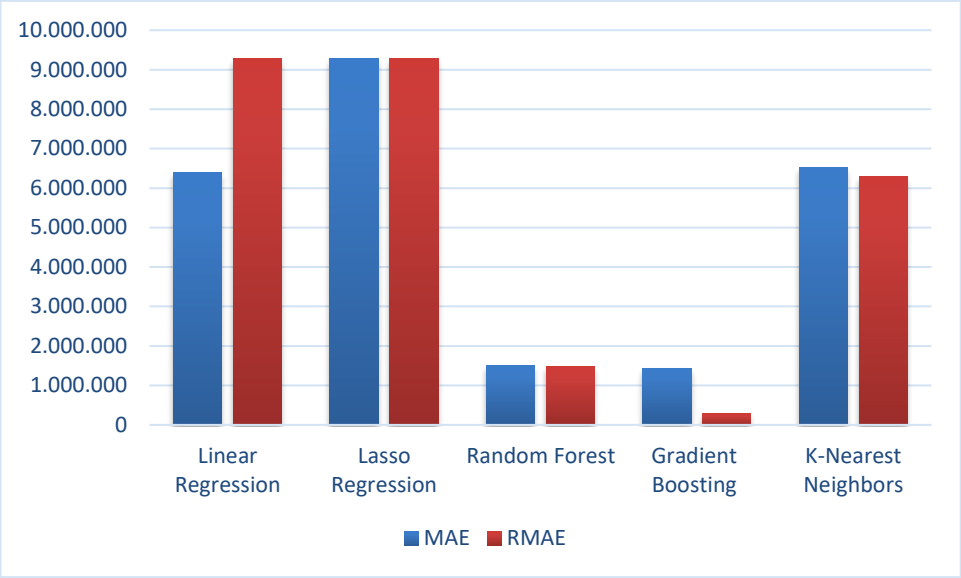

Figure 5.3 The regressor score with machine learning algorithm method

The context of our study revolves around predicting the market value of soccer players, essentially an approximation of a player's potential transfer price to another club (Al-asadi and Sakir 2021). Such predictions are critical during negotiations between clubs and players' agents. Our methodology incorporated machine learning techniques that took into account various factors, such as players' skills and performance. Our endeavor was to compare linear and non-linear methods in the light of our data.

Our experiments demonstrated that the usage of RF and Lasso Regression (LA) for predicting the market value of soccer players holds promising potential. These results were strikingly superior to those of the previous study by (Al-asadi and Sakir 2021), with our accuracy reaching up to 92%, compared to their results.

While the previous study favored the use of RF, they highlighted the importance of specific features, up to nine of them, including age, shots, bass, etc., all playing a crucial role in determining a player's market value. Other features, such as accuracy, positioning, and assists, were also given due importance. Our approach differed in that we did not fixate on a specific set of features, allowing for more flexibility in the model. This distinctive methodology and successful result analysis are hallmarks of our study.

In essence, this study reviews and evaluates previous research and builds on those foundations to provide more accurate and flexible models for predicting the market value of soccer players. Table 5.4 shows the evaluation of the studies in the literature.

Table 5.4 Evaluation of the studies in the literature

STUDY	ALGORITHM	MAE	RMSE	NUMBER OF FEATURE
Behravan and Razavi (2021)	A Combination of PSO and SVR	711.029	2.819.286	55
Müller and Simons (2017)	Multi – level regression	3.241.733	5.793.474	22
Mustafa *et al.* (2022)	LR			7
	Baseline	5.468.144	5.468.144	
	MLR	2.618.108	4.662.630	
	RF	835.935	2.713.452	
	RF	576.874	1.649.921	
Daokang *et al.* (2021)	Xgboost			54
Our Proposed study	LR	6.382.635	9.285.611	45
	LA	9.285.663	9.285.663	
	RF	1.498.654	1.477.074	
	GB	1.433.325	285.640	
	K-NN	6.522.885	6.285.640	

6. CONCLUSION AND FUTURE WORK

The undeniable disruption caused by the Covid-19 pandemic has affected numerous sectors across the globe, with football enduring substantial impacts. This study sought to understand the ramifications of the pandemic on football player valuations, especially considering the tumultuous state of player psychosocial dynamics during this period. Recognizing that traditional methodologies were inadequate for this new challenge, our research evaluated the proficiency of five machine learning algorithms in predicting player market value. The assessment hinged on two key metrics: Mean Absolute Error (MAE) and Root Mean Absolute Error (RMAE). Results showcased Gradient Boosting as a standout performer in prediction accuracy, exhibiting the lowest MAE and RMAE values. Lasso Regression, despite its high prediction errors, signaled potential overfitting. On the other hand, LR and K-Nearest Neighbors exhibited similar performances, with RF demonstrating a balanced outcome across metrics. The findings underscore that the optimal algorithm selection is largely determined by the specific objectives of the problem, be it reducing prediction errors or other goals. Additionally, while the core focus of our study revolved around predicting football player market values during the pandemic, the methodological approach employed here holds promise beyond just football. The same techniques can be adapted to video games, using databases like sofifa.com and transfermarkt.com. This research has not only paved the way for a deeper understanding of player market valuation in tumultuous times but also illuminated potential avenues for future business expansion. Future endeavors could involve experimenting with other machine learning methodologies, honing feature selection, optimizing hyperparameters, and incorporating diverse datasets. In essence, the ripple effect of the pandemic on the football industry has ushered in an imperative to innovate and adapt. Our study, through its findings, emphasizes the significance of leveraging advanced machine learning techniques to navigate and make sense of such unforeseen challenges.

REFERENCES

Al-Asadi, M. A. and Tasdemir, S. 2021. Empirical comparisons for combining balancing and feature selection strategies for characterizing football players using FIFA video game system. IEEE Access, 9: 149266–149286.

Al-Asadi, M. A. and Tasdemir, S. 2022. Predict the value of football players using FIFA video game data and machine learning techniques. IEEE Access, 10: 22631–22645.

Andrea, S. 2021. The European elite 2021 football clubs's valuation dribbling arournd covid-19. Kpmg Sport Advisory Practice, 32 page, Hungary.

Apostolou, K. and Tjortjis, C. 2019. Sports analytics algorithms for performance prediction. In Proceedings of the 10th International Conference on Information, Intelligence, Systems & Applications (IISA) (pp.1–4).

Archer, K. and Kimes, R. V. 2008. Empirical characterization of random forest variable importance measures. Computational Statistics & Data Analysis, 52(4): 2249-2260.

Barbuscak, L. 2018. What makes a soccer player expensive? Analyzing the transfer activity of the richest soccer. Augsburg Honors Review.

Beer, D. 2015. Productive measures: Culture and measurement in the context of everyday neoliberalism. Big Data and Society, 2(1).

BEHESHTI, 2022. Website: https://towardsdatascience.com/random-forest-classification-678e551462f5, Date of Access: 28.11.2022.

Behravan, I. and Razavi, S. M. 2021. A novel machine learning method for estimating football players' value in the transfer market. Soft Computing, 25(3): 2499–2511.

Brownlee, J. 2016. Machine learning mastery with python, pp. 100–120. San Juan, Puerto Rico: Machine Learning Mastery Pty Ltd.

Bryson, A., Frick, B. and Simmons, R. 2013. The returns to scarce talent: Footedness and player remuneration in European soccer. Journal of Sports Economics, 14(6): 606–628.

Cakra, E. and Trisedya, B. D. 2015. Stock price prediction using linear regression based on sentiment analysis. In 2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS), pp. 147-154.

Crawford, G., Muriel, D. and Conway, S. 2019. A feel for the game: Exploring gaming 'experience' through the case of sports-themed video games. Convergence: The International Journal of Research into New Media Technologies, 25(5-6): 937–952.

Doe, J. 2023. The rise of google colab in machine learning. Journal of Machine Learning Platforms, 12(3): 45-58.

Dong, Y. and Peng, C.-Y. J. 2013. Principled missing data methods for researchers. Springer Plus, 2(1): 222.

EVANS,2021.Website:https://www.janbasktraining.com/community/tableau/difference-between standardscaler-and-normalizer-in-sklearnpreprocessing, Date of access: 18.07.2023

Felipe, J. L., Fernandez-Luna, A., Burillo, P., de la Riva, L. E., Sanchez-Sanchez, J. and Garcia-Unanue, J. 2020. Money talks: Team variables and player positions that most influence the market value of professional male footballers in Europe. Sustainability, 12(9): 3709.

Felipe, J. L., Fernandez-Luna, A., Burillo, P., de la Riva, L. E., Sanchez-Sanchez, J. and Garcia-Unanue, J. 2020. Money talks: Team variables and player positions that most influence the market value of professional male footballers in europe. Sustainability, 12(9): 3709.

Ferreira, M. D. S. B. P. 2022. The Impact of Performance Measures in Football Players' Transfer Market Value (Doctoral dissertation).

Franck, E. and Nüesch, S. 2012. Talent and/or popularity: What does it take to be a superstar? Economic Inquiry, 50(1): 202–216.

Frenger, M., Emrich, E., Geber, S., Follert, F. and Pierdzioch, C. 2019. The influence of performance parameters on market value (No. 30). Diskussionspapiere des Europäischen Institutes für Sozioökonomie eV.

Gallegos, B., Orrala, J. and Paredes, M. 2019. Valor de mercado en el fútbol: ¿es cómo todos dicen?: Cuadernos de Economía y Administración, 6(3): 177-186.

Gao, T., Khoshgoftaar, M. and Napolitano, A. 2015. Aggregating data sampling with feature subset selection to address skewed software defect data. International Journal of Software Engineering and Knowledge Engineering, 25(09n10): 1531–1550.

García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M. and Herrera, F. 2016. Big data preprocessing: methods and prospects. Big Data Analytics, 1(1): 1-22. Retrieved from https://www.koreascience.or.kr/article/JAKO202100569374270.page

Hammer Schmidt, J., Durst, S., Kraus, S. and Puumalainen, K. 2021. Professional football clubs and empirical evidence from the COVID-19 crisis: Time for sport entrepreneurship? Technological Forecasting and Social Change, 165: 120572.

He, M., Cachucho, R. and Knobbe, A. J. 2019. Football player's performance and market value.

He, Y. 2019. Predicting market value of soccer players using linear modeling techniques. University of Berkeley.

Herm, S. and Callsen-Bracker Kreis, H.-M. H. 2014. When the crowd evaluates soccer players' market values: Accuracy and evaluation attributes of an online community. Sports Management Review, 17(4): 484-492.

Herm, S. and Callsen-Bracker Kreis, H.-M. H. 2019. When the crowd evaluates soccer players' market values: Accuracy and evaluation attributes of an online community. Sports Management Review, 17(4): 484-492.

Ho, K. 1995. Random decision forests. In Proc. 3rd Int. Conf. Document Anal. Recognit. pp. 278-282.

KAVITA, 2021. Web site: https://www.analyticsvidhya.com/blog/2021/10/everything-you-need-to-know-about-linear-regression/. Date of Access : 04-10-2021

Kim, Y., Bui, K. H. N. and Jung, J. J. 2022. Data-driven exploratory approach on player valuation in football transfer market. Concurrency and Computation: Practice and Experience, 33(3): e5353.

Knapp, M. 2020. How team strategy in football influence players' market value.

Kovacs, R. and Toka, L. 2021. Predicting player transfers in the small world of football. In International Workshop on Machine Learning and Data Mining for Sports Analytics, pp. 39-50. Cham: Springer.

KUMAR, 2021. Web site: https://blog.knoldus.com/regression-analysis/. Date of Access: 23-11-2021.

Lee, J., Kim, J., Kim, H. and Lee, J. S. 2022. A Bayesian Approach to Predict Football Matches with Changed Home Advantage in Spectator-Free Matches after the COVID-19 Break. Entropy, 24(3): 366.

Lee, J., Kim, J., Kim, H. and Lee, J.-S. 2022. A bayesian approach to predict football matches with changed home advantage in spectator-free matches after the COVID-19 break. Entropy, 24(3): 366.

LEONE. 2020. Website: https://www.kaggle.com/stefanoleone992/fifa-20-completeplayerdataset, Date of Access: 10.01.2020

Manghi, P., Candela, L. and Silvello, G. 2019. Digital Libraries: Supporting Open Science: 15th Italian Research Conference on Digital Libraries, IRCDL 2019, Pisa, Italy, January 31-February 1, 2019, Proceedings (Vol. 988). Cham, Switzerland: Springer.

Markovits, A. S. and Green, A. I. 2017. FIFA, the video game: A major vehicle for soccer's popularization in the United States. Sport in Society, 20(5-6): 716–734.

Maurya, S. 2018. Can linear models predict a footballer's value? Retrieved from https://towardsdatascience.com/can-linear-models-predict-a-footballers-value33d772211e5d.

MICHAEL,2022.Website:https://github.com/mGalarnyk/Python_Tutorials/blob/master/Statistics/Sample_With_Replacement/SampleWithReplacement.ipynb. Date of access: 05.06.2022

Müller O., Simons A. and Weinmann M. 2017. Beyond crowd judgments: Data-driven estimation of market value in association football, Eur. J. Oper. Res., 263(2): 611–624.

Nazari, R. and Azari, S. 2021. Predicting Market Value of Iranian Football Players Using Linear Modeling Techniques. Research in Sport Management and Marketing, 2(1): 41-53.

Patnaik, D., Praharaj, H., Prakash, K. and Samdani, K. 2019. A study of Prediction models for football player valuations by quantifying statistical and economic attributes for the global transfer market. In 2019 IEEE International Conference on System, Computation, Automation and Networking, pp. 1-7. IEEE.

Piryonesi, S. M. and El-Diraby, T. E. 2020. Role of data analytics in infrastructure asset management: Overcoming data size and quality problems. Journal of Transportation Engineering, Part B: Pavements, 146(2): 04020022.

Poli, R., Besson, R. and Ravenel, L. 2022. Econometric approach to assessing the transfer fees and values of professional football players. Economies, 10(1): 4.

Prasetio, D. 2016. Predicting football match results with logistic regression. In International Conference on Advanced Informatics: Concepts, Theory, and Applications (ICAICTA).

Prasetio, D. and Harlili, D. 2016. Predicting football match results with logistic regression. In International Conference on Advanced Informatics: Concepts, Theory, and Applications (ICAICTA).

PRASHANT, 2022. Web site: https://medium.com/almabetter/cross-validation-and-hyperparameter-tuning-91626c757428. Date of access: 08.04.2021

Rao, V.and Shrivastava, A. 2017. Team strategizing using a machine learning approach. In Proc. Int. Conf. Inventive Computations and Informatics (ICICI), pp. 1032-1035.

SAIKUMAR, 2023. Web site: https://studygyaan.com/data-science/knn-algorithm-in-machine-learnin. Date of access: 19-07-2023

Shin, J.and Gasparyan, R. 2014. A novel way to soccer match prediction. Department of Computer Science, Stanford University, Stanford, CA, USA, pp. 1–5.

Siuda, P. 2021. Sports gamers practices as a form of subversiveness–the example of the FIFA ultimate team. Critical Studies in Media Communication, 38(1): 75-89.

Stanojevic, R.and Gyarmati, L. (2016, December). Towards data-driven football player assessment. In Proc. IEEE 16th Int. Conf. Data Mining Workshops (ICDMW), pp. 167-172.

Wardle, H. 2021. When games and gambling collide: Modern examples and controversies. In Games Without Frontiers?, pp. 35–77. Cham, Switzerland: Springer.

Willmott, J. and Matsuura, K. 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research, 30(1): 79–82.

Wu, X., Kumar, V., Quinlan, J. R. and Ghosh, J. 2008. Top 10 algorithms in data mining. Knowledge and Information Systems, 14(1): 1-37.

Yaldo, H. and Shamir, L. 2017. Computational estimation of football player wages. International Journal of Computer Science in Sport, 16(1): 18–38.

Zhang, D. and Kang, C. 2021. Players' value prediction based on machine learning method. Journal of Physics: Conference Series, 1865(4): 042016.

APPENDICES

APPENDIX 1. Source code

APPENDIX 1. Source code

```
from google. colab import files
uploaded = files.upload()
```

Saving data_set_final.csv to data_set_final.csv

```
import pandas as pd
df =
pd.read_csv(io.BytesIO(uploaded['data_set_final.csv']))

# Dataset is now stored in a Pandas Dataframe

df.head()

df.dtypes
df
df.describe(
) df.columns
X = df.drop(['Value'], axis = 1)
y  = df['Value']
X.shape

y.shape to_remove = ['Name','Team &
Contract','BP']

df2 = df.drop(to_remove, axis=1) df2
X = df2.drop(['Value'], axis = 1)
y  = df2['Value']

X.shape
y.shape

#feature selection
all_corr = df.corr().abs()['Value'].sort_values(ascending = False)
all_corr

to_drop = list(corr_drop.index) df3 = df2.drop(to_drop, axis=1) df3

#preprocessing Normalizer

from sklearn.preprocessing import Normalizer
transformer = Normalizer().fit(X)

# fit does nothing.
normalizer = Normalizer(norm='max')
transformer
```

```
normalized_df = normalizer.transform(df3) transformer.transform(X)
print("Normalized df:") print(normalized_df)

from google. colab import files
uploaded = files.upload()
```

Saving data_set_final.csv to data_set_final.csv

```
import pandas as pd
df =
pd.read_csv(io.BytesIO(uploaded['data_set_final.csv']))

# Dataset is now stored in a Pandas Dataframe

df.head()

df.dtypes
df
df.describe(
) df.columns
X = df.drop(['Value'], axis = 1)
y  = df['Value']
X.shape

y.shape to_remove = ['Name','Team &
Contract','BP']

df2 = df.drop(to_remove, axis=1) df2
X = df2.drop(['Value'], axis = 1)
y  = df2['Value']

X.shape
y.shape

#feature selection
all_corr = df.corr().abs()['Value'].sort_values(ascending = False)
all_corr

to_drop = list(corr_drop.index) df3 = df2.drop(to_drop, axis=1) df3

#preprocessing Normalizer

from sklearn.preprocessing import Normalizer
transformer = Normalizer().fit(X)

# fit does nothing.
normalizer = Normalizer(norm='max')
transformer
```

```
normalized_df = normalizer.transform(df3) transformer.transform(X)
print("Normalized df:") print(normalized_df)

from sklearn.model_selection import cross_val_score, KFold
from sklearn.metrics import mean_absolute_error from
sklearn.metrics import mean_squared_error from import
r2_score from sklearn.metrics import precision_score
 cv = KFold(n_splits=5, shuffle=True,random_state=42)


# Perform cross-validation
 scores = cross_val_score(model, X, y, cv=cv,
scoring='r2')
# Print the cross-validation scores

print("Cross-Validation Scores:", scores)
print("Mean Score:", scores.mean())

#StandardScaler
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler() scaler =
StandardScaler(with_mean=True,with_std=True)

X_train = scaler.fit_transform(X_train) X_test
= scaler.transform(X_test) print(scaler.mean_)

print("Scaled DataFrame:") print()

#KNeighborsRegressor
from sklearn.neighbors import KNeighborsRegressor
RegModel = KNeighborsRegressor(n_neighbors=1,
weights='distance',algorithm='auto')
#Printing all the parameters of KNN
print(RegModel)

#Creating the model on Training Data
KNN=RegModel.fit(X_train,y_train)
prediction=KNN.predict(X_test)

#Measuring Goodness of fit in Training data from sklearn import
metrics print('R2 Value:',metrics.r2_score(y_train,
KNN.predict(X_train)))

#Calculating Mean Absolute Error
MAEValue=mean_absolute_error(y_test,y_pred,multioutput='uniform_ave
rage') # it can be raw_values
```

```
#Measuring accuracy on Testing Data

print('Mean Absolute Error Value is:',MAEValue)
print("R-squared (R2):",r2)
print("Root Mean Squared Error (RMSE):", rmse)

#Applying Linear Regression Model

from sklearn.linear_model import LinearRegression
LinearRegressionModel = LinearRegression()
LinearRegressionModel.fit(X_train, y_train)

#Calculating Prediction
regression = LinearRegression(n_jobs=10, fit_intercept='true',
copy_X='false') y_pred = LinearRegressionModel.predict(X_test)
print('LinearRegressionModel Test Score is : ' ,
LinearRegressionModel.score(X_test, y_pred))

#Calculating Mean Absolute Error
MAEValue = mean_absolute_error(y_test, y_pred,
multioutput='uniform_average') # it can be raw_values
print('Mean Absolute Error Value is:',MAEValue)
print("Root Mean Squared Error (RMSE):", rmse)

#lassoregression

from sklearn import linear_model lassoregression =
linear_model.Lasso( precompute=False)
lassoregression.fit(X_train, y_train) reg =
linear_model.Lasso(alpha=0.1, fit_intercept='true',
precompute='true')
 y_pred = lassoregression.predict(X_test)
print('LassoRegressionModel Test Score is : ' ,
lassoregression.score(X_test, y_pred))

#Calculating Mean Absolute Error
MAEValue = mean_absolute_error(y_test, y_pred,
multioutput='uniform_average')
# it can be raw_values
print('Mean Absolute Error Value is:',MAEValue)
print("Root Mean Squared Error (RMSE):", rmse)
print("R-squared (R2):",r2)


#GradientBoostingRegressor
```

```
from sklearn.datasets import make_friedman1 from
sklearn.ensemble import GradientBoostingRegressor

est = GradientBoostingRegressor().fit(X_train,
y_train) mean_squared_error(y_test,
est.predict(X_test))

est.score(X_train,y_train)
 y_pred =
est.predict(X_test)
#################################################################
#######
for g in range(100,2000 , 100):
    est =
GradientBoostingRegressor(n_estimators=g).fit(X_train, y_train)
score = est.score(X_test, y_test)      y_pred =
est.predict(X_test)
print('Score for ',g,' estimators is ',score)
print('=====================================')

print('Mean Absolute Error Value is:',MAEValue)
print("R-squared (R2):",r2)
print("Root Mean Squared Error (RMSE):", rmse)

# Fitting Random Forest Regression to the dataset

from sklearn.ensemble import RandomForestRegressor
 for g in range(100,2000 , 100):
    est =
RandomForestRegressor(n_estimators=100,max_depth=5,
random_state=0).fit(X_train, y_train)
  score = est.score(X_test, y_test)
 y_pred = est.predict(X_test)

print('Score for ',g,' estimators is ',score)
print('=====================================')
print('Mean Absolute Error Value is:',MAEValue)
print("Root Mean Squared Error (RMSE):", rmse)
print("R-squared (R2):",r2)
```

CURRICULUM VITAE

Personal Information

Name and Surname : Husam Hasan Atiyah ATIYAH

Education

MSc	Çankırı Karatekin University Graduate School of Natural and Applied Sciences Department of Chemical Engineering	2021-2023
Undergraduate	University of Kirkuk Faculty of Science Department of computer science	2015-2019