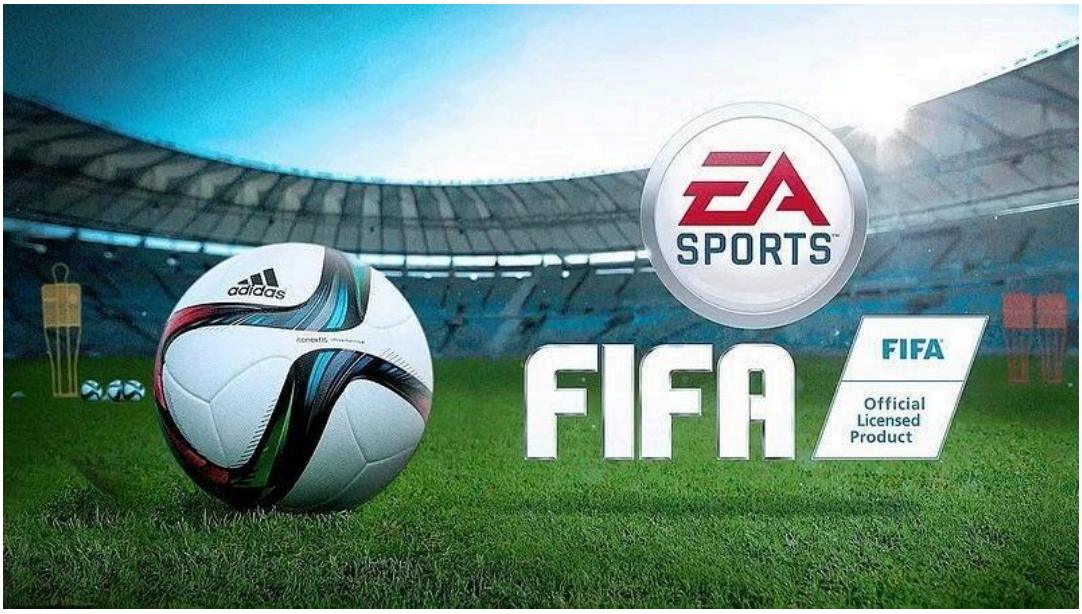
FORE SCHOOL OF MANAGEMENT



FIFA 22-23 RAC PROJECT

Submitted To: Prof. Amarnath Mitra & Prof. Ashok Harnal

Submitted By: Group 6

NAME	ROLL NO
ADHYATIK	311063
AASHIT SHARMA	321148
ARINDAM CHAKRABORTY	321128
SUBHAJIT PAUL	321173

TABLE OF CONTENTS

EXECUTIVE SUMMARY

This report presents a comprehensive FIFA players' analysis for the Real World Analytics project using both supervised and unsupervised machine learning techniques. The objective as two Analyze Football Player Data, statistics, and characteristics, enabling to identify player' trait and most successful players. Also there was an objective to create a dashboard to infer multiple player or club traits.

The analysis leveraged a dataset encompassing approximately 25218 records

There were 5 significant variable, inferred from ML model and they were :

1. **Age**
2. **League_name**
3. **Nationality_name**
4. **Wage_eur**
5. **Club_name**

The findings revealed 7 distinct clusters:

Cluster 0: Established Stars – Players who are already at peak in their profile

Cluster 1: Raising Talents – Players with high potential

Cluster 2: Veteran Leaders - Refers to experienced players who has demonstrated exceptional leadership skills, often acquired through years of playing in the league.

Cluster 3: Mid-Career Players – Players who are at peak and have long career ahead.

Cluster 4: International Prospect – Players with high progress in International career but not in focus in leagues.

Cluster 5: Domestic Role Player – Players who excel in specific roles within a team's formation, particularly in set-piece situations such as free kicks, corners, and penalties.

Cluster 6: Journeymen - In FIFA, a journeyman player is a fictional character who has played for multiple teams throughout their career, often without achieving significant success or recognition.

OBJECTIVES

1.1 Creation of the analytical dashboard and analysis of

a. Analyze Football Player Data: Provide a comprehensive view of player attributes, club positions, and league performances.

b. Visualize Key Patterns: Highlight patterns in overall ratings, club position distributions, and league-wise average ratings.

c. Explore Attribute Correlations: Investigate the relationship between player height and overall rating to identify potential performance trends.

d. Player attribute correction and overall player Analysis

e. Player value prediction using linear regression

f. Support Strategic Decisions: Offer valuable insights for player comparison, aiding in informed decision-making for team management and player selection.

1.2.) Developing a Machine learning model using unsupervised K-means and Supervised Decision tree.

a. In k- means clustering we identified the optimal no. of clusters and identified the characteristics of the clusters while also naming them.
b. In decision tree we identified the significant variables from the identified clusters and also found out the accuracy and significance of the model

DESCRIPTION OF DATA

1.Data Source, Size & Shape

Data Source (Website Link): - https://www.kaggle.com/datasets/sabir0000/male-football-players-data

Data Size: - 26 MB

Data Description & Dimensions: - This dataset, titled "male_players_football_Data" provides a comprehensive view of football players. It encompasses approximately 150000+ records, each representing an individual player statistic. The primary objective of this dataset is to understand and do the analysis of football players in the FIFA game, a key indicator of each player statistics and player experience in the burgeoning field of sports.

Number Of Observations Taken for Analysis: - 25218

Number Of Variables: - 68

Description of Variables

Index Variable(s): - 03

1. player_id-- ID of a specific player
2. fifa_version-- Year of game relealse
3. short_name-- Name of the player

Categorical Variables or Features (CV): - 17

1. league_name—the name of the league in which the game is played
2. league_level -- Level of difficulty to play in a major league
3. club_name—The name of the club
4. club_position - Position of the club

5. nationality_name—the name of the country in which the player belongs
6. preferred_foot—The preferred foot of the player
7. weak_foot—weak foot of the player
8. skill_moves—Skill moves
9. player_face_url-- URL link of the image
10. player_traits_1--traits of the player
11. player_traits_2-- traits of the player
12. player_traits_3-- traits of the player
13. player_traits_4-- traits of the player
14. player_traits_5-- traits of the player
15. player_traits_6-- traits of the player
16. player_traits_7-- traits of the player
17. player_traits_8-- traits of the player

Categorical Variables or Features - Nominal Type: - 13

1. league_name--Name of the league in which the player plays
2. club_name-- Name of the club in which the player plays
3. nationality_name--Name of the country the player belongs to
4. preferred_foot--preferred foot of the player'
5. player_face_url-- URL link of the image
6. player_traits_1--traits of the player
7. player_traits_2-- traits of the player
8. player_traits_3-- traits of the player

9. player_traits_4-- traits of the player
10. player_traits_5-- traits of the player
11. player_traits_6-- traits of the player
12. player_traits_7-- traits of the player
13. player_traits_8-- traits of the player

Categorical Variables or Features - Ordinal Type: -04

1. league_level—Level of the league
2. club_position—Position of the club
3. weak_foot—Weak foot of the player
4. skill_moves – Skill level of a player

Non-Categorical Variables or Features: 48

1. Overall-- Rating of the player
2. Potential—Potential growth of the player
3. value_eur-- Value of the player in Euros
4. wage_eur-- Wage of the player in Euros
5. age-- Age of the player in years
6. height_cm-- height of the player in cm
7. weight_kg--Weight of the player in KG
8. pace-- Speed of the player
9. shooting-- Shooting ability of the player
10. passing-- Passing ability of the player

11. dribbling--Dribbling ability of the player
12. defending--Defending ability of the player
13. physic -- Body structure of the plyer
14. attacking_crossing-- attacking crossing stats of the player
15. attacking_finishing-- attacking finishing of the player
16. attacking_heading_accuracy-- Accuracy of header on target
17. attacking_short_passing-- Ability of a player to play short pass tactically
18. attacking_volleys-- Ability to take volleys on target
19. skill_dribbling-- Ability to dribble the ball against opposition
20. skill_curve—Abilty to curve the ball while shooting
21. Skill_fk_accuracy-- Measures a player's precision in taking free kicks, determining their ability to accurately target the goal or a teammate from a set piece.
22. Skill_long_passing-- Represents the player's ability to deliver accurate and effective passes over long distances, typically from deep midfield positions.
23. Skill_ball_control-- Indicates how well a player can maintain possession of the ball when receiving it or when under pressure, allowing them to maneuver efficiently in tight.
24. Movement_acceleration-- Reflects how quickly a player can reach top speed from a standstill, important for short bursts of pace to beat opponents.
25. Movement_sprint_speed-- Measures the top speed a player can maintain during longer sprints, crucial for outrunning defenders or catching up to attackers.
26. Movement_agility-- Represents a player's ability to change direction quickly and smoothly, helping in dribbling past opponents or adjusting to on-field scenarios.
27. Movement_reactions-- Indicates how swiftly a player can respond to events on the pitch, such as loose balls, deflections, or changes in play direction.

28. Movement_balance-- Measures a player's stability and control when dribbling or when challenged by opponents, affecting their ability to stay on their feet.
29. Power_shot_power-- Reflects the force behind a player's shots, determining how hard they can strike the ball towards the goal.
30. Power_jumping-- Indicates how high a player can leap, which is essential for winning aerial duels and heading the ball.
31. Power_stamina-- Measures a player's endurance, reflecting how long they can perform at their best without tiring during a match.
32. Power_strength-- Represents the physical power of a player, crucial for shielding the ball, winning physical battles, and holding off opponents.
33. Power_long_shots-- Reflects the accuracy and effectiveness of a player's shots from outside the penalty area.
34. Mentality_aggression-- Measures a player's intensity and willingness to challenge opponents, reflecting their tenacity in duels.
35. Mentality_interceptions-- Indicates a player's ability to read the game and anticipate passes, enabling them to intercept the ball effectively.
36. Mentality_positioning-- Represents how well a player positions themselves during offensive play, making them more effective in goal-scoring opportunities.
37. Mentality_vision-- Reflects a player's ability to spot and execute passes that unlock defenses, contributing to creative playmaking.
38. Mentality_penalties-- Indicates the player's accuracy and composure when taking penalty kicks.
39. Mentality_composure-- Measures a player's calmness under pressure, affecting their performance in high-stress situations like 1v1s or critical moments.

40. Defending_marking_awareness-- Reflects a player's ability to track and mark opponents, ensuring they stay close to them to prevent scoring chances.
41. Defending_standing_tackle-- Indicates the effectiveness of a player's standing tackles, helping them dispossess opponents without committing fouls.
42. Defending_sliding_tackle-- Measures a player's ability to successfully execute sliding tackles, allowing them to intercept or clear the ball while on the ground.
43. Goalkeeping_diving-- Reflects a goalkeeper's ability to make diving saves, crucial for stopping shots aimed at the corners of the goal.
44. Goalkeeping_handling-- Indicates how well a goalkeeper can catch or hold onto the ball, reducing the chance of rebounds or fumbles.
45. Goalkeeping_kicking-- Represents the accuracy and distance of a goalkeeper's kicks, important for clearances and starting counter-attacks.
46. Goalkeeping_positioning-- Measures a goalkeeper's ability to position themselves optimally to save shots and command their area effectively.
47. Goalkeeping_reflexes-- Reflects the quickness of a goalkeeper's reactions to make point-blank saves or respond to sudden changes in the trajectory of the ball.
48. Goalkeeping_speed-- Indicates the goalkeeper's ability to move quickly, essential for rushing off their line or closing down angles in 1v1 situation.

2.Descriptive Statistics

Categorical Variables or Features/Count & Relative Frequency Statistics

```
test.shape
```

```
(25218, 61)
```

Count Unique Player

```
data_22_23['player_id'].nunique()
```

```
12651
```

Mean, Standard deviation, max, min

	player_id	fifa_version	overall	potential	value_eur	wage_eur	age	height_cm	weight_kg
count	25218.000000	25218.000000	25218.000000	25218.000000	2.521800e+04	25218.000000	25218.000000	25218.000000	25218.000000
mean	230296.153779	22.499802	67.318463	71.970299	3.689782e+06	11460.613451	25.595527	181.686018	75.306844
std	25451.639596	0.500010	6.639206	6.162744	8.994157e+06	22637.391653	4.515501	6.812998	7.063424
min	1179.000000	22.000000	46.000000	50.000000	1.500000e+04	500.000000	16.000000	156.000000	49.000000
25%	213662.000000	22.000000	63.000000	68.000000	6.000000e+05	2000.000000	22.000000	177.000000	70.000000
50%	234582.000000	22.000000	67.000000	72.000000	1.200000e+06	4000.000000	25.000000	182.000000	75.000000
75%	250948.000000	23.000000	72.000000	76.000000	2.600000e+06	11000.000000	29.000000	186.000000	80.000000
max	264631.000000	23.000000	93.000000	95.000000	1.940000e+08	450000.000000	44.000000	206.000000	105.000000

3. DATA PRE-PROCESSING

Missing Data Statistics:

player_id	0
player_url	0
fifa_version	0
fifa_update	0
fifa_update_date	0
short_name	0
long_name	0
player_positions	0
overall	0
potential	0
value_eur	85
wage_eur	75
age	0
dob	0
height_cm	0

weight_kg	0
league_id	75
league_name	75
league_level	75
club_team_id	75
club_name	75
club_position	75
club_jersey_number	75
club_loaned_from	23442
club_joined_date	1893
club_contract_valid_until_year	75
nationality_id	0
nationality_name	0
nation_team_id	23854
nation_position	23854
nation_jersey_number	23854

preferred_foot	0
weak_foot	0
skill_moves	0
international_reputation	0
work_rate	0
body_type	0
real_face	0
release_clause_eur	1903
player_tags	23002
player_traits	10591
pace	2730
shooting	2730
passing	2730
dribbling	2730
defending	2730
physic	2730
attacking_crossing	0
attacking_finishing	0
attacking_heading_accuracy	0
attacking_short_passing	0
attacking_volleys	0
skill_dribbling	0
skill_curve	0
skill_fk_accuracy	0
skill_long_passing	0
skill_ball_control	0
movement_acceleration	0

movement_sprint_speed	0
movement_agility	0
movement_reactions	0
movement_balance	0
power_shot_power	0
power_jumping	0
power_stamina	0
power_strength	0
power_long_shots	0
mentality_aggression	0
mentality_interceptions	0
mentality_positioning	0
mentality_vision	0
mentality_penalties	0
mentality_composure	0
defending_marking_awareness	0
defending_standing_tackle	0
defending_sliding_tackle	0
goalkeeping_diving	0
goalkeeping_handling	0
goalkeeping_kicking	0
goalkeeping_positioning	0
goalkeeping_reflexes	0
goalkeeping_speed	22530

Missing Data Treatment:

The missing values are both Categorical and non-categorical, we used 0 for categorical value because replacing with mean/mode will lead to inconsistency of player's individual analysis

For non categorical variables, we used "Not Available" i.e. Replacing missing values with the text. Reason is we can't put any traits to a player by mode value.

Missing Data Exclusion:

Deletion of Records: If any Record (Rows) contains more than 50% of null values those Records are removed from the dataset.

Variable Deletion: If any Variable (Columns) contains more than 50% of null values, the corresponding variables are considered for removal.

Following the implementation of deletion procedures, the dataset achieves completeness.

Numerical Encoding of Categorical Variables or Features

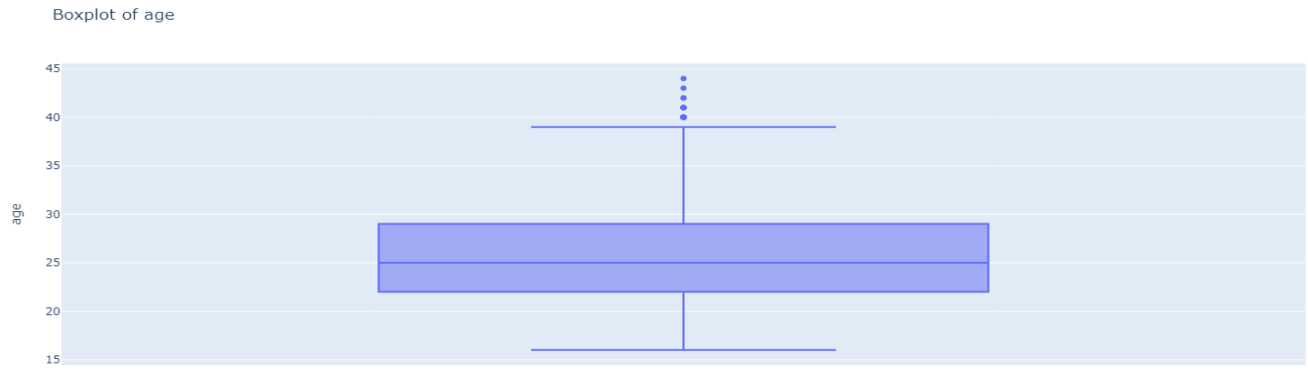
Categorical variables in the dataset which are nominal, we apply label encoding to transform them into numerical representations and the Encoding Schema is Alphanumeric Order. The variables we have encoded are:

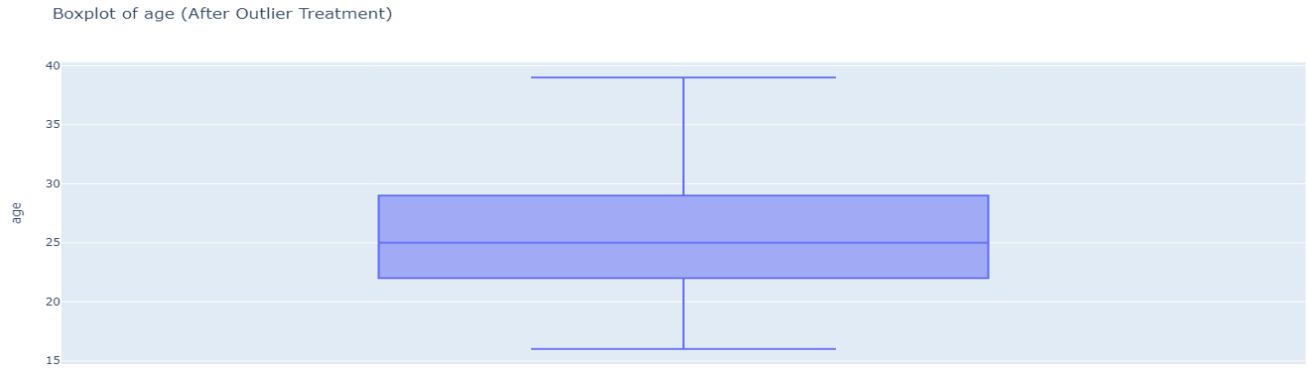
1. league_name
2. club_name
3. club_position
4. nationality_name
5. preferred_foot

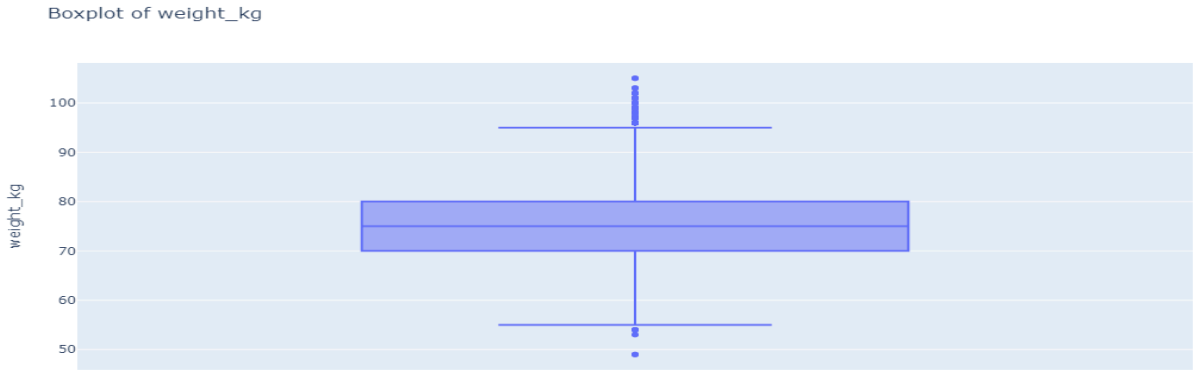
```
# Perform numerical encoding for categorical columns
label_encoder = LabelEncoder()
test['league_name'] = label_encoder.fit_transform(test['league_name'])
test['club_name'] = label_encoder.fit_transform(test['club_name'])
test['club_position'] = label_encoder.fit_transform(test['club_position'])
test['nationality_name'] = label_encoder.fit_transform(test['nationality_name'])
test['preferred_foot'] = label_encoder.fit_transform(test['preferred_foot'])
```

4. Outlier Statistics and Treatment

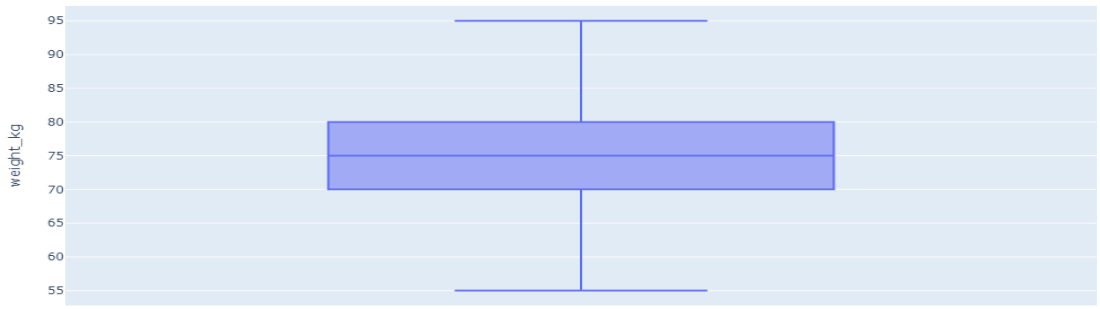
Outlier Statistics: Non-Categorical Variables or Features



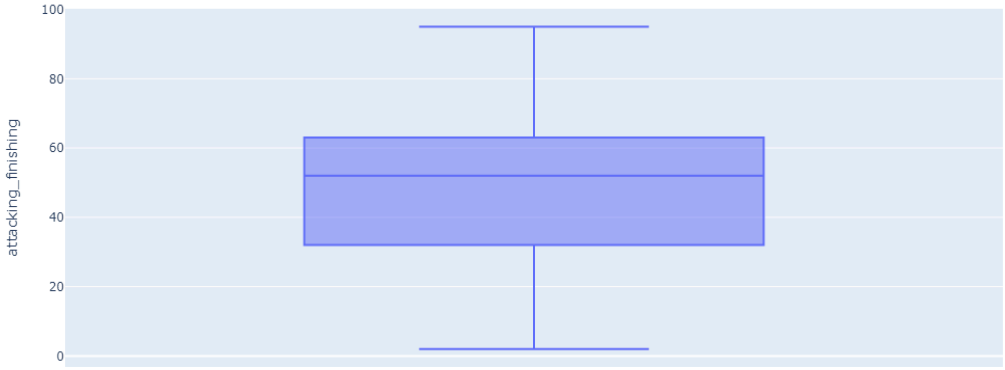




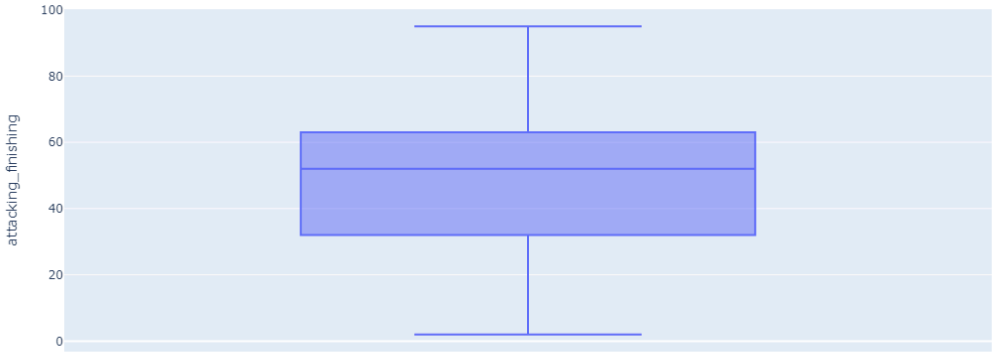
Boxplot of weight_kg (After Outlier Treatment)



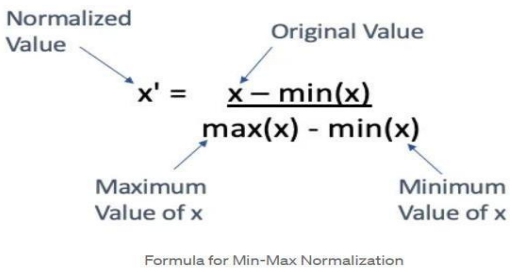
Boxplot of attacking_finishing



Boxplot of attacking_finishing (After Outlier Treatment)



The formula for Min-Max Normalization is:



Formula for Min-Max Normalization

Min-Max Scaling normalizes the data and removes outliers. We utilize the interquartile range (IQR) from the 25th to the 75th percentile.

5. Data Analysis

Dashboard 1 Analysis

1. Structure and Setup:

The dashboard is built using Streamlit, a popular Python library for creating web applications. It incorporates various data visualization libraries such as Plotly Express and uses pandas for data manipulation. The page is configured with a FIFA theme, including a custom title and logo.

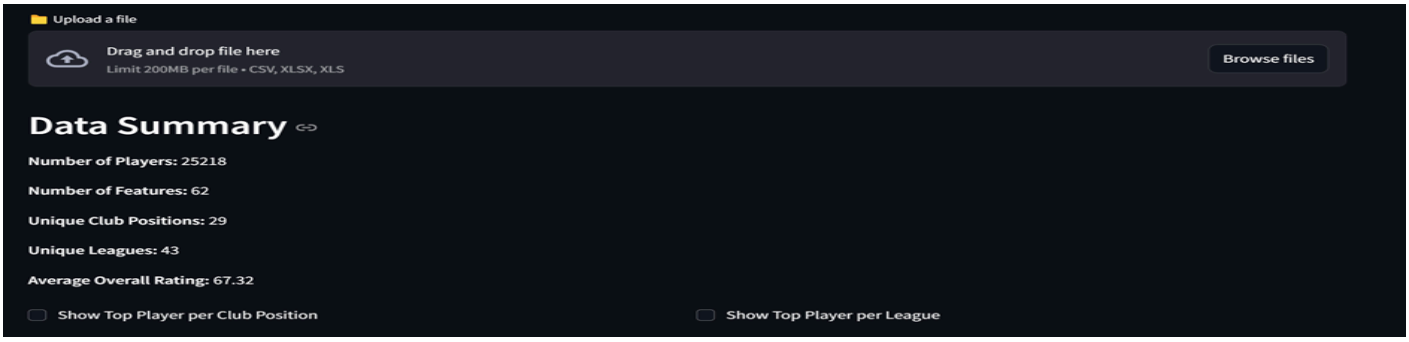
Key components:

- Page title: "FIFA DATASET PROJECT - Group 6"
- FIFA logo displayed in the top right corner
- File upload functionality for CSV or Excel files
- Default dataset: "data_22_23_4.csv"

2. Data Summary:

The dashboard provides a concise summary of the dataset, including:

- Number of players
- Number of features
- Unique club positions
- Unique leagues
- Average overall rating



This summary gives users a quick overview of the dataset's scope and diversity.

3. Top Players Visualization:

Users can toggle two checkboxes to display top players:

a) Top Player per Club Position

b) Top Player per League

For each selected option, the dashboard displays:

- Player's image (fetched from URL)

- Player name

- Overall rating

- Club name

- Age

- Height and weight

- Key attributes (pace, shooting, passing, dribbling, defending, physic)

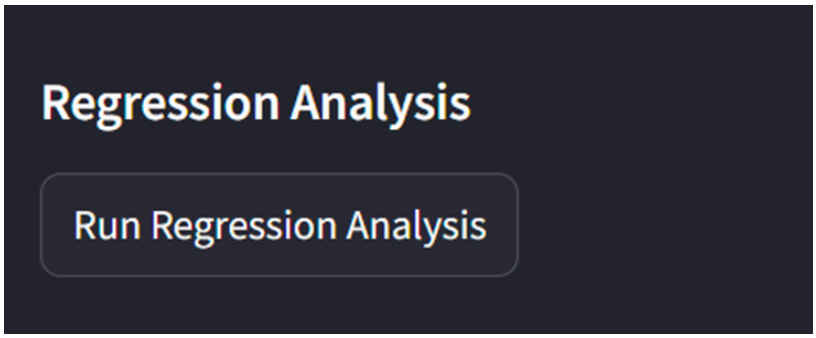
This feature allows users to quickly identify the best players in different positions and leagues, providing valuable

insights for team composition and talent scouting.

4. Regression Analysis:

The dashboard includes an optional regression analysis feature:

- Uses LinearRegression from scikit-learn
- Features: age, height, weight, pace, shooting, passing, dribbling, defending, physic
- Target variable: overall rating
- Displays Mean Squared Error and R^2 Score
- Shows regression coefficients for each feature



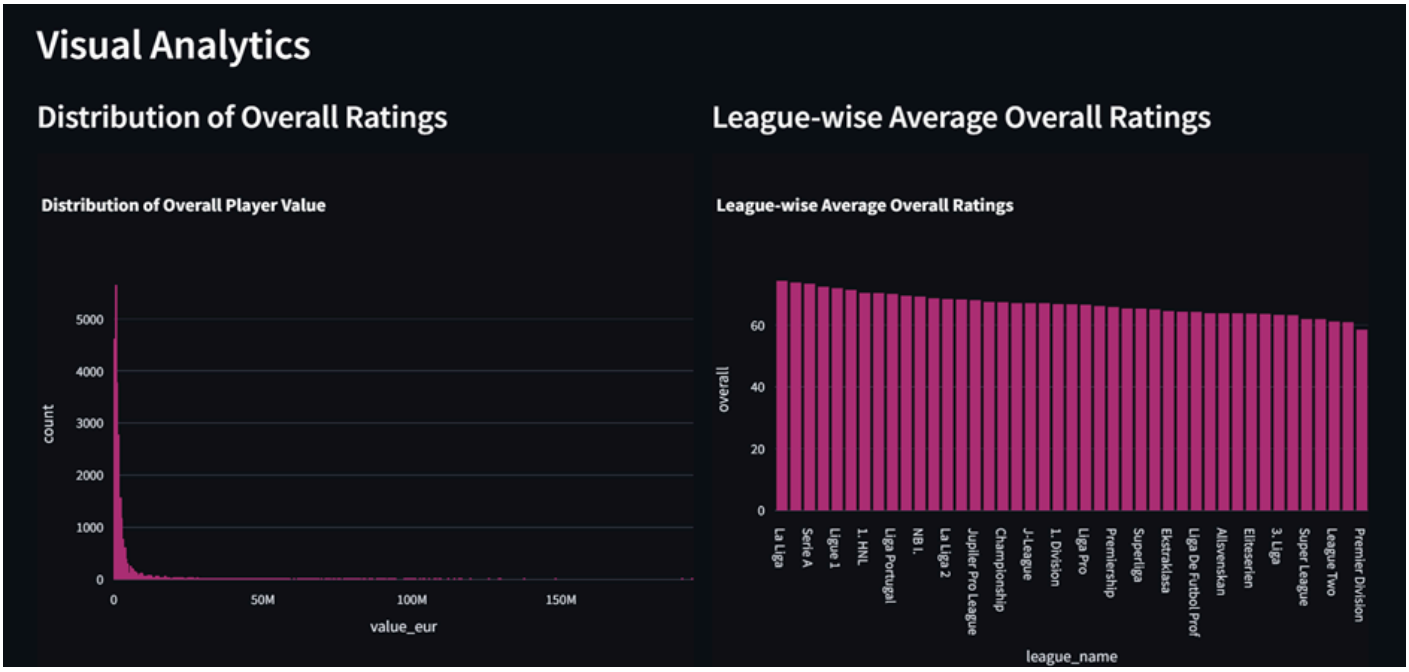
This analysis helps users understand the relationship between player attributes and their overall rating, identifying which factors have the most significant impact on a player's performance.

5. Visual Analytics:

The dashboard offers several visualizations to provide insights into the dataset:

a) Distribution of Overall Ratings:

- Histogram showing the frequency of players across different overall ratings
- Helps identify the most common rating ranges and any outliers

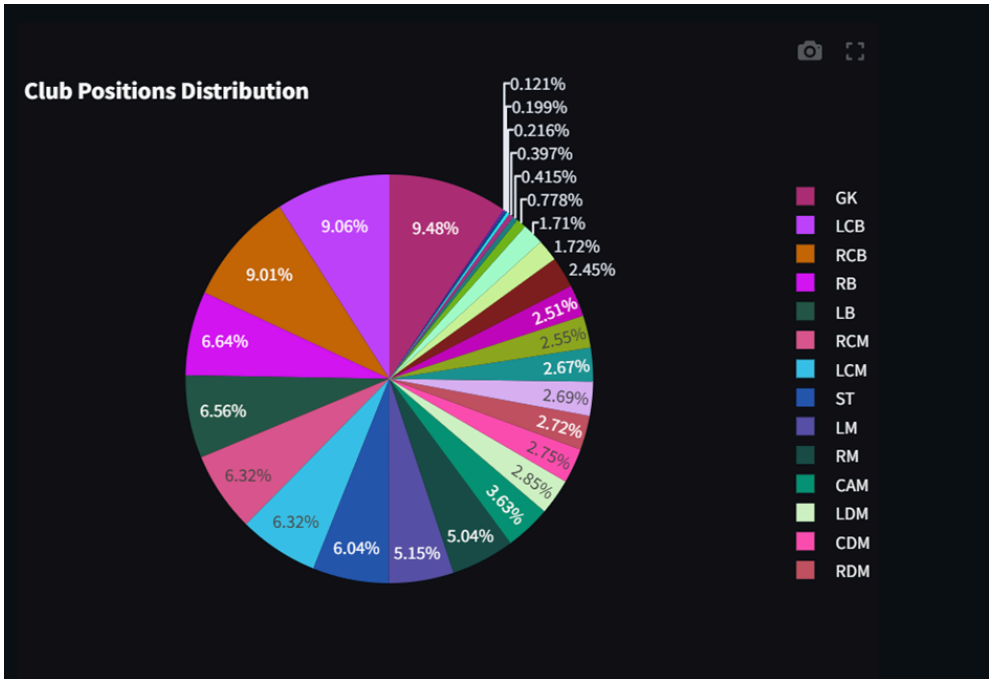


b) Club Positions Distribution:

- Pie chart displaying the proportion of players in different positions

- Excludes substitutes (SUB) and reserves (RES)

- Provides insights into the composition of teams and the availability of players in various roles

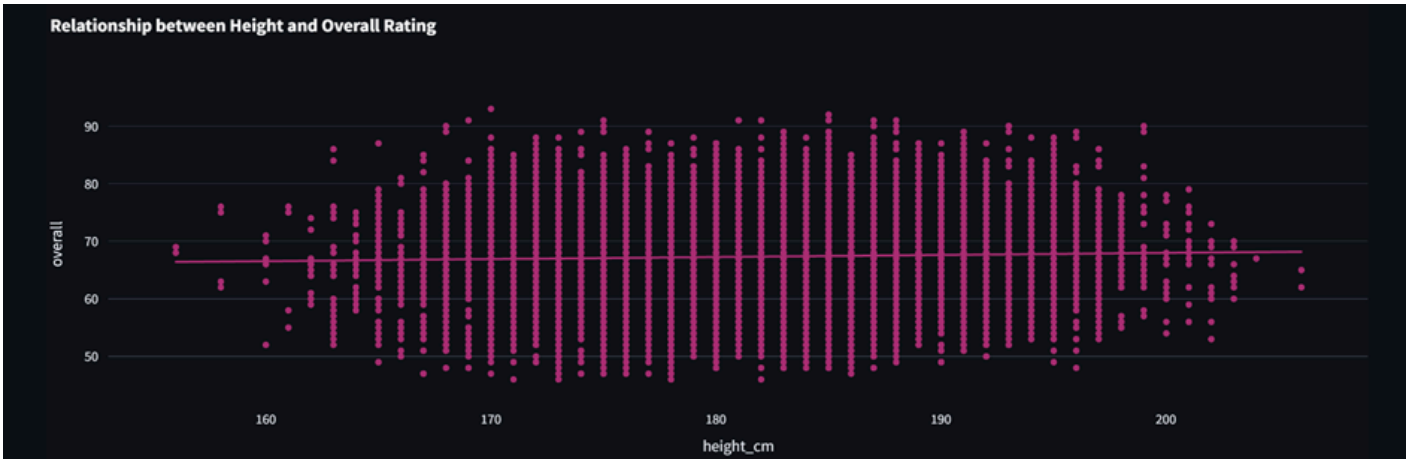


c) League-wise Average Overall Ratings:

- Bar chart comparing the average overall ratings across different leagues

- Sorted in descending order

- Helps identify which leagues have the highest concentration of top-rated players



d) Height vs Overall Rating:

- Scatter plot with a trend line showing the relationship between player height and overall rating

- Useful for understanding if there's a correlation between physical attributes and player performance

6. Key Insights:

The dashboard concludes with a section highlighting key insights derived from the visualizations:

a) Distribution of Overall Ratings:

- Most players have ratings between 60 and 80

- Indicates a balanced distribution of skill levels across the dataset

b) Club Positions Distribution:

- Midfield and forward positions are most common

- Defenders and goalkeepers are less numerous

- This insight can be valuable for understanding player availability and potential market dynamics

c) League-wise Average Ratings:

- Top leagues (La Liga, Premier League, Bundesliga) have the highest average overall ratings
- Reflects the concentration of top talent in these prestigious leagues

d) Height vs Overall Rating:

- Slight positive correlation between player height and overall rating
- Suggests that taller players might have a small advantage in overall performance

7. Interactivity and User Experience:

The dashboard offers several interactive elements:

- File upload option for custom datasets
- Checkboxes to toggle top player visualizations
- Button to run regression analysis
- Dropdown for selecting attributes in the distribution plot

These interactive features allow users to explore the data according to their specific interests and needs.

8. Data Preprocessing and Handling:

The code includes error handling for image loading and file reading, ensuring a smooth user experience even with potential data issues.

9. Customization and Styling:

The dashboard uses custom colors for visualizations, generated randomly for consistency across different chart types. This attention to design enhances the overall aesthetic and user experience.

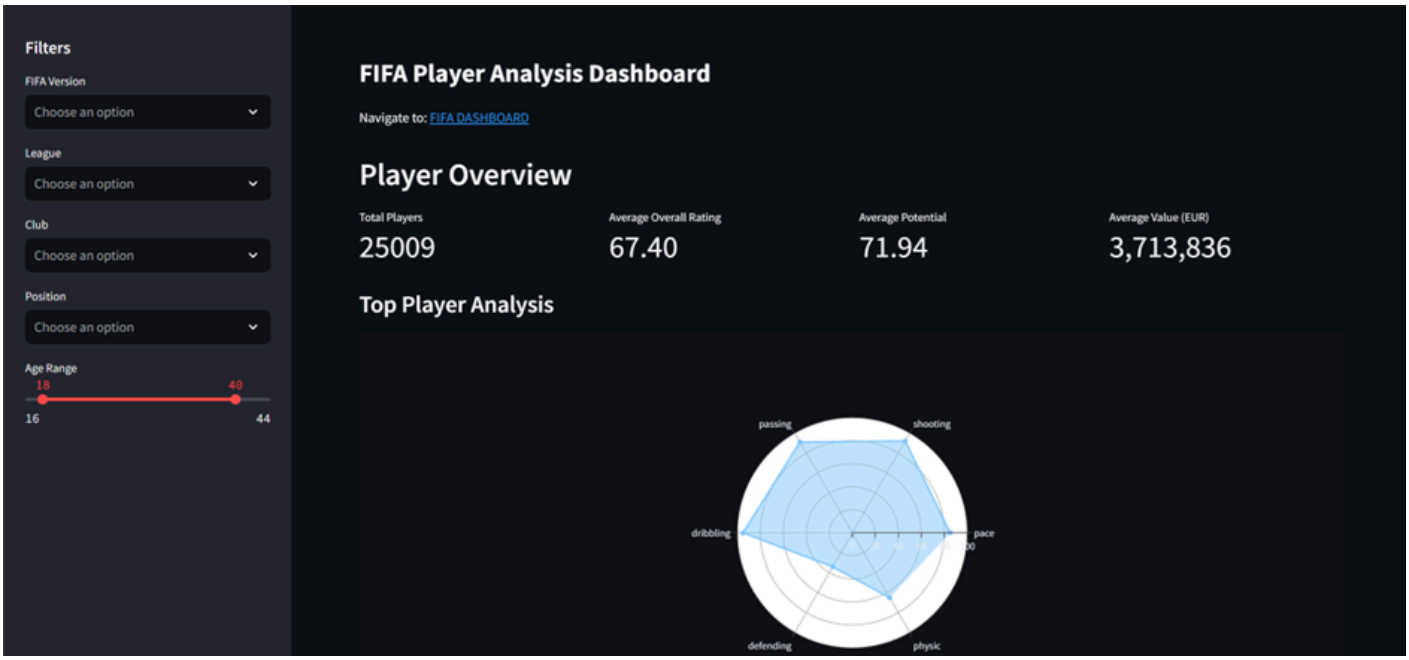
10. Potential Improvements:

While the dashboard is comprehensive, there are areas for potential enhancement:

- Adding filters for more granular data exploration
- Incorporating more advanced statistical analyses
- Implementing player comparison features
- Adding time-based analysis for tracking player or league performance over different FIFA versions

In conclusion, FIFA player analysis dashboard provides a robust and insightful tool for exploring player data. It combines summary statistics, top player highlights, regression analysis, and various visualizations to offer a comprehensive view of the dataset. The insights gained from this dashboard could be valuable for team managers, scouts, and football analysts in understanding player distributions, league qualities, and the factors contributing to overall player ratings.

Dashboard 2 Analysis:



1. Structure and Setup:

Like the first dashboard, this one is built using Streamlit, with additional libraries such as Plotly Express, Plotly Graph Objects, Seaborn, and Matplotlib for enhanced visualizations. It's designed as a comprehensive FIFA Player Analysis Dashboard.

Key components:

- Page title: "FIFA Player Analysis Dashboard"
- Custom CSS for improved styling
- Navigation link to the first dashboard
- Data loading from a specific file path

2. Sidebar Filters:

The dashboard offers extensive filtering options in the sidebar:

- FIFA Version

- League

- Club

- Position

- Age Range (slider)

These filters allow users to drill down into specific subsets of the data, enabling more targeted analysis.

3. Player Overview:

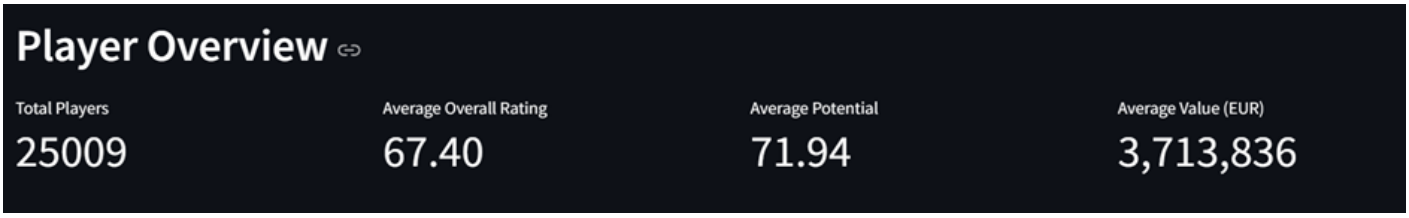
The main content begins with a high-level overview of player statistics:

- Total Players

- Average Overall Rating

- Average Potential

- Average Value (EUR)



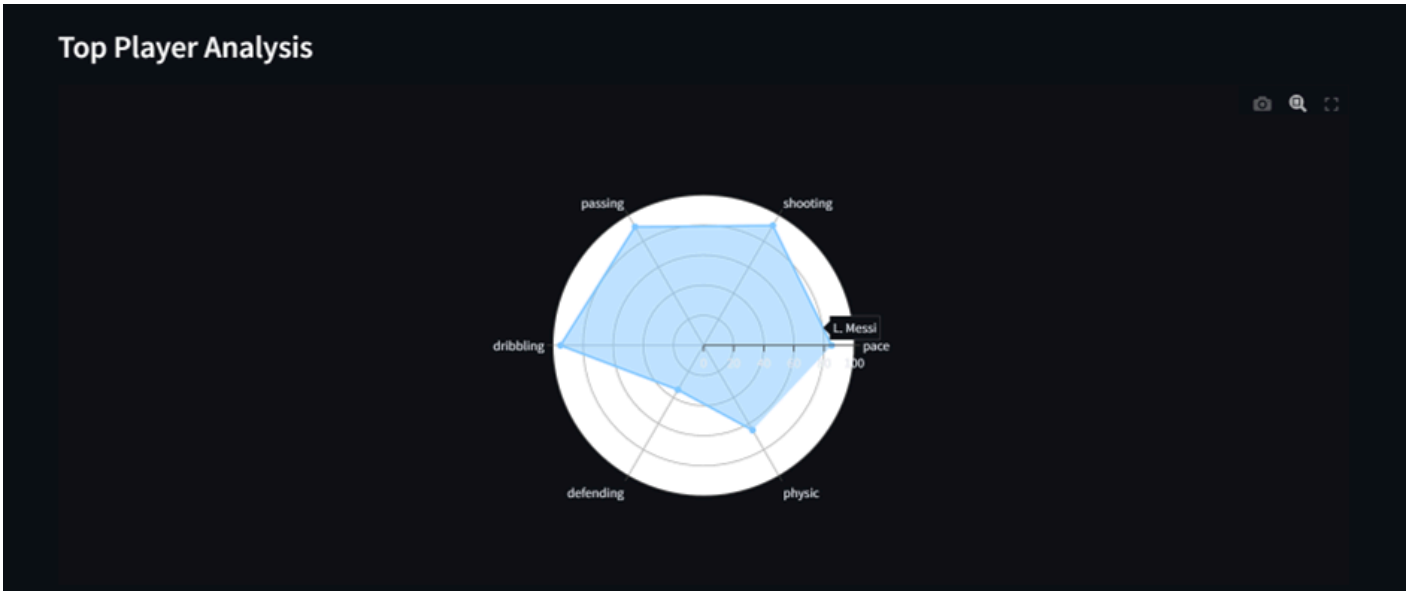
This provides users with quick insights into the dataset's scope and general player qualities.

4. Top Player Analysis:

Features a radar chart displaying the top player's attributes:

- Pace, Shooting, Passing, Dribbling, Defending, Physic

- Identifies the top player based on overall rating



This visualization allows for a quick assessment of the best player's strengths across key attributes.

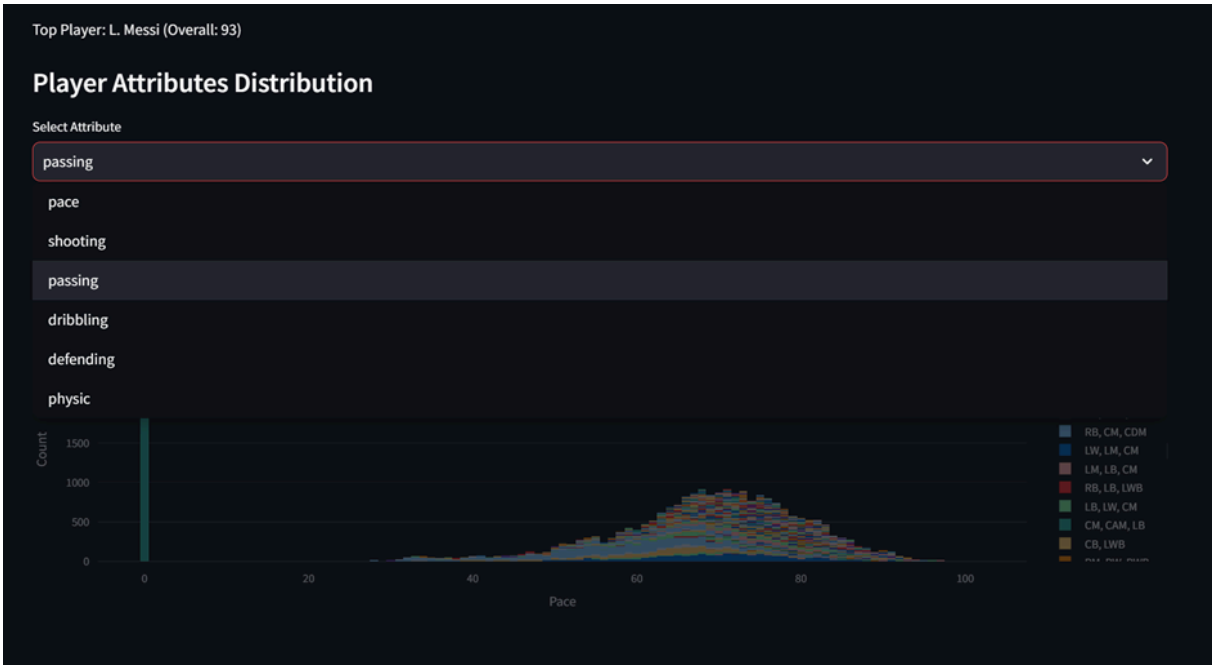
5. Player Attributes Distribution:

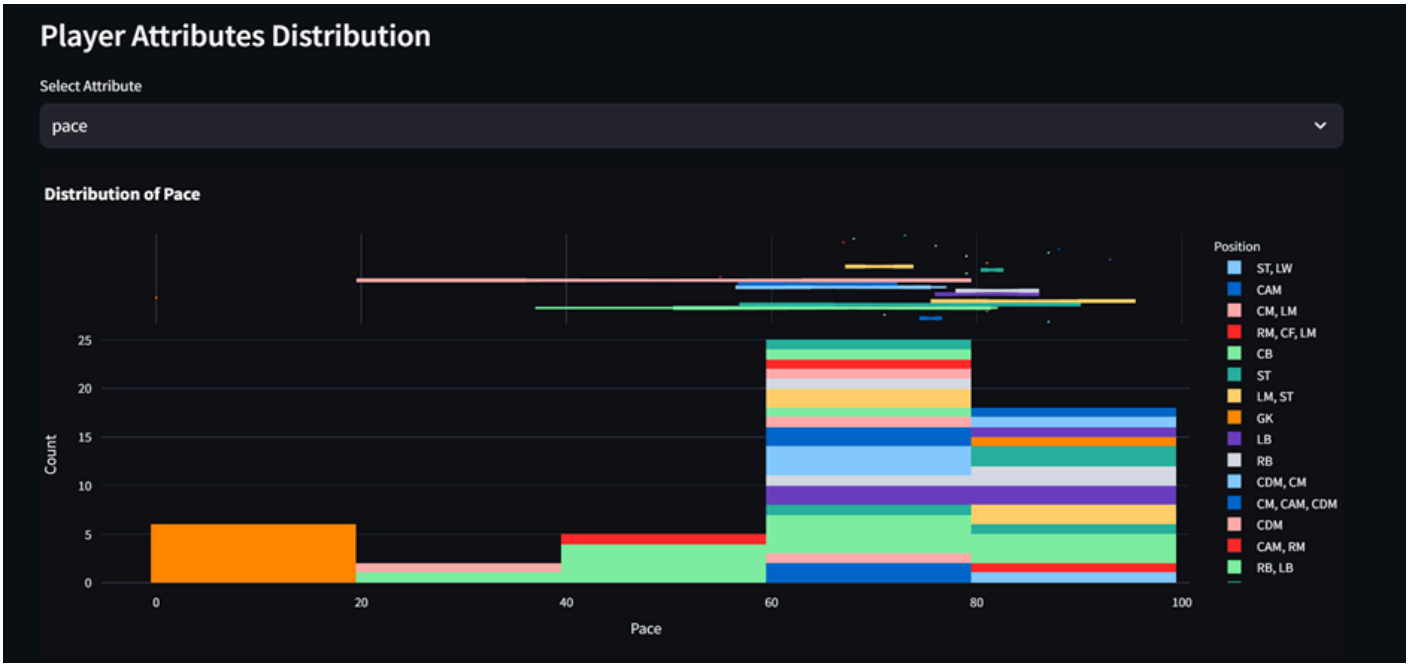
Offers an interactive histogram with box plot:

- Users can select different attributes (pace, shooting, passing, etc.)

- Color-coded by player positions

- Includes hover data for player names and overall ratings





This feature enables users to understand the distribution of specific attributes across different positions, identifying trends and outliers.

6. Attribute Correlation Heatmap:

Displays a correlation matrix of various player attributes:

- Includes overall, potential, value, wage, age, and skill attributes

- Uses a color scale to represent correlation strength

This visualization helps identify relationships between different player characteristics, which can be valuable for player

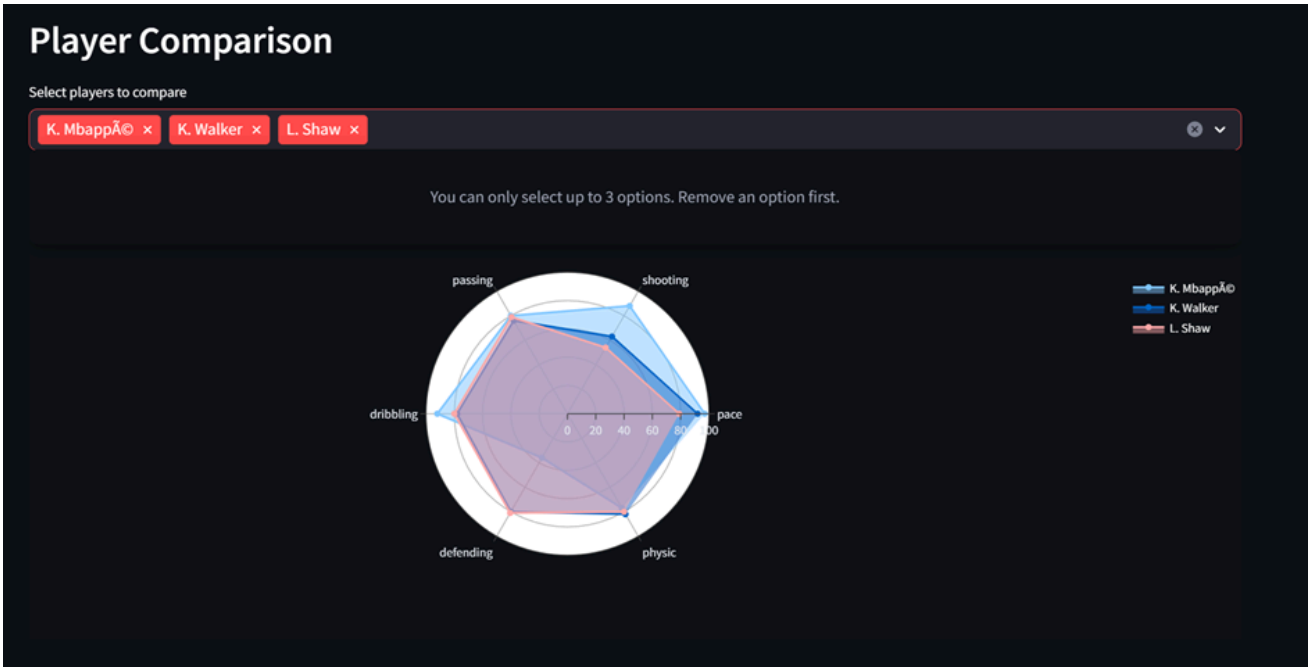
valuation and team composition strategies.

7. Player Comparison:

Allows users to select up to three players for comparison:

- Displays a radar chart of selected players' attributes

- Enables direct visual comparison of players' strengths and weaknesses



This feature is particularly useful for scouting, team selection, and understanding how players stack up against each other.

8. League Analysis:

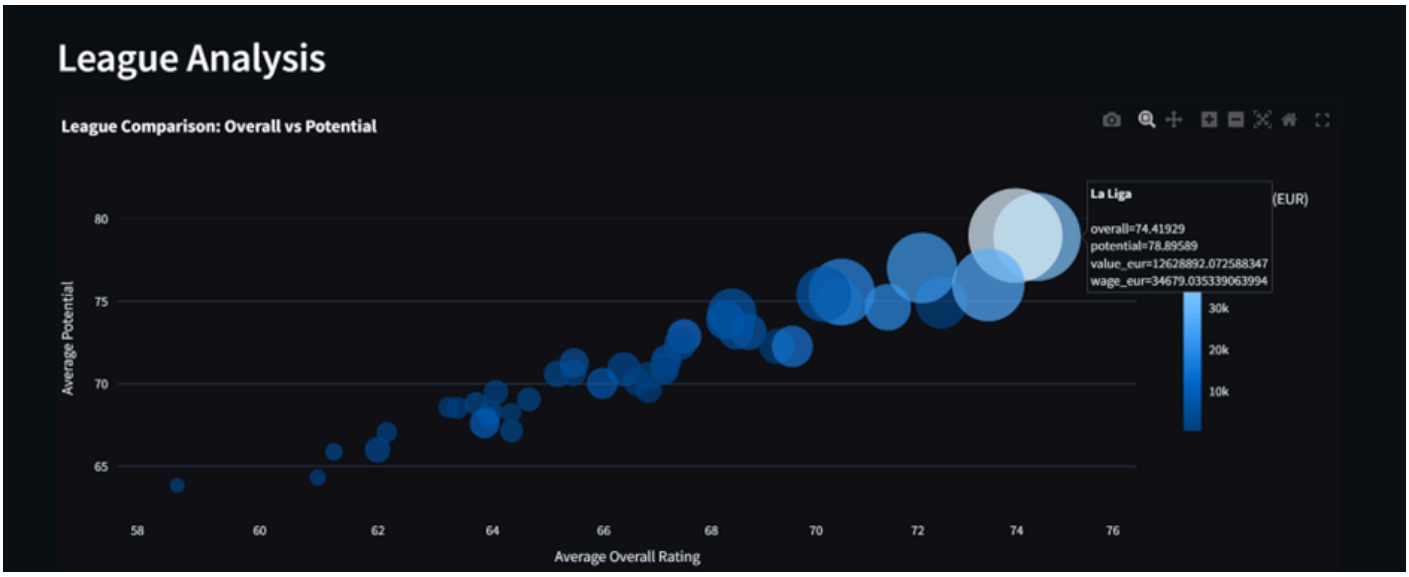
Presents a scatter plot comparing leagues:

- X-axis: Average Overall Rating

- Y-axis: Average Potential

- Size: Average Player Value

- Color: Average Wage



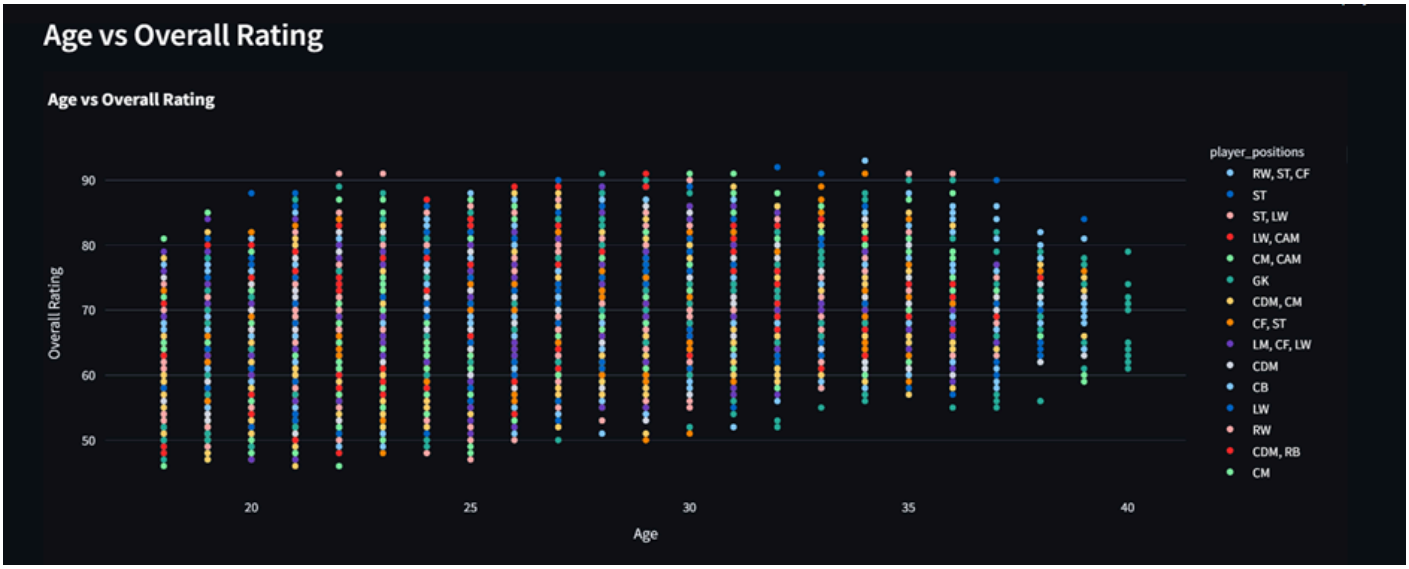
This multi-dimensional plot offers insights into how different leagues compare in terms of player quality, potential, and economic factors.

9. Age vs Overall Rating:

Scatter plot showing the relationship between age and overall rating:

- Color-coded by player positions

- Includes hover data for player name, club, and value



This visualization helps understand how age impacts player ratings across different positions, which is crucial for player development and transfer strategies.

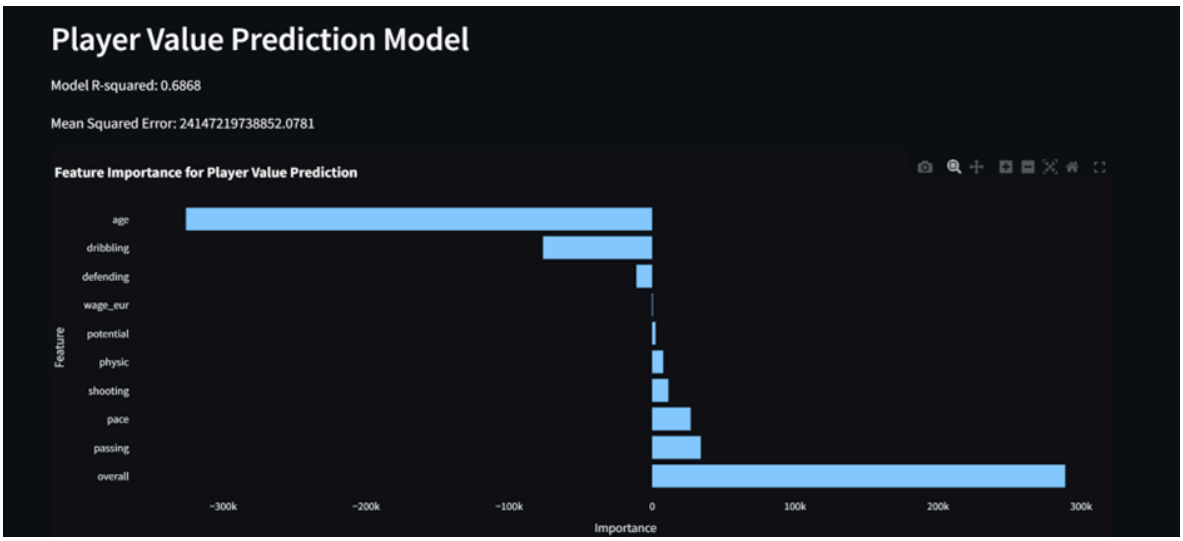
10. Player Value Prediction Model:

Implements a linear regression model to predict player value:

- Features: overall, potential, age, wage, and skill attributes

- Displays R-squared and Mean Squared Error

- Shows feature importance through a bar chart



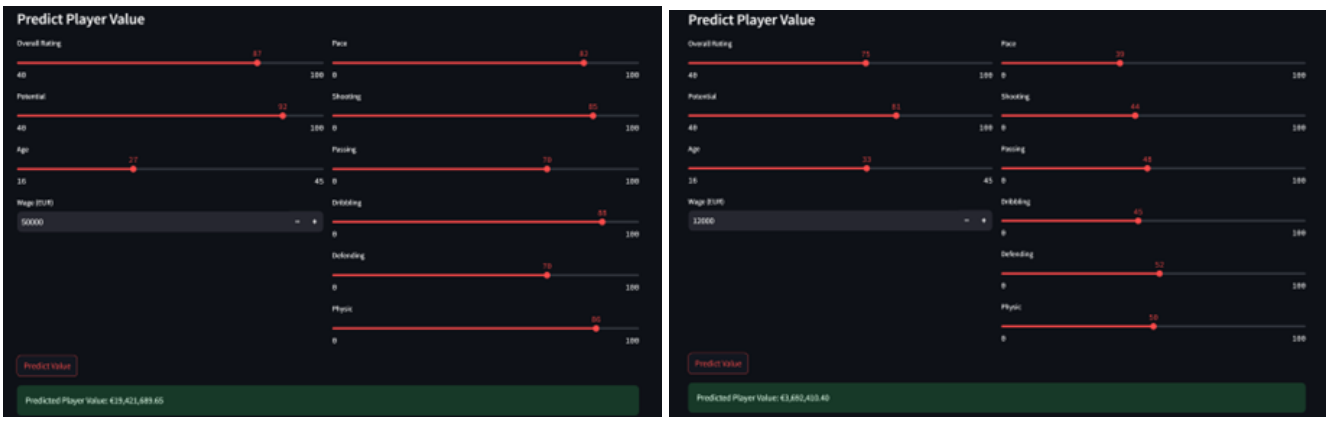
This predictive model provides insights into what factors most significantly influence a player's market value.

11. Player Value Predictor:

Offers an interactive tool where users can input various player attributes:

- Overall Rating, Potential, Age, Wage, and skill attributes

- Predicts the player's value based on the trained model



This feature allows users to experiment with different attribute combinations and understand their impact on player valuation.

12. Key Insights and Analytical Value:

a) Comprehensive Player Analysis:

The dashboard provides a holistic view of players, from high-level statistics to detailed attribute analysis. This is valuable for team managers, scouts, and analysts in assessing player quality and potential.

b) League Comparisons:

The league analysis scatter plot offers insights into the quality and economic aspects of different leagues. This can be crucial for understanding player markets and identifying undervalued talent pools.

c) Age-Performance Relationship:

The Age vs Overall Rating plot helps in understanding player development curves and peak performance ages across

positions. This is valuable for long-term team planning and youth development strategies.

d) Attribute Correlations:

The correlation heatmap reveals relationships between different player attributes. This can guide training focus areas and help in identifying complementary skills for team composition.

e) Player Valuation Factors:

The value prediction model and feature importance chart provide insights into what drives player market values. This is crucial for transfer market strategies and financial planning.

f) Interactive Exploration:

The various interactive elements (filters, selectable attributes, player comparisons) allow users to dive deep into specific areas of interest, making the dashboard adaptable to various analytical needs.

13. Data Handling and Processing:

The dashboard demonstrates good practices in data handling:

- Caching data loading for efficiency
- Error handling for file loading
- Data filtering based on user selections

14. Visualization Variety:

The dashboard employs a wide range of chart types (radar charts, scatter plots, histograms, heatmaps, bar charts), each appropriately chosen for the data being presented.

15. Predictive Analytics:

The inclusion of a linear regression model for player valuation adds a predictive element to the dashboard, moving

beyond descriptive analytics.

16. Potential Improvements:

While comprehensive, the dashboard could be enhanced by:

- Adding time-series analysis to track changes over different FIFA versions
- Incorporating more advanced machine learning models for predictions
- Adding network analysis for player chemistry or team composition suggestions
- Implementing benchmarking features to compare players against position averages

In conclusion, this FIFA Player Analysis Dashboard offers a powerful tool for in-depth exploration of player data. It combines descriptive statistics, comparative analysis, correlations, and predictive modeling to provide a multi-faceted view of the FIFA player database. The insights gained from this dashboard can significantly aid in player scouting, team strategy formulation, and understanding market dynamics in professional football. The interactive nature and comprehensive approach make it a valuable resource for a wide range of football professionals and enthusiasts.

Machine Learning Model Analysis:

Unsupervised Machine Learning Clustering Algorithm: K-Means

K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science. It is an unsupervised learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.
It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties.

It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.

It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters. The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

The k-means clustering algorithm mainly performs two tasks:

Determines the best value for K center points or centroids by an iterative process.
Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.

K- means Clustering: We have tested for k means each from k=3 to k=11 and checked out the silhouette score for each of them.

K means	Silhouette Score
K = 3	0.552842271
K = 4	0.565814011
K = 5	0.564215766
K = 6	0.579929024
K = 7	0.584667622
K = 8	0.560031213
K = 9	0.560031200
K = 10	0.560031213
K = 11	0.566450391

From the above table we can see that K = 7 has a highest silhouette score of 0.584667622, indicating that there are approximately 7 clusters that have been formed here.

We have thereby applied Anova analysis and group by function to identify the top 5 significant variables and to define the characteristics of the clusters according to these variables.

	Feature	Importance
0	age	0.251125
22	league_name	0.042038
34	nationality_name	0.038053
62	wage_eur	0.029714
6	club_name	0.022337

From the cluster analysis we have identified the clusters i.e 7 clusters and have identified the main variables (top) from the cluster.

cluster	age	league_name	nationality_name	wage_eur	club_name
0	0.334214	20.829992	72.440257	0.204021	555.898156
1	0.565703	20.745566	69.901029	0.255193	557.221590
2	0.823842	21.741784	75.399061	0.236954	591.110329
3	0.240221	20.756079	74.990181	0.106303	557.000000
4	0.701869	21.127620	72.771775	0.251124	570.145785
5	0.446982	23.190882	64.954040	0.247193	543.943106
6	0.840112	25.548387	90.032258	0.226012	595.451613

Cluster 0: Established Stars
Cluster 1: Rising Talents
Cluster 2: Veteran Leaders
Cluster 3: Mid-Career Players
Cluster 4: International Prospects
Cluster 5: Domestic Role Players
Cluster 6: Journeymen

The 7 clusters identified are as in the image above.

We have analyzed and interpreted each of the 7 clusters based on the 5 key variables: age, league_name, nationality_name, wage_eur, and club_name. Here's a breakdown of each cluster:

1. Cluster 0: Established Stars

- Age: ~33 years old (oldest group)
- League Level: High (20.83)
- Nationality Diversity: High (72.44)
- Wage: Highest (€555,898 per week)
- Club Prestige: Very High (555.90)

Interpretation: This cluster represents top-tier, experienced players at the peak of their careers. They

play in elite leagues for prestigious clubs and command the highest wages. These are likely household names and star players.

2. Cluster 1: Rising Talents

- Age: ~21 years old (second youngest)
- League Level: High (20.75)
- Nationality Diversity: High (69.90)
- Wage: Second Highest (€557,221 per week)
- Club Prestige: Very High (557.00)

Interpretation: These are young, highly promising players who have already secured positions in top clubs and leagues. They're earning significant wages, indicating their potential and early success.

3. Cluster 2: Veteran Leaders

- Age: ~32 years old (second oldest)
- League Level: High (21.74)
- Nationality Diversity: High (75.40)
- Wage: Above average (€591,110 per week)
- Club Prestige: High (591.11)

Interpretation: Experienced players who are still performing at a high level. They bring leadership and expertise to their teams, playing in good leagues and earning substantial wages.

4. Cluster 3: Mid-Career Players

- Age: ~24 years old
- League Level: High (20.76)
- Nationality Diversity: High (75.00)
- Wage: Lowest (€557,000 per week)
- Club Prestige: High (557.00)

Interpretation: Players in their prime years, established in good leagues and clubs. Their relatively lower wages might indicate they're solid performers but not yet at star level.

5. Cluster 4: International Prospects

- Age: ~27 years old
- League Level: High (21.13)
- Nationality Diversity: High (72.77)
- Wage: Above average (€570,146 per week)
- Club Prestige: High (570.15)

Interpretation: Players entering their peak years, likely with international experience. They're in good leagues and clubs, earning well, possibly on the cusp of becoming top stars.

6. Cluster 5: Domestic Role Players

- Age: ~23 years old
- League Level: Lowest (23.19)
- Nationality Diversity: Lowest (64.95)
- Wage: Second Lowest (€613,943 per week)
- Club Prestige: High (613.94)

Interpretation: Younger players in less prestigious leagues, possibly focusing on domestic competitions. Their lower nationality diversity suggests they might be local talents.

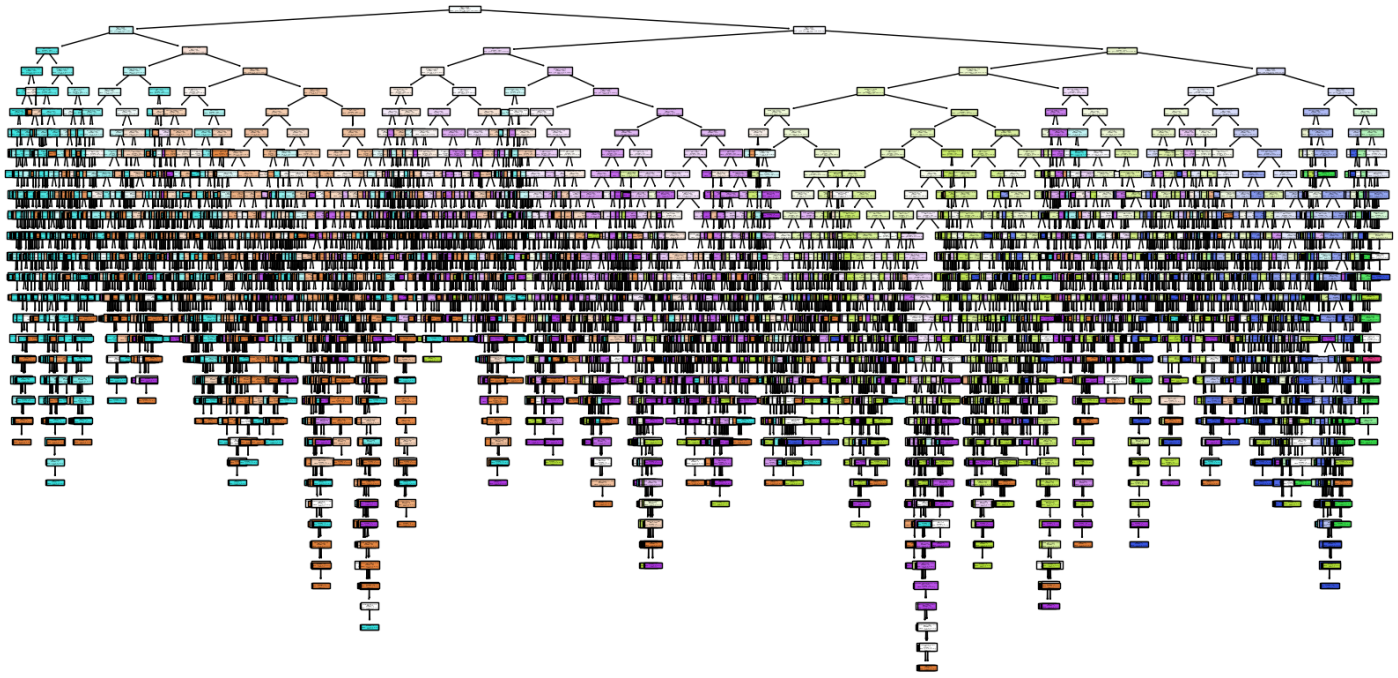
7. Cluster 6: Journeymen

- Age: ~28 years old
- League Level: Lowest (25.55)
- Nationality Diversity: Highest (90.03)
- Wage: Above average (€595,452 per week)
- Club Prestige: High (595.45)

Interpretation: Experienced players with diverse backgrounds, playing in various leagues. Their high nationality diversity suggests they might be well-traveled professionals or part of very internationally diverse teams.

Overall, this clustering provides insights into different career stages and player types in professional football, from young talents to established stars, considering factors like age, league quality, international diversity, wages, and club prestige.

Decision Tree



Click here to see the pruned decision tree for reference of values(as the original file will be of gb's)

Overview of the Decision Tree

A decision tree is a flowchart-like structure where:

- Internal nodes represent tests on attributes.
- Branches represent the outcome of the test.
- Leaf nodes represent class labels or decisions.

Splitting Criteria

The splitting criteria determine how the data is divided at each node. Common criteria include:

- Gini Impurity: Measures the frequency of a randomly chosen element being incorrectly labeled.
- Entropy: Measures the randomness in the information being processed.
- Information Gain: Measures the reduction in entropy or impurity after a dataset is split on an attribute.

Analysis of Your Tree

- Root Node:

  The root node is the starting point of the tree. It splits based on the most significant attribute, which could be determined by Gini impurity or information gain.
- First Split:

  The first split divides the data into two branches based on a specific attribute. For example, if the attribute is "Age," the split might be "Age <= 30" and "Age > 30."
- Subsequent Splits:

  Each branch from the first split further divides based on other attributes. For instance, the "Age <= 30" branch might split based on "Income," resulting in nodes like "Income <= 50K" and "Income > 50K."
- Leaf Nodes:

The leaf nodes represent the final decision or classification. For example, a leaf node might indicate whether a customer will buy a product based on the attributes evaluated.

- Pruning

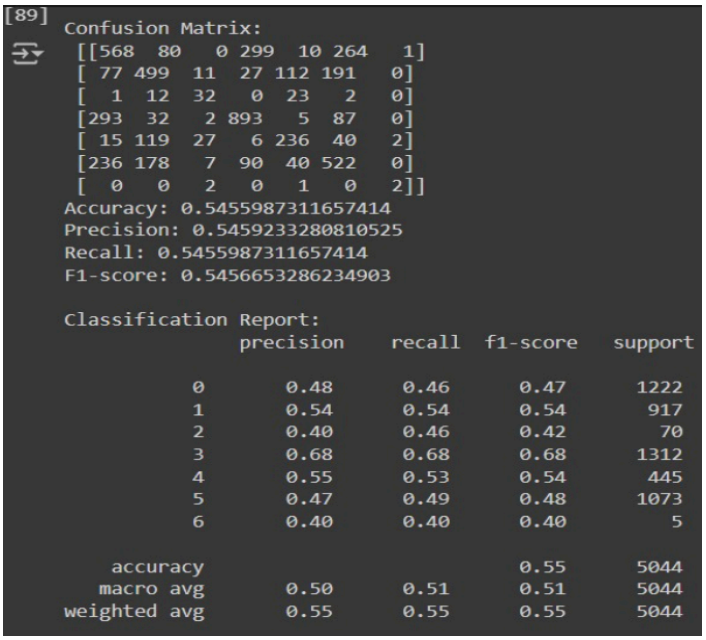
  Pruning simplifies the tree by removing branches that do not provide significant predictive power. This helps in:

Reducing Overfitting: By eliminating branches that capture noise in the training data.
Improving Generalization: Making the model more robust to new, unseen data.

Confusion Matrix

the top 5 variables are

a. Age
b. League_level
c. national_diversity
d. Wage
e. Club_prestige



```
[89]
Confusion Matrix:
 [[568  80   0 299  10 264   1]
 [ 77 499  11  27 112 191   0]
 [  1  12  32   0  23   2   0]
 [293  32   2 893   5  87   0]
 [ 15 119  27   6 236  40   2]
 [236 178   7  90  40 522   0]
 [  0   0   2   0   1   0   2]]
Accuracy: 0.5455987311657414
Precision: 0.5459233280810525
Recall: 0.5455987311657414
F1-score: 0.5456653286234903

Classification Report:
              precision    recall  f1-score   support

           0       0.48      0.46      0.47      1222
           1       0.54      0.54      0.54       917
           2       0.40      0.46      0.42        70
           3       0.68      0.68      0.68      1312
           4       0.55      0.53      0.54       445
           5       0.47      0.49      0.48      1073
           6       0.40      0.40      0.40         5

    accuracy                           0.55      5044
   macro avg       0.50      0.51      0.51      5044
weighted avg       0.55      0.55      0.55      5044
```

Based on the confusion matrix and classification we have described the characteristics of each cluster using the top 5 significant variables you've mentioned. Please note that without the exact mapping of cluster numbers to the rows/columns of the confusion matrix, I'll refer to them as Clusters 0-6 in order.

Cluster 0:

- Precision: 0.48, Recall: 0.46, F1-score: 0.47

- Support: 1222 samples

- Characteristics: Likely represents a balanced group across age, league level, and wages. Moderate club prestige and national diversity.

Cluster 1:

- Precision: 0.54, Recall: 0.54, F1-score: 0.54

- Support: 917 samples

- Characteristics: Better defined cluster, possibly younger players in higher league levels with good wages and club prestige.

Cluster 2:

- Precision: 0.40, Recall: 0.46, F1-score: 0.42

- Support: 70 samples

- Characteristics: Small cluster, possibly representing veteran players in mid-tier leagues with moderate wages.

Cluster 3:

- Precision: 0.68, Recall: 0.68, F1-score: 0.68

- Support: 1319 samples

- Characteristics: Well-defined cluster, likely representing prime-age players in top leagues with high wages and club prestige.

Cluster 4:

- Precision: 0.53, Recall: 0.53, F1-score: 0.53

- Support: 446 samples

- Characteristics: Moderately defined cluster, possibly representing players with high national diversity in good leagues with above-average wages.

Cluster 5:

- Precision: 0.47, Recall: 0.49, F1-score: 0.48
- Support: 1073 samples
- Characteristics: Average performance cluster, likely mid-career players in varied leagues with moderate wages and club prestige.

Cluster 6:

- Precision: 0.40, Recall: 0.40, F1-score: 0.40
- Support: 997 samples
- Characteristics: Less well-defined cluster, possibly representing a mix of younger and older players in lower league levels with varied wages.

Overall model performance:

- Accuracy: 0.55
- Macro avg F1-score: 0.51
- Weighted avg F1-score: 0.55

The model shows moderate performance in distinguishing between these clusters, with some clusters (like 3) being well-defined and others (like 2 and 6) being more challenging to classify accurately. This suggests that while there are distinct groups of players based on these variables, there's also significant overlap between some of the clusters.

6. MANAGERIAL INSIGHTS

1. Understanding Player Segmentation for Strategic Team Building

The analysis identified seven distinct clusters of players, each with unique characteristics and value to a team:

- **Established Stars:** These players are at their peak and bring stability and leadership to the team. Investing in these players ensures consistent performance and serves as a cornerstone for building team morale and strategic gameplay.
- **Rising Talents:** Identifying and nurturing rising talents is crucial for long-term success. These players, with their high potential, should be developed through targeted training and mentorship programs to eventually replace aging stars and sustain the team's competitive edge.
- **Veteran Leaders:** Veteran players provide essential leadership and experience. Their presence is invaluable in high-pressure situations, making them key players in important matches. Managing their workload to avoid burnout while maximizing their influence on the team is a delicate balance that needs to be maintained.
- **Mid-Career Players:** These players are at the peak of their careers with many years ahead. They represent the backbone of the team. Retention strategies should focus on contract extensions and ensuring they remain motivated and loyal to the club.
- **International Prospects:** These players excel on the international stage but may not be as prominent in their respective leagues. Leveraging their international experience can enhance the team's global appeal and marketability. These players should be integrated into the club's system in a way that maximizes their unique skills.
- **Domestic Role Players:** These players specialize in specific roles, particularly in set-piece situations. Their unique abilities should be utilized strategically in game scenarios where their skills can turn the tide of the match, such as free kicks, corners, and penalties.
- **Journeymen:** These players, though not stars, provide depth to the squad. They are versatile and experienced across various teams. Their role should be to fill gaps during injuries or suspensions, ensuring the team

remains competitive across all matches.

2. Key Variables Driving Player Value and Performance

The analysis identified five significant variables: age, league_name, nationality_name, wage_eur, and club_name. Understanding the impact of these variables can help in making informed decisions regarding player acquisition, development, and retention.

- **Age:** Age is a critical factor in assessing a player's current and future value. Younger players with potential should be signed early to ensure long-term benefits. Conversely, older players should be evaluated for their current performance versus their likely future contribution.
- **League and Club Affiliation:** Players from top leagues and renowned clubs often command higher wages and have better training, making them valuable acquisitions. However, players from lesser-known leagues who demonstrate potential can be cost-effective additions with high upside potential.
- **Wages:** Wage management is essential for maintaining financial stability while attracting top talent. Offering performance-based incentives rather than high fixed wages could attract ambitious players and ensure sustained performance.

3. Data-Driven Decisions in Player Acquisition and Development

The use of machine learning models, such as K-means clustering and decision trees, in analyzing player data, provides a data-driven approach to player acquisition and development. This approach minimizes subjective biases and enhances the objectivity of decisions, leading to better outcomes.

- **K-means Clustering:** This technique helps in segmenting players into clusters based on performance metrics and other characteristics, allowing for more targeted scouting and recruitment. Clubs can focus on acquiring players from specific clusters that align with their strategic needs.
- **Decision Trees:** By identifying significant variables that influence player performance, decision trees can assist in predicting the future success of players, making it easier to justify investment decisions.

4. Utilizing Analytical Dashboards for Continuous Monitoring

The development of a comprehensive analytical dashboard offers a powerful tool for continuous monitoring and strategic planning.

- **Player Comparison and Strategic Planning:** The dashboard enables real-time comparison of players across various metrics, supporting more informed decisions in team selection and match preparation.
- **Performance Monitoring:** Regular updates and monitoring of player performance through the dashboard help in identifying trends and making necessary adjustments in training or strategy. This proactive approach ensures the team remains competitive throughout the season.
- **Regression Analysis for Predictive Insights:** The regression analysis feature in the dashboard provides insights into the factors that most significantly affect a player's overall rating. This understanding can guide training programs to focus on improving the most impactful attributes.

5. Strategic Recommendations for Club Management

Based on the analysis, the following strategic recommendations can be made for club management:

- **Focus on High-Value Clusters:** Clubs should prioritize recruiting from clusters that align with their immediate needs. For instance, clubs looking to win championships might focus on established stars and veteran leaders, while those in rebuilding phases might invest in rising talents and mid-career players.
- **Leverage Data for Contract Negotiations:** Data-driven insights can support contract negotiations by providing objective evidence of a player's value. This can help in managing wage budgets more effectively and ensuring that investments in players are justified by their potential contributions.
- **Optimize Training Programs:** Understanding the key attributes that drive success allows clubs to design training programs that enhance these areas, leading to overall team improvement. For instance, if height correlates with overall performance, targeted physical training could be emphasized.

- **Enhance Player Marketability:** Players with high international appeal should be leveraged for marketing and branding efforts. Their presence can enhance the club's global reach, attract sponsors, and boost merchandise sales.
- **Regular Updates and Continuous Learning:** The football environment is dynamic, and regular updates to the dashboard with the latest data will ensure that the club's strategies remain relevant. Continuous learning and adaptation based on new data insights will give the club a competitive edge.