FIFA 23
ML Project

Supervised By: Dr. Doaa Mahmoud

Supervised By: Eng. Mohamed Abdullah

Presented By:

Mira Ehab

LinkedIn

Github

Kaggle

CONTENTS

01
Problem Definition and Introduction to the Data

Problem Definition and Introduction to the Data

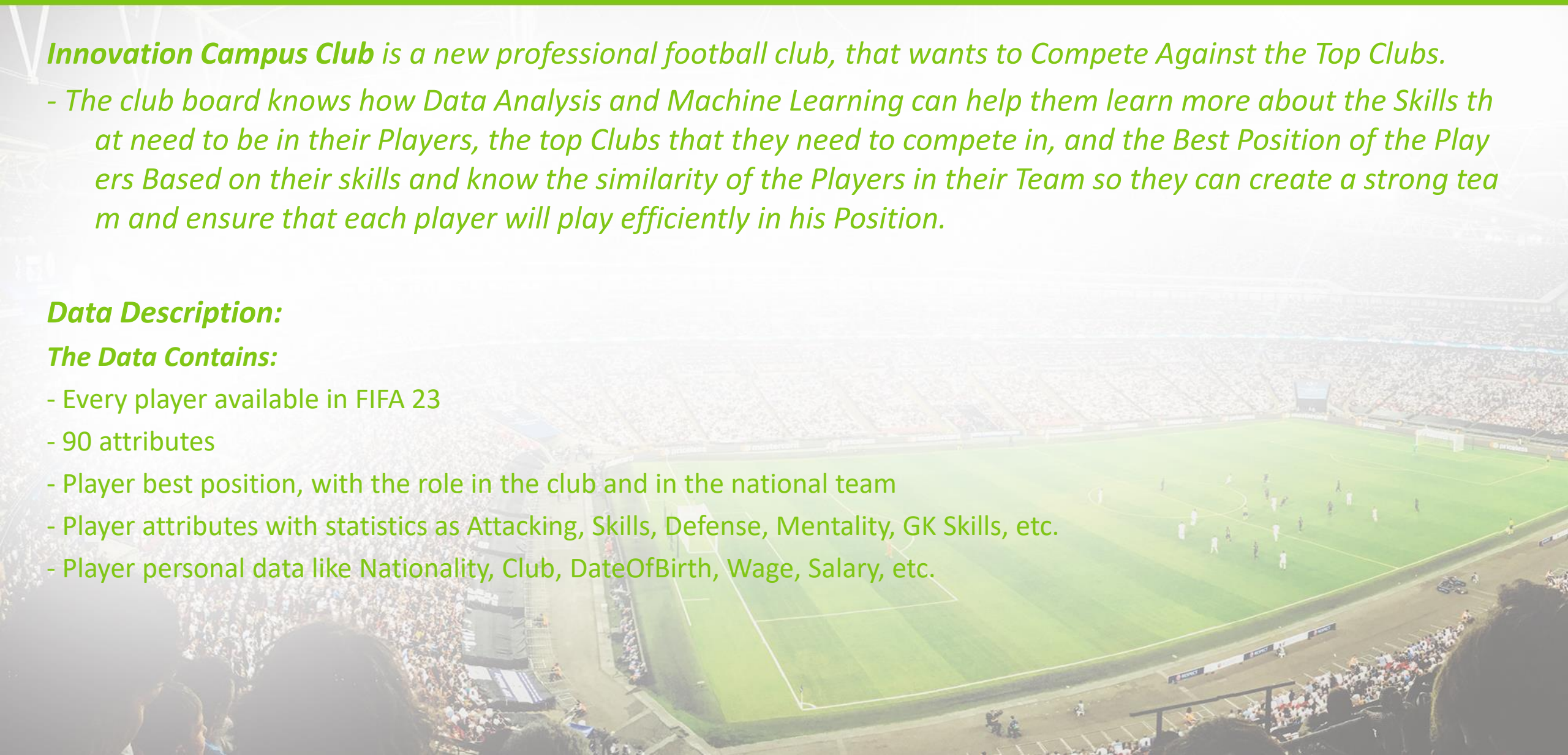
Innovation Campus Club *is a new professional football club, that wants to Compete Against the Top Clubs.*

- The club board knows how Data Analysis and Machine Learning can help them learn more about the Skills th at need to be in their Players, the top Clubs that they need to compete in, and the Best Position of the Play ers Based on their skills and know the similarity of the Players in their Team so they can create a strong tea m and ensure that each player will play efficiently in his Position.

Data Description:

The Data Contains:

- Every player available in FIFA 23
- 90 attributes
- Player best position, with the role in the club and in the national team
- Player attributes with statistics as Attacking, Skills, Defense, Mentality, GK Skills, etc.
- Player personal data like Nationality, Club, DateOfBirth, Wage, Salary, etc.



02
Objectives

Objectives

*Help the **Innovation Campus Club** by:*

1- Helping the club board know the best players in the different Clubs.

2- Helping them Understand their competitor's Clubs.

3- Knowing the skills that need to be in their players.

4- Helping them put the players in their suitable Position.

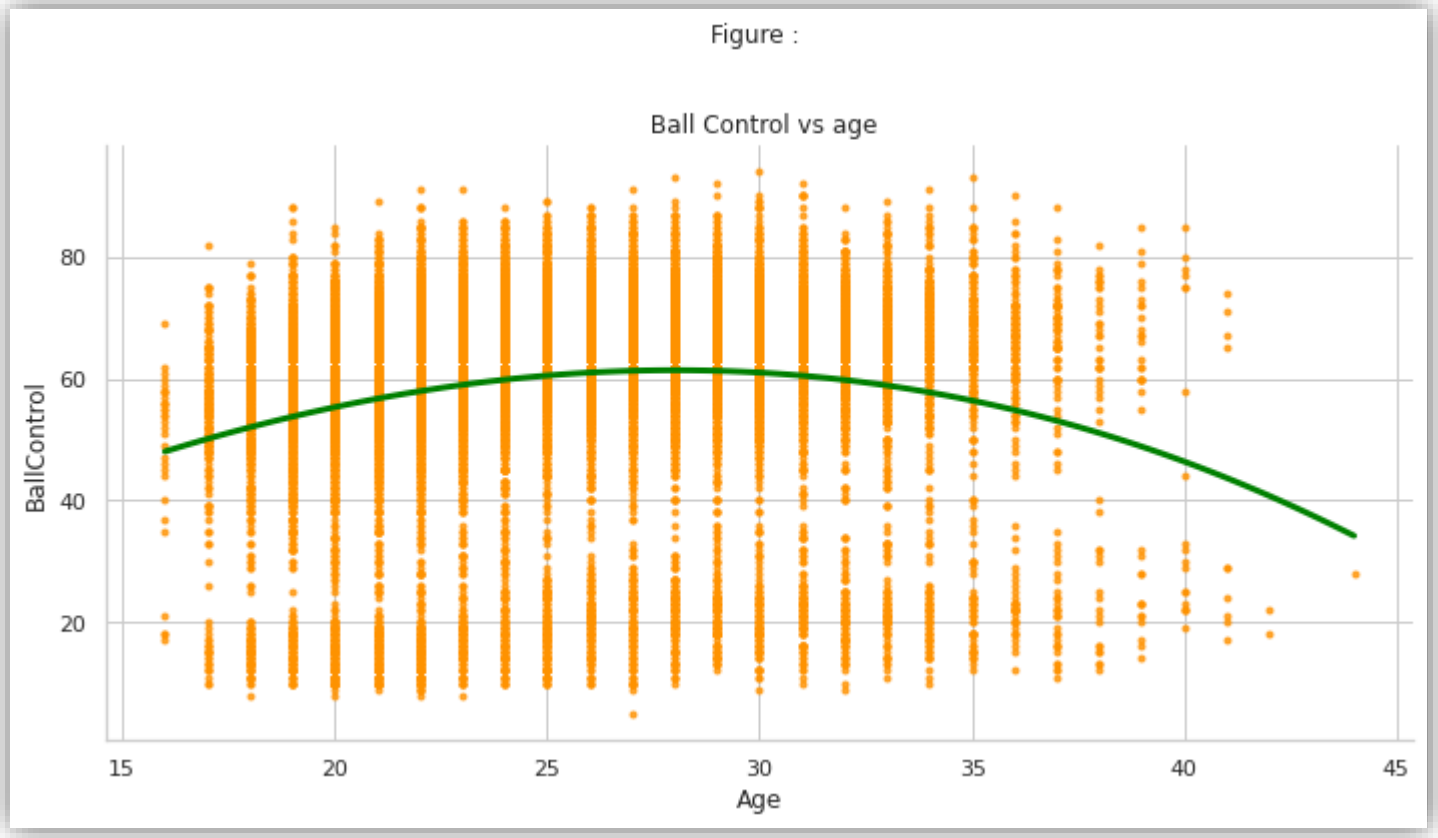
5- Grouping the Club Players in Groups.

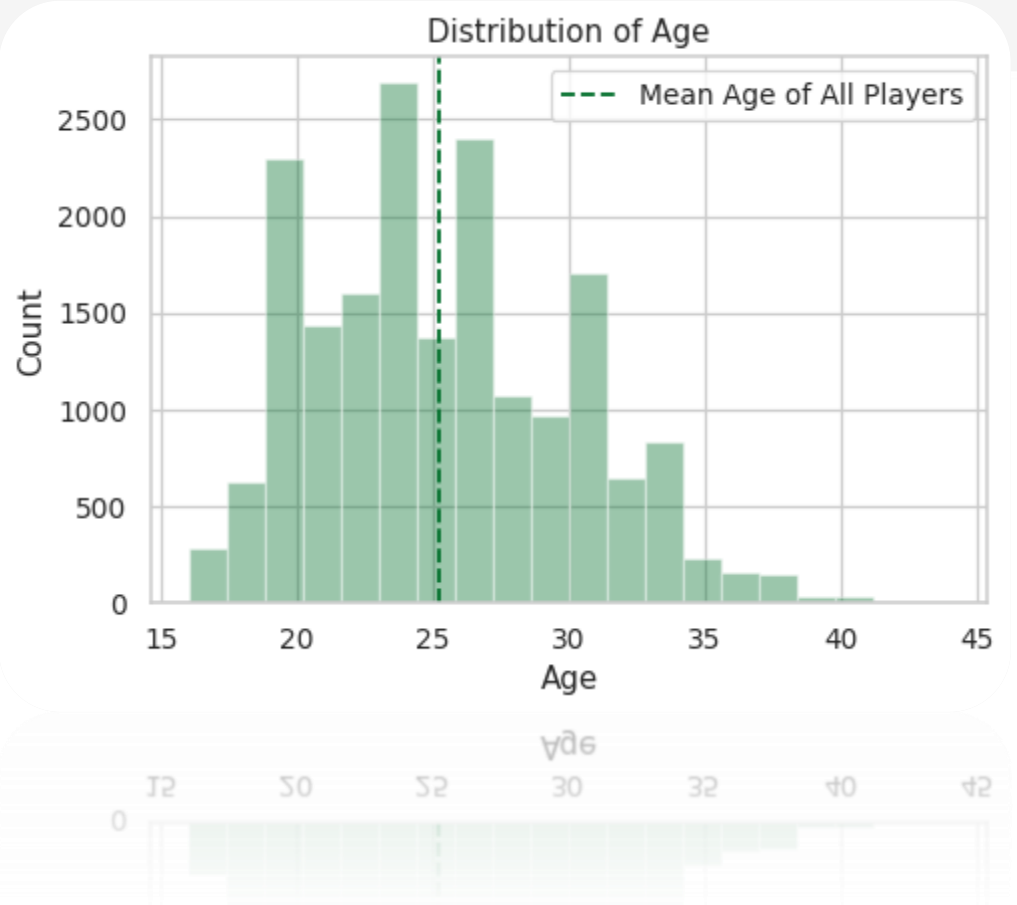
03
Data Analysis

Questions to be Answered :

1. Does the Age of the Player Affect on his Ball Control Performance?
2. How Height affects different factors like stamina, dribbling, pace, passing and HeadingAccuracy
3. Show if there is a relation between Wage and Overall of the Players
4. Show the top Fastest Players
5. Determine if their is a relation between the Position of the Player and his Wage and Value
6. See the Nationality of the Players that got the highest Wages
7. Show the effect of the Age on the Potential of the Players
8. View the Top 50 Players and their Clubs

Does the Age of the Player Affect on his Ball Control Performance?

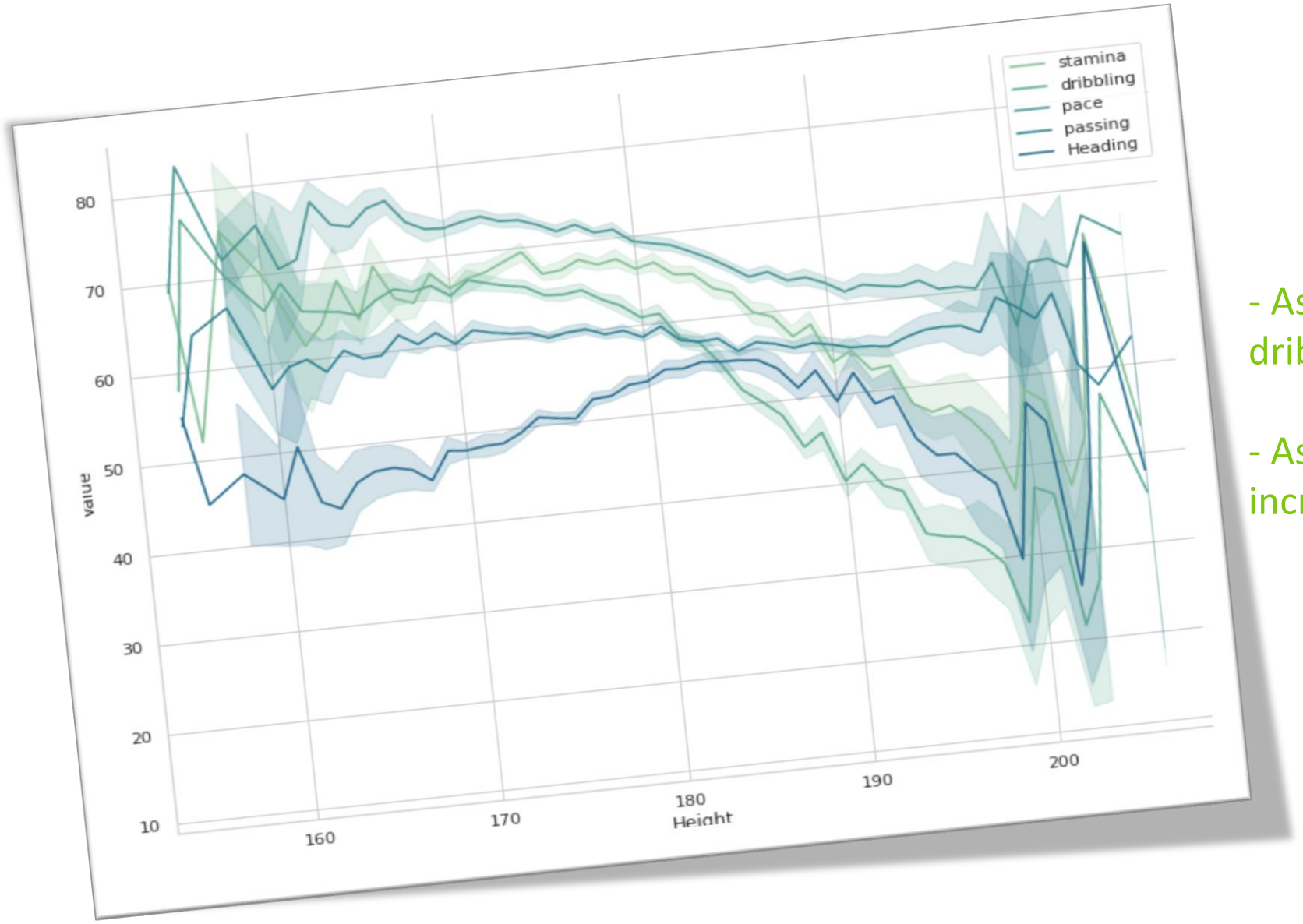




- So We can deduce that the age has an effect on the Player's Ball Control

- While the Age is increasing, the Ball Control decreases.

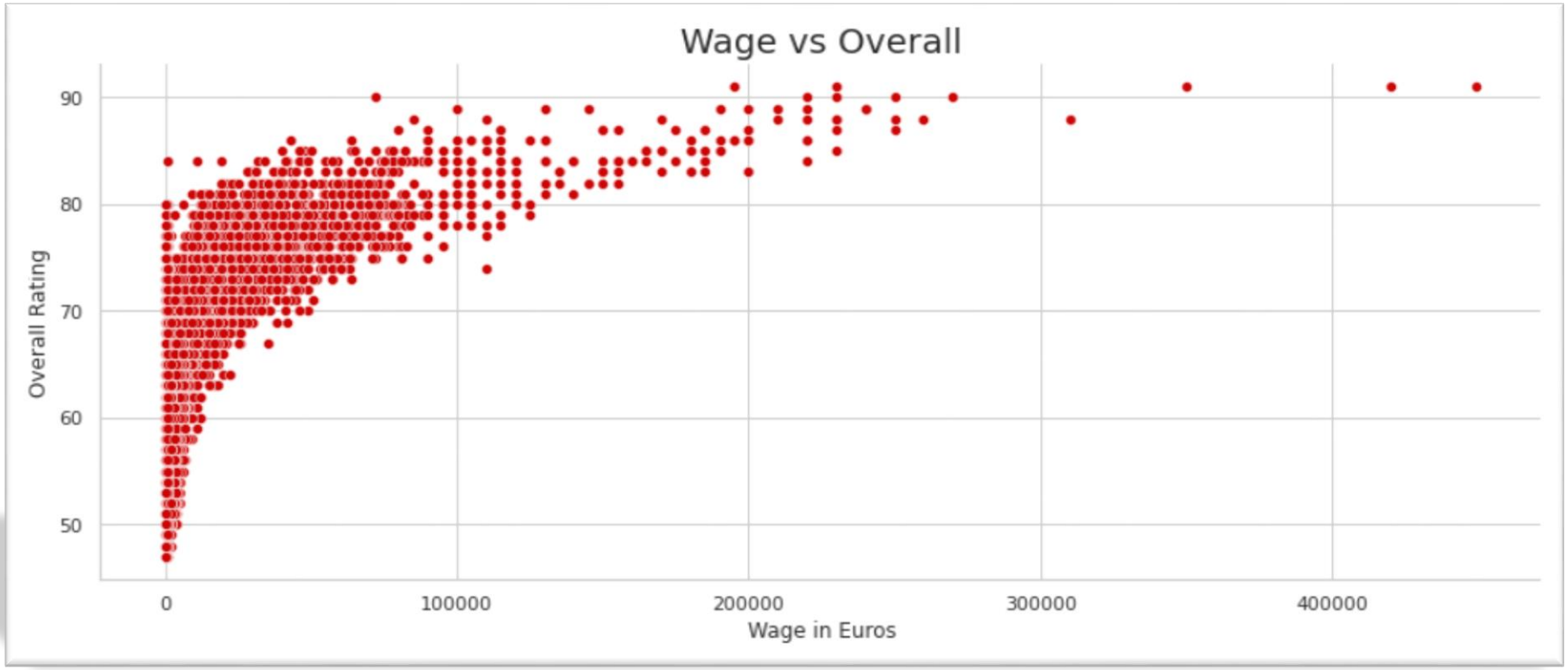
How Height affects different factors like stamina, dribbling, pace, passing and HeadingAccuracy ?



- As height increases, features like stamina, dribbling, pace, passing decreases.

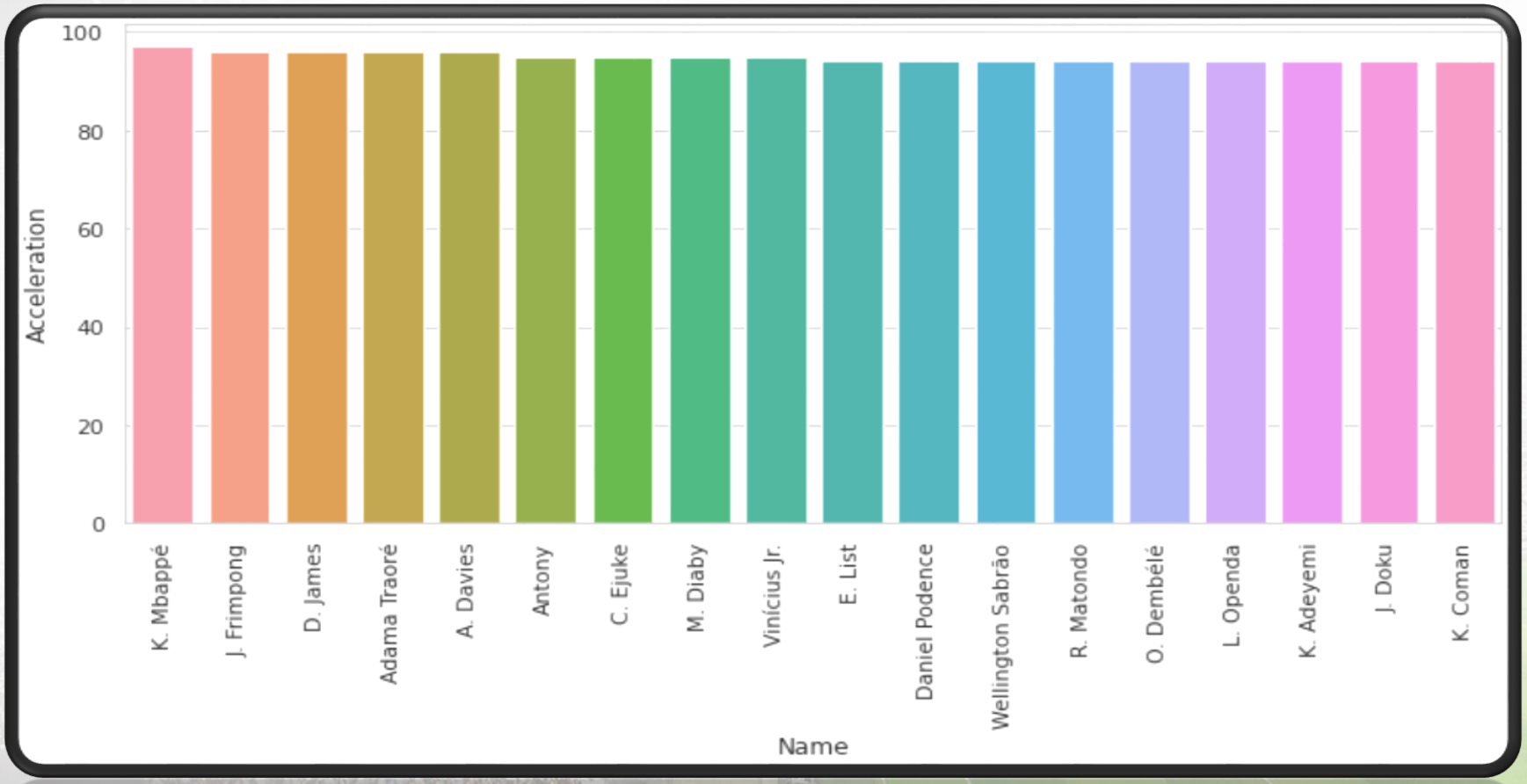
- As height increases, features like Heading increase.

Show if there is a relation between Wage and Overall of the Players

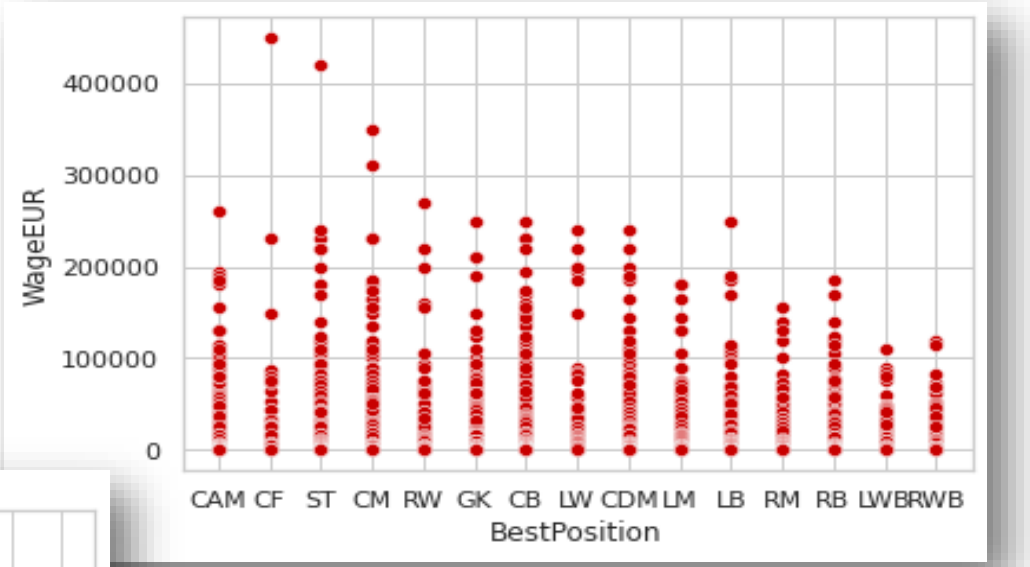


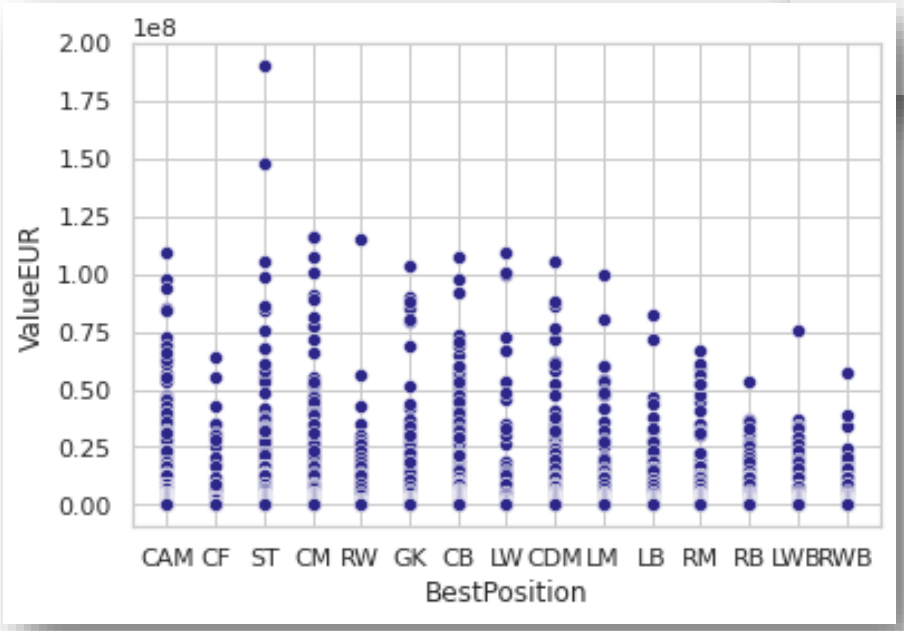
As the Overall Rating Increase, the Wage of the Player Increases too.

Show the top Fastest Players



Determine if their is a relation between the Position of the Player and his Wage and Value

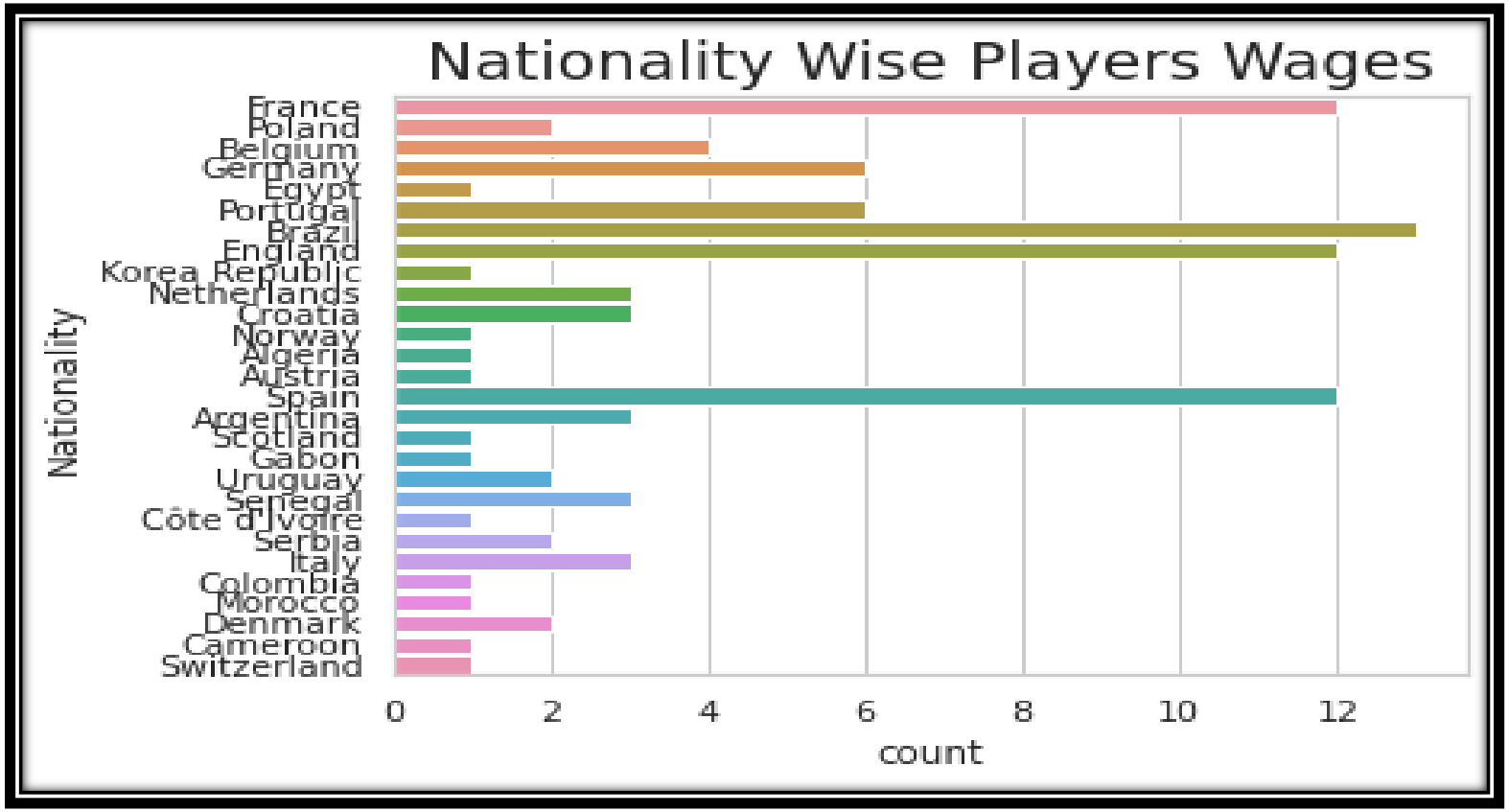




- So we can see that the Players in Positions LM, RM, RB, LWB, RWB got the lowest Wages.

- And the Players With Positions LB, RB, LWB, RWB, CF , RW have the lowest Values.

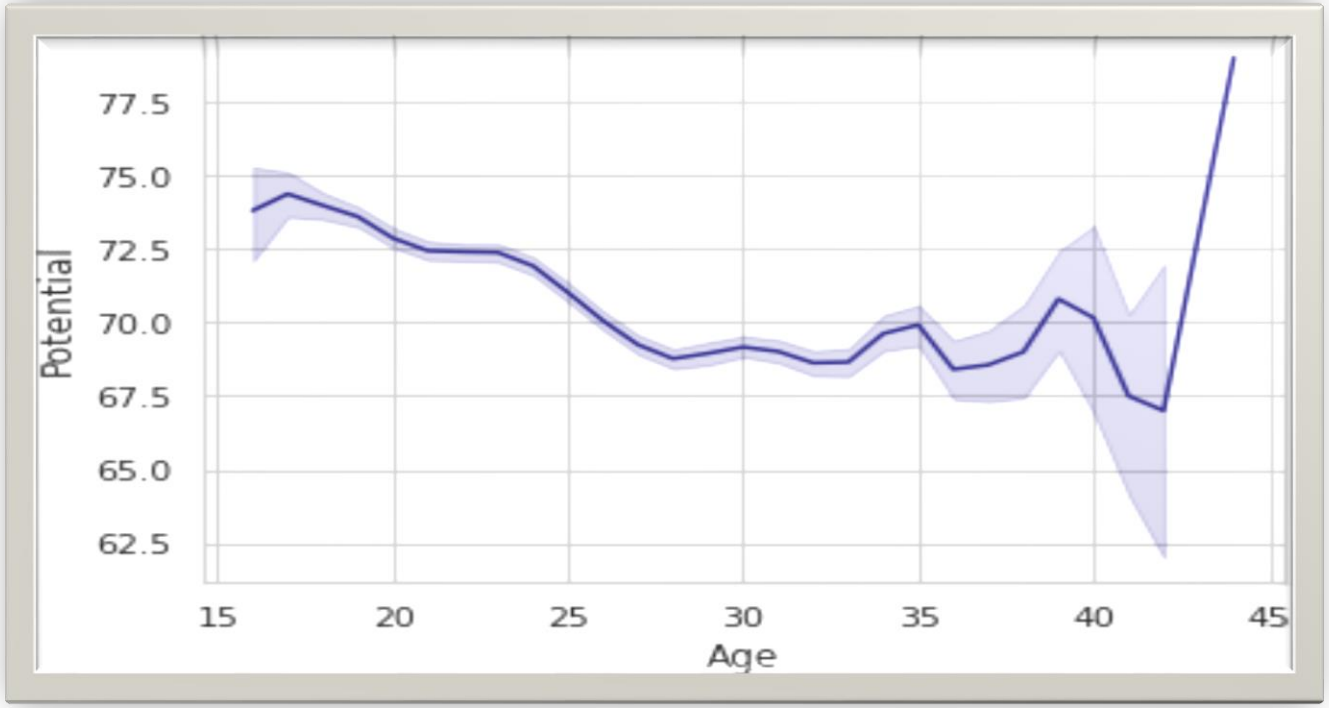
See the Nationality of the Players that got the highest Wages



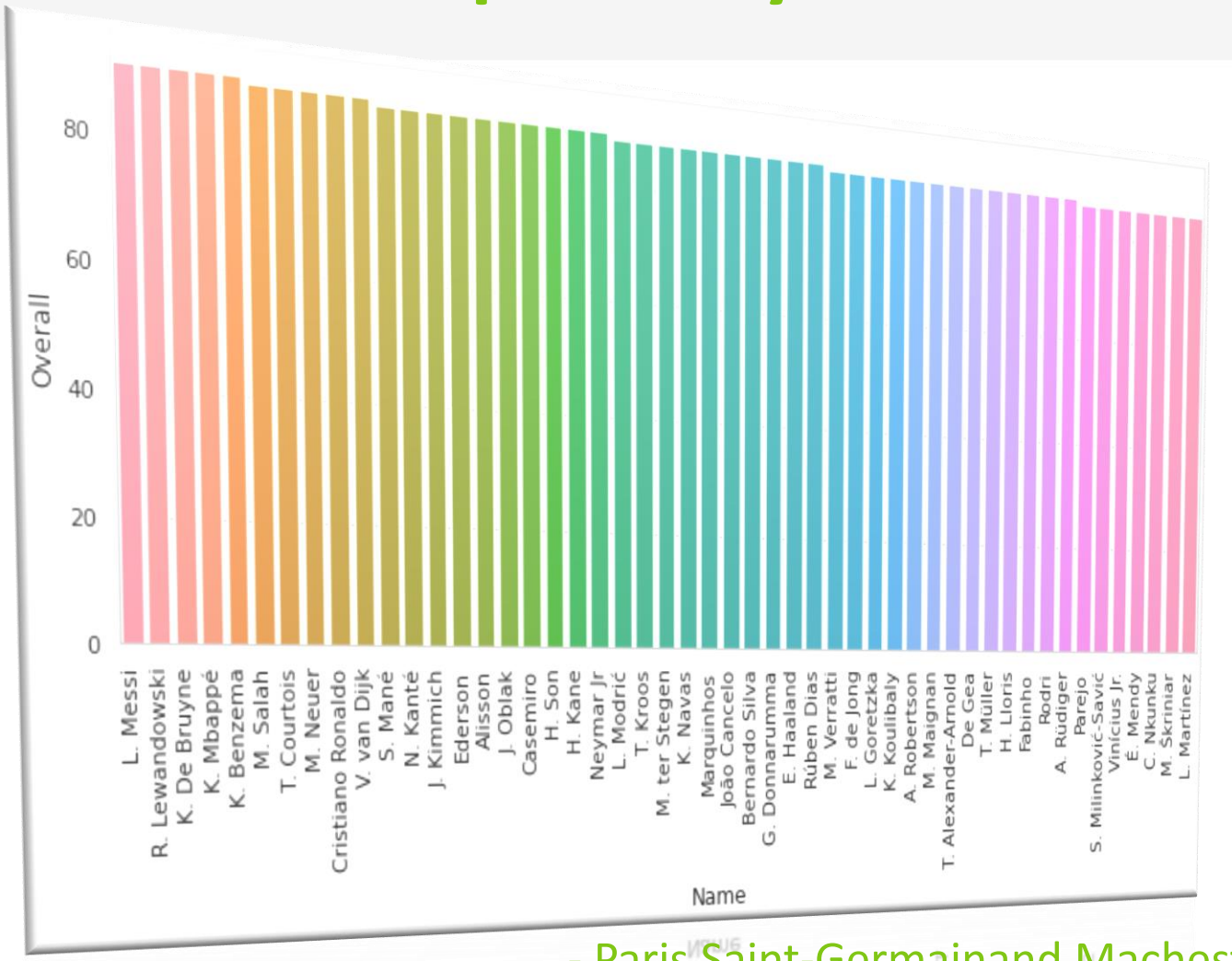
So we can deduce that the Players that got the Maximum Wage are from Brazil , France, England and Spain.

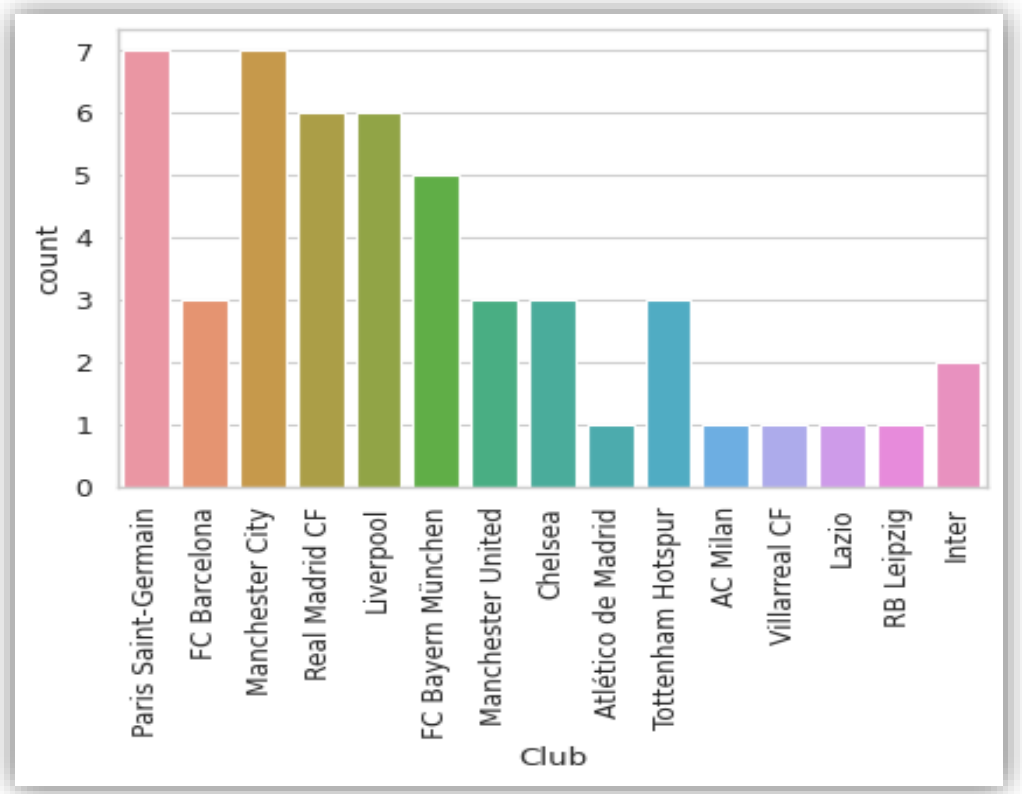
Show the effect of the Age on the Potential of the Players

While the Age Increases the Potential of the Player Decreases.



View the Top 50 Players and their Clubs





- Paris Saint-Germainand Machester City have the maximum top Playe rs numbers
- Liverpool and Real Madrid have the second Maximum top Players numbers.

04
Data Preprocessing

Steps :

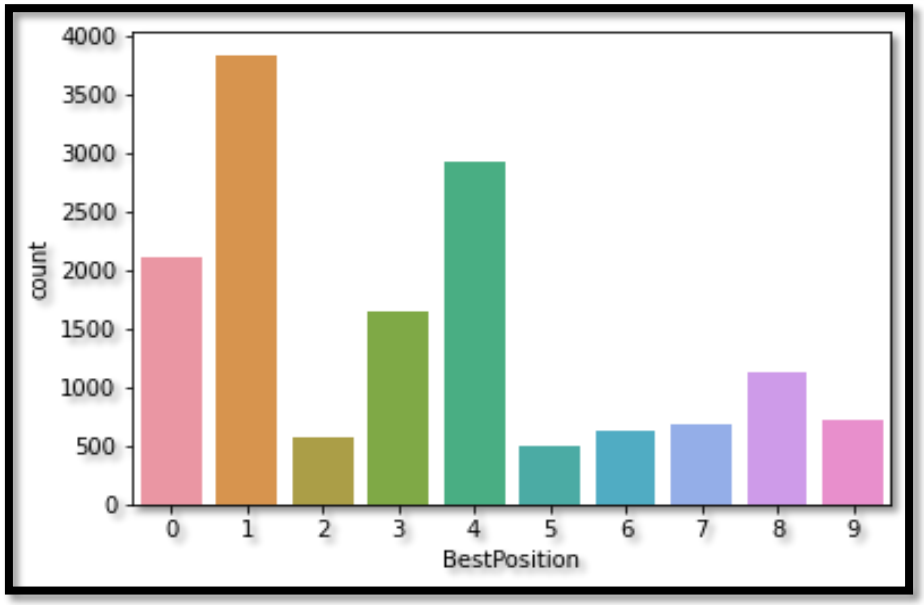
1. Handle the missing values

2. Handle The Categorical Columns

3. Handle the Imbalanced Data

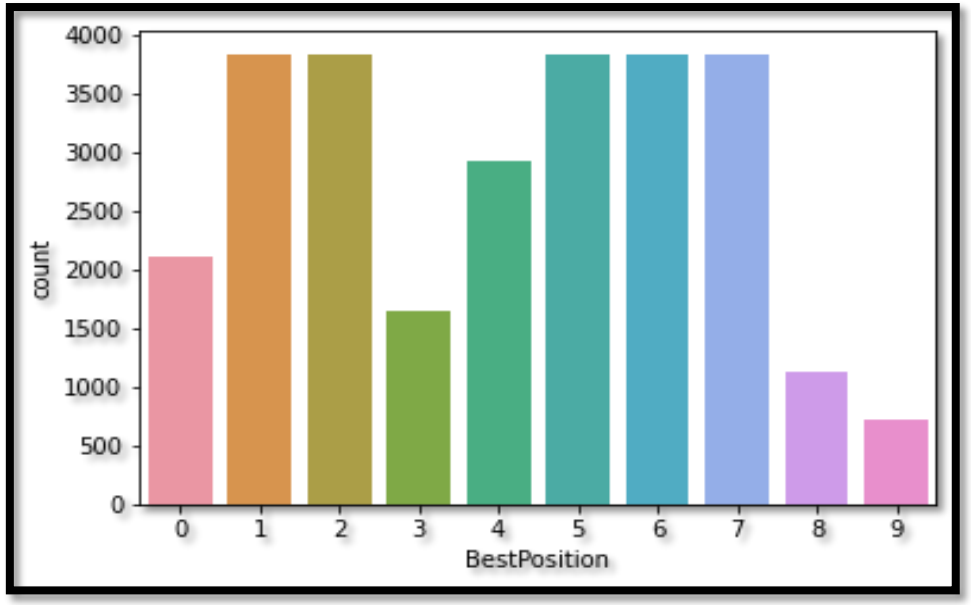
4. Feature Scaling

Handle the Imbalanced Data



As We can see Here the Data is Imbalanced s o we need to fix this issue.

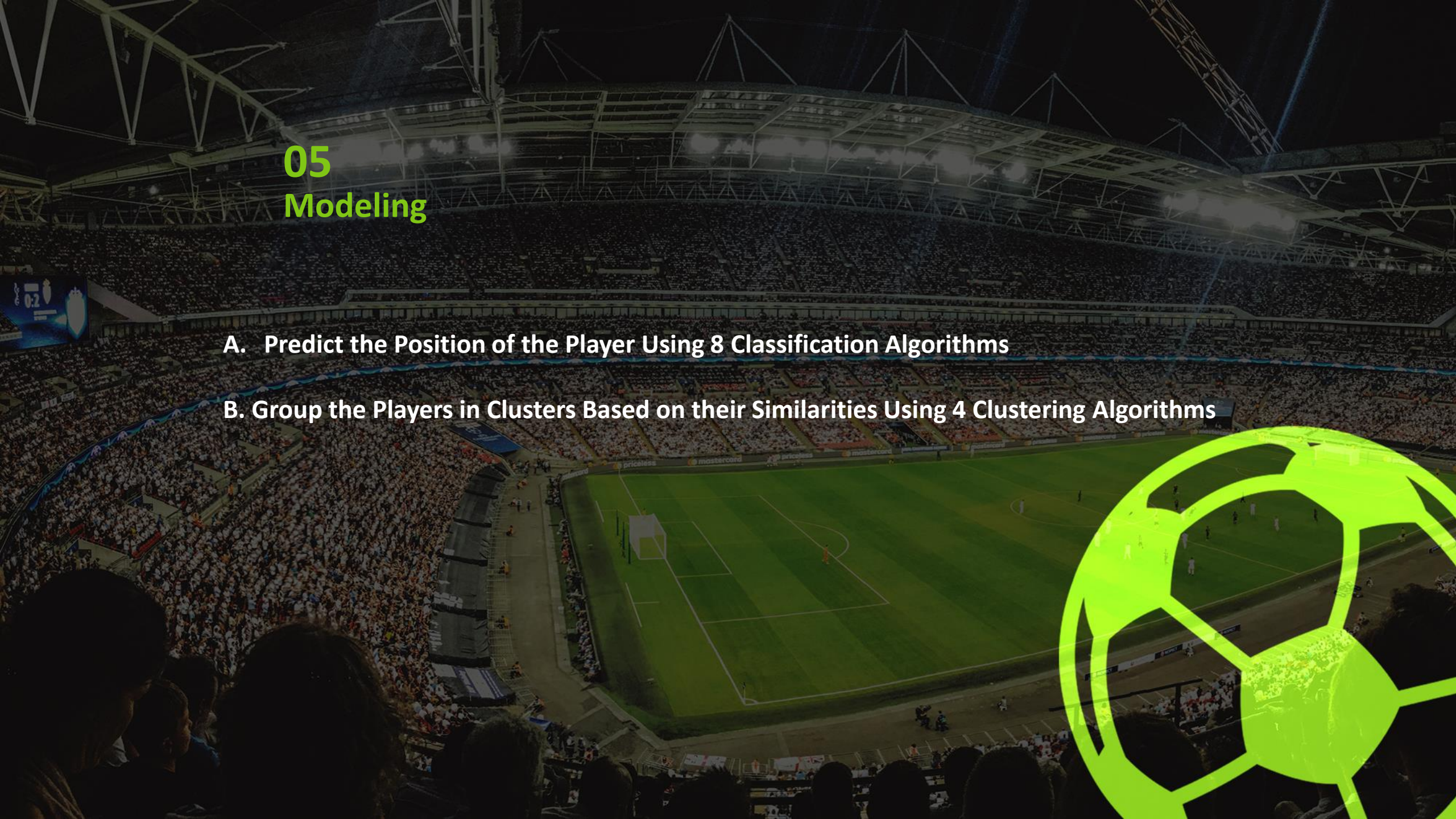
Used the Over Sampling method to Balance the classes 2, 5, 6, 7 So the model would not be biased.



05
Modeling

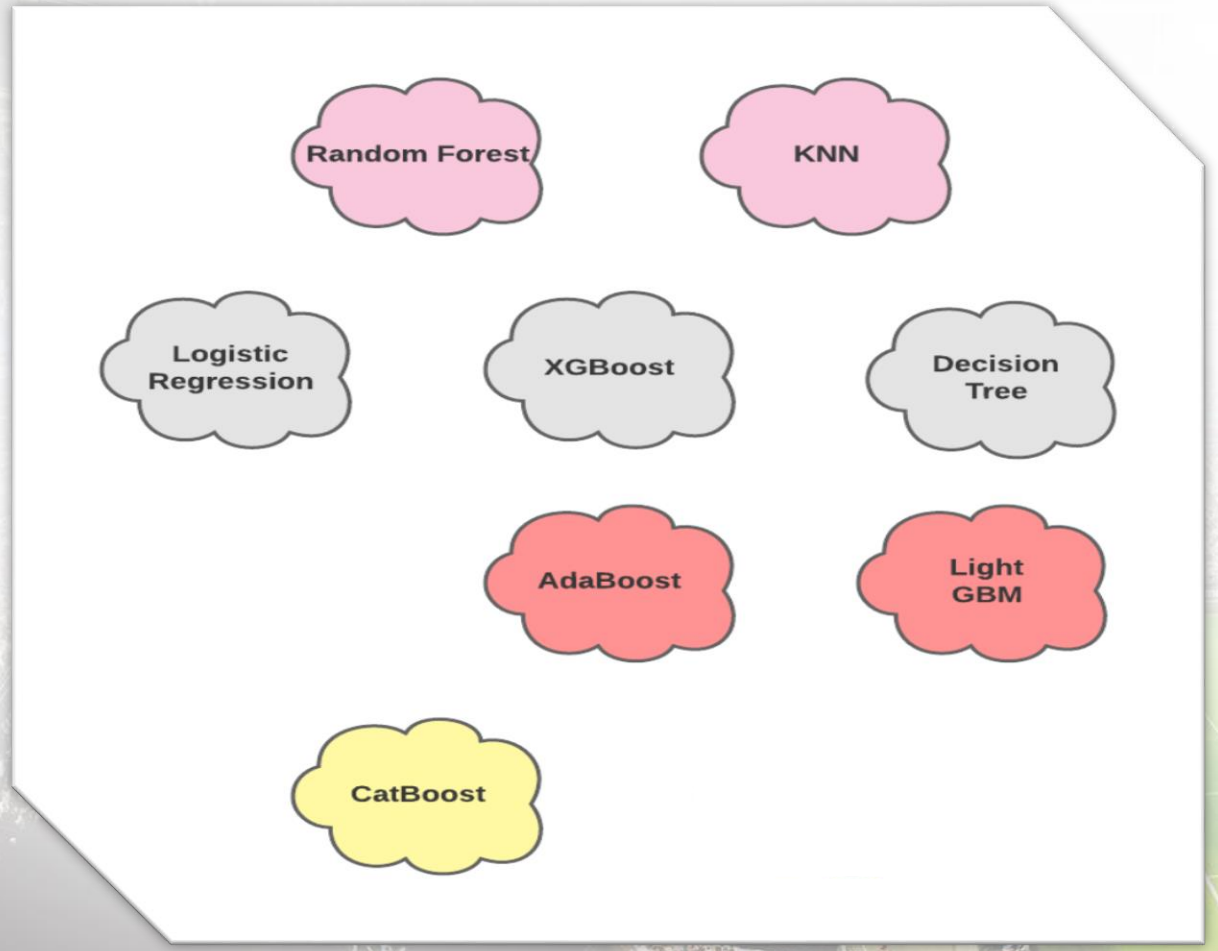
A. Predict the Position of the Player Using 8 Classification Algorithms

B. Group the Players in Clusters Based on their Similarities Using 4 Clustering Algorithms



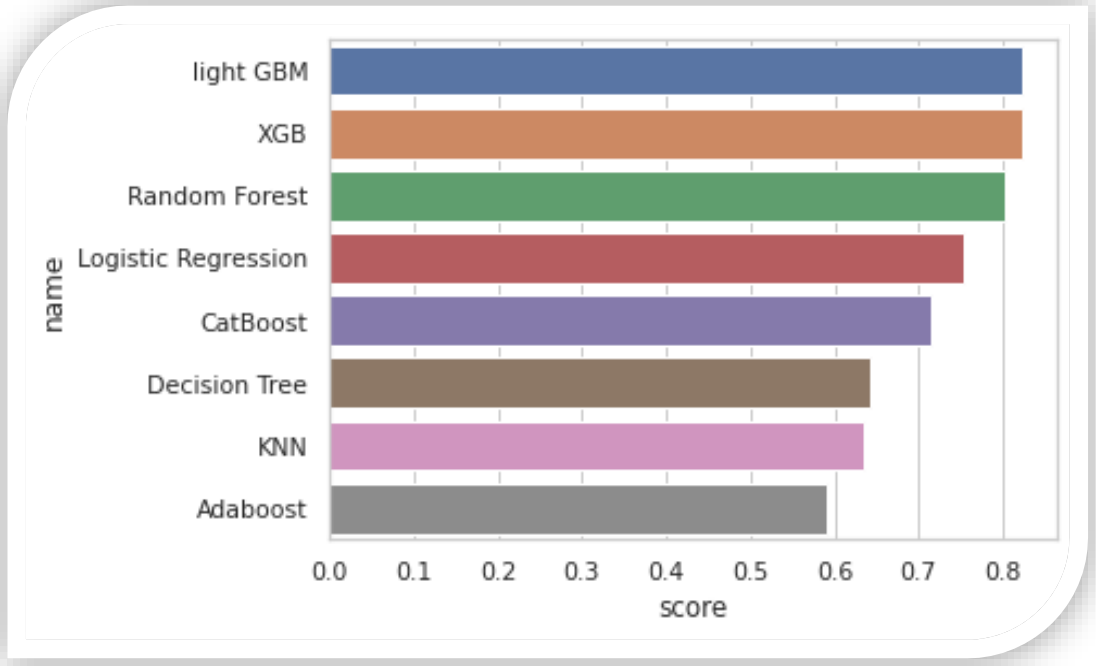
A.Predict the Position of the Player

Models Used:



Positions Predicted:

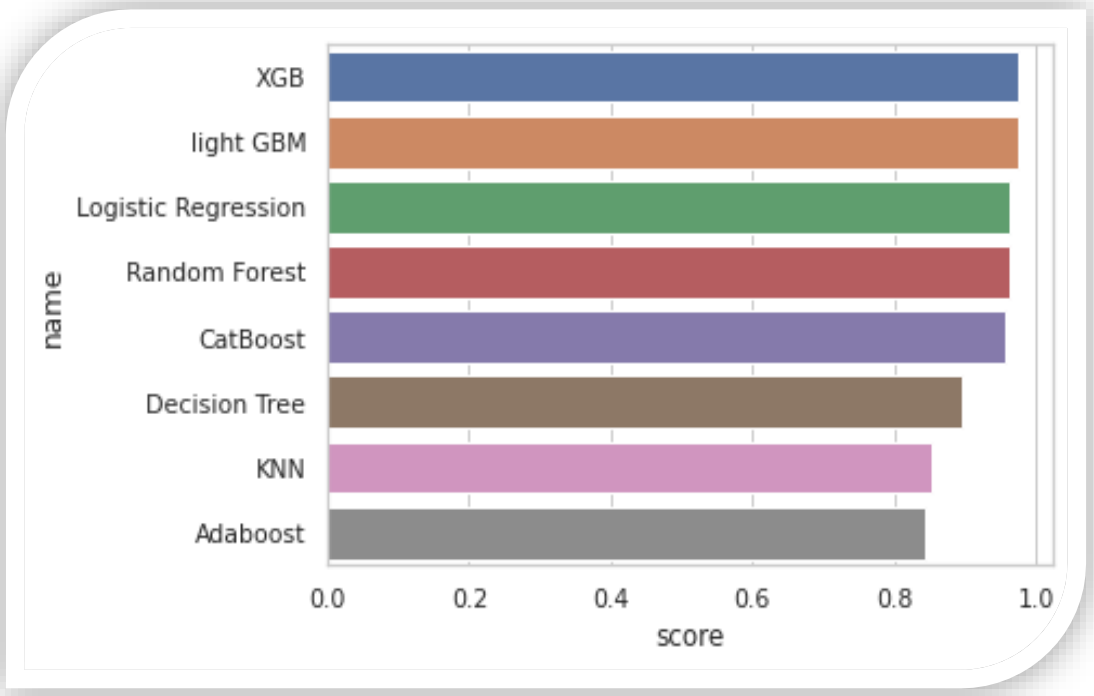
Comparing the test accuracy of the 8 Algorithms



name	score
Logistic Regression	0.752033
Random Forest	0.803252
XGB	0.821680
Decision Tree	0.642276
Adaboost	0.591599
light GBM	0.822222
CatBoost	0.714634
KNN	0.633604

So We Can Say that the light GBM and the XGB Algorithms are the best 2 Algorithms for that problem based on the accuracy.

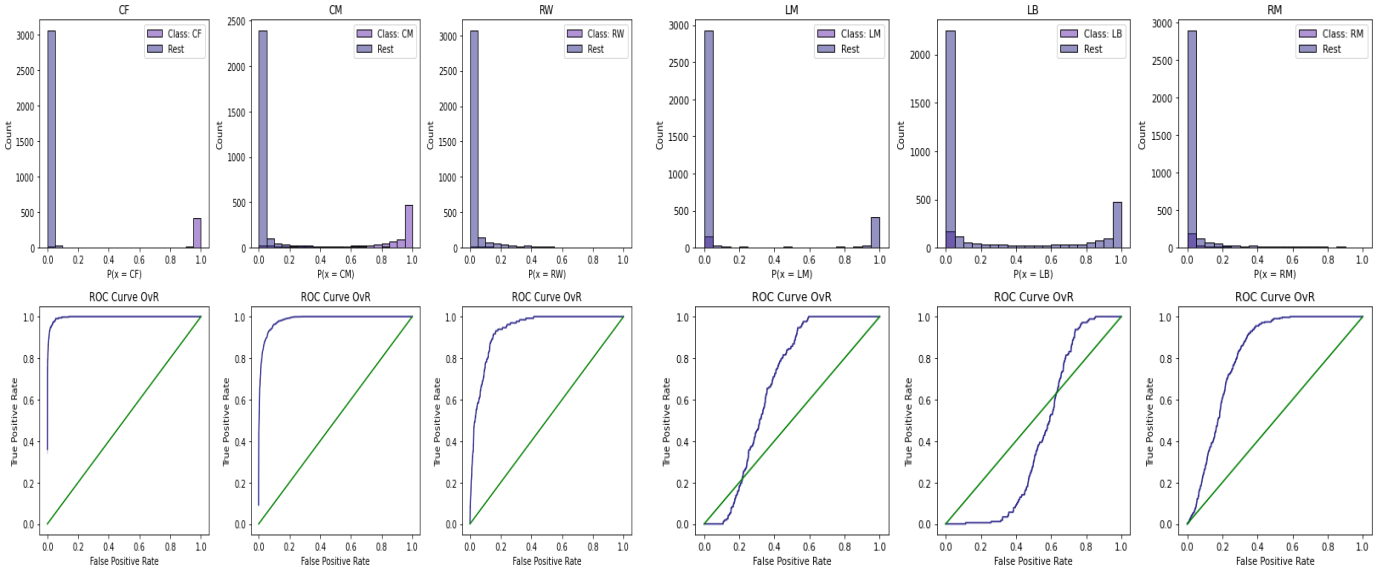
Comparing the ROC AUC Score of the 8 Algorithms

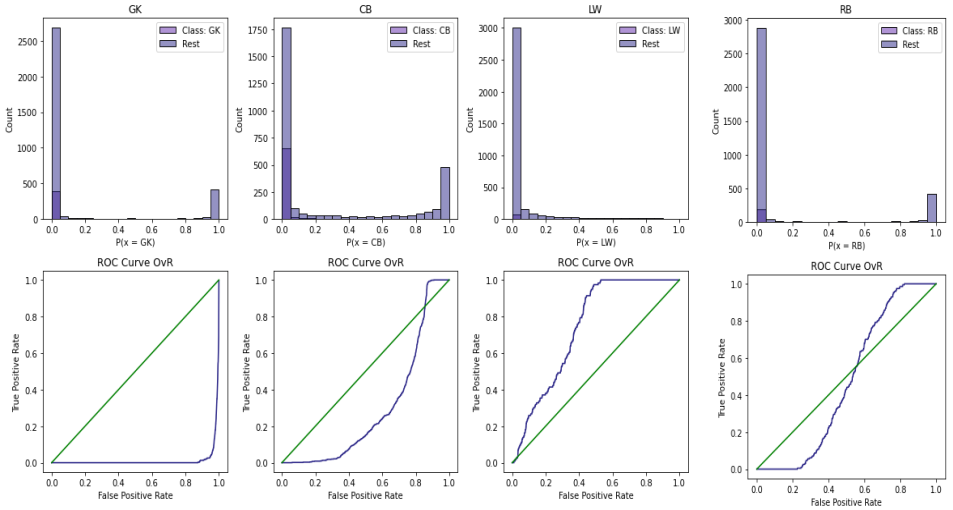


name	score
Logistic Regression	0.964473
Random Forest	0.964400
XGB	0.974716
Decision Tree	0.895296
Adaboost	0.842583
light GBM	0.974000
CatBoost	0.956526
KNN	0.854147

So We Can Say that the light GBM and the XGB Algorithms are the best 2 Algorithms for that problem based on the ROC AUC Score.

Light GBM





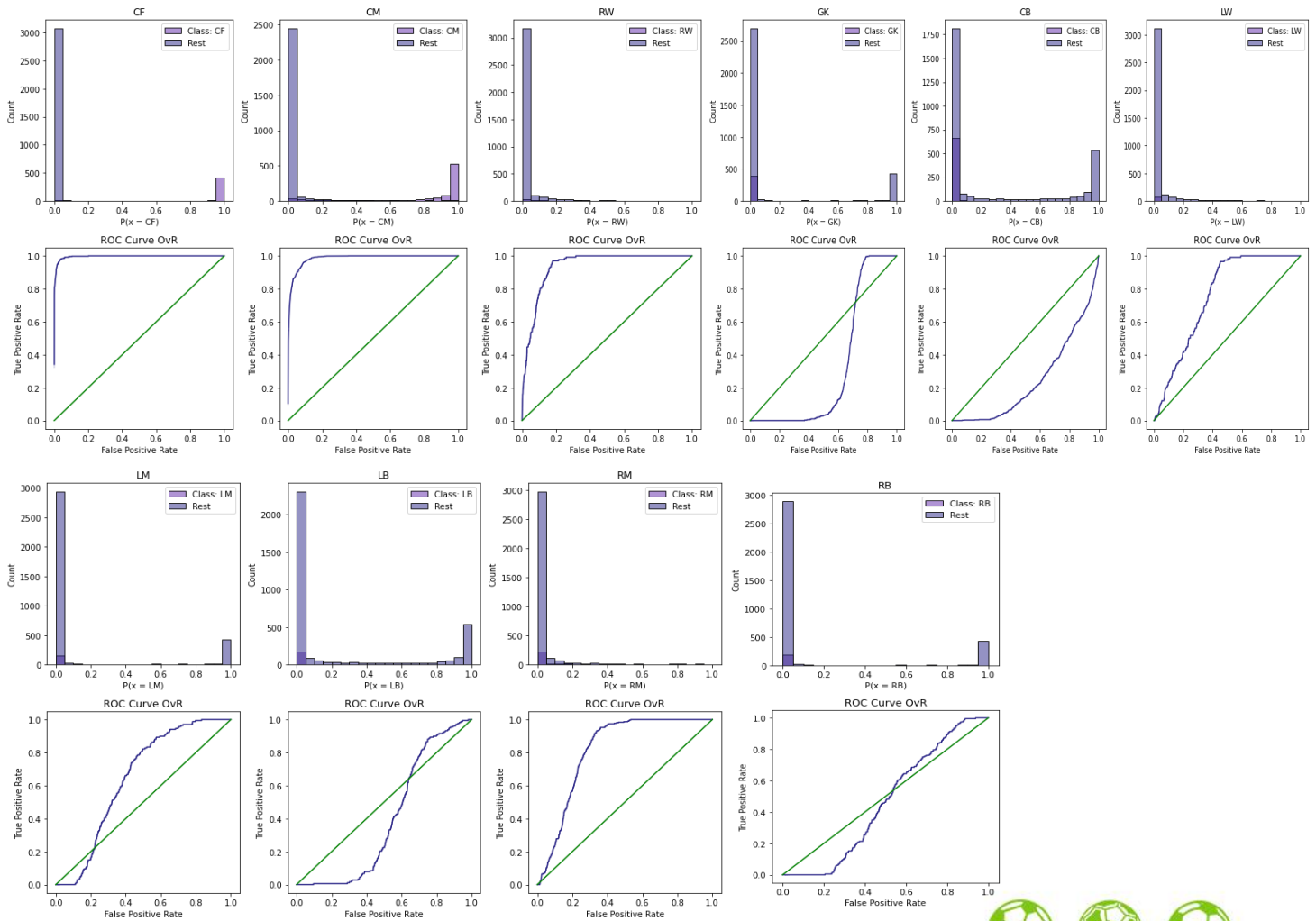
The Classification Report for light GBM Classifier:

	precision	recall	f1-score	support
0	0.91	0.92	0.92	522
1	0.90	0.85	0.88	962
2	0.35	0.39	0.37	132
3	1.00	1.00	1.00	391
4	0.91	0.95	0.93	711
5	0.37	0.32	0.34	116
6	0.58	0.58	0.58	168
7	0.71	0.74	0.73	178
8	0.67	0.67	0.67	313
9	0.68	0.71	0.70	197
accuracy			0.82	3690
macro avg	0.71	0.71	0.71	3690
weighted avg	0.82	0.82	0.82	3690

XGB

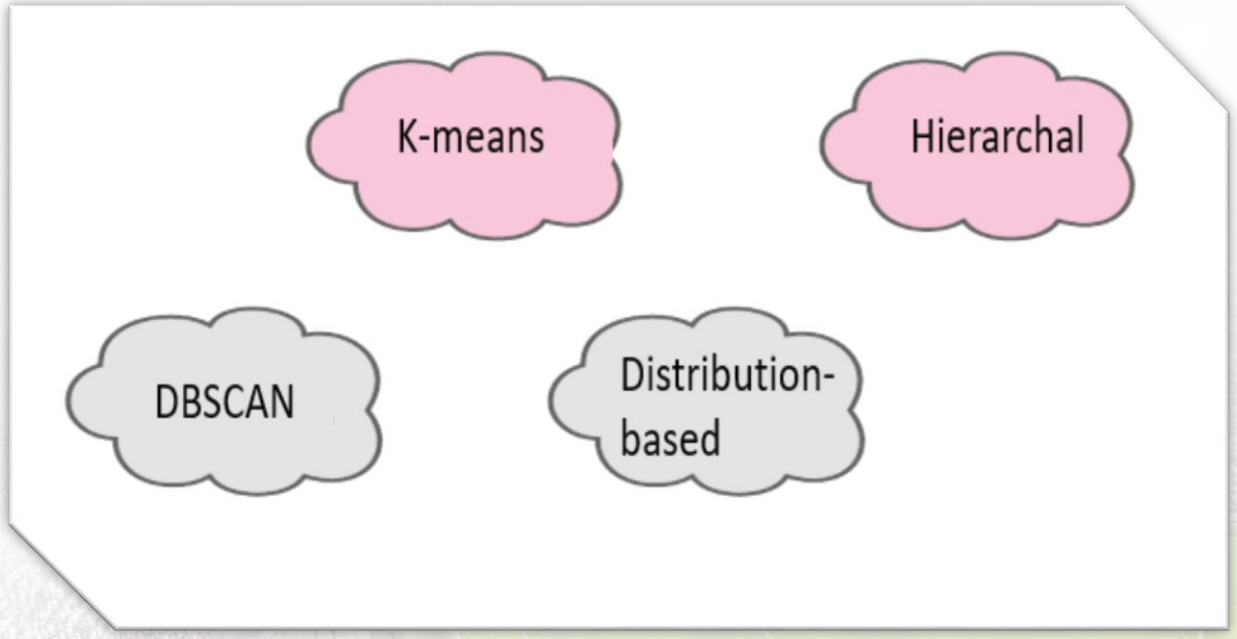
The Classification Report for XGB Classifier:

	precision	recall	f1-score	support
0	0.91	0.92	0.92	522
1	0.90	0.84	0.87	962
2	0.39	0.37	0.38	132
3	1.00	1.00	1.00	391
4	0.92	0.95	0.94	711
5	0.40	0.34	0.36	116
6	0.52	0.51	0.52	168
7	0.71	0.74	0.73	178
8	0.64	0.69	0.66	313
9	0.67	0.75	0.71	197
accuracy			0.82	3690
macro avg	0.71	0.71	0.71	3690
weighted avg	0.82	0.82	0.82	3690

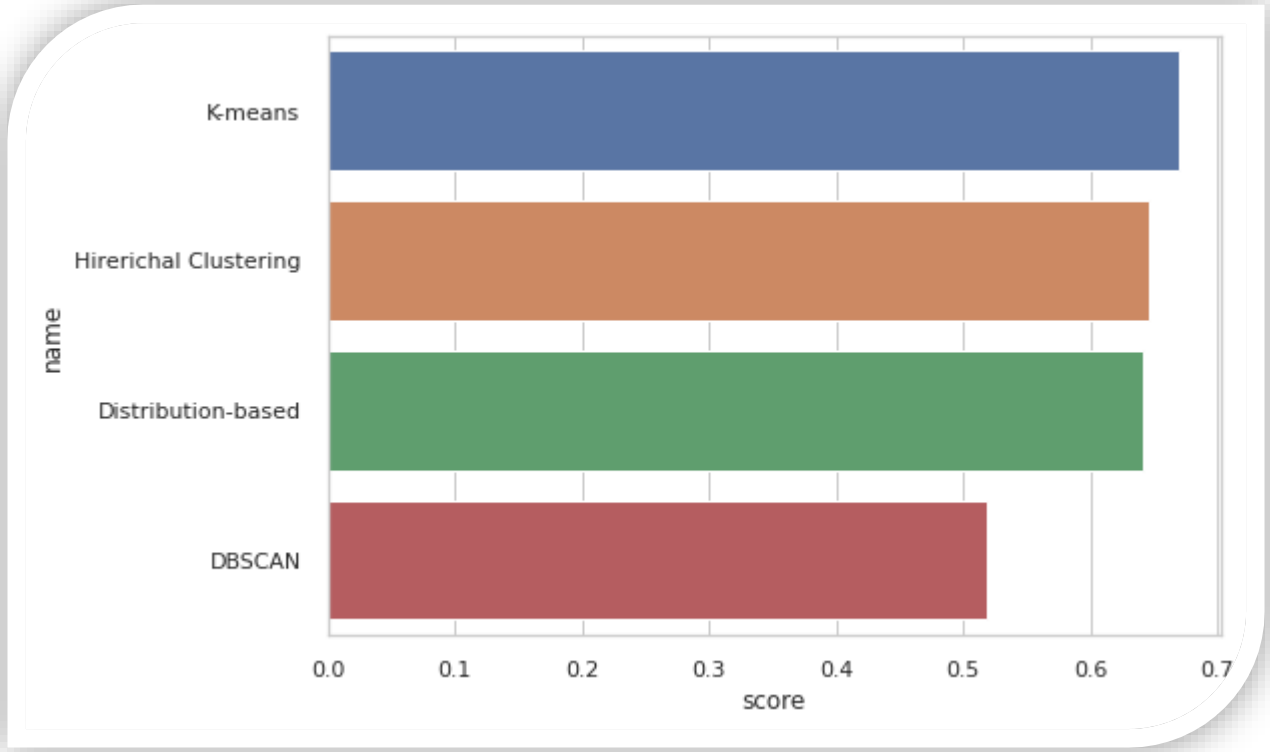


B. Group the Players in Clusters (with Overall > 86)

Models Used:



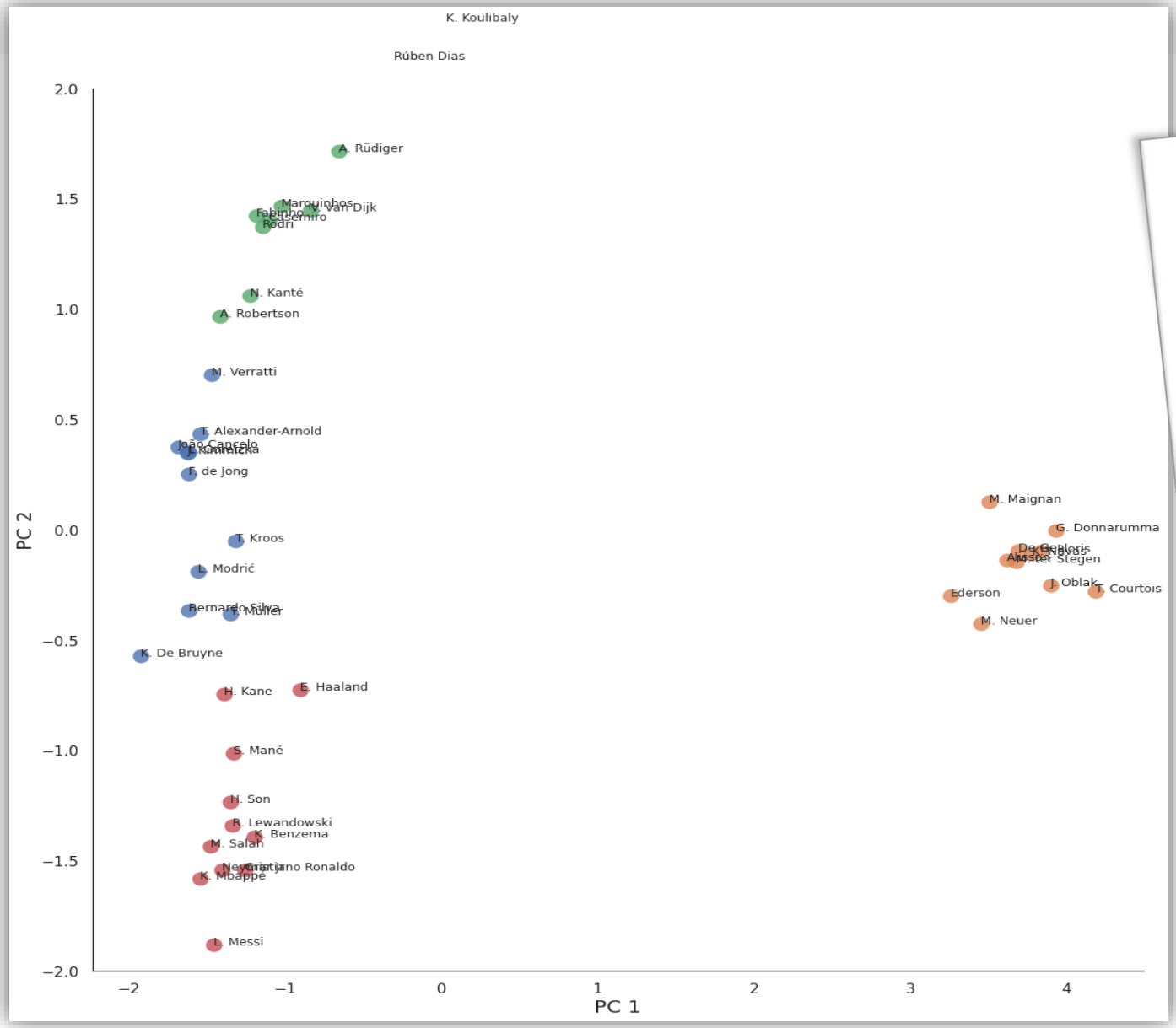
Comparing the 4 Algorithms based on the Silhouette Score

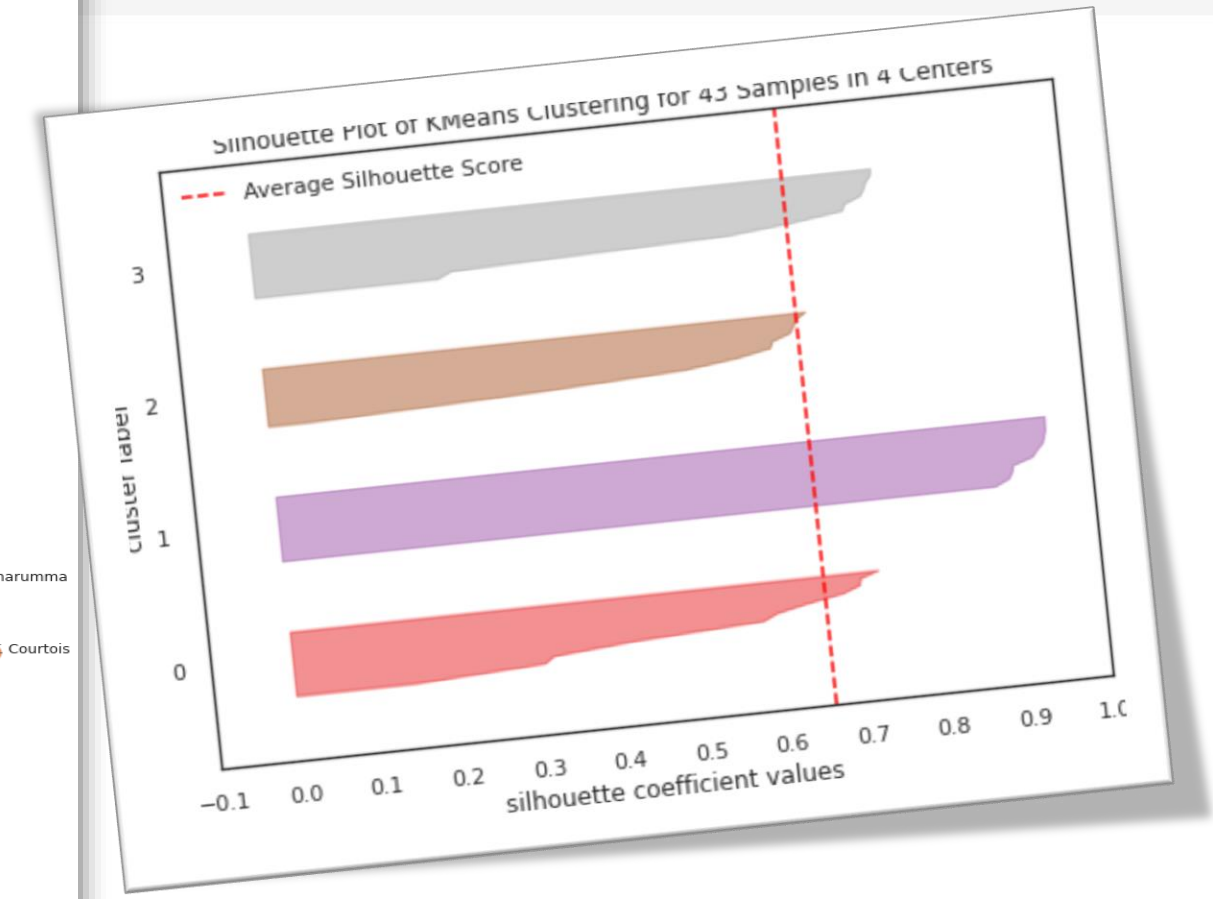


name	score
K-means	0.669469
Hirerichal Clustering	0.646751
DBSCAN	0.518514
Distribution-based	0.641061

So We Can Say that the Hierarchal Clustering and the K-means Algorithms are the best 2 Algorithms for that problem.

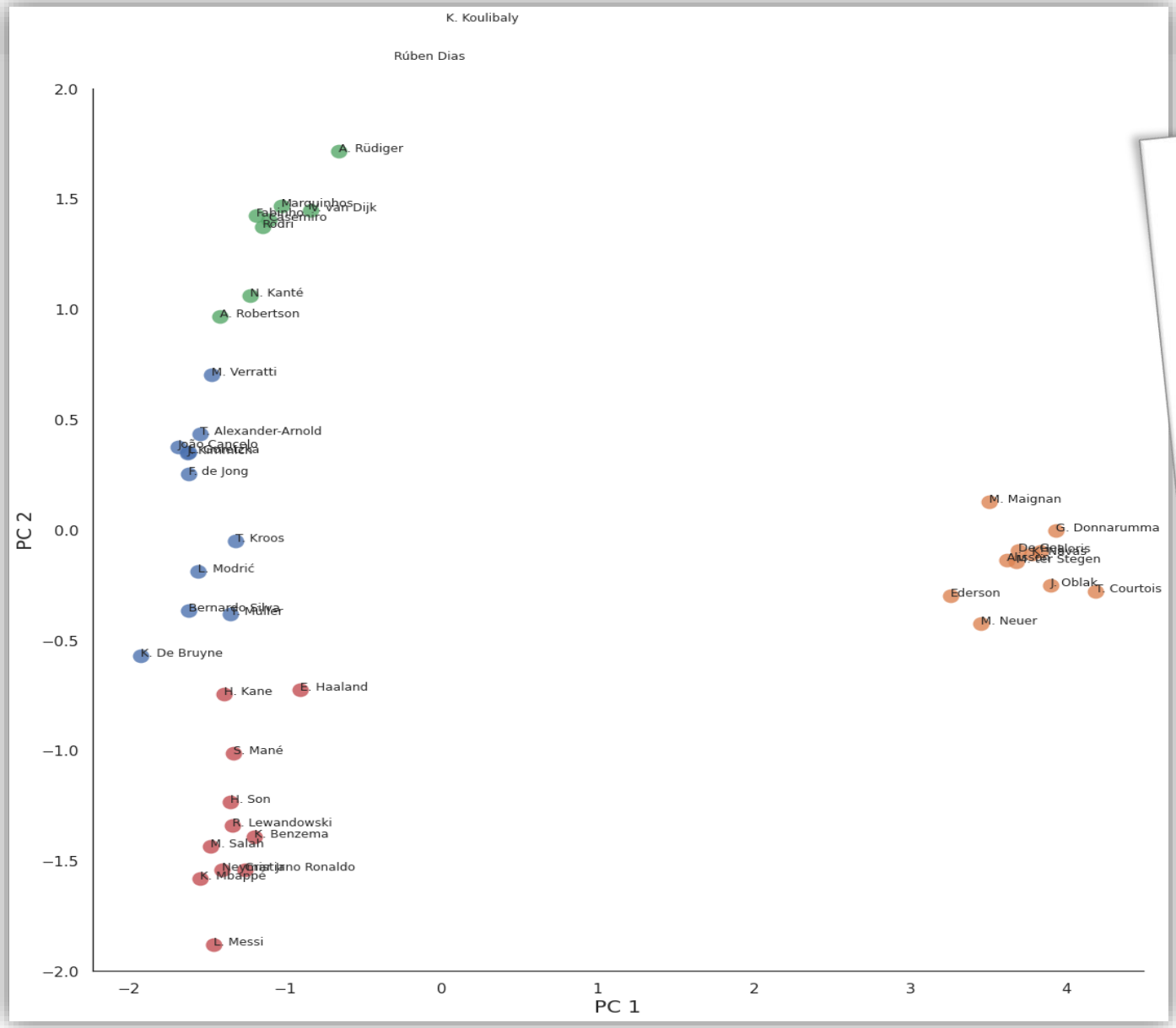
K-means

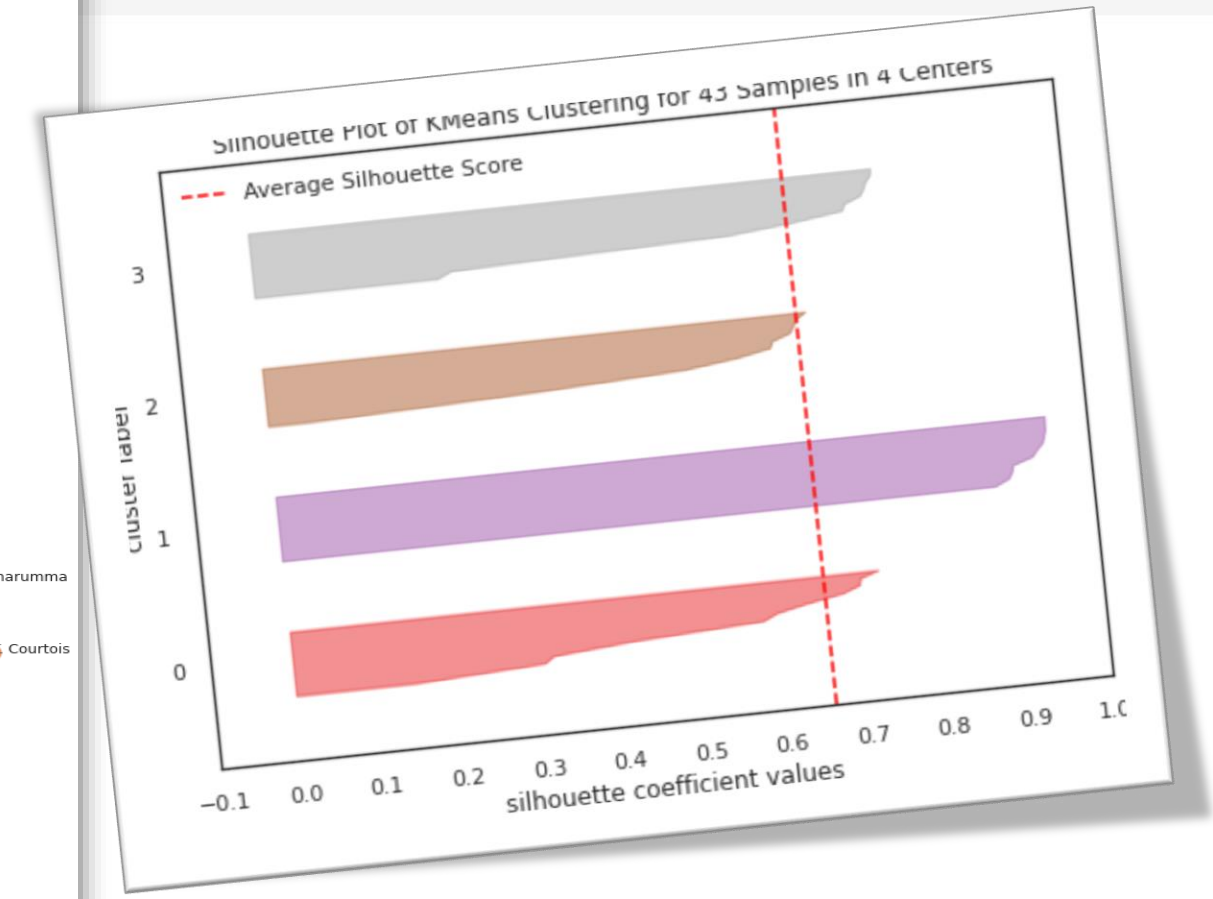

K. Koulibaly
Rúben Dias
A. Rüdiger
Marquinhos
Fabinho
Casemiro
Rodri
N. Kanté
A. Robertson
M. Verratti
T. Alexander-Arnold
João Cancelo
F. de Jong
T. Kroos
L. Modrić
Bernardo Silva
K. De Bruyne
H. Kane
E. Haaland
S. Mané
H. Son
R. Lewandowski
K. Benzema
M. Salah
K. Mbappé
L. Messi
M. Maignan
G. Donnarumma
Ederson
J. Oblak
T. Courtois
M. Neuer
PC 1
PC 2


Silhouette Plot of KMeans Clustering for 43 Samples in 4 Centers
Average Silhouette Score
Cluster Label
silhouette coefficient values

Hierarchal Clustering

K-means


K. Koulibaly
Rúben Dias
A. Rüdiger
Marquinhos
Fabinho
Casemiro
Rodri
N. Kanté
A. Robertson
M. Verratti
T. Alexander-Arnold
João Cancelo
F. de Jong
T. Kroos
L. Modrić
Bernardo Silva
K. De Bruyne
H. Kane
E. Haaland
S. Mané
H. Son
R. Lewandowski
K. Benzema
M. Salah
K. Mbappé
L. Messi
M. Maignan
G. Donnarumma
Ederson
J. Oblak
T. Courtois
M. Neuer
PC 1
PC 2


Silhouette Plot of KMeans Clustering for 43 Samples in 4 Centers
Average Silhouette Score
Cluster Label
silhouette coefficient values

06
Deployment

Deployment

Deployed the Classification Model and make an HTML page to test the predictions.


THANK YOU