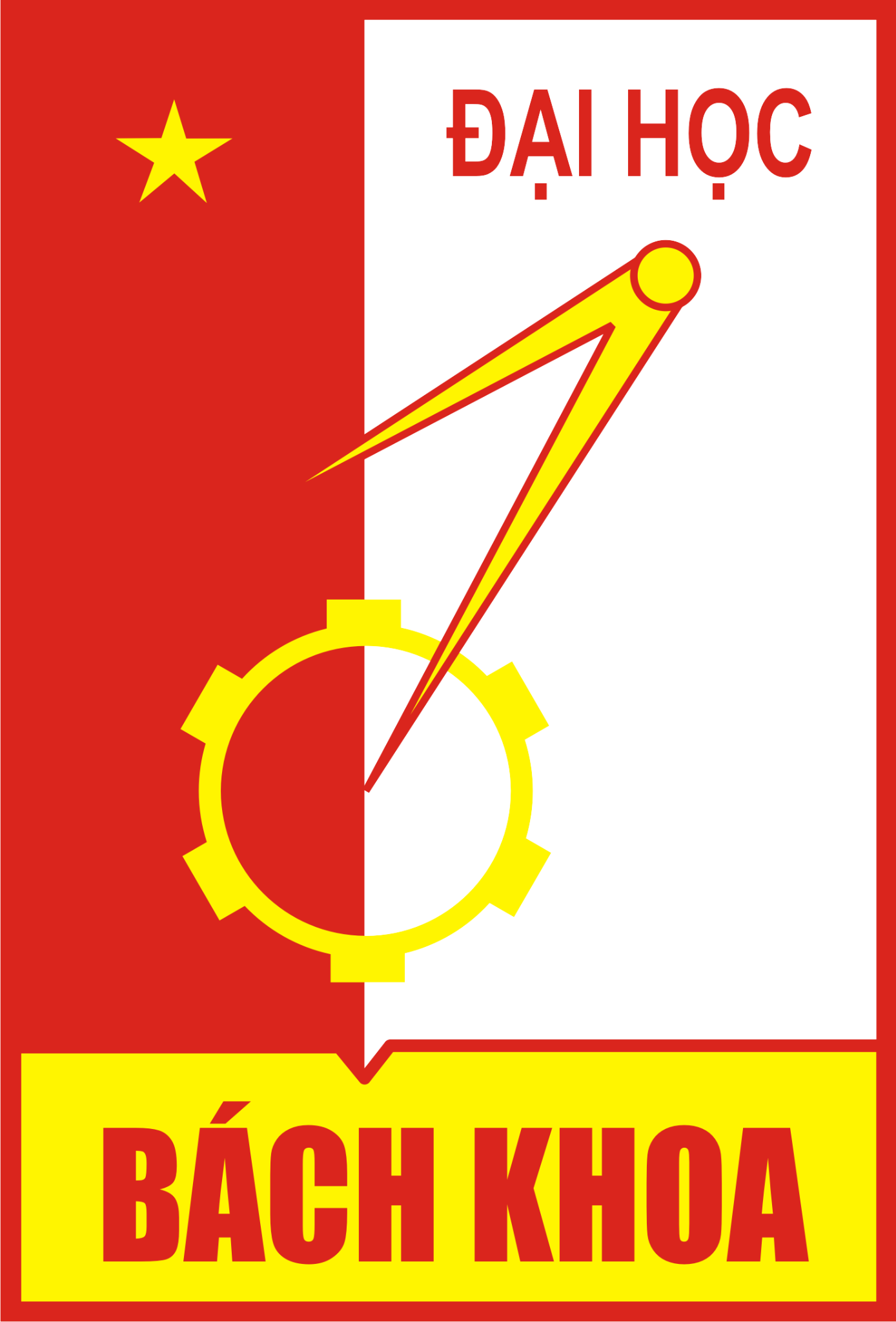
**HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY**

**SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY**



INTRODUCTION TO DATA SCIENCE – IT4142E

PROJECT: FOOTBALL PLAYER PERFORMANCE

ANALYSIS & PREDICTION

Instructor: Assoc Prof. Than Quang Khoat

TA Quang Hieu Pham

Students: Nguyen Quang Minh 20204884

Bui Manh Cuong 20204871

Hoang Anh Chung 20204901

**Table of contents**

[**Assignment** 1](#_Toc127131187)

[**Abstract** 2](#_Toc127131188)

[**1. Introduction** 2](#_Toc127131189)

[**2. Problem formulation** 3](#_Toc127131190)

[**3. Data collection** 3](#_Toc127131191)

[3.1. Data crawling 3](#_Toc127131192)

[3.2. Data scraping 7](#_Toc127131193)

[**4. Data development** 9](#_Toc127131194)

[**5. Data visualization and EDA** 11](#_Toc127131195)

[5.1. General Data 11](#_Toc127131196)

[5.2. Each position data 14](#_Toc127131197)

[**6. Data analysis** 18](#_Toc127131198)

[6.1. Introduction 18](#_Toc127131199)

[6.2. Clustering analysis 18](#_Toc127131200)

[6.2.1. Motivation 18](#_Toc127131201)

[6.2.2. Execution 19](#_Toc127131202)

[6.2.3. Analysis 20](#_Toc127131203)

[6.3. Quantile analysis 28](#_Toc127131204)

[6.4. Mutual information analysis 31](#_Toc127131205)

[6.4.1. Motivation 31](#_Toc127131206)

[6.4.2. Definition 32](#_Toc127131207)

[6.4.3. Analysis 33](#_Toc127131208)

[6.5. Summary 36](#_Toc127131209)

[**7. Experiments** 40](#_Toc127131210)

[**8. Detect outlier by SelectFromModel method** 40](#_Toc127131211)

[**9. Conclusion** 41](#_Toc127131212)

# **Assignment**

Nguyen Quang Minh 20204884: EDA and Artificial Neural Network

Bui Manh Cuong 20204871: Data visualization and Random Forest

Hoang Anh Chung 20204901: Data collection & database development & Data analysis & Ridge regression

# **Abstract**

Football is one of the most popular sports in the world. The need for data science method to analyze player performance is becoming more and more vital for many tasks in the sports. Many football teams in the world have depend on data science to make decisions on signing, selling players, analyze player performance of the team to man-manage players, find the best methods to improve player quality. In this project we are looking to deep dive into player statistics from around the world, investigate them, summarize their characteristics and use multiple machine learning models to predict the rating for each player.

# **1. Introduction**

Nowadays, football teams and organizations are always interested in player performance data to predict the quality of a player to make plans for that player. We want to make that prediction task less complicated in this project based on our knowledge on data science.

We will first collect the data…

Through data preparation and the use of machine learning models like Linear Regression, Decision Tree, and Random Forest, we will provide practical solutions to the aforementioned problem and data set. We will be able to evaluate data, visualize graphs, and create machine learning models quickly and easily with the help of these models.

We shall describe the process of problem solving in this report. We start by dig through the available documentation to identify the problem. Then, we take the readily available data set and concentrate on in-depth mining and analysis. Finally, using machine learning models, we will develop a solution whose main objective is to assess the characteristics of our data set and predict the rating of football players

# **2. Problem formulation**

Our first step in this project and one of the most important one is to define the problem we want to solve in order to plan the correct techniques and appropriate solutions.

Our problem is a supervised regression problem as we want to obtain the rating of football players based on their past performances to correctly predict the rating for future performance

Our problem can be described as a triple (P, T, E)

* Task (T): Predict the correct rating of football players
* Performance (P): The loss error of the prediction, which is the difference between the real rating and predicted rating
* Experience (E): List of the rating of previous performances’ statistics of the players

The main aim after we have identified the problem definition is to find the best solutions of the problem. We believe our models can tackle the problem efficiently, help the players improve their performance and give the teams more chance to make correct man-management decisions on those players.

# **3. Data collection**

## 3.1. Data crawling

Data is crawled from whoscored.com website, with a total of more than 20 different leagues. Because in the new seasons from 2018, the tournaments include VAR technology, along with more modern television technologies, the parameters will be saved more fully, so we will collect data from 3 seasons. The last tournament from 2019 - 2022, the season 2022 - 2023 has not ended, so we do not have enough player data.

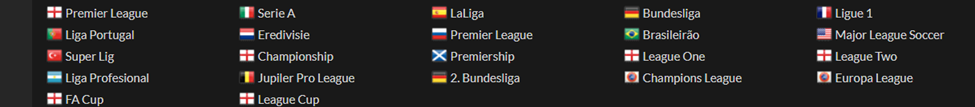


Figure 1. More than 20 prestigious tournaments in Whoscored.com

To be able to access the statistics saved from the tournament, we proceed to access each tournament one by one. We immediately select detailed tab to see detailed statistics

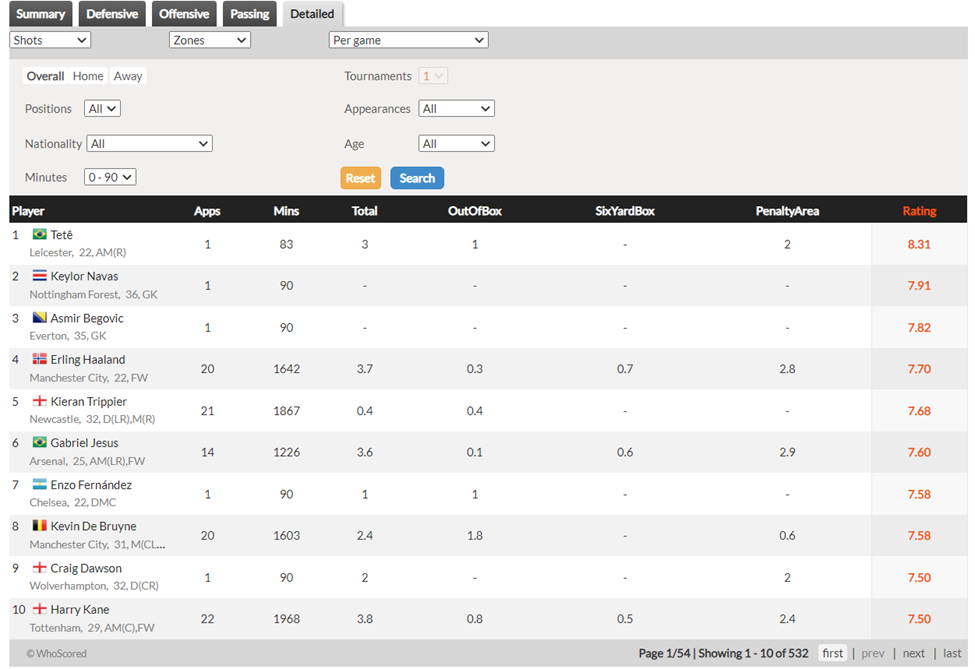


Figure 2. Detail tab of a tournament

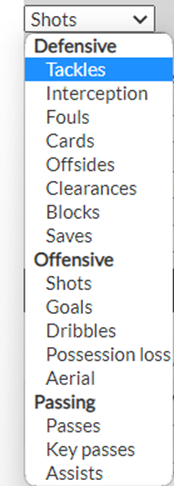


Figure 3. Categories of statistics table

In each tournament, we always have 14 main categories. When selecting a category, 4 different subcategories will appear for the user to choose to export to the statistics table.

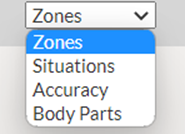


Figure 4. Subcategories of the "shots" category

However, the data here is quite difficult to collect fully, because the website source code is dynamic java-script, so we will find a way to retrieve the original link to the table.

By using google chrome's network tracking we can easily trace the origin. To do this, we will perform a switch between subcategory so that network tracking can detect the access links. As shown below, it can be the original link to the data of the table.

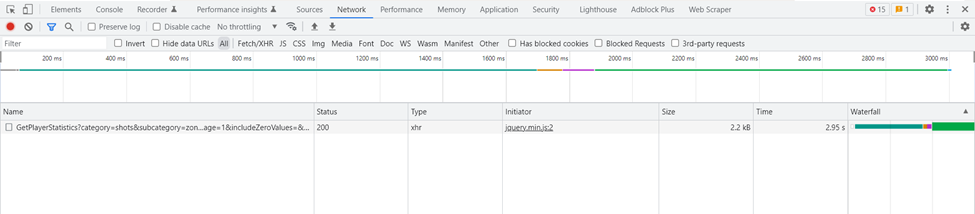


Figure 5. Network tracking table when switching category.

Here's a link to the statistics table https://1xbet.whoscored.com/StatisticsFeed/1/GetPlayerStatistics?category=shots&subcategory=zones&statsAccumulationType=0&isCurrent=true&playerId=&teamIds=&matchId=&stageId=20934&sortAssort=20934&tournamentOptions=ating =&ageComparisonType=0&appearances=&appearancesComparisonType=0&field=&nationality=&positionOptions=%27FW%27,%27AML%27,%27AMC%27,%27AMR%27,%27ML%27,%27MC%27,%27MR%27,%27DMC %27,%27DL%27,%27DC%27,%27DR%27,%27GK%27,%27Sub%27&timeOfTheGameEnd=5&timeOfTheGameStart=0&isMinApp=&page=1&includeZeroValues=&numberOfPlayersToPick=10

Notice that this link contains, elements such as category, and subcatory as shown in Figure 3, so we will scrappe this information, to be able to access all the remaining table sources.

Besides, by accessing each tournament , we see that each season of tournaments will have its own stageId, so our task is to access each season of each tournament to scrape all the stageIds, along with there, element position to change position. However, position is not counted as a feature in the data, so we have to manually add this feature to the data.

All information about stageId, category, subcategory, position is saved in json files, serving to generate url.

All implemented using scrappy\_playwright framework library

The reason for using it is because scrapy\_playwright is a library that supports the headless browser

o Faster speed than selenium

o Support reading "Asynchronous JavaScript" effectively, when we see on the web page there are many elements in the form of java-script.

o Below, is the data stored in a web page. It can be seen that the data is saved as a java-script, so we need to parse it to be able to use it.



Figure 6. Json text in whoscored statistics source link

To be able to collect stageId, tournaments, we need to understand the html text structure of the web page, then know the css selector address of the elements

o First to find the stageId of the Premier League season 2021 – 2022, we look for the stageId component of the website, we find that the stageId is stored in a javascript script, this code is very difficult to extract, so we find code 19793

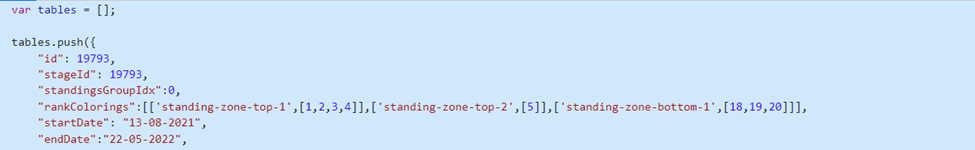
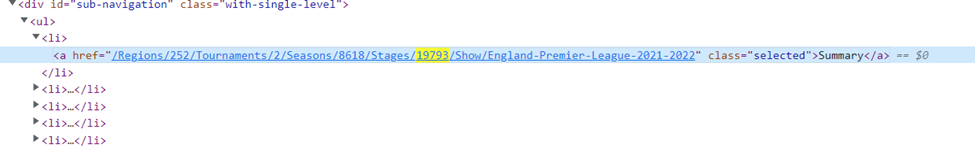


Figure 7. Javascript text contains stageId

That way we can find the stageId of the Premier League season 2021 – 2022, we need to

inspect the website, and find the css address: head > link:nth-child(34).

Figure 8. Link url text contains stageId

After getting content from that element, we will save to json file.Elements such as position, category are also crawled similarly.

After having enough data needed for generating url, we proceed to do this by saying above is to take sample url, to make a sample.

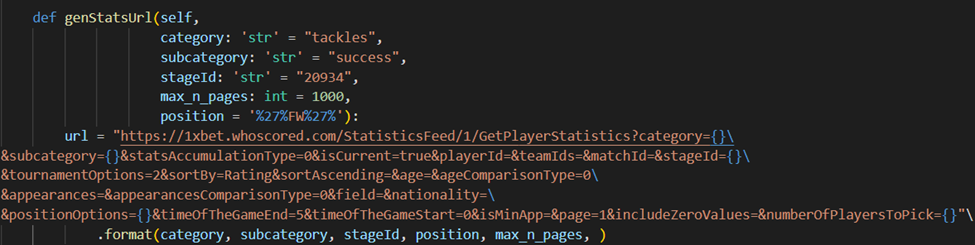


Figure 9. Code for generating url

## 3.2. Data scraping

After getting the set of links, scraped from the above step, we proceed to access each link to get all the data.

Here we tweak a bit in the parameter “&numberOfPlayersToPick”, we tweak it to 1000 to get the number of players right in 1 web page, to speed up the scrape time.

In this step, we extract the data using the scrapy wright framework json library to convert the data in the form of a dict, thereby converting it to a dataframe.

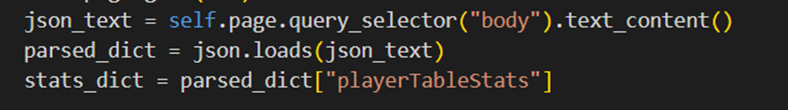
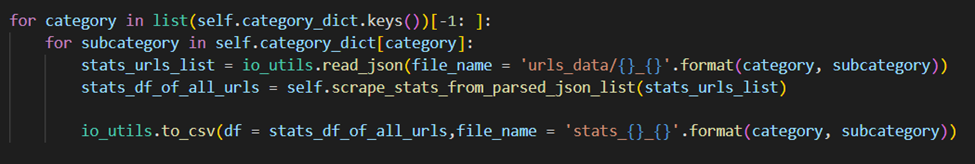


Figure 10. Demonstration code for parsing javascript

Each category and subcategory pair will have its own statistics table, so we will create a separate csv file for each of these pairs.

Figure 11. Demonstration code of converting statistics to csv.

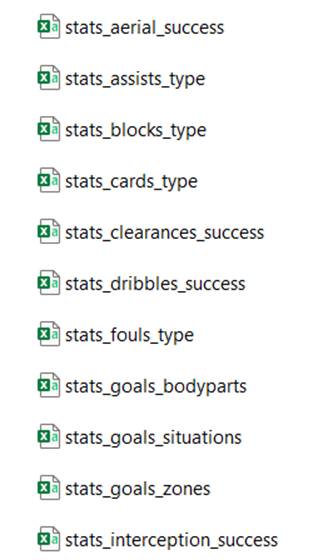


Figure 12. File csv for statistics

**3.3 Database development**

There are a total of 23 data tables, but each data has duplicate attributes, so we have to normalize the table to make query easier and remove redundant attributes. Here are the stored stats files:

o stats\_aerial\_success.csv stats\_key-passes\_type.csv

o stats\_assists\_type.csv stats\_offsides\_type.csv

o stats\_blocks\_type.csv stats\_passes\_length.csv

o stats\_cards\_type.csv stats\_passes\_type.csv

o stats\_clearances\_success.csv stats\_possession-loss\_type.csv

o stats\_dribbles\_success.csv stats\_saves\_shotzone.csv

o stats\_fouls\_type.csv stats\_shots\_accuracy.csv

o stats\_goals\_bodyparts.csv stats\_shots\_bodyparts.csv

o stats\_goals\_situations.csv stats\_shots\_situations.csv

o stats\_goals\_zones.csv stats\_shots\_zones.csv

o stats\_interception\_success.csv stats\_tackles\_success.csv

o stats\_key-passes\_length.csv

# **4. Data development**

Observing the attributes of the statistics tables, we see that all csv files contain common attributes:

o height

o weight

o age

o isManOfTheMatch

o isActive

o playedPositions

o playedPositionsShort

o teamRegionName

o regionCode

o tournamentShortName

o apps

o subOn

o name

o firstName

o lastName

o playerId

o positionText

o teamId

o teamName

o seasonId

o seasonName

o isOpta

o tournamentId

o tournamentRegionId

o tournamentRegionCode

o tournamentRegionName

o tournamentName

o rating

o minsPlayed

o ranking

o positionCode

So we will separate these features from all csv files, creating a new csv file named info

In which features are personal information of players, height

* weight
* age
* playedPositions
* playedPositionsShort
* name
* firstName
* lastName
* playerId

Features regarding the team:

* teamId
* teamName
* teamRegionName

Features regarding the season:

* seasonId
* seasonName
* tournamentId
* tournamentRegionId
* tournamentRegionCode
* tournamentRegionName
* § tournamentName
* § tournamentShortName

Diagram

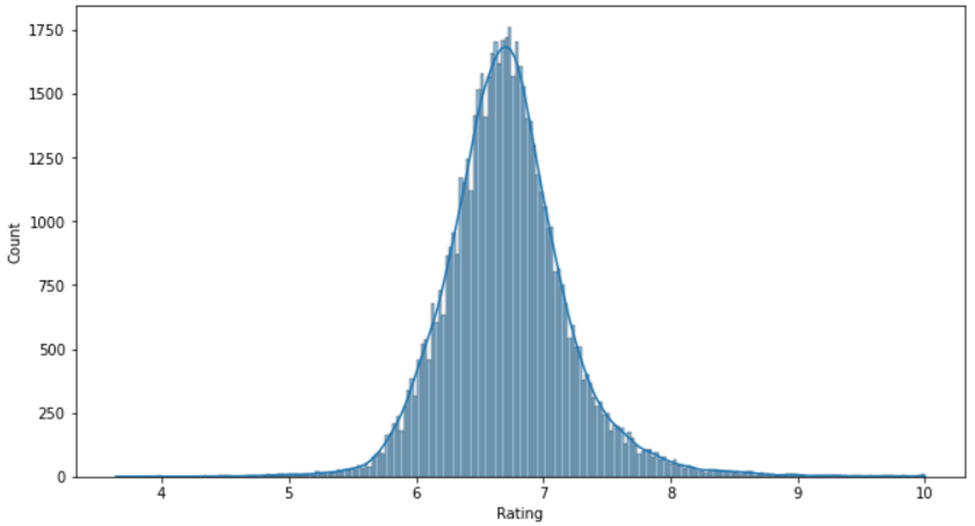
Description automatically generated

**Fig.1. Rating distribution plot of all data**

# **5. Data visualization and EDA**

## 5.1. General Data

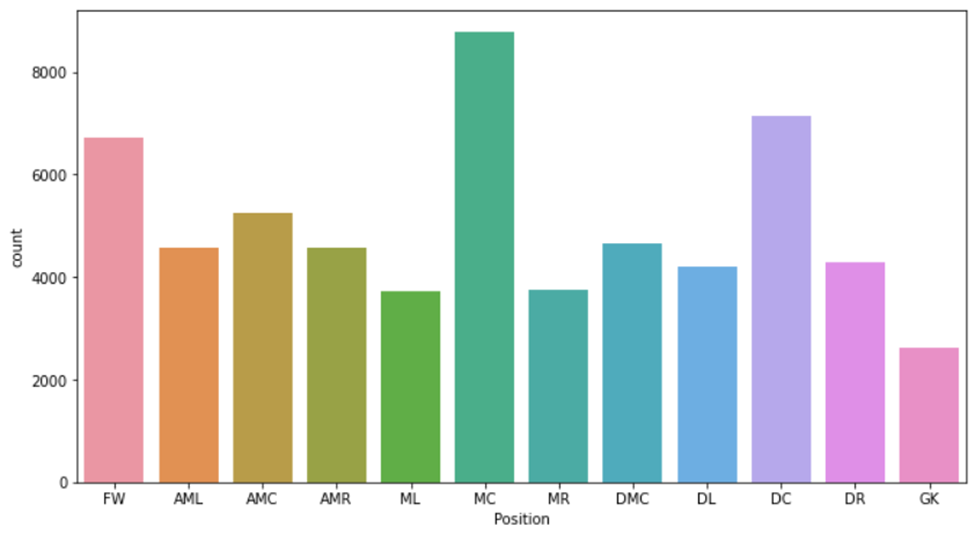
An important part of machine learning's problem solving is data visualization. We will give specific views through the analysis of each chart.



**Fig.2. Rating distribution plot of all data**

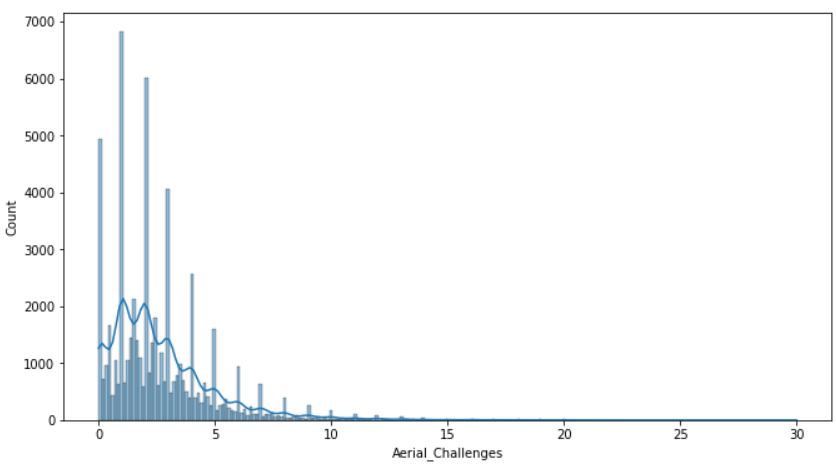
Here we have the distribution plot of the target variable ‘Rating’. We can clearly see that the distribution of ‘Rating’ has a bell shape and the highest amount of rating is around 6.5 to 6.8.

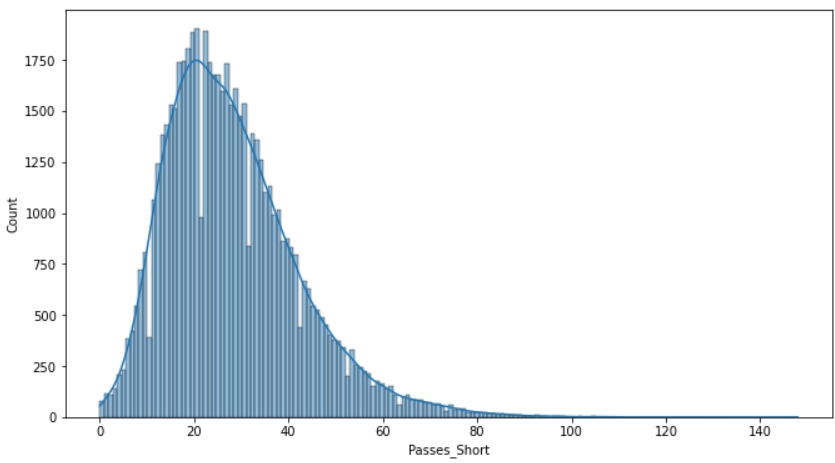
Next we move on to categorical feature.



**Fig.3. Positions count plot for all data**

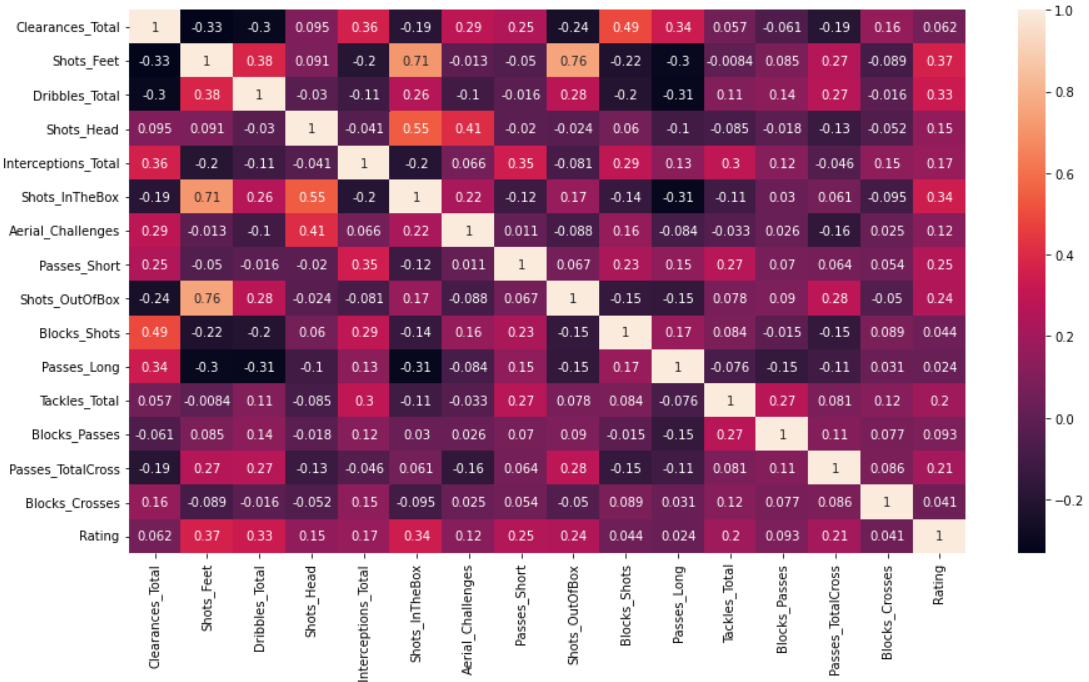
The ‘Position’ variable has a total amount of 12 positions. The highest amount comes from MC position with around over 8000 players being observed play in this positon, almost 3 times more than the lowest, which is GK, at around 2500 players. The following are the distribution plot of some continuous variables





**Fig. 4. Distribution plot for Aerial\_Challenges and Passes\_Short**

We can clearly see that an average player make around under 5 aerial challenges in a match and around 30 short passes.

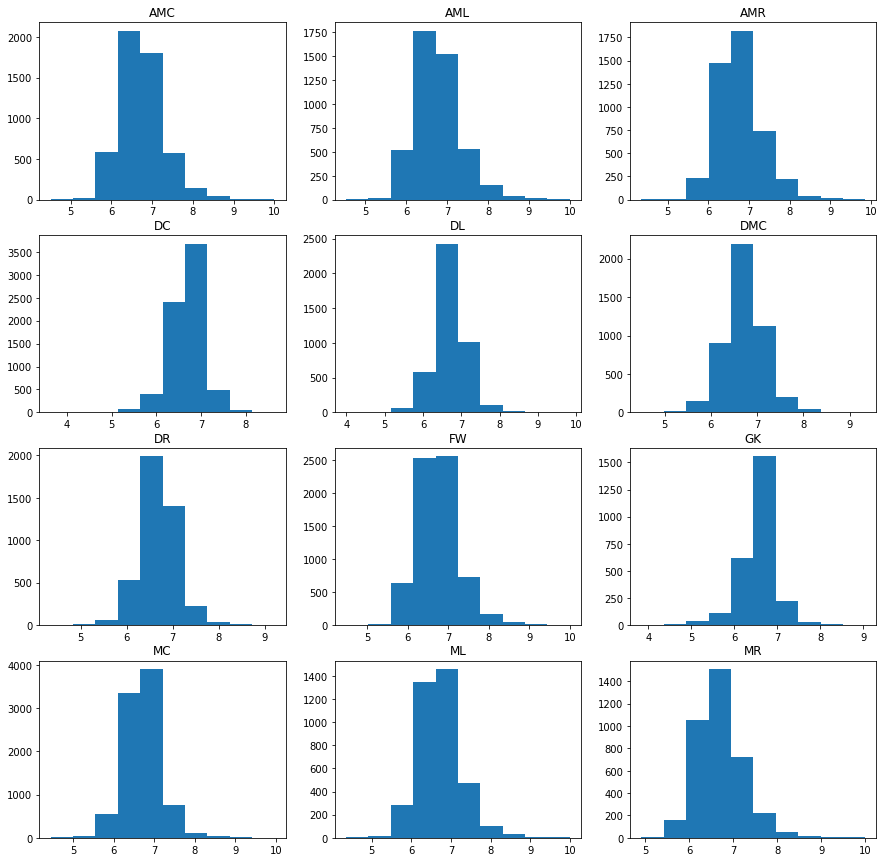


**Fig. 5. Correlation coefficient matrix for all features in data sets**

Here we have the correlation plot of all the continuous variable in our data. Regarding the relationship of the features with our target variable, our target variable has the strongest relationship with Shots\_Feet and Shots\_InTheBox, with the correlation coefficient being 0.37 and 0.34 respectively. It seems like the higher amount of shots a player takes in a match the higher the rating of that player performance

## 5.2. Each position data

The rating plot for each position has little differences with the rating distribution plot for all data. Unlike the plot of all data, which are almost normally distributed, the AMC,AML, AMR, ML, MR have vertices skewed to the left, while the graphs DC, DL, DMC, GK have vertices slanted to the right.



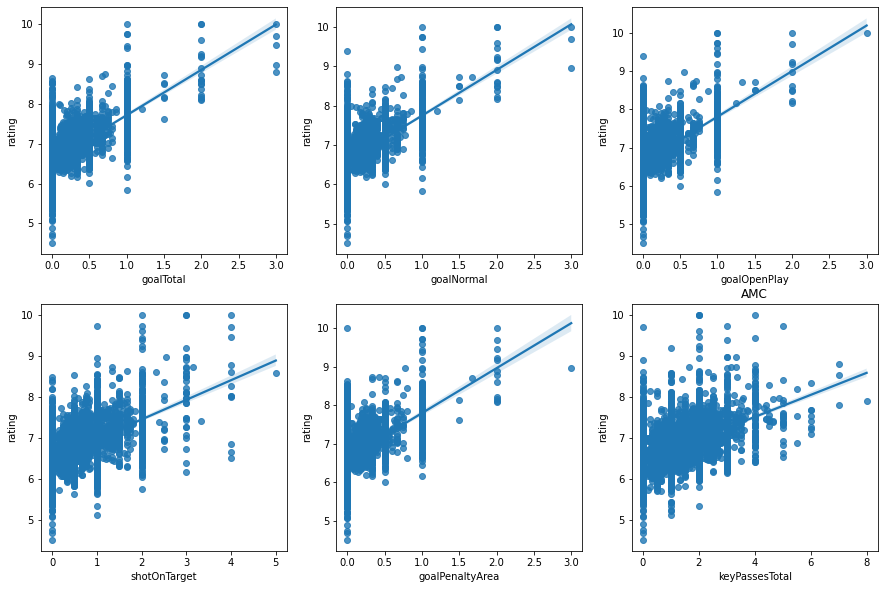
**Fig. 6. Rating distribution plots for each position**

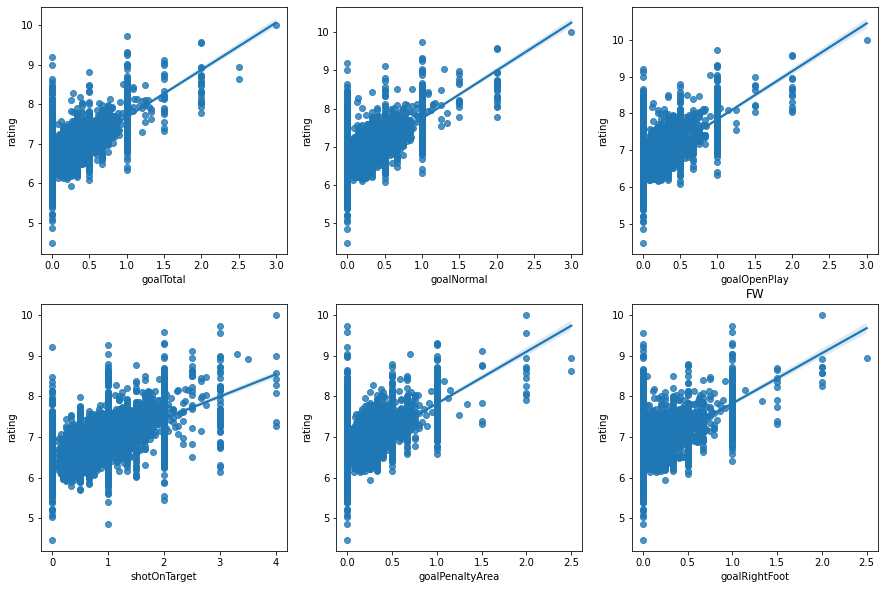
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Mean** | **Standard deviation** | **Variance** | **Median** |
| **AMC** | **6.7461** | **0.5479** | **0.3002** | **6.69** |
| **AML** | **6.7756** | **0.5663** | **0.3207** | **6.71** |
| **AMR** | **6.7574** | **0.5479** | **0.3002** | **6.705** |
| **DC** | **6.6632** | **0.4007** | **0.1605** | **6.6882** |
| **DL** | **6.6904** | **0.4603** | **0.2119** | **6.6887** |
| **DMC** | **6.7130** | **0.4338** | **0.1882** | **6.7063** |
| **DR** | **6.6641** | **0.4277** | **0.1829** | **6.6680** |
| **FW** | **6.7494** | **0.5151** | **0.2653** | **6.7125** |
| **GK** | **6.5558** | **0.4312** | **0.1859** | **6.5896** |
| **MC** | **6.7167** | **0.4463** | **0.1992** | **6.6979** |
| **ML** | **6.7154** | **0.5481** | **0.3004** | **6.6866** |
| **MR** | **6.7027** | **0.5354** | **0.2866** | **6.66** |
| **All positions** | **6.7238** | **0.5028** | **0.2528** | **6.6931** |

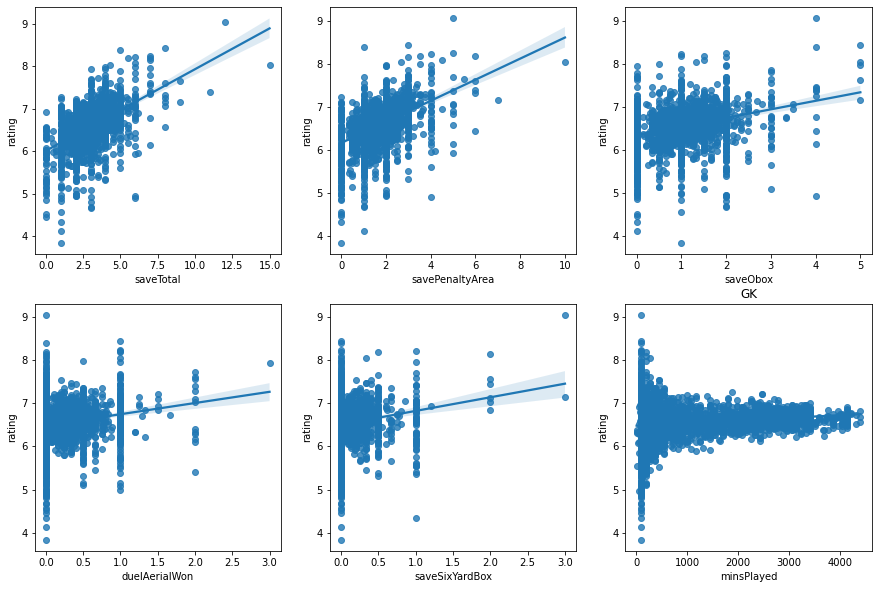
**Table.1. Mean, Standard deviation, Variance, Median of each position and All positions**

In Table.1, we can see that the rating mean and rating value distribution of Attacking positions (FW, AMC, AMR, AML) are higher than general, while the figures for Defending positions ( DC, DL, GK,...) are smaller.

In each position, we plot regression between rating and six attributes have highest correlation with rating. Here is some plot in prominent positions:

**Fig. 7. Regression plots for AMC position**



**Fig.8 . Regression plots for FW position**

**Fig. 9. Regression plot for GK position**

# **6. Data analysis**

## 6.1. Introduction

There are all 82 features to evaluate a player in this set of stats, but each position has its own indicator to evaluate the player. For example, a goalkeeper who compares goal stats with strikers will probably be lame, or a defender who compares defensive stats such as tackles will generally be much better than a goalkeeper. a striker.

## 6.2. Clustering analysis

### 6.2.1. Motivation

Usually a group of players in the same position will have outstanding stats, for example a group of goalkeepers will have a more prominent Save parameter than the others. With a striker player will have more prominent Goal or shot parameters than other positions. Suppose, when comparing 2 players in the 2 groups above, we will see that the sum of the parameter differences between the 2 players will be larger than comparing 2 players in a group.

Therefore, from here we think now to the Euclidean distance between the two players and use the K-means clustering algorithm to analyze the difference between the players.

In a nutshell, we use k-means clustering to calculate the index similarity between players.

### 6.2.2. Execution

Techique here we do it by dividing the group of players into 12 clustering equivalent to 12 positions on the data.

After the clustering is done, there will be players belonging to different positions sharing a clustering. The reason is because, there are a few main indicators of the player class that have a very large difference, while the secondary parameters are smaller, that is, some players participate in secondary stats more than the main role. For a specific example, there are some AMC players participating in the attack, maybe they will have many shots on par with one FW, then these AMCs are mixed in a group with many FWs.

During algorithm training, initialization centroids are random, and are tested by running it over and over again.

After running the algorithm, since there are 82 attributes, we take the first 10 attributes as a sample, we have the following table:

Graphical user interface

Description automatically generated

***Fig 10. Centroids table at specific attributes.***

Dựa vào clustering và vị trí cầu thủ ta có bảng couting giữa vị trí và số cầu thủ trong 1 clustering

* A picture containing table

  Description automatically generated

*Fig 11. Counting table for 12 clusters and positions*

### 6.2.3. Analysis

Taking an attribute as analysis, we will have a detailed look at how to analyze, to find out how to generalize the method to analyze then other attributes will be similar,

Considering the duelAerialLost attribute, look at Figure 14:

* Cluster 1th is dominant with its centroid > 5, outperforming other clusters, which shows that in this first cluster, players will tend to dispossess the ball on the aerial a lot.
* Next, we look at the Figure 15 statistics table, we see that many FW players are included in this group with up to 798 players, this tells us that there are many FW players with large duelAerialLost parameters, so duelAerialLost will be one of the indicators to evaluate FW. In addition to FW, there are other types of players such as AMC, AMR, DC > 100. So we also temporarily leave these parameters as parameters to evaluate the types of players mentioned above.
* Continuing from Table Figure 14, we observe that the 2nd and 3rd highest clusters are 5 and 6, by the same analysis as above, we know this parameter can be used in all areas. all positions except GK. Because in cluster 5, in this cluster, only GK has a small number of players in this group (80).
* There are also types, we, notice that there are 11th and 8th clusters, with very low centroids in this parameter, especially cluster 11, it is close to 0. Which in cluster 11, accounts for most of the majority. number is Goalkeeper. Therefore, it is certain that this parameter will not be selected for GK evaluation.
* Note that the threshold to select a positon at full is 100, if the number of positions is > 100, we will choose that position corresponding to the selection or not of taking that position

Completing the same, we evaluate for other attributes, we have the following table

|  |  |  |  |
| --- | --- | --- | --- |
| FW | AML | AMC | AMR |
| passCrossBlockedDefensive keyPassCorner assist goalHead shotBlocked assistOther dribbleWon shotCounter passCrossInaccurate saveTotal dribbleTotal foulCommitted goalOpenPlay shotOnPost shotSetPiece dribbleLost shortPassInaccurate redCard shotsTotal goalObox shotSixYardBox penaltyTaken duelAerialTotal goalCounter assistCorner shotLeftFoot keyPassOther keyPassShort goalTotal passCrossAccurate shotObp keyPassFreekick goalNormal dispossessed outfielderBlockedPass shotOffTarget goalObp assistFreekick goalLeftFoot savePenaltyArea turnover duelAerialWon keyPassLong goalSixYardBox offsideGiven saveObox duelAerialLost penaltyScored shotHead keyPassesTotal shotOboxTotal saveSixYardBox assistCross shotPenaltyArea foulGiven keyPassThrowin shotRightFoot shotOnTarget goalRightFoot assistThrowin passCornerInaccurate shotOpenPlay goalOwn goalSetPiece passCornerAccurate keyPassCross goalPenaltyArea | passCrossBlockedDefensive keyPassCorner assist goalHead shotBlocked assistOther dribbleWon shotCounter assistThroughball passCrossInaccurate saveTotal dribbleTotal foulCommitted goalOpenPlay shotOnPost dribbleLost shotsTotal redCard goalObox shotSixYardBox penaltyTaken goalCounter assistCorner shotLeftFoot keyPassOther keyPassShort challengeLost goalTotal passCrossAccurate shotObp keyPassFreekick goalNormal dispossessed outfielderBlockedPass shotOffTarget goalObp assistFreekick goalLeftFoot savePenaltyArea turnover keyPassLong tackleWonTotal goalSixYardBox offsideGiven saveObox duelAerialLost penaltyScored shotHead keyPassesTotal shotOboxTotal saveSixYardBox assistCross shotPenaltyArea foulGiven keyPassThrowin shotRightFoot shotOnTarget goalRightFoot assistThrowin passCornerInaccurate passFreekickInaccurate shotOpenPlay goalOwn tackleTotalAttempted passCornerAccurate keyPassCross goalPenaltyArea | passCrossBlockedDefensive keyPassCorner assist goalHead shotBlocked assistOther dribbleWon shotCounter assistThroughball passCrossInaccurate saveTotal dribbleTotal foulCommitted goalOpenPlay shotOnPost shotSetPiece dribbleLost shortPassInaccurate redCard shotsTotal goalObox clearanceTotal shotSixYardBox penaltyTaken goalCounter duelAerialTotal assistCorner shotLeftFoot keyPassOther keyPassShort challengeLost goalTotal passCrossAccurate shotObp keyPassFreekick goalNormal dispossessed outfielderBlockedPass shotOffTarget goalObp assistFreekick goalLeftFoot savePenaltyArea turnover duelAerialWon keyPassLong tackleWonTotal goalSixYardBox offsideGiven saveObox duelAerialLost penaltyScored yellowCard shotHead keyPassesTotal shotOboxTotal saveSixYardBox assistCross shotPenaltyArea foulGiven keyPassThrowin shotRightFoot shotOnTarget goalRightFoot assistThrowin outfielderBlock passCornerInaccurate passFreekickInaccurate shotOpenPlay goalOwn tackleTotalAttempted goalSetPiece passCornerAccurate keyPassCross goalPenaltyArea | passCrossBlockedDefensive keyPassCorner assist goalHead shotBlocked assistOther dribbleWon shotCounter assistThroughball passCrossInaccurate saveTotal dribbleTotal foulCommitted goalOpenPlay shotOnPost shotSetPiece dribbleLost shortPassInaccurate redCard shotsTotal goalObox shotSixYardBox penaltyTaken goalCounter duelAerialTotal assistCorner shotLeftFoot keyPassOther keyPassShort challengeLost goalTotal passCrossAccurate shotObp keyPassFreekick goalNormal dispossessed outfielderBlockedPass shotOffTarget goalObp assistFreekick goalLeftFoot savePenaltyArea turnover duelAerialWon keyPassLong tackleWonTotal goalSixYardBox offsideGiven saveObox duelAerialLost penaltyScored shotHead keyPassesTotal shotOboxTotal saveSixYardBox assistCross shotPenaltyArea foulGiven keyPassThrowin shotRightFoot shotOnTarget goalRightFoot assistThrowin passCornerInaccurate passFreekickInaccurate shotOpenPlay goalOwn tackleTotalAttempted goalSetPiece passCornerAccurate keyPassCross goalPenaltyArea |

|  |  |  |  |
| --- | --- | --- | --- |
| ML | MC | MR | DMC |
| passCrossBlockedDefensive keyPassCorner assist goalHead shotBlocked assistOther dribbleWon shotCounter assistThroughball passCrossInaccurate saveTotal dribbleTotal foulCommitted goalOpenPlay shotOnPost dribbleLost shotsTotal redCard goalObox shotSixYardBox penaltyTaken goalCounter assistCorner shotLeftFoot keyPassOther keyPassShort challengeLost goalTotal passCrossAccurate shotObp keyPassFreekick goalNormal dispossessed outfielderBlockedPass shotOffTarget goalObp assistFreekick goalLeftFoot savePenaltyArea turnover keyPassLong tackleWonTotal goalSixYardBox offsideGiven saveObox duelAerialLost penaltyScored shotHead keyPassesTotal shotOboxTotal saveSixYardBox assistCross shotPenaltyArea foulGiven keyPassThrowin shotRightFoot shotOnTarget goalRightFoot assistThrowin passCornerInaccurate passFreekickInaccurate shotOpenPlay goalOwn tackleTotalAttempted passCornerAccurate keyPassCross goalPenaltyArea | passCrossBlockedDefensive assist assistOther assistThroughball saveTotal foulCommitted redCard goalTotal shotOffTarget savePenaltyArea duelAerialWon goalSixYardBox saveSixYardBox foulGiven goalRightFoot assistThrowin shotOpenPlay passFreekickAccurate shotBlocked shotCounter shotSetPiece shotsTotal goalObox assistCorner keyPassShort passCrossAccurate passLongBallInaccurate goalObp shortPassAccurate keyPassLong tackleWonTotal penaltyScored shotHead shotPenaltyArea shotRightFoot tackleTotalAttempted goalOwn goalSetPiece keyPassCross passTotal goalPenaltyArea dribbleWon passCrossInaccurate goalOpenPlay dribbleLost shortPassInaccurate shotSixYardBox goalCounter duelAerialTotal shotLeftFoot keyPassOther challengeLost keyPassFreekick goalNormal dispossessed outfielderBlockedPass passLongBallAccurate interceptionAll saveObox duelAerialLost keyPassesTotal shotOboxTotal assistCross keyPassThrowin passCornerAccurate keyPassCorner goalHead dribbleTotal shotOnPost clearanceTotal penaltyTaken shotObp assistFreekick goalLeftFoot turnover offsideGiven yellowCard shotOnTarget passCornerInaccurate outfielderBlock passFreekickInaccurate keyPassThroughball | passCrossBlockedDefensive keyPassCorner assist goalHead shotBlocked assistOther dribbleWon shotCounter assistThroughball passCrossInaccurate saveTotal dribbleTotal foulCommitted goalOpenPlay shotOnPost dribbleLost shotsTotal redCard goalObox shotSixYardBox penaltyTaken goalCounter assistCorner shotLeftFoot keyPassOther keyPassShort challengeLost goalTotal passCrossAccurate shotObp keyPassFreekick goalNormal dispossessed outfielderBlockedPass shotOffTarget goalObp assistFreekick goalLeftFoot savePenaltyArea turnover keyPassLong tackleWonTotal goalSixYardBox offsideGiven saveObox duelAerialLost penaltyScored shotHead keyPassesTotal shotOboxTotal saveSixYardBox assistCross shotPenaltyArea foulGiven keyPassThrowin shotRightFoot shotOnTarget goalRightFoot assistThrowin passCornerInaccurate passFreekickInaccurate shotOpenPlay goalOwn tackleTotalAttempted passCornerAccurate keyPassCross goalPenaltyArea | passCrossBlockedDefensive assist assistOther assistThroughball saveTotal foulCommitted redCard goalTotal shotOffTarget savePenaltyArea duelAerialWon goalSixYardBox saveSixYardBox foulGiven goalRightFoot assistThrowin shotOpenPlay passFreekickAccurate shotBlocked shotCounter shotSetPiece shotsTotal goalObox assistCorner keyPassShort passCrossAccurate passLongBallInaccurate goalObp shortPassAccurate keyPassLong tackleWonTotal penaltyScored shotHead shotPenaltyArea shotRightFoot tackleTotalAttempted goalOwn goalSetPiece keyPassCross passTotal goalPenaltyArea dribbleWon passCrossInaccurate goalOpenPlay dribbleLost shortPassInaccurate shotSixYardBox goalCounter duelAerialTotal shotLeftFoot keyPassOther challengeLost keyPassFreekick goalNormal dispossessed outfielderBlockedPass passLongBallAccurate interceptionAll saveObox duelAerialLost keyPassesTotal shotOboxTotal assistCross keyPassThrowin passCornerAccurate keyPassCorner goalHead dribbleTotal shotOnPost clearanceTotal penaltyTaken shotObp assistFreekick goalLeftFoot turnover offsideGiven yellowCard shotOnTarget passCornerInaccurate outfielderBlock passFreekickInaccurate keyPassThroughball |

|  |  |  |  |
| --- | --- | --- | --- |
| DL | DC | DR | GK |
| passCrossBlockedDefensive keyPassCorner assist goalHead shotBlocked assistOther dribbleWon shotCounter assistThroughball passCrossInaccurate saveTotal dribbleTotal foulCommitted goalOpenPlay shotOnPost dribbleLost shotsTotal redCard goalObox clearanceTotal shotSixYardBox penaltyTaken goalCounter passTotal assistCorner shotLeftFoot keyPassOther keyPassShort challengeLost goalTotal passCrossAccurate passLongBallInaccurate keyPassFreekick goalNormal dispossessed outfielderBlockedPass shotObp shotOffTarget goalObp shortPassAccurate interceptionAll assistFreekick goalLeftFoot savePenaltyArea turnover keyPassLong tackleWonTotal goalSixYardBox offsideGiven saveObox duelAerialLost penaltyScored yellowCard shotHead keyPassesTotal shotOboxTotal saveSixYardBox assistCross shotPenaltyArea foulGiven keyPassThrowin shotRightFoot shotOnTarget goalRightFoot assistThrowin outfielderBlock passCornerInaccurate passFreekickInaccurate shotOpenPlay goalOwn tackleTotalAttempted goalSetPiece passCornerAccurate keyPassCross keyPassThroughball goalPenaltyArea | passCrossBlockedDefensive assist assistOther assistThroughball saveTotal foulCommitted redCard goalTotal shotOffTarget savePenaltyArea duelAerialWon goalSixYardBox saveSixYardBox foulGiven goalRightFoot assistThrowin shotOpenPlay passFreekickAccurate shotBlocked shotCounter shotSetPiece shotsTotal goalObox assistCorner keyPassShort passCrossAccurate passLongBallInaccurate goalObp shortPassAccurate keyPassLong tackleWonTotal penaltyScored shotHead shotPenaltyArea shotRightFoot tackleTotalAttempted goalOwn goalSetPiece keyPassCross passTotal goalPenaltyArea dribbleWon passCrossInaccurate goalOpenPlay dribbleLost shortPassInaccurate shotSixYardBox goalCounter duelAerialTotal shotLeftFoot keyPassOther challengeLost keyPassFreekick goalNormal dispossessed outfielderBlockedPass passLongBallAccurate interceptionAll saveObox duelAerialLost keyPassesTotal shotOboxTotal assistCross keyPassThrowin passCornerAccurate keyPassCorner goalHead dribbleTotal shotOnPost clearanceTotal penaltyTaken shotObp assistFreekick goalLeftFoot turnover offsideGiven yellowCard shotOnTarget passCornerInaccurate outfielderBlock passFreekickInaccurate keyPassThroughball | passCrossBlockedDefensive keyPassCorner assist goalHead shotBlocked assistOther dribbleWon shotCounter assistThroughball passCrossInaccurate saveTotal dribbleTotal foulCommitted goalOpenPlay shotOnPost dribbleLost shotsTotal redCard goalObox clearanceTotal shotSixYardBox penaltyTaken goalCounter passTotal assistCorner shotLeftFoot keyPassOther keyPassShort challengeLost goalTotal passCrossAccurate passLongBallInaccurate keyPassFreekick goalNormal dispossessed outfielderBlockedPass shotObp shotOffTarget goalObp shortPassAccurate interceptionAll assistFreekick goalLeftFoot savePenaltyArea turnover keyPassLong tackleWonTotal goalSixYardBox offsideGiven saveObox duelAerialLost penaltyScored yellowCard shotHead keyPassesTotal shotOboxTotal saveSixYardBox assistCross shotPenaltyArea foulGiven keyPassThrowin shotRightFoot shotOnTarget goalRightFoot assistThrowin outfielderBlock passCornerInaccurate passFreekickInaccurate shotOpenPlay goalOwn tackleTotalAttempted goalSetPiece passCornerAccurate keyPassCross keyPassThroughball goalPenaltyArea | passFreekickAccurate saveSixYardBox passLongBallAccurate savePenaltyArea passFreekickInaccurate saveTotal saveObox passLongBallInaccurate |

|  |  |  |  |
| --- | --- | --- | --- |
| DL | DC | DR | GK |
| passCrossBlockedDefensive keyPassCorner assist goalHead shotBlocked assistOther dribbleWon shotCounter assistThroughball passCrossInaccurate saveTotal dribbleTotal foulCommitted goalOpenPlay shotOnPost dribbleLost shotsTotal redCard goalObox clearanceTotal shotSixYardBox penaltyTaken goalCounter passTotal assistCorner shotLeftFoot keyPassOther keyPassShort challengeLost goalTotal passCrossAccurate passLongBallInaccurate keyPassFreekick goalNormal dispossessed outfielderBlockedPass shotObp shotOffTarget goalObp shortPassAccurate interceptionAll assistFreekick goalLeftFoot savePenaltyArea turnover keyPassLong tackleWonTotal goalSixYardBox offsideGiven saveObox duelAerialLost penaltyScored yellowCard shotHead keyPassesTotal shotOboxTotal saveSixYardBox assistCross shotPenaltyArea foulGiven keyPassThrowin shotRightFoot shotOnTarget goalRightFoot assistThrowin outfielderBlock passCornerInaccurate passFreekickInaccurate shotOpenPlay goalOwn tackleTotalAttempted goalSetPiece passCornerAccurate keyPassCross keyPassThroughball goalPenaltyArea | passCrossBlockedDefensive assist assistOther assistThroughball saveTotal foulCommitted redCard goalTotal shotOffTarget savePenaltyArea duelAerialWon goalSixYardBox saveSixYardBox foulGiven goalRightFoot assistThrowin shotOpenPlay passFreekickAccurate shotBlocked shotCounter shotSetPiece shotsTotal goalObox assistCorner keyPassShort passCrossAccurate passLongBallInaccurate goalObp shortPassAccurate keyPassLong tackleWonTotal penaltyScored shotHead shotPenaltyArea shotRightFoot tackleTotalAttempted goalOwn goalSetPiece keyPassCross passTotal goalPenaltyArea dribbleWon passCrossInaccurate goalOpenPlay dribbleLost shortPassInaccurate shotSixYardBox goalCounter duelAerialTotal shotLeftFoot keyPassOther challengeLost keyPassFreekick goalNormal dispossessed outfielderBlockedPass passLongBallAccurate interceptionAll saveObox duelAerialLost keyPassesTotal shotOboxTotal assistCross keyPassThrowin passCornerAccurate keyPassCorner goalHead dribbleTotal shotOnPost clearanceTotal penaltyTaken shotObp assistFreekick goalLeftFoot turnover offsideGiven yellowCard shotOnTarget passCornerInaccurate outfielderBlock passFreekickInaccurate keyPassThroughball | passCrossBlockedDefensive keyPassCorner assist goalHead shotBlocked assistOther dribbleWon shotCounter assistThroughball passCrossInaccurate saveTotal dribbleTotal foulCommitted goalOpenPlay shotOnPost dribbleLost shotsTotal redCard goalObox clearanceTotal shotSixYardBox penaltyTaken goalCounter passTotal assistCorner shotLeftFoot keyPassOther keyPassShort challengeLost goalTotal passCrossAccurate passLongBallInaccurate keyPassFreekick goalNormal dispossessed outfielderBlockedPass shotObp shotOffTarget goalObp shortPassAccurate interceptionAll assistFreekick goalLeftFoot savePenaltyArea turnover keyPassLong tackleWonTotal goalSixYardBox offsideGiven saveObox duelAerialLost penaltyScored yellowCard shotHead keyPassesTotal shotOboxTotal saveSixYardBox assistCross shotPenaltyArea foulGiven keyPassThrowin shotRightFoot shotOnTarget goalRightFoot assistThrowin outfielderBlock passCornerInaccurate passFreekickInaccurate shotOpenPlay goalOwn tackleTotalAttempted goalSetPiece passCornerAccurate keyPassCross keyPassThroughball goalPenaltyArea | passFreekickAccurate saveSixYardBox passLongBallAccurate savePenaltyArea passFreekickInaccurate saveTotal saveObox passLongBallInaccurate |

## 6.3. Quantile analysis

During clustering analysis, figure out which features are strongly or lightly related to each player position by dividing quantiles.

This quantile means how many % index is greater than 50% quantile, 75% quantile, or 90% quantile. When the values are sorted in ascending order.

In this sub-chapter, an analysis of the positions of the players which are in highest tenth (90%-100%), twenty-fifth (75%-100%) and fiftieth (50%-100%) for 83 each one of the variables that were considered as possible key performance indicators will be produced.

Similar to subchapter B, we will choose 1 attribute for analysis, the other attributes are completely similar. Here, we choose assistOtherChart, bar chart

Description automatically generated

*Fig 12. Quantile stackbar chart for attributes assistOther*

Table

Description automatically generated

*Fig 13. Quantile table of 'assistOther' attribute*

* Observing the graph, we see, the value domain that accounts for 50 - 100% of the quantile of GK in this attribute is very low, only about 3%, which proves that very few large values are in the GK, so we will not select. Quite the contrary, in other positions, the quantile value ranges from 50 to 100% of the values.
* Looking at the quantile table, we also see the same thing, when the total percentage of GK in the attribute from 50 - 100% is only 3%, while in other positions there are very large percentages, especially with many values. have quantile from 90-100% large. Therefore, from here we can see that all positions are heavily influenced by the assistOther stat, while with GK, it is not a necessary stat.
* Continuing the analysis with all other attributes, matching the attributes of each positon found in the clustering analysis, we can add and remove positions for each position as follows:

Continuing the analysis with all other attributes, matching the attributes of each positon found in the clustering analysis, we can add and remove positions for each position as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| FW | AML | AMC | AMR |
| passFreekickAccurate shortPassAccurate passLongBallAccurate interceptionAll clearanceTotal passTotal tackleWonTotal outfielderBlock passFreekickInaccurate assistThroughball tackleTotalAttempted challengeLost passLongBallInaccurate yellowCard keyPassThroughball | passFreekickAccurate shortPassAccurate shortPassInaccurate interceptionAll passLongBallAccurate clearanceTotal duelAerialTotal duelAerialWon passTotal outfielderBlock goalSetPiece passLongBallInaccurate yellowCard keyPassThroughball shotSetPiece | passFreekickAccurate shortPassAccurate passLongBallAccurate interceptionAll passTotal passLongBallInaccurate keyPassThroughball | passFreekickAccurate shortPassAccurate passLongBallAccurate interceptionAll clearanceTotal passTotal outfielderBlock passLongBallInaccurate yellowCard keyPassThroughball |

|  |  |  |  |
| --- | --- | --- | --- |
| ML | MC | MR | DMC |
| passFreekickAccurate shortPassAccurate shortPassInaccurate interceptionAll passLongBallAccurate clearanceTotal duelAerialTotal duelAerialWon passTotal outfielderBlock passLongBallInaccurate yellowCard keyPassThroughball shotSetPiece | **Remain stable** | passFreekickAccurate shortPassAccurate shortPassInaccurate interceptionAll passLongBallAccurate clearanceTotal duelAerialTotal duelAerialWon passTotal outfielderBlock passLongBallInaccurate yellowCard keyPassThroughball shotSetPiece | **Remain stable** |

|  |  |  |  |
| --- | --- | --- | --- |
| DL | DC | DR | GK |
| passFreekickAccurate shortPassInaccurate passLongBallAccurate duelAerialTotal duelAerialWon goalSetPiece shotSetPiece | **Remain stable** | passFreekickAccurate shortPassInaccurate passLongBallAccurate duelAerialTotal duelAerialWon goalSetPiece shotSetPiece | shortPassAccurate clearanceTotal keyPassLong yellowCard keyPassFreekick passTotal |

## 6.4. Mutual information analysis

6.4.1. Motivation

When looking through jointplot visualizations of some data, we see that the plots of some attributes are quite different

For example:

* With position AMC, attribute assistThrowin, most of its scatterplot lies on a straight line, with attribute values lying at zero most of this position, while rating still changesChart

  Description automatically generated

*Fig 14. Joint-plot between assistThrown and rating of AMC*

* We understand that here, the correlation between this attribute and the rating is quite low,

Consider the assist attribute vs rating of GK:

Chart, scatter chart

Description automatically generated

*Fig 15. Joint-plot between assist and rating of GK*

* + Looking at the distribution of ratings, we see that it is concentrated at 6 and 7, also there, there are attributes of assist > 0 from 0.02 to 0.25 appearing and changing, which suggests that there is still some amount certain information between assist and rating. If the amount of information of assist and rating is mostly at 8-9, then we understand that the amount of information is insignificant, because 8 and 9 are quite few players located there.
  + This reminds us of using mutual information to evaluate how good feature is.

6.4.2. Definition

Mutual information is a quantity that measures a relationship between two random variables that are sampled simultaneously. It measures the amount of communication between two different variables. In simple terms, if a variable A asks for a value, how much information will variable B answer for variable A?

For example, if X represents the roll of one fair die, and Z represents the roll of another fair die, then X and Z share no mutual information. The roll of one die does not contain any information about the outcome of the other die. An theorem from information theory says that the mutual information between two variables is 0 if and only if the two variables are statistically independent.

The general formular:

- Diagram, venn diagram

Description automatically generated

*Fig 16. Ven-diagram of mutual information*

* With is the entropy of event X and event with condition .
* Following that, is mutual information between entropy of H(X) and H(Y).

6.4.3. Analysis

Applying the formula for calculating mutual information above, we choose threshold 0.1 or higher to select features with mutual information score

Note that although mutual information score is used for discrete variables, here the attributes are initially calculated by the number of actions and then divided by the number of appearances in the season. Therefore, these numbers are also rational numbers, and they are also discrete.

Applying the mutual information table for each score pair, position and rating , we can extract new features from

* GK duelAerialLost 1.7878029533279636
* GK duelAerialTotal 4.09052915396218
* GK duelAerialWon 4.041464394974207
* GK assistCross 0.0
* GK assistCorner 0.0
* GK assistThroughball 0.010158884332080688
* GK assistFreekick 0.035549702154880215
* GK assistThrowin 0.0
* GK assistOther 0.24301870377378976
* GK assist 0.27846513806858225
* GK passCrossBlockedDefensive 0.01947917401355944
* GK outfielderBlockedPass 0.8125851682503229
* GK outfielderBlock 0.025394472075560108
* GK yellowCard 2.3178740887036087
* GK redCard 0.27667243821464904
* GK clearanceTotal 4.5975283906141335
* GK dribbleLost 0.15884370328413067
* GK dribbleTotal 1.3923022817255333
* GK dribbleWon 1.336770308028976
* GK foulGiven 3.5333735529160464
* GK foulCommitted 1.496602937250116
* GK goalRightFoot 0.019479174013559444
* GK goalLeftFoot 0.015237778968060289
* GK goalHead 0.005079624602527025
* GK goalObp 0.0
* GK goalOpenPlay 0.005079624602527025
* GK goalCounter 0.0
* GK goalSetPiece 0.018326068636213767
* GK penaltyScored 0.01947917401355944
* GK goalNormal 0.023404232422175857
* GK goalTotal 0.03779939407520243
* GK goalOwn 0.3443945224210745
* GK goalSixYardBox 0.0
* GK goalPenaltyArea 0.032722328225457384
* GK goalObox 0.010158884332080688
* GK interceptionAll 1.809143226390317
* GK keyPassLong 1.5531448664155079
* GK keyPassShort 0.17164150010731316
* GK keyPassesTotal 1.6235138699463232
* GK keyPassCross 0.005079624602527025
* GK keyPassCorner 0.0
* GK keyPassThroughball 0.042876093502150756
* GK keyPassFreekick 0.42920400034241324
* GK keyPassThrowin 0.0
* GK keyPassOther 1.4423457259675456
* GK offsideGiven 0.020316308289598024
* GK passTotal 6.187385194891042
* GK passLongBallAccurate 5.703404234459635
* GK passLongBallInaccurate 6.008678629442208
* GK shortPassAccurate 6.118365783831129
* GK shortPassInaccurate 3.691470410351635
* GK passCornerAccurate 0.0
* GK passCornerInaccurate 0.005079624602527025
* GK passFreekickAccurate 4.94510450218104
* GK passFreekickInaccurate 5.089886466591358
* GK passCrossInaccurate 0.01947917401355944
* GK passCrossAccurate 0.015237778968060289
* GK turnover 1.895382898774057
* GK dispossessed 0.36493329993364015
* GK saveSixYardBox 3.565211067962646
* GK savePenaltyArea 4.896046141931029
* GK saveObox 4.729817131323564
* GK saveTotal 5.16627071228947
* GK shotOnTarget 0.09315157064003401
* GK shotOffTarget 0.1525254871724007
* GK shotOnPost 0.0
* GK shotBlocked 0.07024432615206701
* GK shotRightFoot 0.06516946218170838
* GK shotLeftFoot 0.0728074195325728
* GK shotHead 0.14861038280378086
* GK shotObp 0.0
* GK shotOpenPlay 0.02539447207556011
* GK shotCounter 0.0
* GK shotSetPiece 0.22707423454944003
* GK penaltyTaken 0.01947917401355944
* GK shotSixYardBox 0.033875433602803054
* GK shotPenaltyArea 0.17618668519509645
* GK shotOboxTotal 0.08039295058346574
* GK shotsTotal 0.24914200216583604
* GK tackleWonTotal 1.6530068002314895
* GK tackleTotalAttempted 2.7175089215929074
* GK challengeLost 2.155123084031281

Choose threshold 0.25, we choose features:

* + dribbleWon
  + passFreekickInaccurate
  + foulGiven
  + outfielderBlockedPass
  + goalOwn
  + clearanceTotal
  + saveTotal
  + yellowCard
  + keyPassOther
  + challengeLost
  + keyPassesTotal
  + interceptionAll
  + turnover
  + shortPassAccurate
  + duelAerialWon
  + duelAerialTotal
  + tackleWonTotal
  + dribbleTotal
  + duelAerialLost
  + saveSixYardBox
  + passFreekickAccurate
  + shortPassInaccurate
  + foulCommitted
  + passLongBallInaccurate
  + passTotal
  + keyPassLong
  + saveObox
  + tackleTotalAttempted
  + assist
  + dispossessed
  + savePenaltyArea
  + keyPassFreekick
  + passLongBallAccurate
  + redCard

- That is, compared to the feature combined from the above 2 analysis, we can add:

* outfielderBlockedPass
* shortPassInaccurate
* tackleWonTotal
* duelAerialTotal
* keyPassOther
* turnover
* keyPassesTotal
* dispossessed
* dribbleWon
* interceptionAll
* assist
* redCard
* foulGiven
* dribbleTotal
* challengeLost
* foulCommitted
* tackleTotalAttempted
* goalOwn
* duelAerialWon
* duelAerialLost

## 6.5. Summary

To sum it all up, all the attributes used are:

|  |  |  |
| --- | --- | --- |
| GK | Other positions | |
| clearanceTotal interceptionAll passFreekickAccurate saveSixYardBox duelAerialLost outfielderBlockedPass passTotal shortPassAccurate shortPassInaccurate dribbleWon saveTotal tackleTotalAttempted goalOwn duelAerialWon challengeLost yellowCard passLongBallInaccurate keyPassesTotal assist foulCommitted savePenaltyArea duelAerialTotal dispossessed tackleWonTotal redCard foulGiven turnover passFreekickInaccurate keyPassOther keyPassLong keyPassFreekick dribbleTotal passLongBallAccurate saveObox | clearanceTotal goalSetPiece keyPassShort interceptionAll dribbleLost passCrossInaccurate duelAerialLost passTotal shotOffTarget shotRightFoot dribbleWon goalNormal assistOther goalOwn duelAerialWon offsideGiven yellowCard penaltyTaken shotSetPiece penaltyScored keyPassesTotal shotOnPost tackleWonTotal redCard shotCounter keyPassCorner keyPassLong keyPassFreekick shotLeftFoot assistThrowin goalOpenPlay goalTotal shotObp assistThroughball shortPassAccurate shortPassInaccurate shotOboxTotal passCornerInaccurate goalCounter assistCross assistFreekick passCornerAccurate | passLongBallInaccurate foulCommitted shotOpenPlay goalSixYardBox goalPenaltyArea foulGiven keyPassThrowin shotHead passLongBallAccurate shotOnTarget passFreekickAccurate outfielderBlockedPass shotsTotal keyPassCross keyPassThroughball challengeLost assistCorner outfielderBlock passCrossBlockedDefensive assist dispossessed dribbleTotal saveSixYardBox passCrossAccurate goalLeftFoot goalHead shotPenaltyArea saveTotal shotBlocked tackleTotalAttempted goalRightFoot shotSixYardBox savePenaltyArea duelAerialTotal goalObox goalObp turnover passFreekickInaccurate keyPassOther saveObox |

**7.2 Random Forest**

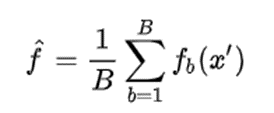
Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of overfitting to their training set.

The training algorithm for random forests applies the general technique of bootstrap aggregating, or bagging, to tree learners. Given a training set X = x1, ..., xn with responses Y = y1, ..., yn, bagging repeatedly (B times) selects a random sample with replacement of the training set and fits trees to these samples:

For b = 1, ..., B:

1. Sample, with replacement, n training examples from X, Y; call these Xb, Yb.
2. Train a regression tree fb on Xb, Yb.

After training, predictions for unseen samples x' can be made by averaging the predictions from all the individual regression trees on x':



This bootstrapping procedure leads to better model performance because it decreases the variance of the model, without increasing the bias. This means that while the predictions of a single tree are highly sensitive to noise in its training set, the average of many trees is not, as long as the trees are not correlated. Simply training many trees on a single training set would give strongly correlated trees (or even the same tree many times, if the training algorithm is deterministic); bootstrap sampling is a way of de-correlating the trees by showing them different training sets.

Since each decision tree in the Random Forest algorithm does not use all the training data, nor does it use all the attributes of the data to build the tree, each tree may make a bad prediction, then each decision tree model can be underfitting, in other words, the model has high bias. However, the final result of the Random Forest algorithm is aggregated from many decision trees, so the information from the trees will complement each other, leading to a model with low bias and low variance, or a model with good predictive results.

**6.2 Artificial Neural Network (ANN):**

The problem we are facing is a complex non-linear problem, which is why we are in need of a technique that focus on the complex relationships of the features and target variable. One of which is Artificial Neural Network.

Artificial neural networks (ANNs), also known as neural networks, is a group of interconnected node, that resemble the neurons in a biological brain. ANNs consist of an input layer, one or more hidden layers and an output layer. Each node connects with another and has a specific weight to it. When processing samples that each have a known "input" and "output," neural networks learn (or are trained) by creating probability-weighted associations between the two that are then stored within the net's data structure.

Learning algorithm in neural networks specifies the way to adapt the weight weights of the connections in the ANN from training data , given a fixed network structure (Parameter learning) or to learn the network structure, including the number of neurons and the types of connections between them, and the weights (Structure learning). In parameter learning, the goal is to minimize an empirical loss function:

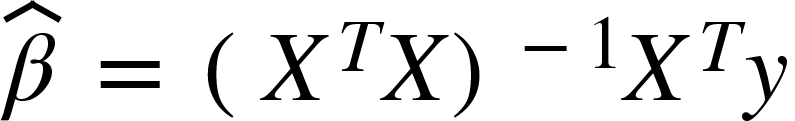
In this problem we are going to use two hidden layers with ten neurons each and one output layer with one neuron. The activation function for the first hidden layer is ReLU and the second is hyperbolic tangent

***6.3. Ridge Regression***

Ridge regression is a method of estimating the coefficients of multiple-regression models in scenarios where the independent variables are highly correlated. It has been used in many fields including econometrics, chemistry, and engineering.

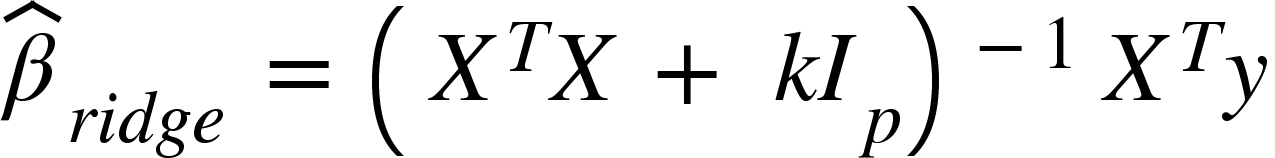
Ridge regression was developed as a possible solution to the imprecision of least square estimators when linear regression models have some multicollinear (highly correlated) independent variables—by creating a ridge regression estimator (RR). This provides a more precise ridge parameters estimate, as its variance and mean square estimator are often smaller than the least square estimators previously derived.

In standard linear regression, an n × 1 column vector y is to be projected onto the column space of the n × p design matrix X (typically p ≪ n) whose columns are highly correlated. The ordinary least squares estimator of the coefficients β ∈ ℝp×1 by which the columns are multiplied to get the orthogonal projection Xβ is



where XT is the transpose of X.

In situations where the dependent variables of the regression problem (columns of X) are highly correlated, the inverse above may be difficult to compute. So, ridge regression might be used, in which the regression coefficients are computed using the alternate formula:



where Ip is the p × p identity matrix and k > 0 is small. The name 'ridge' refers to the shape along the diagonal of *I*.

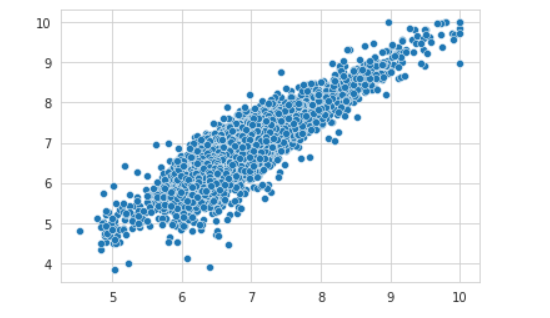
# **7. Experiments**

Now, we will come up with a comparison of the potential of these algorithms. According to the evaluation, we use mean absolute error to calculate the mean difference between the predicted value and the real value. The objective to evaluate is the training set and testing set. The following table is the result we collected.

For Random forest, we use set of hyperparameters with n\_estimator=100 and max\_depth of a tree is 8. For ANN, we use 2 hidden layers with 10 nodes for each layers.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Train | | | Test | | |
| MAE | MSE | RMSE | MAE | MSE | RMSE |
| Random forest | 0.1957 | 0.070 | 0.2644 | 0.2091 | 0.0821 | 0.2866 |
| ANN | 0.2585 | 0.1330 | 0.3647 | 0.2672 | 0.1433 | 0.3785 |
| Ridge Regression  with chosen features each position | 0.1387 | 0.038 | 0.1949 | 0.14 | 0.04 | 0.2000 |

# **8. Detect outlier by SelectFromModel method**

****

**Fig. 17. Find some outliers**

* As we can see in Fig 10 that, it isn’t significantly seen that there are any clear outliers since the plot is quite at the state of linearity. The false prediction is not far from the diagonal line. Therefore, no need to delete any row from data.

# **9. Conclusion**

In this report, we introduced and solved our prediction problems using different methods in data analysis and machine learning models. Overall, the best model that we conclude is Ridge regression, which is the model with the most stable performance through and MAE evaluation. It has the smallest MAE.

In order to be able to give each type of model and practical results as above, we had to carefully select the most optimal algorithm types as well as the algorithm with the best value of error. Therefore, if we have more time to research, we will expand the scope of the algorithm and study some new models along with processing techniques. different from the given data set.

**Reference**

[1] Scikit-learn;. Available from: <https://scikit-learn.org/stable/>.

[2] Pandas;. Available from: https://pandas.pydata.org/.

[3] Seaborn; 2021. Available from: https://seaborn.pydata.org/. 4

[4] Using Data Science to Analyse Player Performance in Football: https://towardsdatascience.com/using-data-science-to-analyse-player-performance-in-football-4dd04e579cfa