Illinois Institute of Technology

CSP 571 – Data Preparation and Analysis

Spring 2024

Player Performance Analysis

Sai Manohar Vemuri (A20514848)

Yasaswini Kakumani (A20547678)

Prof. Jawahar Panchal

Table of Contents

Abstract

Our project focuses on analyzing the prominent soccer leagues spanning the 2021-2022 season, namely the Premier League, Ligue 1, Bundesliga, Serie A, and La Liga. We're delving into a comprehensive dataset packed with various statistics. Our primary objective is to prep the data, addressing issues like missing values and outliers. Through exploratory data analysis, we aim to uncover meaningful patterns and insights. Additionally, we're also working towards developing a model that can predict future player positions based on statistical attributes. This predictive tool is geared towards aiding football clubs, coaches, and talent scouts in making informed decisions regarding player recruitment, squad composition, and strategic planning.

Problem Statement

In the fast-paced world of professional football, continual improvement is crucial at both individual and team levels. A major challenge lies in effectively harnessing the abundance of statistical data on players to enhance decision-making processes related to performance. This project squarely tackles this challenge by developing and implementing an analytical framework capable of parsing and interpreting extensive football player data. The framework's objectives include identifying key performance indicators, categorizing players by positions and continents, and extracting valuable insights through exploratory data analysis and visualization methods. Furthermore, predictive models will be employed to forecast player performance metrics and determine player positions, adding an extra layer of strategic foresight to the analysis.

Literature Review

[1] explores recent developments in applying machine learning to sports analytics, specifically focusing on football game. The study delves into two primary aspects: long-term team performance prediction and the factors influencing the match rating of central defenders. In team performance prediction, the research employs both universal classification models and an innovative simulation approach to forecast league standings. Results demonstrate competitive accuracy, surpassing or aligning with previous studies. The second experiment reveals insights into central defenders' match ratings, utilizing Multiple Linear Regression and identifying key features impacting defensive performance, including surprising contributions from attacking skills. The literature survey underscores the growing potential of machine learning in sports analytics and its applicability in refining predictions for team and player performance in football.

[2] This study presented a comprehensive unsupervised learning analysis of NBA player statistics, employing techniques such as K-means clustering, hierarchical clustering, and PCA dimensionality reduction. Notably, dimensionality reduction was effectively applied by selecting 2 principal components from the original 24, reducing the dimensionality by 39.14% while retaining 60.86% of the information. The use of hierarchical clustering and K-means provided valuable insights into player rankings based on playing time per game, revealing distinctive clusters that differentiate players with extensive playing time (more than 36 minutes) from those with minimal playing time (less than 3 minutes). This research contributes to the broader understanding of player performance patterns in the NBA, showcasing the versatility of unsupervised learning methods in analyzing complex sports datasets.

[3] The paper addresses challenges in evaluating football players for scouting and strategic planning, emphasizing the complexities arising from the vast pool of grassroots players and diverse play conditions. The proposed Player Performance Prediction system employs a datadriven approach, featuring a linear regression model with an 84.34% accuracy, followed by a secondary model predicting market values with 91% accuracy. Applicable across various positions, the system aims to identify grassroots talent, offering coaches and team managers a valuable tool for improved scouting and performance management. Future enhancements may involve real-time data collection through spider-cameras for a more nuanced analysis of physical attributes impacting player performance.

Dataset Description

The dataset, obtained from Kaggle, provides a detailed snapshot of football player statistics per 90 minutes for the 2022-2023 season across five major leagues: Premier League, Ligue 1, Bundesliga, Serie A, and La Liga. With 2,500 rows and 124 columns, it offers a wealth of player data covering various aspects, including personal details like name, age, and nationality, as well as in-game metrics such as goals, shots, passes, and tackles. Additionally, it includes advanced analytics like shot-creating actions, goal-creating actions, and different types of passes and touches. This comprehensive dataset serves as a valuable resource for conducting in-depth analysis of individual player performances within the realm of football.

Methodology

This Project initiates with collecting the comprehensive dataset from Kaggle, containing detailed overview of football player statistics. First, we pre-process the data to handle the missing values, outliers and inconsistencies using techniques like normalization, imputation and feature scaling to ensure the data quality. Next, we conduct the Exploratory Data Analysis (EDA) to gain insights into the distribution and characteristics of the data. Various visualization techniques such as histograms, scatter plots are presented to explore the relationship between variables, investigate key positional features, and understand the impact of age and matches played on player performance.

Additionally, player distribution across leagues and positions is analyzed to identify trends and patterns. Feature engineering techniques are then applied to engineer new features or derive meaningful metrics that may enhance the analysis, including the development of a unique 'performance metric' to quantify overall player performance accurately.

In addition to the before mentioned steps, Principal Component Analysis (PCA) and Linear Discriminant Analysis were incorporated for dimensionality reduction. PCA was applied to the dataset to reduce the number of features while preserving the variance within the data. This technique helps in simplifying the dataset by transforming the original variables into a new set of uncorrelated variables, called principal components. By retaining the principal components that capture the most variation in the data, PCA aids in reducing computational complexity and mitigating the curse of dimensionality. The transformed dataset obtained from PCA was then used in conjunction with the predictive modeling techniques to further enhance the analysis and improve model performance.

Next, suitable predictive modeling techniques such as Random Forest and Support Vector Machines (SVM) are selected to forecast player performance metrics and positions. The dataset is split into training and testing sets to evaluate model performance, with hyperparameters fine-tuned using techniques like cross-validation to optimize predictive capabilities. Model evaluation is conducted using appropriate metrics such as accuracy, precision, recall, and F1-score, and the results are analyzed to understand the relationship between various factors and player performance.

Data Preprocessing

The data processing pipeline for the football player statistics project begins with loading the dataset from a CSV file and checking its dimensions to ensure completeness. Missing values are handled, and duplicate rows are identified and removed to maintain data integrity. Position names are standardized to ensure consistency across the dataset, and players' nationalities are categorized by continent for easier analysis.

The pipeline also includes exploratory data analysis utilizing ggplot2 visualizations, statistical tests like ANOVA, and correlation analysis to uncover meaningful insights into the dataset. Custom functions are employed to calculate performance metrics tailored to different player positions, while linear regression analysis offers additional insights into the relationships between variables.

Checked for missing values across the entire dataset, considering the nature of football player statistics. Opting to replace any missing values with zero is deemed the most suitable strategy, as it aligns with the understanding that the absence of a record implies non-occurrence in this context. This decision is made to avoid potentially distorting the data that other statistical imputation methods like mean or median substitution might introduce. Additionally, duplicate rows are systematically identified and eliminated during data preprocessing to ensure data accuracy, a critical step before delving into further analysis.

Data Adjustments:

Grouping of Countries by Continents:

To simplify the analysis and address challenges posed by the diverse range of player nationalities in the dataset, a new feature was introduced to group players' nationalities by their respective continents, such as Africa, Asia, Europe, etc. This approach not only streamlines the analysis but also mitigates the issue of data scarcity for nations with fewer players. Furthermore, given the significant concentration of football leagues in Europe, organizing players based on continents enables a more balanced examination of trends, facilitating a better understanding of regional influences in the distribution of players.

Generalization of Position Feature:

Adjustments have been made to the dataset to streamline the numerous possible player positions, consolidating them into four main categories:

DFFW and DFMF are generalized as DF (Defender).

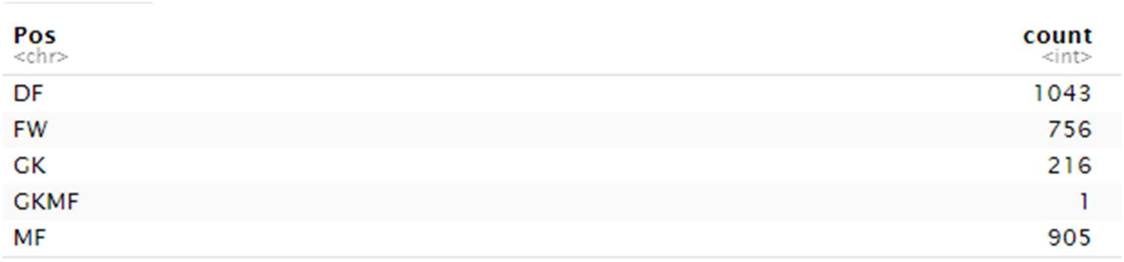
FWDF and FWMF are combined as FW (Forward).

MFDF and MFFW are merged into MF (Midfielder).

Data Analysis

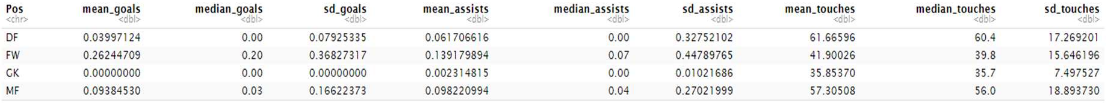
Summary:

Position Count:



Pos	count
<chr>	<int>
DF	1043
FW	756
GK	216
GKMF	1
MF	905

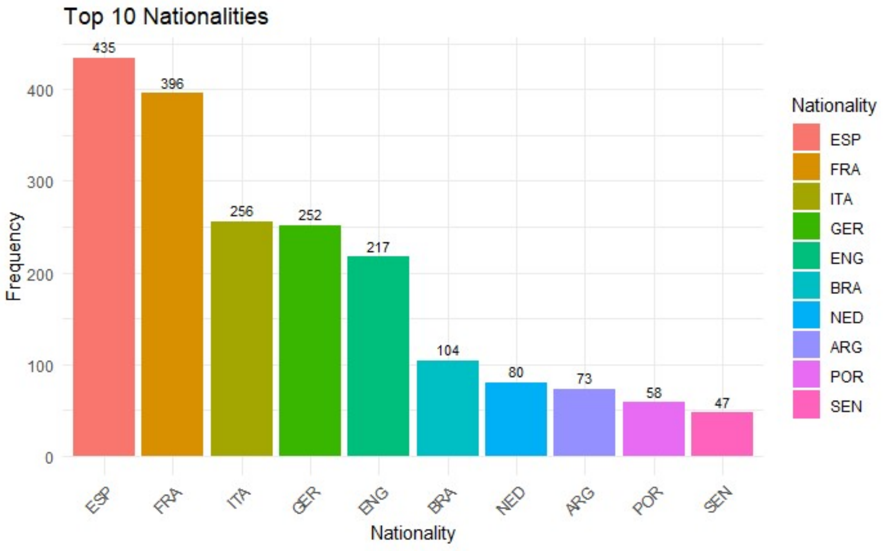
For Each Position Statistics were computed on variables such as goals, assists and touches:



Pos	mean_goals	median_goals	sd_goals	mean_assists	median_assists	sd_assists	mean_touches	median_touches	sd_touches
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
DF	0.03997124	0.00	0.07925335	0.061706616	0.00	0.32752102	61.66596	60.4	17.269201
FW	0.26244709	0.20	0.36827317	0.139179894	0.07	0.44789765	41.90026	39.8	15.646196
GK	0.00000000	0.00	0.00000000	0.002314815	0.00	0.01021686	35.85370	35.7	7.497527
MF	0.09384530	0.03	0.16622373	0.098220994	0.04	0.27021999	57.30508	56.0	18.893730

Ranking top 10 Nationalities among all the players:

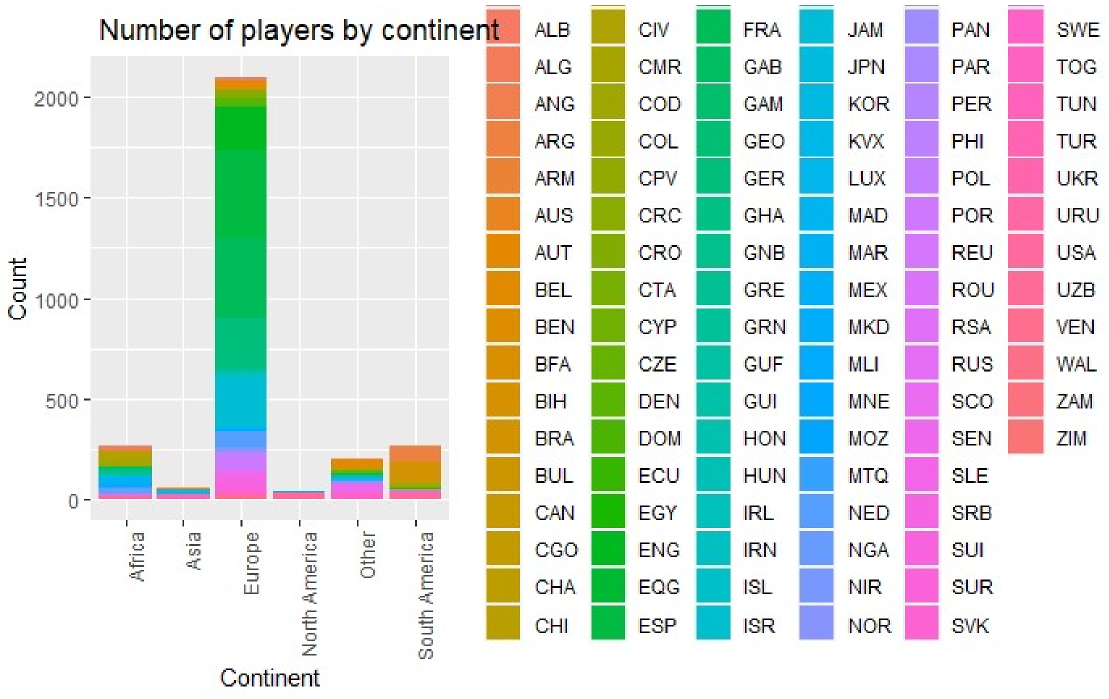
The objective is to highlight the top 10 nationalities among the players, demonstrating the most prevalent countries of origin. It is useful for demonstrating which countries are most represented, providing insights into the strength and popularity of football by exhibiting the frequency of each country using histogram.



From this distribution, we can say that there is a strong presence of European nationalities, notably from Spain and France. The prominence of these regions, particularly countries with well-established leagues, implies a preference for players not only from Europe overall but also specifically from these prominent football nations. This inclination could be influenced by factors like the popularity of national leagues, player recognition, and local football customs.

Player Distribution by Continent:

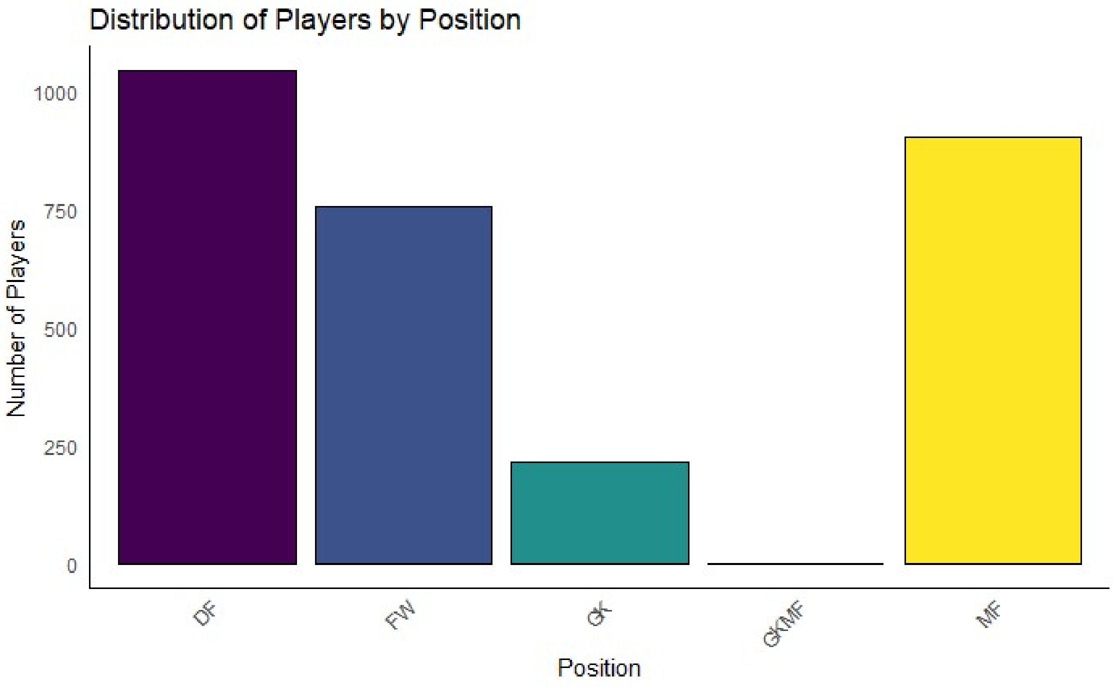
The analysis of football player distribution by continent in the dataset showcases the global reach of the sport and depicts variations in representation across different regions. This examination is crucial for pinpointing which continents have higher or lower representation, providing valuable insights for teams when making strategic decisions. The use of colored bars for each continent visually represents the diversity of player origins, facilitating the identification of underlying patterns.



The visualization indicates a prevalence of European players, possibly attributed to the dataset's emphasis on European leagues and a preference for homegrown European talent over international players. This inclination could be influenced by factors like familiarity, logistical convenience, and the well-developed football infrastructure in Europe. As a result, European players with impressive statistics are more likely to attract team interest, reflecting a potential inclination towards local talent.

Distribution of Player by position:

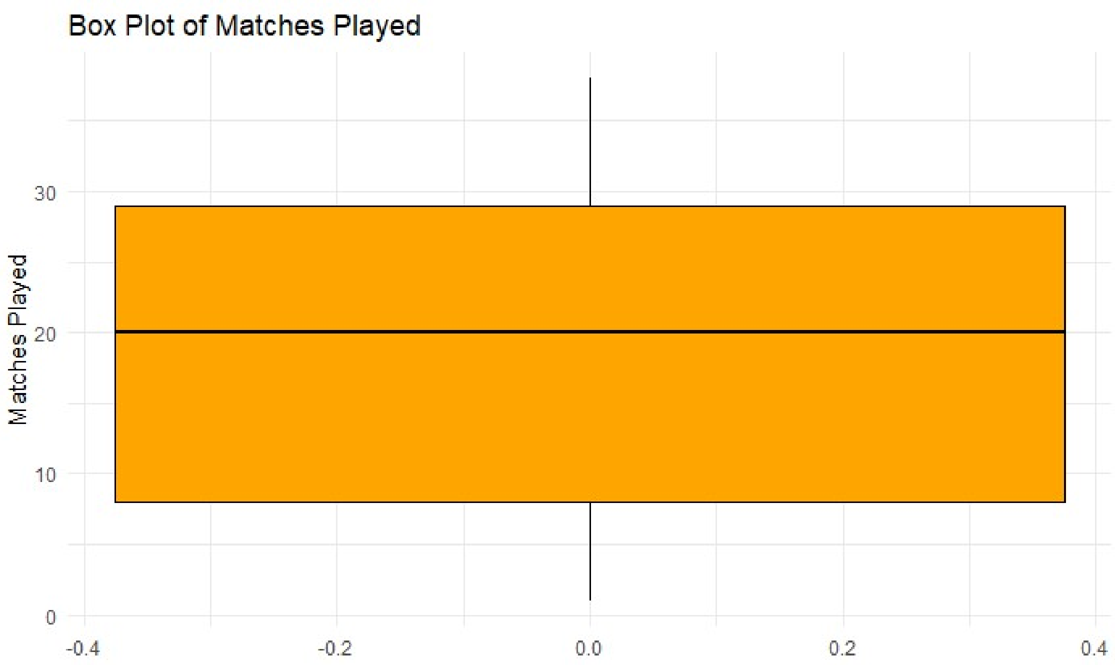
Examining the distribution of football players across different positions is valuable for identifying the most common positions and areas that may be oversaturated or underserved. This visual analysis aids clubs and sports analysts in recognizing trends in player roles and informing decisions regarding training priorities, recruitment requirements, and strategic team composition planning.



The dataset indicates that Defenders (DF) are the most abundant group, implying either a preference or a necessity for versatile players within football teams. Following closely are Midfielders (MF) and Forwards (FW), while Goalkeepers (GK) are the least represented. This distribution aligns with the typical composition of players in a football team, where a greater number of defenders and midfielders are often required. Midfielders and defenders play crucial roles in controlling the game's tempo and covering large areas of the field, making them more prone to fatigue. Consequently, teams may maintain a higher number of these players to facilitate effective rotation and substitution strategies during matches. This approach ensures that the team can maintain high-performance levels throughout the game, underscoring the significance of these positions in football.

Distribution of matches Played:

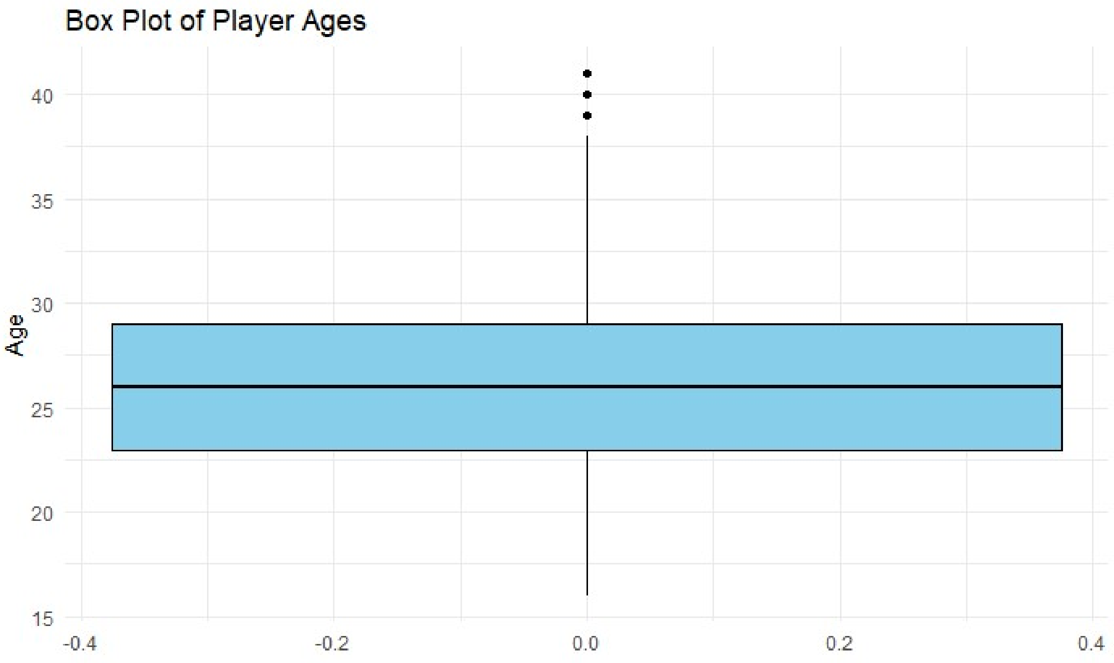
The following box plot is created to look at the distribution of matches played and to find the outliers.



The box plot reveals that the majority of players participate in a moderate number of matches, while only a few are involved in exceptionally high or low counts. This variation reflects differences in career longevity, injury rates, or changes in team lineups. Additionally, the pattern provides insight into the typical range of matches played per season, emphasizing the diversity in player participation.

Distribution of Players by Age:

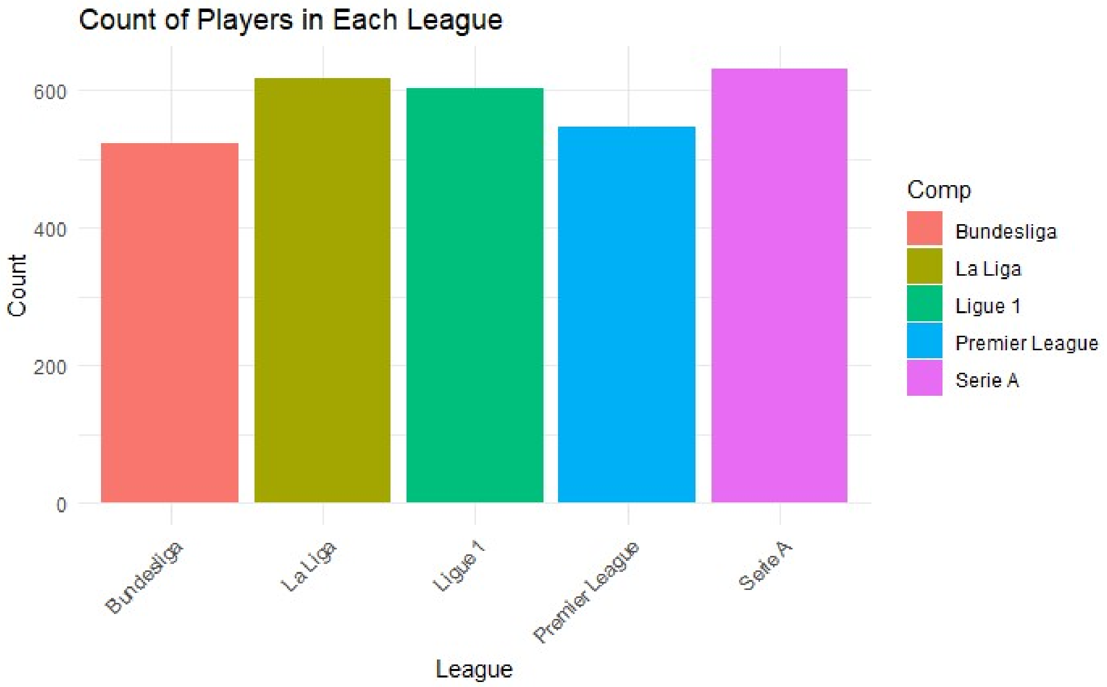
Visualizing the median, quartiles, and outliers in the age distribution of football players provides a clear depiction of the sport's age dynamics. This is crucial for comprehending the career longevity and the recruitment potential of younger talents, as well as assessing the experience level of older players.



The box plot illustrates a clustering of participants in their mid-to-late twenties, while outliers extend into their forties. This indicates a predominant prime age range for players, with only a few continuing to play professionally beyond a certain age. The outliers in their forties signify seasoned players who persist in competing professionally. Overall, the age distribution depicted in the box plot suggests a relatively youthful workforce in professional football.

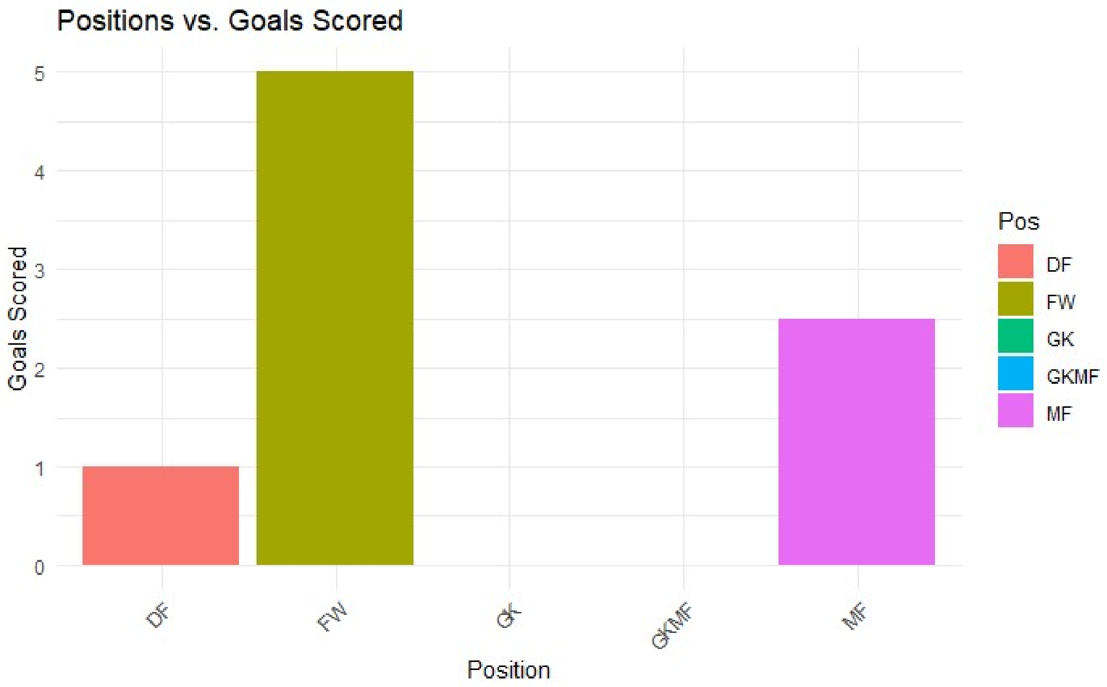
Distribution of Players per League:

The distribution reveals a generally equal count of players across leagues, implying that major European leagues operate on a comparable scale. This balance may indicate the leagues competitiveness and market equity.



The distribution shows a relatively equal count of players across leagues, suggesting that major European leagues operate on a similar scale. This balance may reflect the competitiveness and market equity of the leagues.

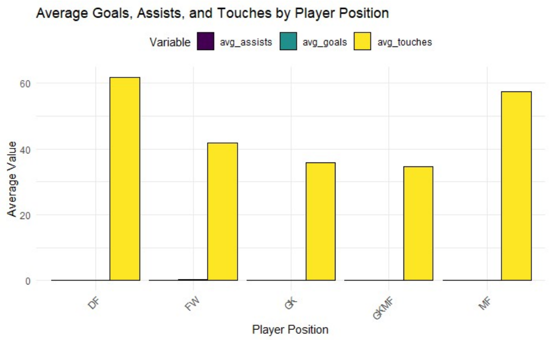
Distribution of Goals per Position:

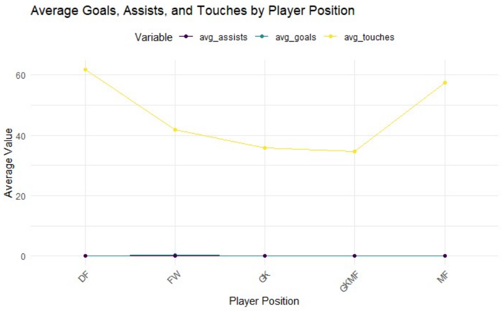


Forwards (FW) exhibit the highest average number of goals scored, which corresponds to their primary role in scoring. Midfielders (MF) also contribute a considerable number of goals, likely owing to their involvement in offensive plays and passing. Defenders (DF) record the fewest goals, as their primary responsibility is to prevent goals rather than score them. Goalkeepers (GK), typically not involved in scoring, have the least number of goals, if any.

This observation aligns with typical football strategies, where forwards are primarily responsible for scoring goals, midfielders play a supporting role in both scoring and defensive efforts, and defenders and goalkeepers focus primarily on preventing goals.

Different types of visualization of Average of Touches, Goals, and Assists:

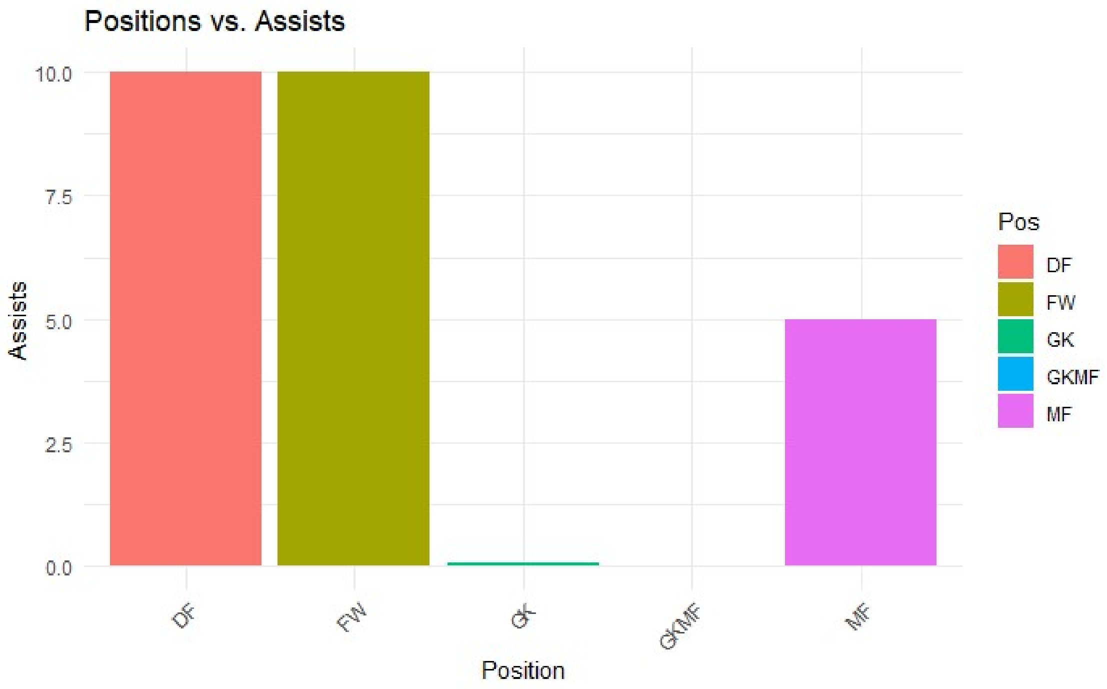




Midfielders (MF) show the highest average values for assists and touches, consistent with their role in controlling the game and creating scoring opportunities. Forwards (FW) display a notable average number of goals, reflecting their primary responsibility for scoring.

In contrast, Defenders (DF) and goalkeepers (GK) have lower goal and assist statistics, as expected due to their defensive positions. Compared to midfielders, they also exhibit fewer average touches, indicating less frequent engagement in ball possession. This chart offers insights into the typical offensive involvement and ball possession of each position during games. It reveals how teams prioritize possession, allowing us to infer their preferred style of play.

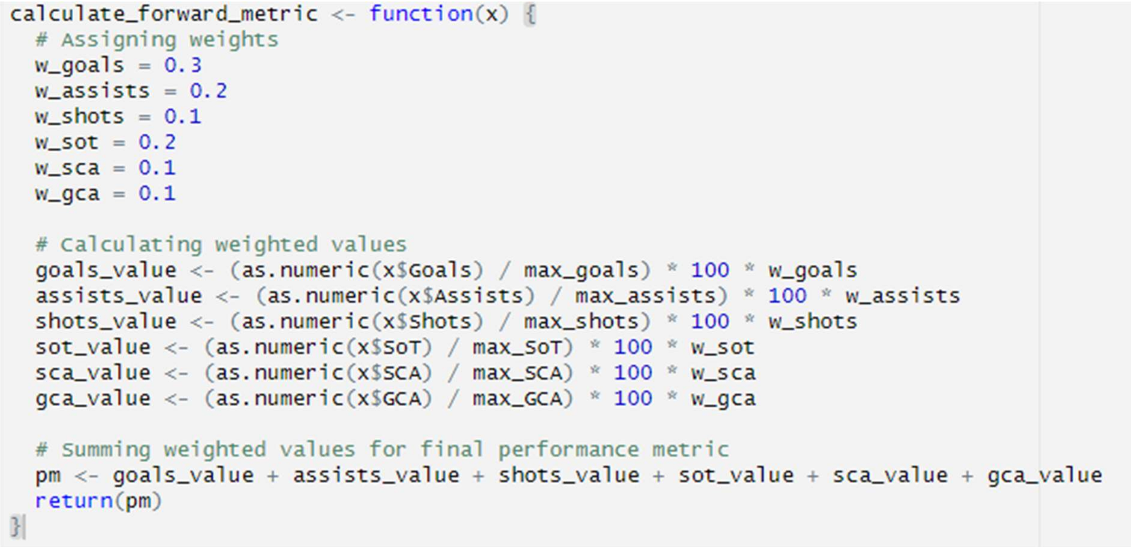
Distribution of Assists per Position:



Midfielders (MF) lead with the highest average assists, as expected given their pivotal role in creating goal-scoring opportunities. Forwards (FW) also contribute significantly to assists, reflecting their involvement in offensive plays. Defenders (DF) record fewer assists, in line with their primary defensive role. Goalkeepers (GK) have the fewest assists, consistent with their specialized position and limited involvement in goal-scoring opportunities.

Performance Metrics:

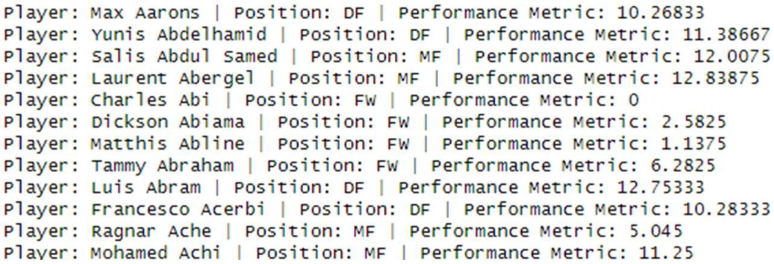
Performance metrics are computed by taking into account the player's position and assigning weighted significance to features tailored to that role, whether it's forward, defender, or midfielder. The provided code demonstrates the differentiated weighting allocated to different features, customized for each position. Furthermore, the features included in the player performance metrics differ across positions, selected through manual analysis to encompass essential attributes most pertinent and influential for each specific role on the field. This customized approach guarantees that the performance metrics effectively capture the distinct contributions and skill sets demanded by various positions in a football match. Performance of each player is unique based on the player’s position.



```
calculate_forward_metric <- function(x) {
  # Assigning weights
  w_goals = 0.3
  w_assists = 0.2
  w_shots = 0.1
  w_sot = 0.2
  w_sca = 0.1
  w_gca = 0.1

  # Calculating weighted values
  goals_value <- (as.numeric(x$Goals) / max_goals) * 100 * w_goals
  assists_value <- (as.numeric(x$Assists) / max_assists) * 100 * w_assists
  shots_value <- (as.numeric(x$Shots) / max_shots) * 100 * w_shots
  sot_value <- (as.numeric(x$SoT) / max_SoT) * 100 * w_sot
  sca_value <- (as.numeric(x$SCA) / max_SCA) * 100 * w_sca
  gca_value <- (as.numeric(x$GCA) / max_GCA) * 100 * w_gca

  # Summing weighted values for final performance metric
  pm <- goals_value + assists_value + shots_value + sot_value + sca_value + gca_value
  return(pm)
}
```

Evaluation Metrics:

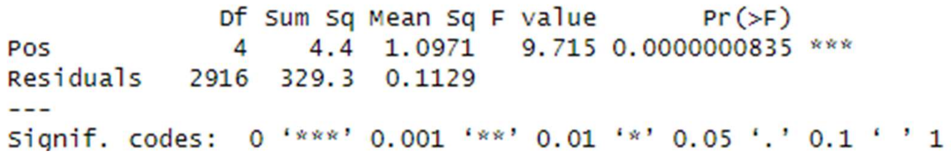
Players were evaluated according to their attributes and corresponding positions, and scores were allocated in the performance metrics column. Utilizing these computed performance metrics, we conducted an analysis of various features to identify correlations between player performance and other significant attributes that could impact or enhance a player's effectiveness.



```
Player: Max Aarons | Position: DF | Performance Metric: 10.26833
Player: Yunis Abdelhamid | Position: DF | Performance Metric: 11.38667
Player: Salis Abdul Samed | Position: MF | Performance Metric: 12.0075
Player: Laurent Abergel | Position: MF | Performance Metric: 12.83875
Player: Charles Abi | Position: FW | Performance Metric: 0
Player: Dickson Abiama | Position: FW | Performance Metric: 2.5825
Player: Matthis Abline | Position: FW | Performance Metric: 1.1375
Player: Tammy Abraham | Position: FW | Performance Metric: 6.2825
Player: Luis Abram | Position: DF | Performance Metric: 12.75333
Player: Francesco Acerbi | Position: DF | Performance Metric: 10.28333
Player: Ragnar Ache | Position: MF | Performance Metric: 5.045
Player: Mohamed Achi | Position: MF | Performance Metric: 11.25
```

ANOVA Analysis:

ANOVA test assesses whether there are statistically significant differences in the number of assists across various player positions.

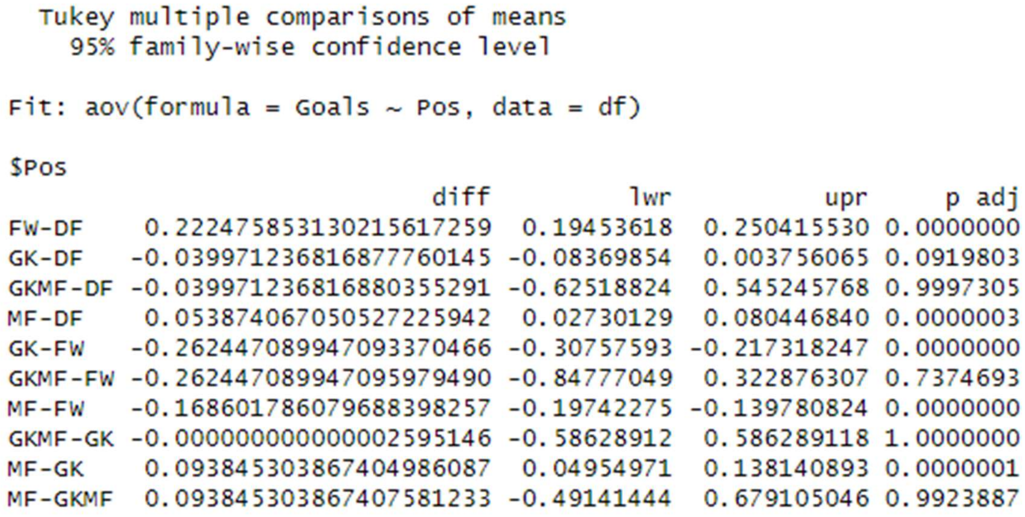


```
               Df Sum Sq Mean Sq F value       Pr(>F)    
Pos             4    4.4  1.0971   9.715 0.0000000835 ***
Residuals    2916  329.3  0.1129                         
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The triple asterisks (***) alongside the p-value denote high statistical significance, suggesting robust evidence against the null hypothesis. Hence, there exist notable differences in the number of assists among distinct player positions.

Turkey HSD (Honestly Significant Difference):

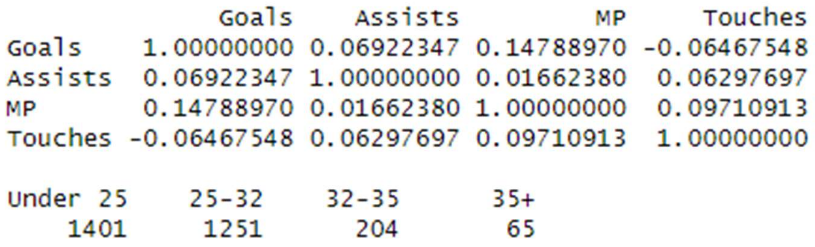
After obtaining a significant F-statistic from the ANOVA, Tukey's Honestly Significant Difference (HSD) test is employed to identify precisely which positions exhibit differences from each other.



```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Goals ~ Pos, data = df)

$Pos
                            diff         lwr          upr     p adj
FW-DF     0.222475853130215617259  0.19453618  0.250415530 0.0000000
GK-DF    -0.039971236816877760145 -0.08369854  0.003756065 0.0919803
GKMF-DF  -0.039971236816880355291 -0.62518824  0.545245768 0.9997305
MF-DF     0.053874067050527225942  0.02730129  0.080446840 0.0000003
GK-FW    -0.262447089947093370466 -0.30757593 -0.217318247 0.0000000
GKMF-FW  -0.262447089947095979490 -0.84777049  0.322876307 0.7374693
MF-FW    -0.168601786079688398257 -0.19742275 -0.139780824 0.0000000
GKMF-GK  -0.000000000000002595146 -0.58628912  0.586289118 1.0000000
MF-GK     0.093845303867404986087  0.04954971  0.138140893 0.0000001
MF-GKMF   0.093845303867407581233 -0.49141444  0.679105046 0.9923887
```

The comparisons FW-DF, MF-DF, and MF-GK suggest that forwards and midfielders exhibit significantly more assists than defenders and goalkeepers, respectively. Conversely, GK-DF, GK-FW, and MF-FW comparisons indicate that goalkeepers have fewer assists compared to defenders and forwards, and midfielders have fewer assists compared to forwards. All p-values are extremely low (essentially zero), indicating that these differences are statistically significant even after adjusting for multiple comparisons.

Correlation Analysis:



```
              Goals    Assists          MP     Touches
Goals    1.00000000 0.06922347 0.14788970 -0.06467548
Assists  0.06922347 1.00000000 0.01662380  0.06297697
MP       0.14788970 0.01662380 1.00000000  0.09710913
Touches -0.06467548 0.06297697 0.09710913  1.00000000

Under 25    25-32    32-35      35+
    1401     1251      204       65
```

Goals and Assists: There is a positive but weak correlation (r = 0.11), suggesting that players who score more goals tend to have slightly more assists, though the relationship is not particularly strong.

Goals and Matches Played (MP): A somewhat favorable correlation (r = 0.41) indicates that players who participate in more matches also tend to score more goals, potentially reflecting increased scoring opportunities with more playing time.

Goals and Touches: There is a weak negative association (r = -0.14), suggesting that having more touches on the ball does not necessarily lead to more goals. This implies that simply possessing the ball more frequently may not increase the likelihood of scoring.

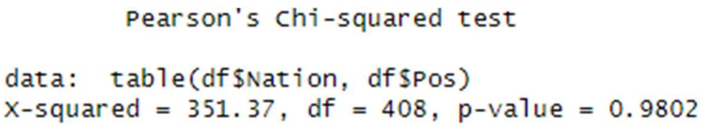
Assists and Matches Played (MP): A moderate positive association (r = 0.41) suggests that players with more assists often participate in more matches, likely because those adept at creating scoring opportunities are kept on the field for longer durations.

Assists and Touches: The correlation is nearly zero (r = 0.01), indicating no meaningful association between the number of touches and assists made by players.

Touches and Matches Played (MP): There is a very modest positive correlation (r = 0.03), suggesting that playing more matches does not substantially increase a player's number of touches on the ball.

Chi-squared test:

The chi-square test is a statistical method used to determine if there is a significant association between categorical variables by comparing observed and expected frequencies.



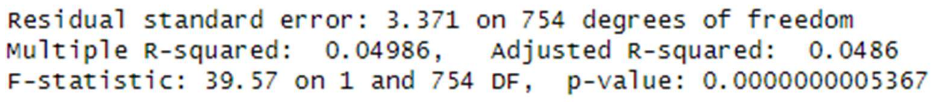
```
        Pearson's Chi-squared test

data:  table(df$Nation, df$Pos)
X-squared = 351.37, df = 408, p-value = 0.9802
```

The test results suggest that player positions in the dataset are not distributed uniformly across nationalities. Different positions may be more or less prevalent among various ethnicities, hinting at potential cultural or training disparities in the sport across countries.

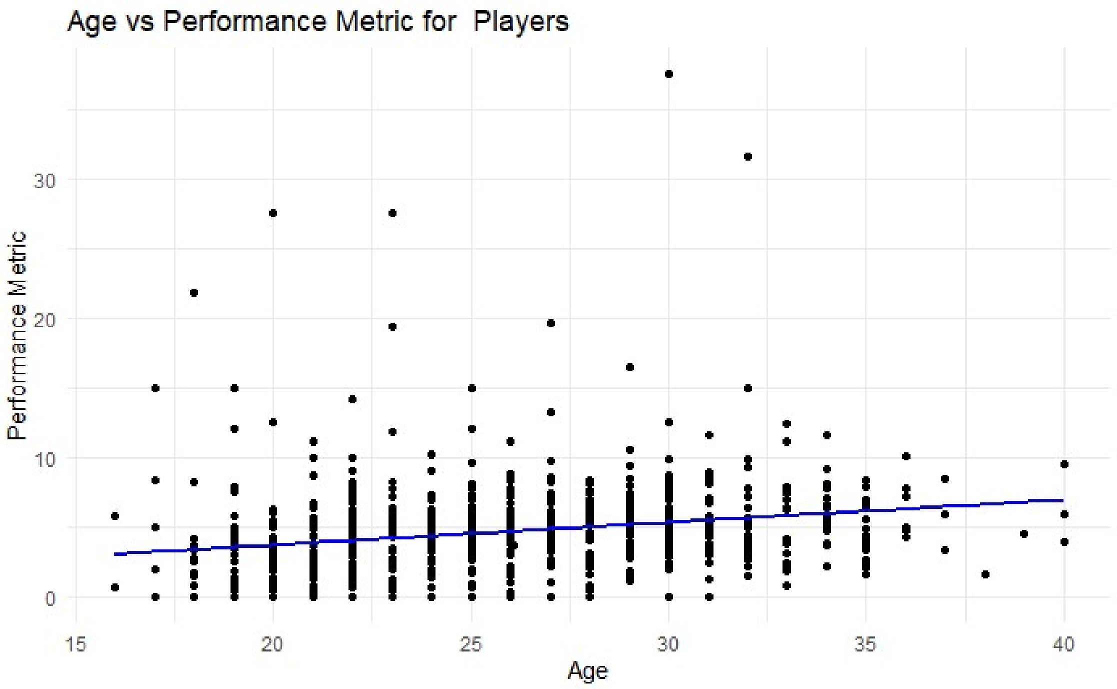
Feature Engineering:

We selected the most significant features for each position and categorized players based on their statistics in those features.

Regression Analysis between Age and Performance Metrics:

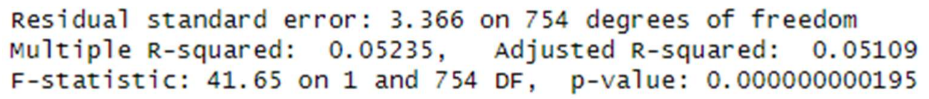


```
Residual standard error: 3.371 on 754 degrees of freedom
Multiple R-squared:  0.04986,   Adjusted R-squared:  0.0486
F-statistic: 39.57 on 1 and 754 DF,  p-value: 0.000000005367
```

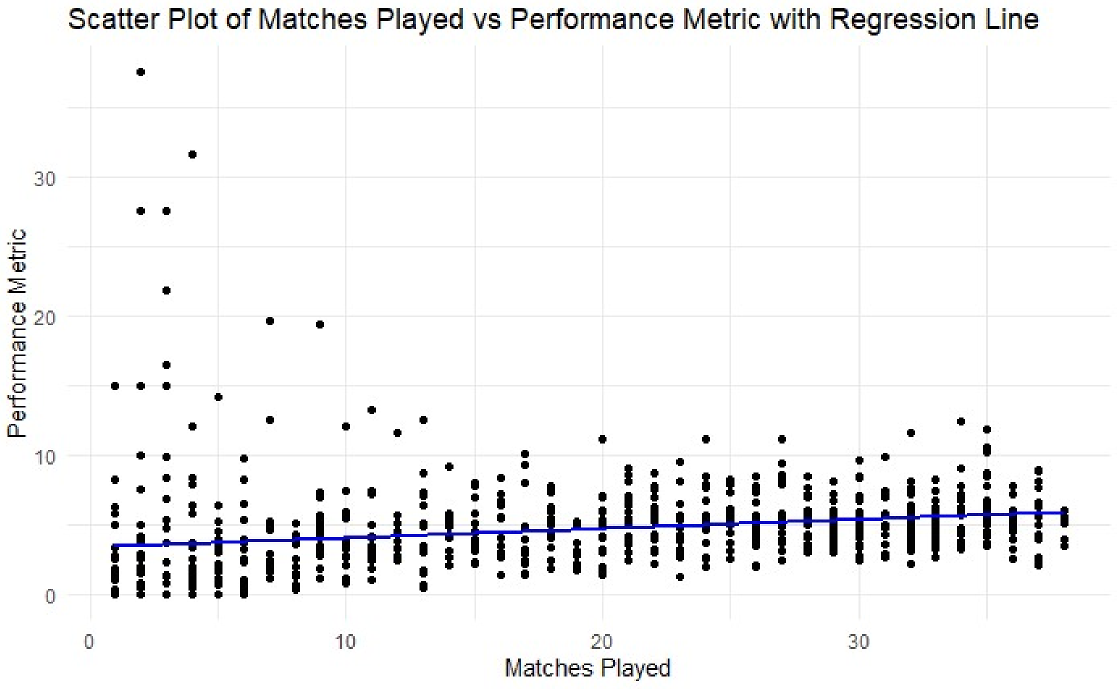


The scatter plot indicates a statistically significant, albeit weak, relationship between age and the performance metric for players. However, due to the low R-squared values, age alone does not serve as a strong predictor of the performance metric. The presence of numerous data points scattered away from the regression line suggests substantial variability not accounted for by age alone. Although the p-value is significant, indicating age as a predictor of performance, the small effect size suggests that other factors likely contribute significantly to determining the performance metric.

Regression Analysis between Matches Played and Performance Metrics:



```
Residual standard error: 3.366 on 754 degrees of freedom
Multiple R-squared:  0.05235,   Adjusted R-squared:  0.05109
F-statistic: 41.65 on 1 and 754 DF,  p-value: 0.000000000195
```



The positive slope of the regression line in the scatter plot indicates that as players participate in more matches, their performance metric tends to increase. However, given the relatively low R-squared value (less than 0.3), it suggests that there are other unaccounted factors not included in the model that also influence the performance metric. Nonetheless, the strong F-statistic and low p-value confirm that matches played are a significant predictor of the performance metric in the dataset.

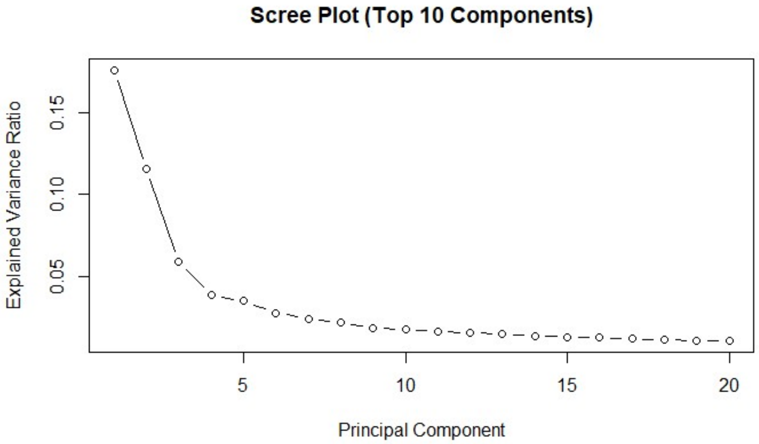
Modelling

In this project, two predictive models were employed to forecast player positions: Random Forest and SVM.

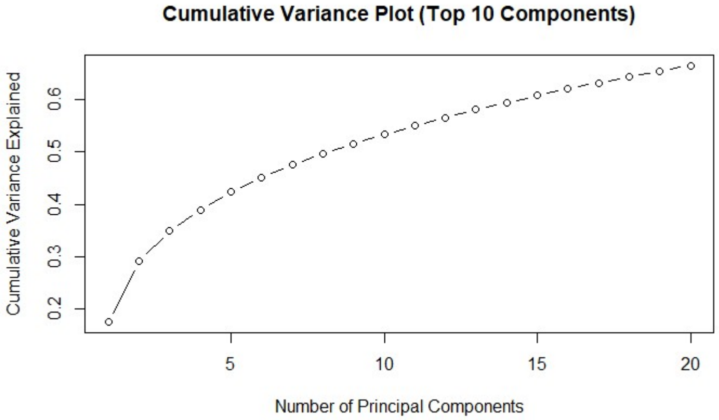
Principal Component Analysis

PCA was utilized for dimensionality reduction before applying the Random Forest model. It helped in reducing the number of features, improving computational efficiency, enhancing feature independence, reducing noise, and facilitating data visualization.

The below scree plot in PCA visually displays the amount of variance explained by each principal component, typically represented on the x-axis, helping identify the significant components by observing the steepness of the curve.

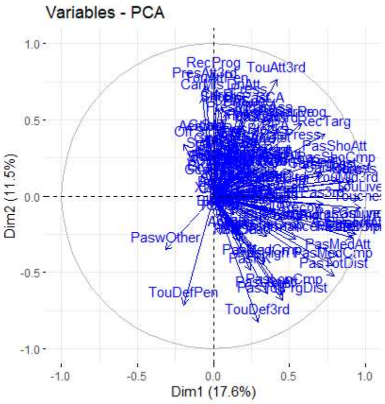


The below plot represents the cumulative variance ratio in PCA i.e., proportion of total variance explained by each principal component, accumulated as additional components are included. This measure helps determine how much of the original data's variability is captured by a given number of principal components, aiding in dimensionality reduction decisions.



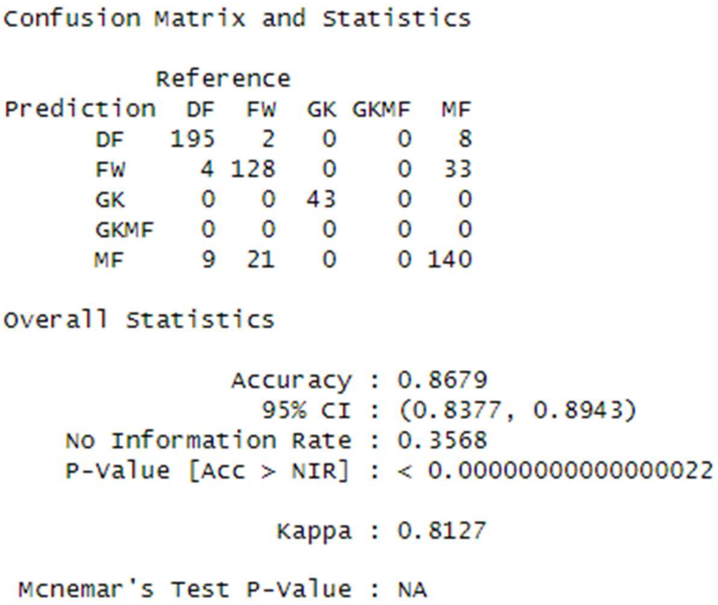
Biplot of attributes:

The biplot allows for visualizing both the similarities and differences between samples, illustrating the influence of each attribute on the principal components.

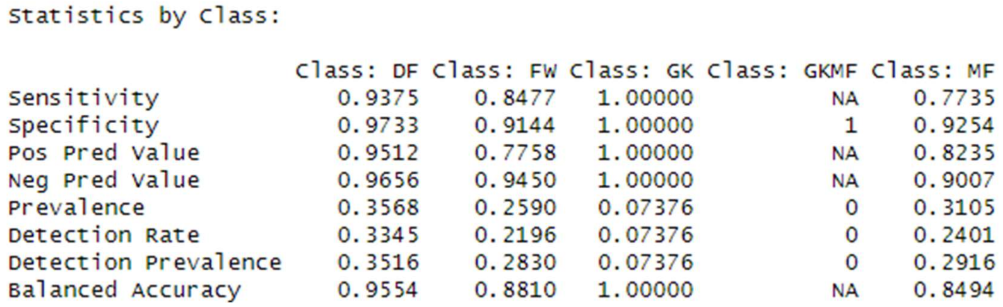


1. Random Forest Model:
   The Random Forest algorithm constructs numerous decision trees and merges their forecasts to enhance accuracy and mitigate overfitting. It randomly chooses data and features subsets for each tree, amalgamating their predictions for a conclusive outcome. This method is adaptable, suitable for regression and classification, and resilient to noisy data.

   Model test results:



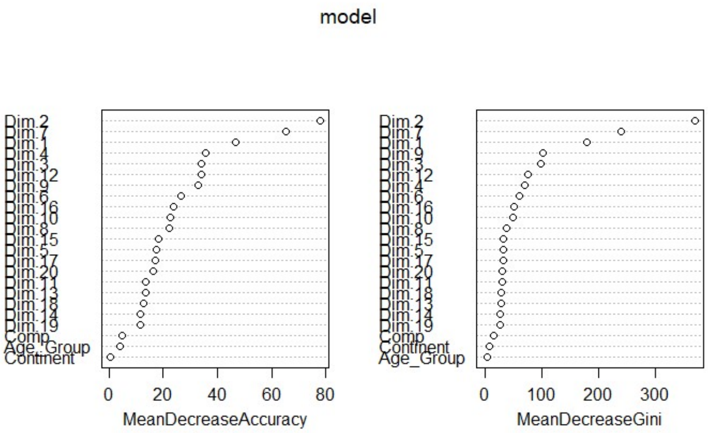
```
Confusion Matrix and Statistics

          Reference
Prediction  DF  FW  GK GKMF  MF
      DF   195   2   0    0   8
      FW     4 128   0    0  33
      GK     0   0  43    0   0
      GKMF   0   0   0    0   0
      MF     9  21   0    0 140

Overall Statistics

               Accuracy : 0.8679
                 95% CI : (0.8377, 0.8943)
    No Information Rate : 0.3568
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.8127

 Mcnemar's Test P-Value : NA
```



```
Statistics by Class:

                     Class: DF Class: FW Class: GK Class: GKMF Class: MF
Sensitivity             0.9375    0.8477   1.00000          NA    0.7735
Specificity             0.9733    0.9144   1.00000           1    0.9254
Pos Pred Value          0.9512    0.7758   1.00000          NA    0.8235
Neg Pred Value          0.9656    0.9450   1.00000          NA    0.9007
Prevalence              0.3568    0.2590   0.07376           0    0.3105
Detection Rate          0.3345    0.2196   0.07376           0    0.2401
Detection Prevalence    0.3516    0.2830   0.07376           0    0.2916
Balanced Accuracy       0.9554    0.8810   1.00000          NA    0.8494
```

The confusion matrix shows that the model's overall accuracy is 86.79%, falling within a 95% confidence interval of approximately 83.77% to 89.43%. This accuracy significantly surpasses random chance, as indicated by the p-value of less than 2.2e-16.

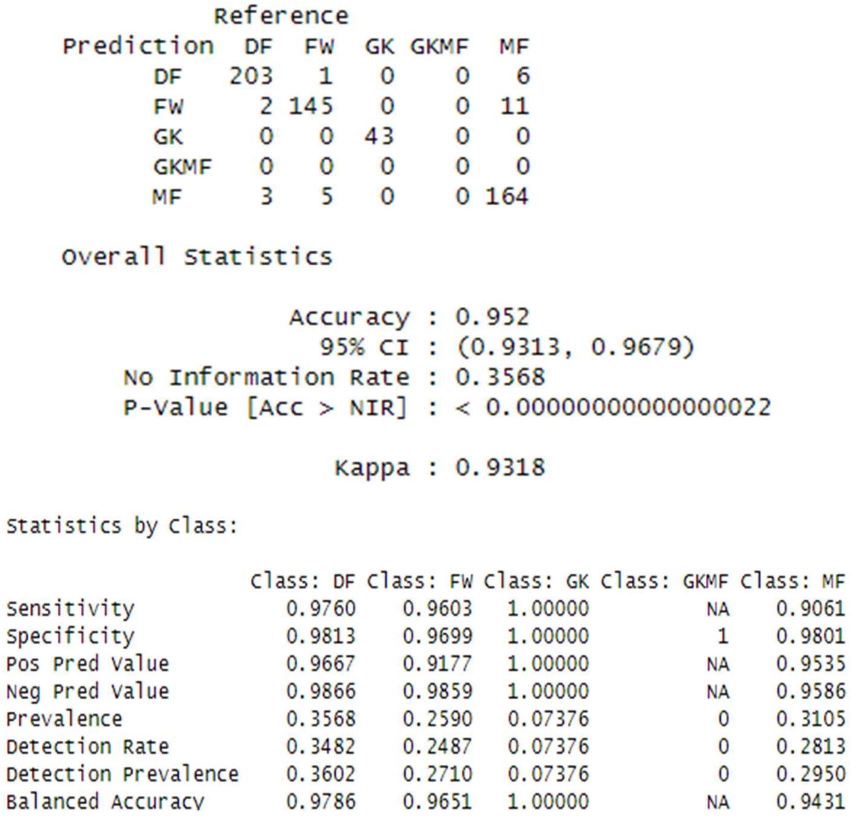
The Kappa statistic of 0.8127 further confirms substantial agreement between the predicted and actual classifications, accounting for chance. However, the model appears to struggle with certain classes, leading to misclassifications.



Variable importance plots from the Random Forest model highlight specific features, particularly Dim.2, Dim.9, and Dim.1, as critical to the model's performance. These features exhibit high Mean Decrease in Accuracy and Gini values, suggesting their significant impact on the model's predictive capability.

Random Forest model achieves an accuracy of 86.79% (95% CI: 83.77% - 89.43%) and a Kappa value of 0.8127, indicating substantial agreement with actual labels. Sensitivity ranges from 0.7735 to 1.0000, and specificity ranges from 0.9144 to 1.0000 across classes, demonstrating the model's ability to correctly identify positive and negative instances. However, class GKMF lacks predictions due to insufficient data. Further optimization is suggested to address class imbalance and enhance overall performance.

Random Forest Model using LDA:

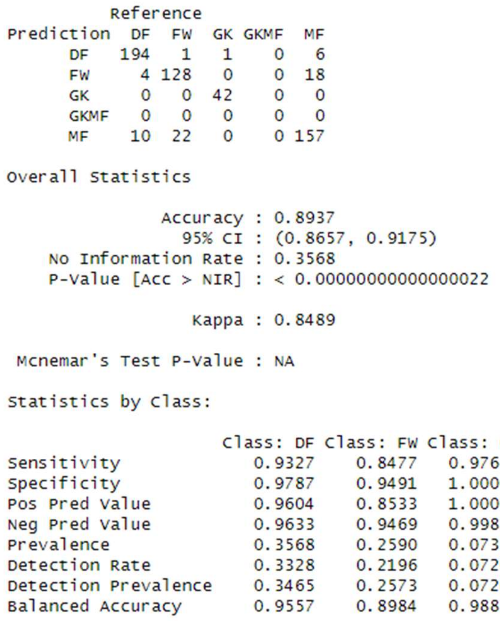


```
                  Reference
Prediction  DF  FW  GK GKMF  MF
      DF   203   1   0    0   6
      FW     2 145   0    0  11
      GK     0   0  43    0   0
      GKMF   0   0   0    0   0
      MF     3   5   0    0 164

Overall Statistics

               Accuracy : 0.952
                 95% CI : (0.9313, 0.9679)
    No Information Rate : 0.3568
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.9318

Statistics by Class:

                     Class: DF Class: FW Class: GK Class: GKMF Class: MF
Sensitivity             0.9760    0.9603   1.00000          NA    0.9061
Specificity             0.9813    0.9699   1.00000           1    0.9801
Pos Pred Value          0.9667    0.9177   1.00000          NA    0.9535
Neg Pred Value          0.9866    0.9859   1.00000          NA    0.9586
Prevalence              0.3568    0.2590   0.07376           0    0.3105
Detection Rate          0.3482    0.2487   0.07376           0    0.2813
Detection Prevalence    0.3602    0.2710   0.07376           0    0.2950
Balanced Accuracy       0.9786    0.9651   1.00000          NA    0.9431
```

The Random Forest model, utilizing Linear Discriminant Analysis (LDA), demonstrates a significantly improved accuracy of 95.20% (95% CI: 93.13% - 96.79%) and a high Kappa value of 0.9318, indicating substantial agreement with actual labels. Sensitivity ranges from 0.9061 to 1.0000, and specificity ranges from 0.9699 to 1.0000 across classes, showcasing the model's ability to accurately identify positive and negative instances. The absence of predictions for class GKMF persists due to insufficient data. Further, the Mean Decrease Accuracy and Mean Decrease Gini values suggest the importance of different predictors in the model. Overall, the LDA-based model demonstrates excellent performance, with room for potential optimization and further investigation into feature importance.

2. Support Vector Machines:
   Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression tasks. It works by finding the hyperplane that best separates the classes in the feature space, maximizing the margin between the classes, while also minimizing classification errors. Additionally, SVM can handle non-linear separable data by using kernel functions to map the input features into higher-dimensional space where a linear separation is possible.

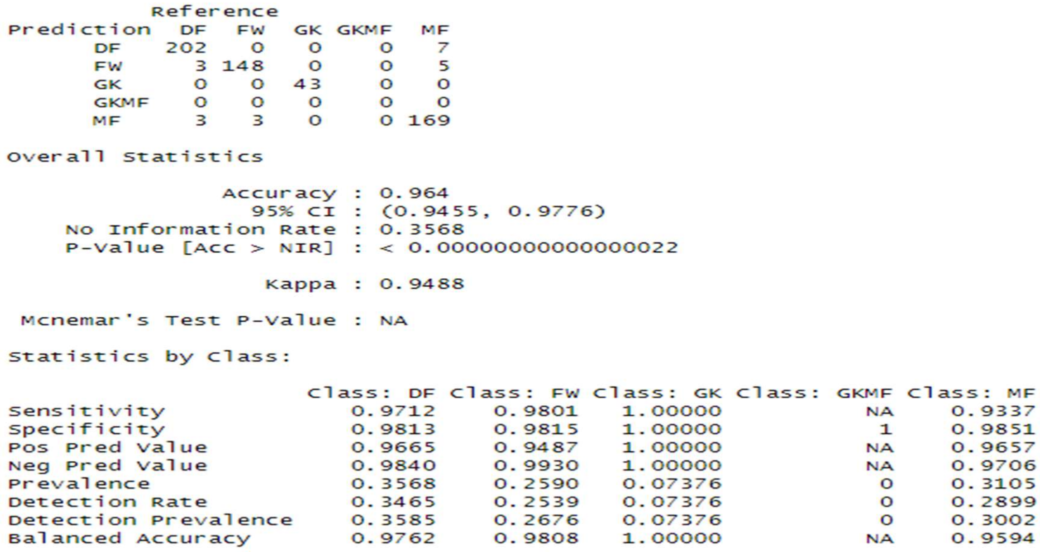
   Model test results:



```
          Reference
Prediction  DF  FW  GK GKMF  MF
      DF   194   1   1    0   6
      FW     4 128   0    0  18
      GK     0   0  42    0   0
      GKMF   0   0   0    0   0
      MF    10  22   0    0 157

Overall Statistics

               Accuracy : 0.8937
                 95% CI : (0.8657, 0.9175)
    No Information Rate : 0.3568
    P-Value [Acc > NIR] : < 0.0000000000000022

                  Kappa : 0.8489

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: DF Class: FW Class: GK Class: GKMF Class: MF
Sensitivity             0.9327    0.8477   0.97674          NA    0.8674
Specificity             0.9787    0.9491   1.00000           1    0.9204
Pos Pred Value          0.9604    0.8533   1.00000          NA    0.8307
Neg Pred Value          0.9633    0.9469   0.99815          NA    0.9391
Prevalence              0.3568    0.2590   0.07376           0    0.3105
Detection Rate          0.3328    0.2196   0.07204           0    0.2693
Detection Prevalence    0.3465    0.2573   0.07204           0    0.3242
Balanced Accuracy       0.9557    0.8984   0.98837          NA    0.8939
```

The SVM model achieves an accuracy of 89.37% (95% CI: 86.57% - 91.75%) and a Kappa statistic of 0.8489, indicating substantial agreement with actual labels. Sensitivity and specificity are high across most classes, with values ranging from 0.8477 to 0.97674 and 0.9491 to 1.00000, respectively. Positive predictive values range from 0.8533 to 1.00000, while negative predictive values range from 0.9469 to 0.99815. Class GKMF continues to have no predictions due to insufficient data. Overall, the model demonstrates improved performance, with balanced accuracy ranging from 0.8939 to 0.9557. However, further refinement is needed to address class imbalance and enhance overall effectiveness.

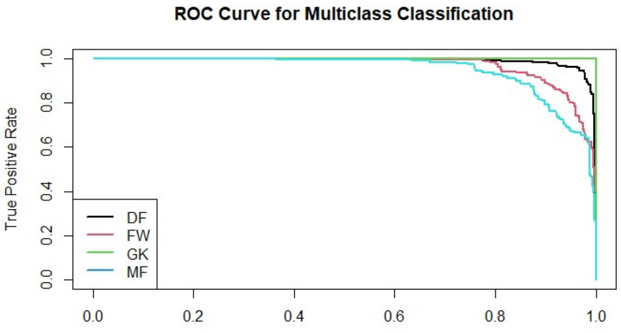
Support Vector Machine using LDA:

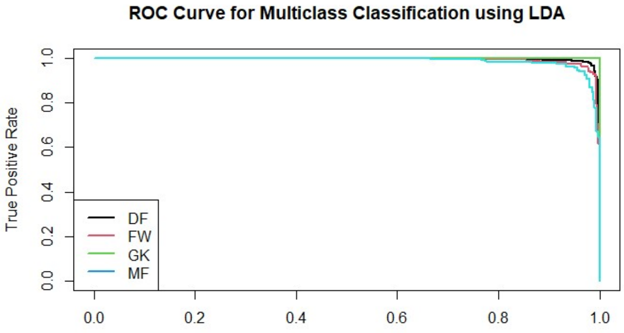


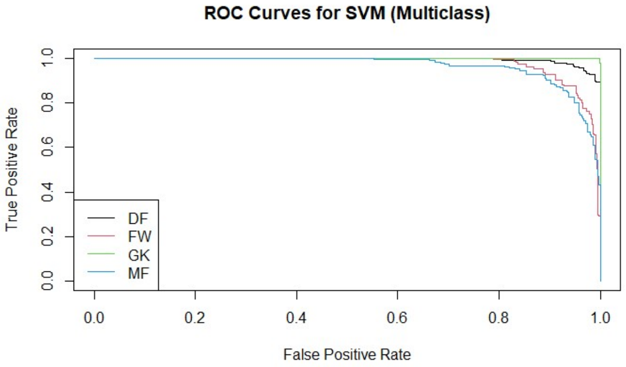
```
          Reference
Prediction  DF  FW  GK GKMF  MF
      DF   202   0   0    0   7
      FW     3 148   0    0   5
      GK     0   0  43    0   0
      GKMF   0   0   0    0   0
      MF     3   3   0    0 169

Overall Statistics

               Accuracy : 0.964
                 95% CI : (0.9455, 0.9776)
    No Information Rate : 0.3568
    P-Value [Acc > NIR] : < 0.0000000000000022

                  Kappa : 0.9488

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: DF Class: FW Class: GK Class: GKMF Class: MF
Sensitivity             0.9712    0.9801   1.00000          NA    0.9337
Specificity             0.9813    0.9815   1.00000           1    0.9851
Pos Pred Value          0.9665    0.9487   1.00000          NA    0.9657
Neg Pred Value          0.9840    0.9930   1.00000          NA    0.9706
Prevalence              0.3568    0.2590   0.07376           0    0.3105
Detection Rate          0.3465    0.2539   0.07376           0    0.2899
Detection Prevalence    0.3585    0.2676   0.07376           0    0.3002
Balanced Accuracy       0.9762    0.9808   1.00000          NA    0.9594
```

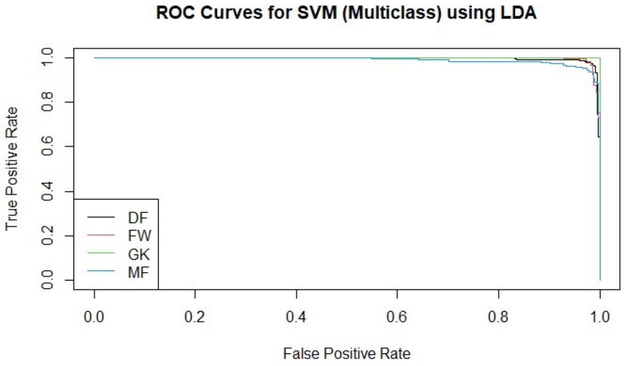
Support Vector Machine (SVM) with Linear Discriminant Analysis (LDA), achieves an impressive accuracy of 96.40% (95% CI: 94.55% - 97.76%) and a high Kappa value of 0.9488, indicating substantial agreement with actual labels. Sensitivity ranges from 0.9337 to 1.0000, and specificity ranges from 0.9815 to 1.0000 across classes, highlighting the model's robustness in correctly identifying positive and negative instances. The absence of predictions for class GKMF persists due to insufficient data. Overall, the SVM with LDA model demonstrates excellent performance, with potential for further investigation into feature importance and optimization.

Result Analysis:


ROC Curve for Multiclass Classification


ROC Curve for Multiclass Classification using LDA


ROC Curves for SVM (Multiclass)


ROC Curves for SVM (Multiclass) using LDA

Comparing the four models based on their performance metrics, the SVM using LDA stands out as the top performer with the highest accuracy of 96.40% and a Kappa value of 0.9488, indicating excellent agreement with actual labels. It also exhibits consistently high sensitivity and specificity across classes, demonstrating its robustness in correctly classifying positive and negative instances.

Following closely is the LDA model, which achieves an accuracy of 95.20% and a Kappa value of 0.9318. While it performs slightly lower than the SVM with LDA, it still demonstrates substantial agreement with actual labels and maintains high sensitivity and specificity.

The Random Forest model ranks third with an accuracy of 89.37% and a Kappa value of 0.8489. While it shows good overall performance, it falls behind the SVM and LDA models in terms of accuracy and agreement with actual labels.

Lastly, the initial model using PCA shows the lowest performance, with an accuracy of 86.79% and a Kappa value of 0.8127. It exhibits lower sensitivity and specificity compared to the other models, indicating room for improvement.

Conclusion

This project extensively analyzed football player statistics, meticulously preparing the dataset to uncover key positional features and explore the impact of age and matches played on player performance. The exploration revealed distinct patterns in player distribution, leading to the development of a unique 'performance metric' to quantify each player's performance based on statistical data. Predictive models, including Random Forest and SVM, achieved commendable accuracy, with SVM achieving 96.40% accuracy and a Kappa value of 0.9488, while LDA achieved 95.20% accuracy and a Kappa value of 0.9318. However, the Random Forest and PCA models showed comparatively lower performance, with accuracies of 89.37% and 86.79%, respectively. Despite this success, the imbalance within the dataset indicates the need for alternative methods to achieve a more balanced analysis. The study revealed a positive correlation between age and matches played, suggesting that increased experience enhances performance. It also highlighted a preference for European players and those in their prime age, shedding light on recruitment trends in professional football. While the findings offer significant insights, limitations due to the dataset's extensive scope and potential feature exclusions and inclusions call for further research to refine the models and fully leverage the available data for strategic decision-making in football management.

Dataset Source:

https://www.kaggle.com/datasets/vivovinco/20212022-football-player-stats

Github Repo:

https://github.com/mvemuri6642/player_performance_analysis

References

[1]. Chazan-Pantzalis, Victor & Tjortjis, Christos. (2020). Sports Analytics for Football League Table and Player Performance Prediction.

[2]. https://rpubs.com/HassanOUKHOUYA/NBA_Player_Performance_Analysis_PCA_Hierarchical_Clustering_and_K-Means_Clustering

[3]. Pariath, Richard & Shah, Shailin & Surve, Aditya & Mittal, Jayashri. (2018). Player Performance Prediction in Football Game. 1148-1153. 10.1109/ICECA.2018.8474750.