Football Data Analysis

Group Members:

111903094 Sejal Sarnaik
111903095 Shalaka Pawar

Introduction:

The project is aimed at studying the FIFA football dataset, to analyse and extract information from it and make predictions based on the data. The objective is used to derive insights and correlations between various attributes and gather information.

Dataset Analysis:

The data is scraped from https://sofifa.com/ website using Beautiful Soup Python library. The scraped data is stored in CSV format.
The dataset includes the information of the players who were a part of FIFA 2021. It is used to derive insights and correlations between various attributes and gather information.

Dataset size is: 11940 rows and 71 columns

Attributes

- Player ID
- Name
- Sprint Speed
- Agility
- Weak Foot
- Skill Moves
- Position
- Dribbling
- Stamina
- Strength
- Penalties
- Defensive awareness
- Standing Tackle
- Age

- Photo
- Nationality
- Height
- Weight
- Flag
- Overall
- Sliding tackle
- Potential
- Club
- Club Logo
- Value
- Wage
- Preferred Foot
- GK Kicking
- GK Positioning
- GK Reflexes

.

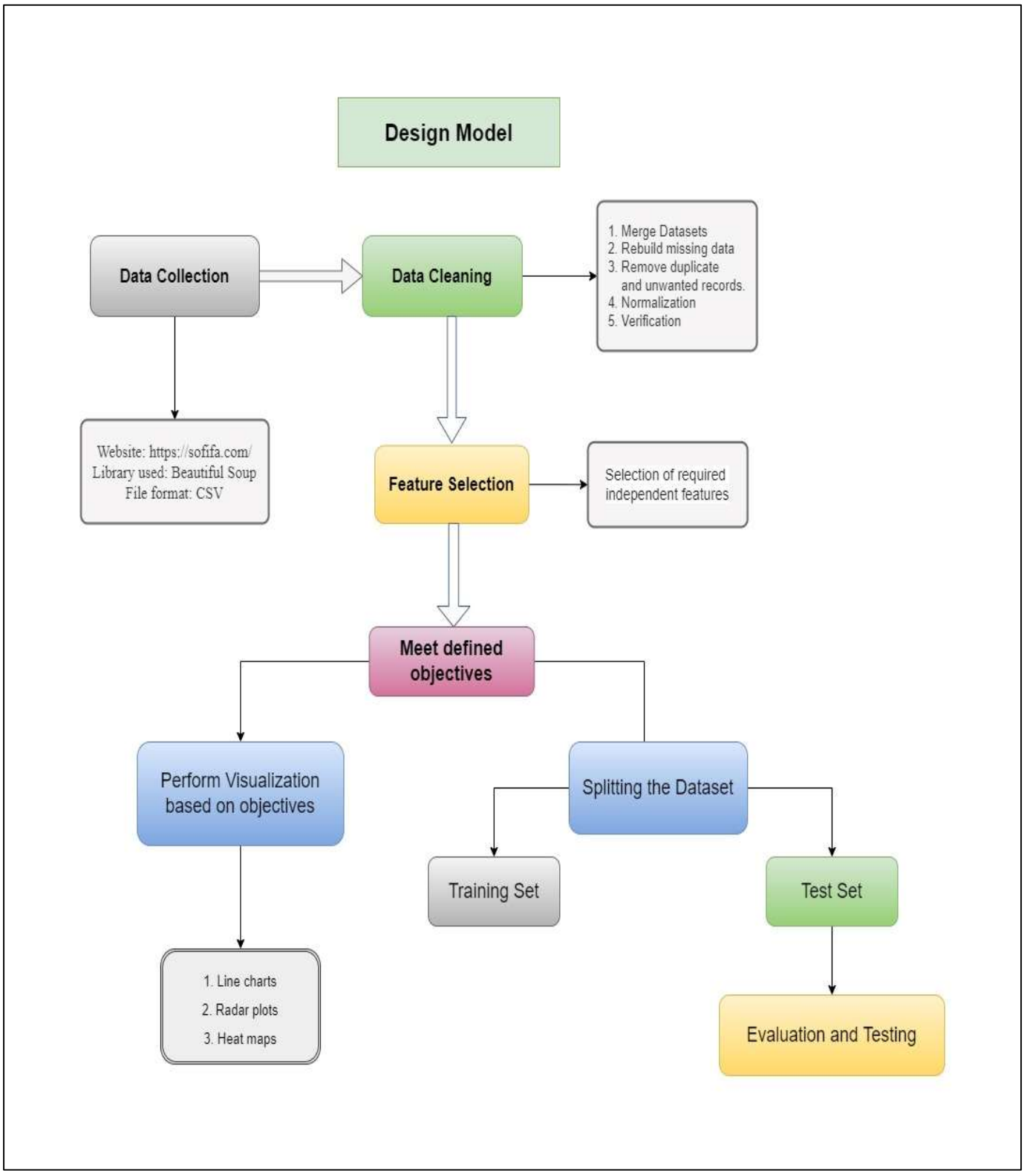
- Positions on field

CAM, CB, CDM, CF, CM, ID, LAM, LB, LCB, LCM, LDM, LF, LM, LS, LW, LWB, Preferred Positions, RAM, RB, RCB, RCM, RDM, RF, RM, RS, RW, RWB, ST

Data Required

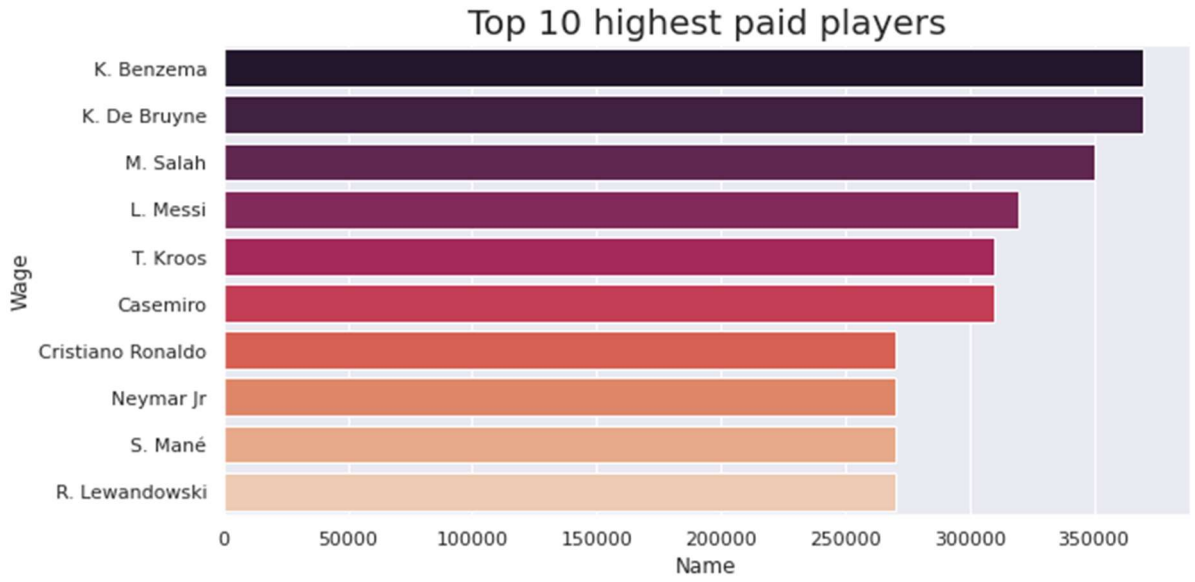
1. Basic information about players featured in FIFA 2021.
2. Personal data like Nationality, Age, Weight, Salary, contract details etc.
3. Player, Club, Flag Images to be used in visualization.
4. Player Position Data based on the FIFA 2021 game.

Design Model:

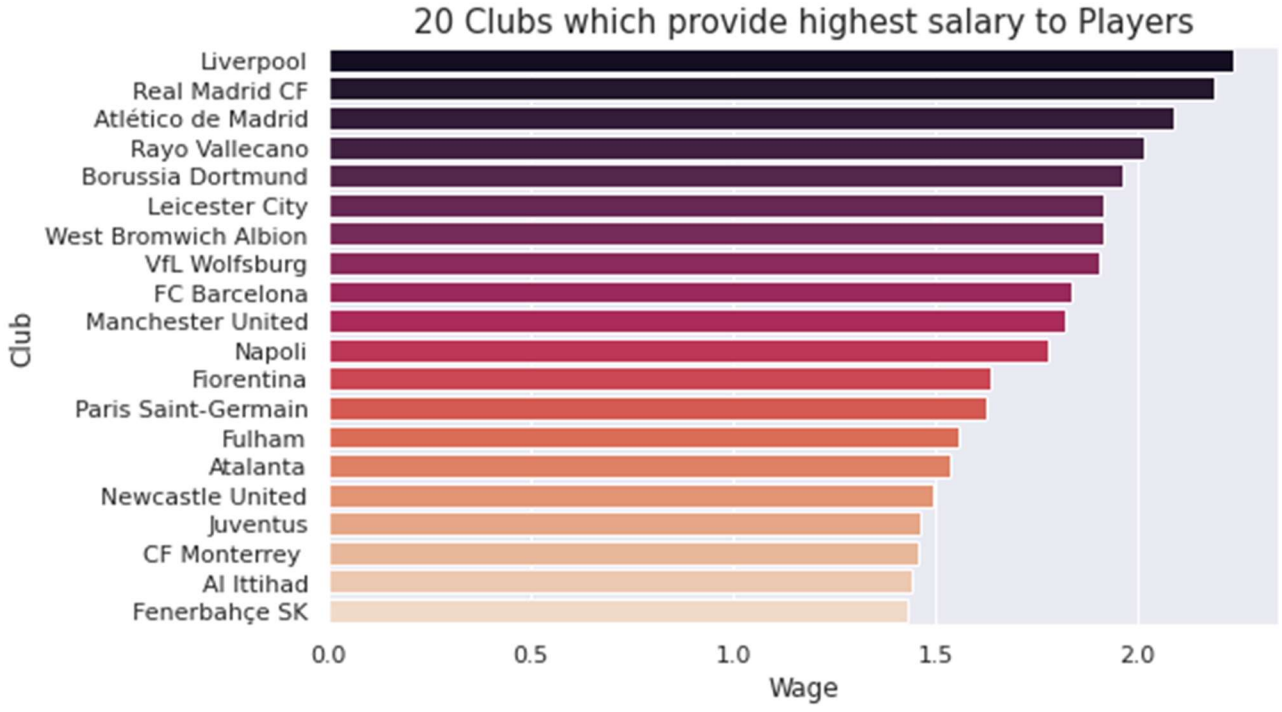

Design Model
Data Collection
Data Cleaning
1. Merge Datasets
2. Rebuild missing data
3. Remove duplicate and unwanted records.
4. Normalization
5. Verification
Website: https://sofifa.com/
Library used: Beautiful Soup
File format: CSV
Feature Selection
Selection of required independent features
Meet defined objectives
Perform Visualization based on objectives
Splitting the Dataset
Training Set
Test Set
1. Line charts
2. Radar plots
3. Heat maps
Evaluation and Testing

Objectives

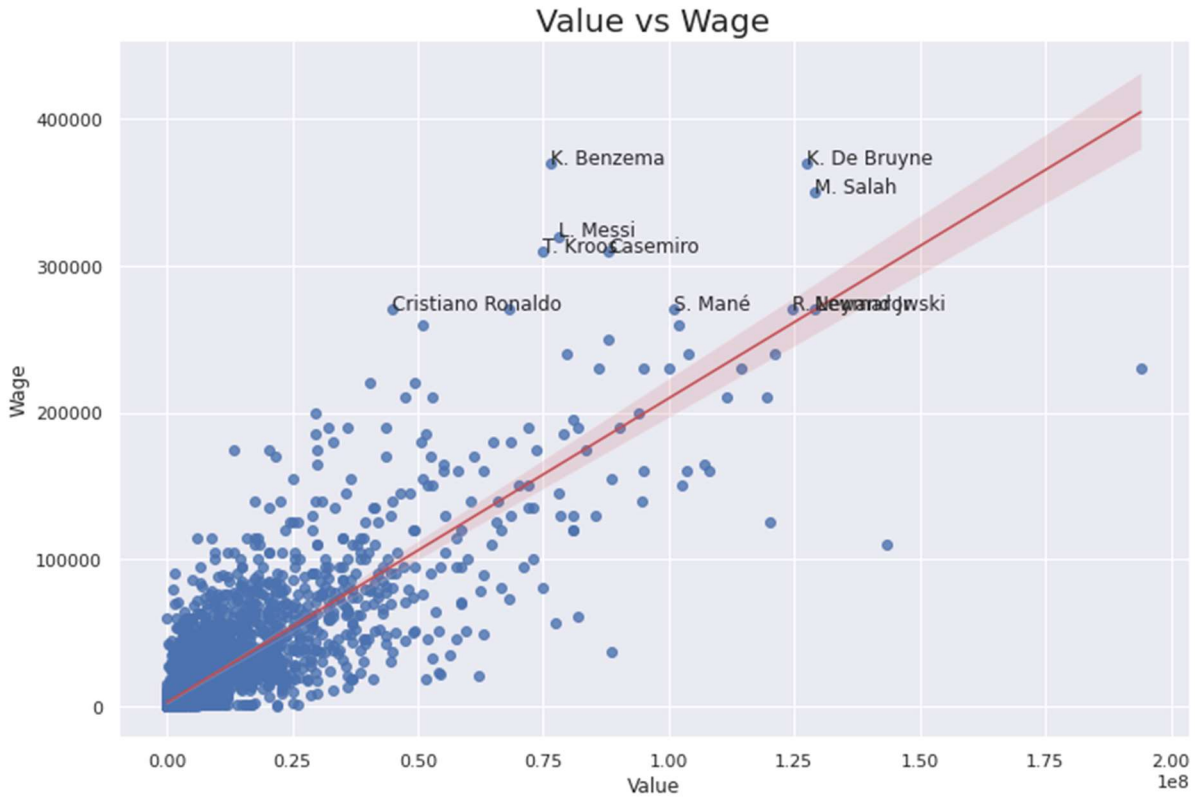
Find the top 10 highest paid players of the dataset.



- **Determine which clubs provide the highest salaries to players.**

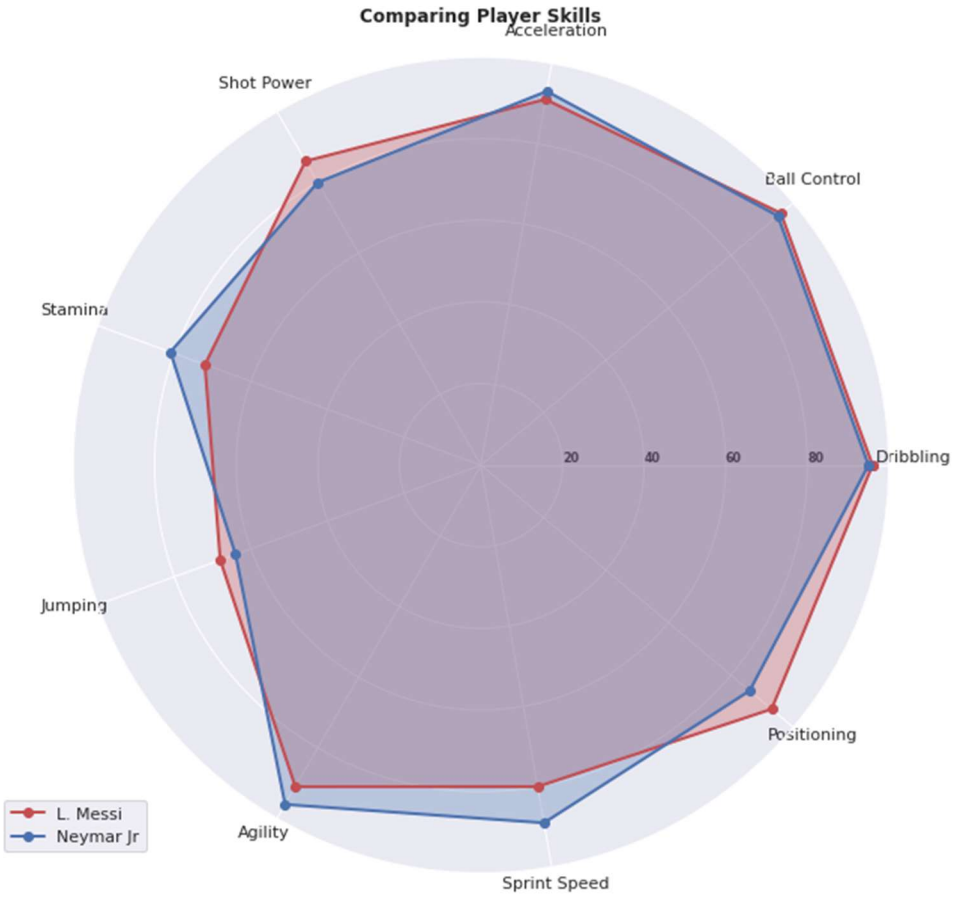


- **An analysis of the players based on the Value and Wages.**



- **Evaluate the co-relation between age and overall rating of the players.**

- **Visualization of the comparison between two or more players based on their attributes or skills.**



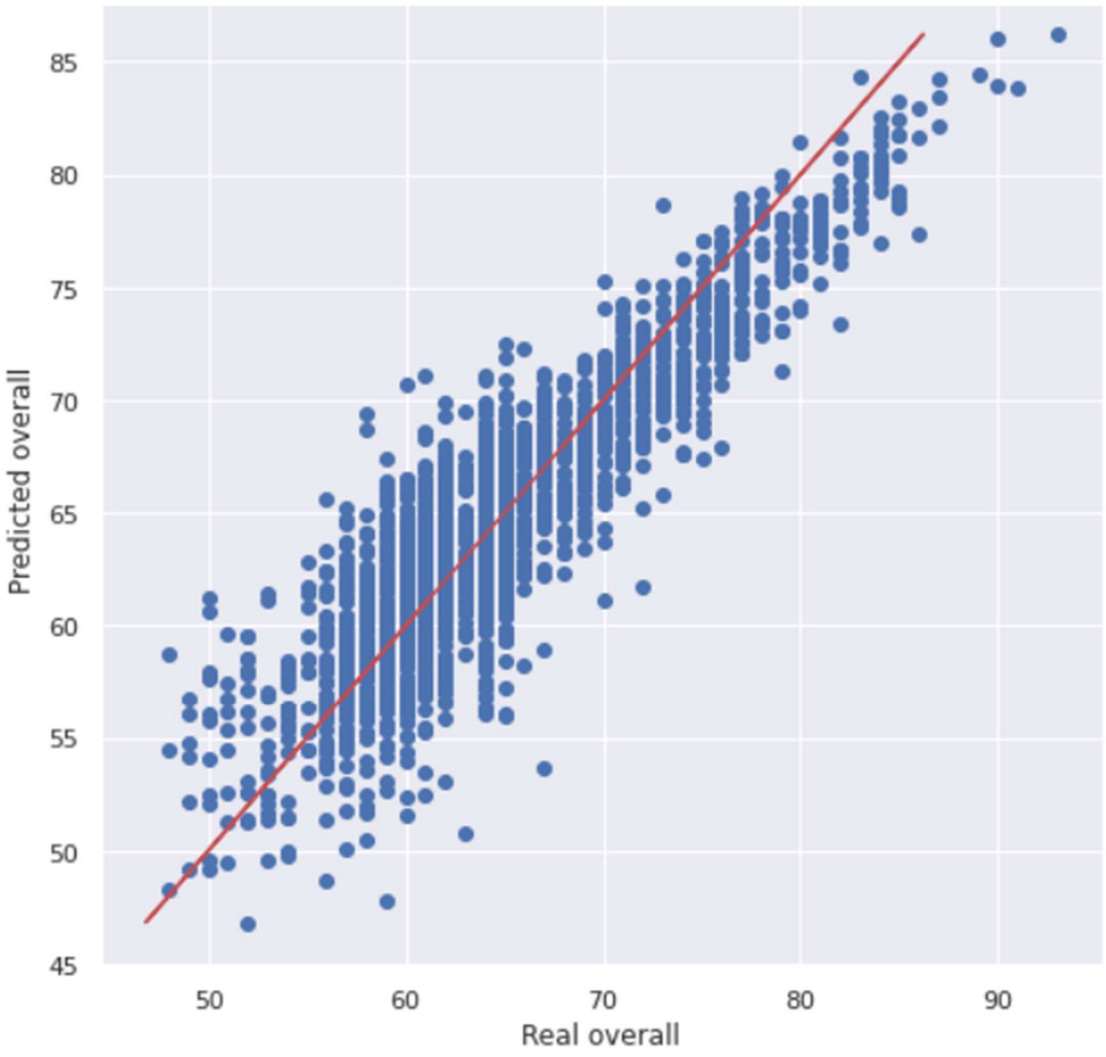
- **Predict player's overall value based on the linear regression and comparison with values given in dataset.**

  Attributes used for predicting the Overall Value:

  'Dribbling', 'Curve', 'FK Accuracy', 'Long Passing', 'Ball Control', 'Acceleration', 'Sprint Speed', 'Agility', 'Reactions', 'Balance', 'Shot Power', 'Jumping', 'Stamina', 'Long Shots', 'Aggression', 'Interceptions', 'Positioning', 'Vision', 'Penalties', 'Defensive awareness', 'Standing Tackle', 'Sliding tackle'

Linear Regression Model:

It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables. It shows a linear relationship between a dependent and one or more independent variables.

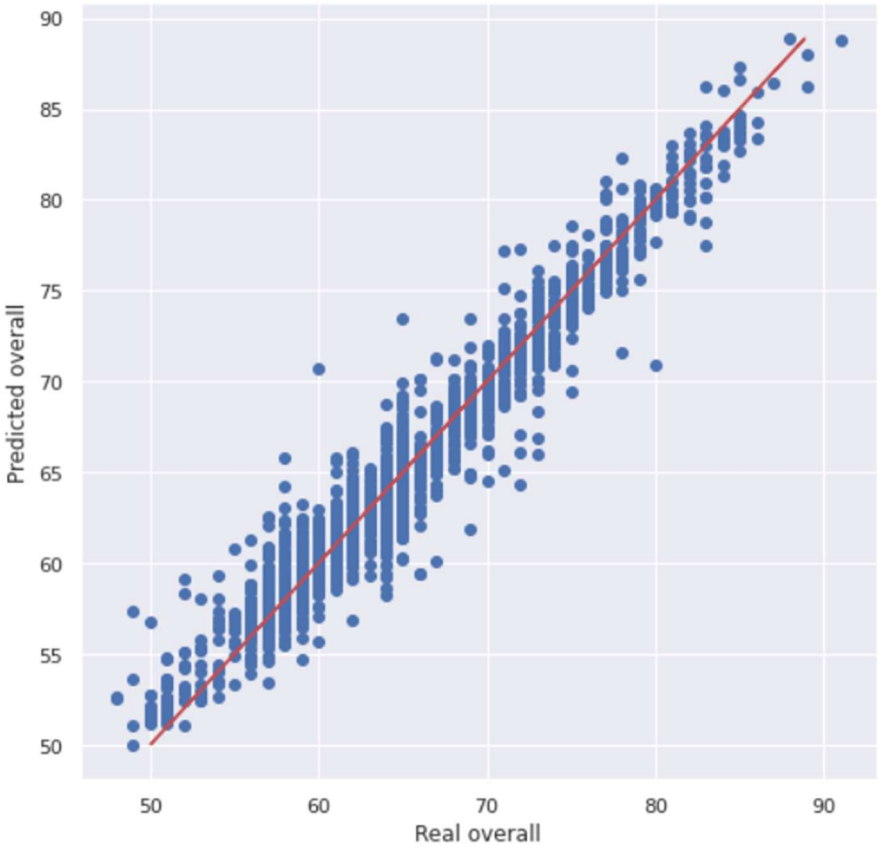


Metric	Value
Mean Aboslute Error	2.3210889290844636
Root Mean Squared Error	8.819653486486967
R2 score	0.8058956227573816

	Real overall	Predicted overall	Difference
0	61	62.094471	1.094471
1	62	63.456390	1.456390
2	65	63.759945	1.240055
3	64	65.902916	1.902916
4	73	75.103370	2.103370
5	63	65.944179	2.944179
6	56	56.369211	0.369211
7	62	64.787400	2.787400
8	64	67.608602	3.608602
9	63	60.240008	2.759992

Random Forest Classifier:

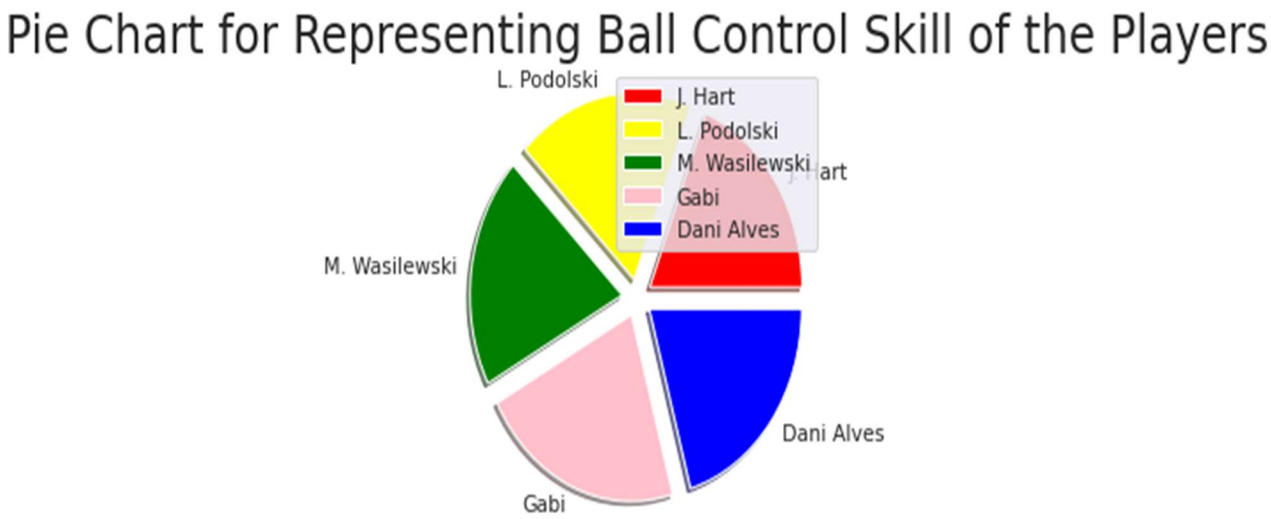
Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression.

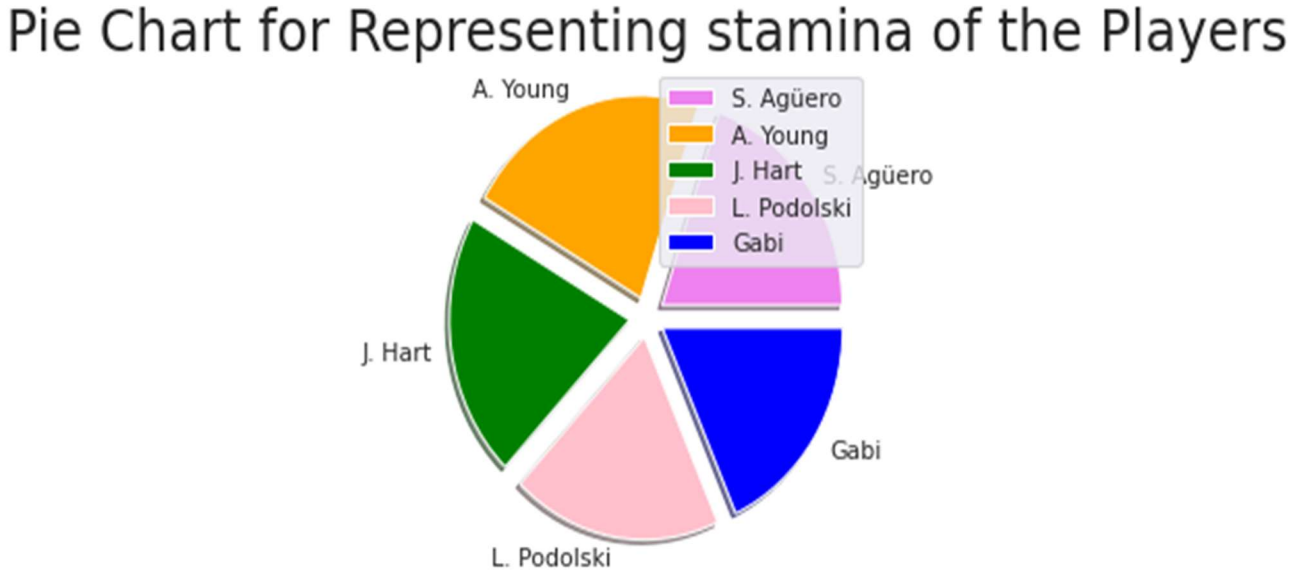


Metric	Value
Mean Aboslute Error	1.1902432551968156
Root Mean Squared Error	2.7089095975232196
R2 score	0.9426600836241934

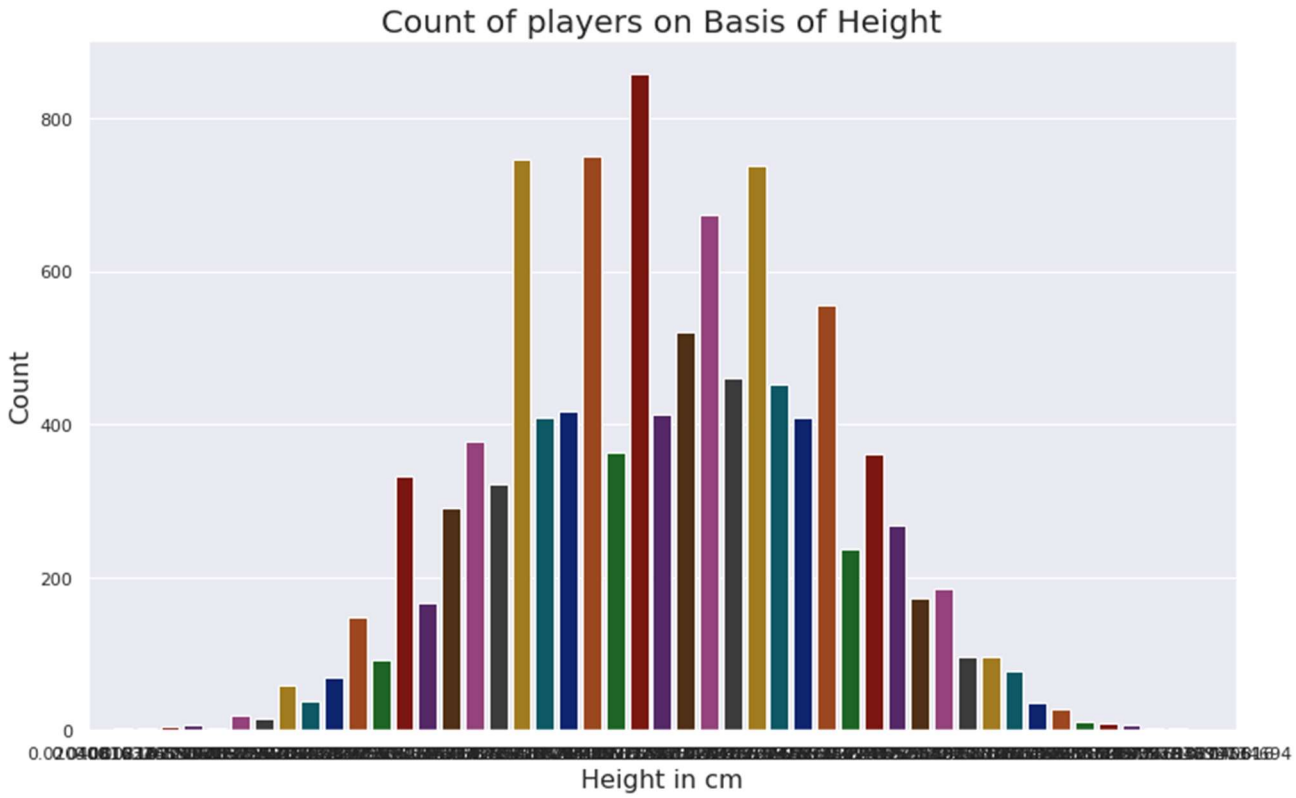
	Real overall	Predicted overall	Difference
0	64	62.68	1.32
1	65	68.59	3.59
2	61	60.89	0.11
3	65	64.63	0.37
4	64	63.13	0.87
5	69	68.50	0.50
6	59	59.97	0.97
7	61	60.62	0.38
8	61	61.24	0.24
9	69	68.45	0.55

Pie chart:


Pie Chart for Representing Ball Control Skill of the Players
L. Podolski
J. Hart
L. Podolski
M. Wasilewski
Gabi
Dani Alves
M. Wasilewski
Dani Alves
Gabi


Pie Chart for Representing stamina of the Players
A. Young
S. Agüero
A. Young
J. Hart
L. Podolski
Gabi
S. Agüero
J. Hart
Gabi
L. Podolski

- **Count of Players based on Height**



Summary:

The project was done with the purpose of finding out factors that affected the performance of the football players, visualizing the data to depict the relation between the attributes, and training a model which accurately predicts the Overall Rating of the Player.

It was observed that the more the R2 score the better the result. More value of R2 was obtained when we used Random Forest Regression.

The important features of the Random Forest Algorithm are Reactions, Ball control, and Positioning. This means that if these factors increase, Overall Rating will increase.

We can conclude that improving the Reaction to situations on the field quality of the player will increase the overall rating of the player. Also, the player should improve their Ball control skills and work on their Positioning for better performance.