FIFA 19 Player Insights: Predicting Wages, Release Clauses, and Positions

A Case Study report submitted in partial fulfillment of the requirements for the award of the degree of

Integrated M.Sc. in Computer Science
with Specialization in
Artificial Intelligence and Machine Learning

Submitted by

Vishnu Prakash J
(NA20PICS15)

Nehru Arts and Science College Kanhangad
Padnekkad P.O., Kasaragod Dt., Kerala - 671314

Affiliated to

Kannur University
Kannur

April 2024

Nehru Arts and Science College Kanhangad

Padnekkad P.O., Kasaragod Dt., Kerala - 671314

CERTIFICATE

This is to certify that report entitled “**FIFA 19 Player Insights: Predicting Wages, Release Clauses, and Positions**” is a bonafide report of the Case Study (8B38ICSC: Lab 11 - Case Study (Data Mining)) presented during VIIIth semester by **Vishnu Prakash J** with Register No. **NA20PICS15** in partial fulfillment of the requirements for the award of the degree of Integrated M.Sc. in Computer Science with Specialization in Artificial Intelligence and Machine Learning.

Faculty In Charge

Head of the Department

External Examiner

Internal Examiner

DECLARATION

I, **Vishnu Prakash J**, VIII Semester Integrated M.Sc. in Computer Science with Specialization in Artificial Intelligence and Machine Learning Student of Nehru Arts and Science College Kanhangad under Kannur University do hereby declare that the case study entitled "**FIFA 19 Player Insights: Predicting Wages, Release Clauses, and Positions**" is original work carried out by me towards the partial fulfillment of the requirement of Integrated M.Sc. in Computer Science with Specialization in Artificial Intelligence and Machine Learning, and no part thereof has been presented for the award of any other degree.

Vishnu Prakash J

ACKNOWLEDGMENT

I would like to express my sincere gratitude to everyone who has contributed to the completion of this case study report. Special thanks to my advisor for their guidance and support throughout the research process.
I am also thankful to my colleagues and friends who provided valuable insights and feedback, enriching the quality of this work. The encouragement and assistance received from Nehru Arts and Science College Kanhangad are greatly appreciated.
Lastly, I extend my thanks to my family for their unwavering support and understanding during this academic endeavor.

Vishnu Prakash J

ABSTRACT

In this case study, I analyzed the FIFA 19 Complete Player dataset to build both classification and regression models. The analysis involved comprehensive data preprocessing and cleaning, including handling missing values and converting non-numeric columns to numeric values. Exploratory Data Analysis (EDA) was conducted to understand data distributions and relationships between variables. Feature engineering was performed to create new attributes encapsulating player abilities such as Attacking, Defensive, Physical, Goalkeeping, Playmaking, and Speed. For model building, I developed a classification model to predict player positions and a regression model to predict the release clause and wage of players. The regression model, a RandomForestRegressor, was evaluated using Mean Squared Error (MSE) and R2-score, showing reasonable accuracy. This case study demonstrates effective use of data preprocessing, EDA, feature engineering, and model building techniques for predictive analytics in sports.

Contents

Chapter 1

Introduction

1.1 Background Information

The FIFA 19 Complete Player dataset contains comprehensive information on football players, encompassing a wide range of attributes related to their performance, demographics, and financial metrics. This dataset presents a valuable opportunity for data-driven analysis and predictive modeling in sports analytics. By leveraging this data, we can derive insights into player characteristics and market values, essential for clubs, managers, and analysts to make informed decisions.

1.2 Problem Statement

In the football industry, accurately predicting key player metrics such as release clauses and wages is crucial for financial planning and talent management. Additionally, predicting player positions based on their attributes is essential for effective scouting and team composition. However, the dataset presents several challenges due to its complexity and the presence of non-numeric data. This complexity requires comprehensive preprocessing, feature engineering, and the development of robust models to achieve reliable predictions.

While many datasets have a straightforward prediction target, this dataset allows for multiple potential predictions. For the time being, the focus is on specific predictions, including the release clause, wage, and player positions. The challenge lies not only in developing accurate models for these predictions but also in managing the dataset's intricacies and ensuring that the predictions provide meaningful insights for decision-making in the football industry.

1.3 Objective

The objective of this project is to apply comprehensive data preprocessing, exploratory data analysis (EDA), and feature engineering techniques to the FIFA 19 Complete Player dataset. Subsequently, we aim to build and evaluate both classification and regression models. The classification model will predict player positions, while the regression model will predict the release clause and wage of players. The goal is to demonstrate the efficacy of these techniques in predictive analytics for sports, providing actionable insights and accurate predictions. This project seeks to illustrate how data-driven approaches can enhance decision-making processes within the football industry.

Chapter 2

Literature Review

The FIFA 19 dataset has been the subject of extensive research and analysis, demonstrating its rich potential for sports analytics and machine learning applications. One of the notable techniques applied to this dataset is K-Means clustering. Various studies have utilized K-Means clustering to segment players based on their attributes, uncovering patterns and similarities among them. For instance, works such as ”K-Means Clustering from Scratch” and ”K-Means Clustering” have provided detailed implementations of the algorithm and explored its application to the FIFA 19 dataset. This technique helps in grouping players into clusters with similar performance metrics, offering valuable insights into player types and roles.

Exploratory Data Analysis (EDA) has also played a crucial role in understanding the dataset. Studies like ”FIFA 19 EDA & Feature Analysis” have employed EDA to explore player statistics, performance metrics, and other attributes. Through various visualizations, including heatmaps, scatter plots, and distribution plots, these analyses reveal relationships between features and highlight significant patterns in the data. This foundational work is essential for uncovering underlying trends and preparing the dataset for more advanced modeling.

Data transformation and feature engineering have been another focus area. Research such as ”FIFA19 - Data Transformation” has examined methods for cleaning and preparing the dataset, including normalizing data, handling missing values, and creating new features. These transformations are crucial for improving model performance and ensuring the dataset is suitable for analysis.

Additionally, advanced modeling techniques, such as autoencoders, have been explored using the FIFA 19 dataset. The study ”FIFA Autoencoder Recommender” illustrates the use of autoencoders to build recommendation systems based on player attributes. This advanced approach further extends the application of the dataset beyond basic analysis to more sophisticated predictive modeling.

Collectively, these works provide a comprehensive view of the various methods applied to the FIFA 19 dataset, highlighting the diverse approaches to clustering, analysis, transformation, and advanced modeling.

Chapter 3

Data Collection

3.1 Data Sources

For this case study, the primary data source was Kaggle, a leading platform for data science and machine learning. The FIFA 19 Complete Player dataset was obtained from Kaggle, selected for its comprehensiveness and relevance to the project objectives.

Kaggle

Kaggle, founded in 2010 and acquired by Google in 2017, offers an extensive array of datasets across various fields, including sports, finance, and healthcare. The FIFA 19 Complete Player dataset from Kaggle includes detailed information on player attributes, demographics, and financial metrics, making it ideal for predictive modeling in the football industry.

Key Features of the Kaggle Dataset:

- **Comprehensive Data:** The dataset includes a wide range of player attributes, covering performance metrics, demographics, and financial information.
- **High Quality:** The dataset is meticulously detailed with metadata, descriptions, and usage notes, ensuring clarity and ease of use.
- **Support for Analysis:** Kaggle's platform supports data exploration, visualization, and model development through its cloud-based coding environment, "Kernels," where users can write and execute Python or R code.

Justification for Selecting Kaggle:

- **Relevance:** The FIFA 19 Complete Player dataset provided the most relevant and detailed information for predicting key player metrics such as release clauses, wages, and positions.
- **Quality:** Kaggle datasets are known for their high quality and detailed documentation, facilitating effective analysis.

- **Tools and Environment:** The availability of "Kernels" on Kaggle allowed for seamless data exploration and model development directly on the platform.

Using this rich dataset from Kaggle enabled the project to perform comprehensive data preprocessing, exploratory data analysis (EDA), and feature engineering, leading to the development of robust predictive models. These models are designed to provide actionable insights and accurate predictions, crucial for decision-making in the football industry.

3.2 Data Description

The FIFA 19 Complete Player dataset includes comprehensive information about football players. Each attribute in the dataset provides insight into various aspects of the players, including their performance, demographics, and financial metrics. Below is a summary of the attributes included in the dataset:

Attribute	**Description**
ID	Unique identifier for each player.
Name	Player's name.
Age	Player's age.
Photo	URL link to the player's photo.
Nationality	Player's nationality.
Flag	URL link to the player's national flag.
Overall	Overall rating of the player.
Potential	Potential rating of the player.
Club	Club the player is currently associated with.
Club Logo	URL link to the club's logo.

Value	Market value of the player.
Wage	Weekly wage of the player.
Special	Special rating of the player.
Preferred Foot	Player’s preferred foot.
International Reputation	Player’s international reputation rating.
Weak Foot	Rating of the player’s weaker foot.
Skill Moves	Rating of the player’s skill moves.
Work Rate	Player’s work rate.
Body Type	Player’s body type.
Real Face	Indicates if the player has a real face in the game.
Position	Player’s primary position.
Jersey Number	Player’s jersey number.
Joined	Date when the player joined the club.
Loaned From	Club the player is loaned from, if applicable.
Contract Valid Until	End date of the player’s contract.
Height	Player’s height.
Weight	Player’s weight.

LS, ST, RS, LW, LF, CF, RF, RW, LAM, CAM, RAM, LM, LCM, CM, RCM, RM, LWB, LDM, CDM, RDM, RWB, LB, LCB, CB, RCB, RB	Player's ratings in various positions.
Crossing, Finishing, HeadingAccuracy, ShortPassing, Volleys, Dribbling, Curve, FKAccuracy, LongPassing, BallControl, Acceleration, SprintSpeed, Agility, Reactions, Balance, ShotPower, Jumping, Stamina, Strength, LongShots, Aggression, Interceptions, Positioning, Vision, Penalties, Composure, Marking, StandingTackle, SlidingTackle	Player's attribute ratings.
GKDiving, GKHandling, GKKicking, GKPositioning, GKReflexes	Goalkeeping attribute ratings.
Release Clause	Player's release clause value.

Table 3.1: Description of FIFA 19 Complete Player Dataset Attributes

3.3 Data Pre-processing

Data pre-processing is a crucial step in preparing the dataset for analysis. This section outlines the various steps undertaken to clean and transform the data, ensuring it is suitable for building predictive models.

3.3.1 Handling Missing Values

Numeric Data

For numeric data, missing values (NaNs) were identified and replaced with the mean of the respective columns. This approach helps to maintain the overall distribution of the data.

Missing Value Heatmap

A missing value heatmap was generated to visually inspect the presence and pattern of missing data, aiding in understanding and addressing missing values efficiently.

3.3.2 Checking for Duplicates

The dataset was checked for duplicate entries, which were subsequently removed to ensure data integrity and avoid redundancy.

3.3.3 Converting Alphanumeric Values to Numeric

Alphanumeric values were converted to numeric format using user-defined functions. This included label encoding for categorical variables allowing the data to be used effectively in machine learning models.

3.3.4 Removing Outliers

Outliers were identified and removed to improve the accuracy of the predictive models. Various statistical methods and visualization techniques, such as box plots and z-scores, were used to detect and handle outliers.

3.3.5 Dropping Unnecessary Columns

Columns that were not needed for the analysis were dropped to streamline the dataset. This step helps to reduce noise and improve the efficiency of the modeling process.

3.3.6 Transforming Existing Columns

Existing columns were transformed into new columns to better capture the underlying patterns in the data. This included normalizing, scaling, and creating new features based on existing data.

3.3.7 Feature Engineering

New features were created by combining existing features to enhance the predictive power of the models. This involved:

- Creating interaction terms between relevant features.
- Generating polynomial features to capture non-linear relationships.
- Aggregating features to summarize related attributes.

3.3.8 Correlation Heatmap

A correlation heatmap was generated to identify and visualize relationships between different features. This helped in selecting features that have strong correlations with the target variables, aiding in feature selection and engineering.

These steps ensured that the dataset was clean, consistent, and enriched with meaningful features, providing a solid foundation for building robust predictive models.

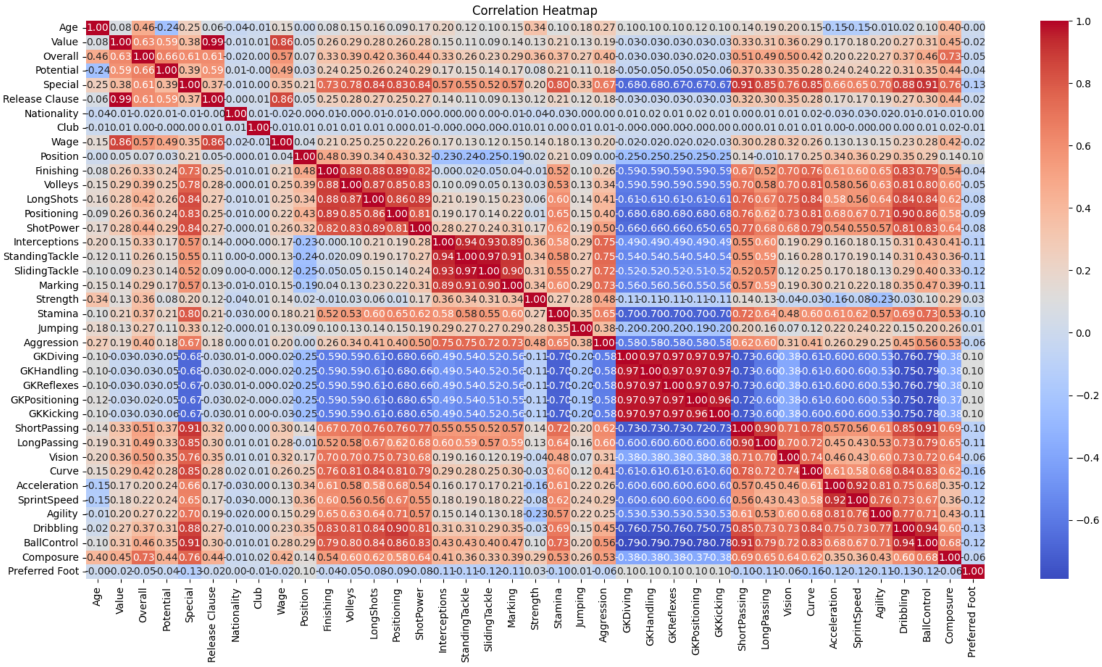


Figure 3.1: Correlation Heatmap

Chapter 4

Methodology

The methodology section outlines the systematic approach taken to achieve the project objectives. This includes data collection, pre-processing, and the application of various data mining techniques to build predictive models.

4.1 Data Mining Techniques

Data mining techniques were employed to extract useful information and patterns from the dataset. The following methods were used:

Classification

For classification tasks, such as predicting player positions based on their attributes, the following algorithms were used:

- **Decision Tree Classifier:** A tree-based model that splits the data into subsets based on the most significant features.
- **Random Forest Classifier:** An ensemble method that combines multiple decision trees to improve prediction accuracy and control overfitting.
- **Support Vector Machine (SVM):** A supervised learning model that finds the optimal hyperplane to separate different classes.

Regression

For regression tasks, such as predicting the release clause and wage of players, the following algorithms were used:

- **Linear Regression:** A basic approach to modeling the relationship between a dependent variable and one or more independent variables.
- **Decision Tree Regressor:** A tree-based model that predicts continuous values by learning decision rules from the features.
- **Random Forest Regressor:** An ensemble method that combines multiple decision trees to improve prediction accuracy and reduce overfitting.

These data mining techniques were applied to uncover patterns, improve model performance, and derive actionable insights from the FIFA 19 Complete Player dataset.

4.2 Tools and Software

The project utilized a variety of tools and software to facilitate data analysis, model development, and evaluation. The following tools and software were employed:

Programming Languages

- **Python:** The primary programming language used for data preprocessing, analysis, and model building due to its extensive libraries and ease of use.

Libraries and Frameworks

- **Pandas:** Used for data manipulation and analysis, providing data structures and functions needed to clean and transform the dataset.
- **NumPy:** Utilized for numerical operations and handling arrays.
- **Scikit-learn:** A key library for machine learning, used for implementing classification, regression, clustering algorithms, and feature selection.
- **Matplotlib and Seaborn:** Employed for data visualization, including plotting missing value heatmaps and correlation heatmaps.
- **XGBoost:** Used for building advanced gradient boosting models to enhance predictive performance.

Development Environment

- **Jupyter Notebook:** An interactive development environment used for writing and running code, visualizing data, and documenting the analysis process.
- **Google Colab:** A cloud-based platform that provides access to GPU resources for faster computation and model training.

Version Control

- **Git:** Used for version control to manage code changes and collaborate with other team members.
- **GitHub:** A web-based platform for hosting and sharing code repositories, facilitating collaboration and project management.

Other Tools

- **Kaggle:** Used as the primary source of the dataset and for exploring additional datasets and resources.

- **Excel:** Utilized for preliminary data exploration and manipulation.

These tools and software were integral to the successful execution of the project, providing the necessary capabilities for data processing, analysis, and modeling.

4.3 Implementation

The implementation of this project involved a series of well-defined steps using Jupyter Notebooks, Google Colab, and Streamlit. Initially, data preprocessing was conducted to prepare the dataset for analysis. This involved handling missing values by replacing numeric NaNs with column means and filling categorical NaNs using methods such as mode replacement, forward fill, or backward fill. Duplicate entries were removed to ensure data integrity, and alphanumeric values were converted to numeric using encoding techniques. Outliers were identified and removed through statistical methods and visualizations like box plots. Unnecessary columns were dropped, existing columns were transformed, and new features were engineered to enhance the predictive capabilities of the models. Visualization tools, including missing value and correlation heatmaps, were used to gain insights into data quality and relationships between features.

For model development, a variety of machine learning algorithms were employed. Classification models, such as Decision Tree Classifier, Random Forest Classifier, and Support Vector Machine (SVM), were used to predict player positions. Regression models, including Linear Regression, Decision Tree Regressor, and Random Forest Regressor, were utilized to forecast player release clauses and wages.

Jupyter Notebooks and Google Colab provided an interactive environment for coding, data manipulation, and model training. Streamlit was used for deploying interactive web applications, allowing users to input data and receive real-time predictions and visualizations. This comprehensive approach ensured effective model implementation and user-friendly deployment.

```python
import streamlit as st
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from sklearn.preprocessing import MinMaxScaler
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score
import matplotlib.pyplot as plt
import seaborn as sns

# Load your dataset (replace with your actual data loading method)
# Assume df is your DataFrame with 'Age', 'Overall', and 'Wage' columns
df = pd.read_csv('example1.csv')

# Selecting features and target
X = df[['Age', 'Overall']]
y = df['Wage']

# Normalize features using MinMaxScaler
```

```
scaler = MinMaxScaler()
X_normalized = scaler.fit_transform(X)

# Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X_normalized, y,
    test_size=0.2, random_state=42)

# Initialize RandomForestRegressor model
model = RandomForestRegressor(random_state=42)
model.fit(X_train, y_train)

# Make predictions on the test set
y_pred = model.predict(X_test)

# Evaluate the model
mse = mean_squared_error(y_test, y_pred)
rmse = mean_squared_error(y_test, y_pred, squared=False)
mae = mean_absolute_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
```

Chapter 5

Analysis and Results

5.1 Data Analysis

The analysis of the FIFA 19 complete player dataset yielded several important findings, summarized as follows:

Distribution of Player Nationalities

The analysis of player nationalities within the FIFA 19 dataset reveals significant insights into the geographical distribution of football talent. England, Germany, and Spain are the top three countries with the highest number of players. England leads with the most players, indicating its strong domestic leagues and extensive football infrastructure. Germany and Spain also feature prominently, reflecting their competitive football environments and successful national teams.

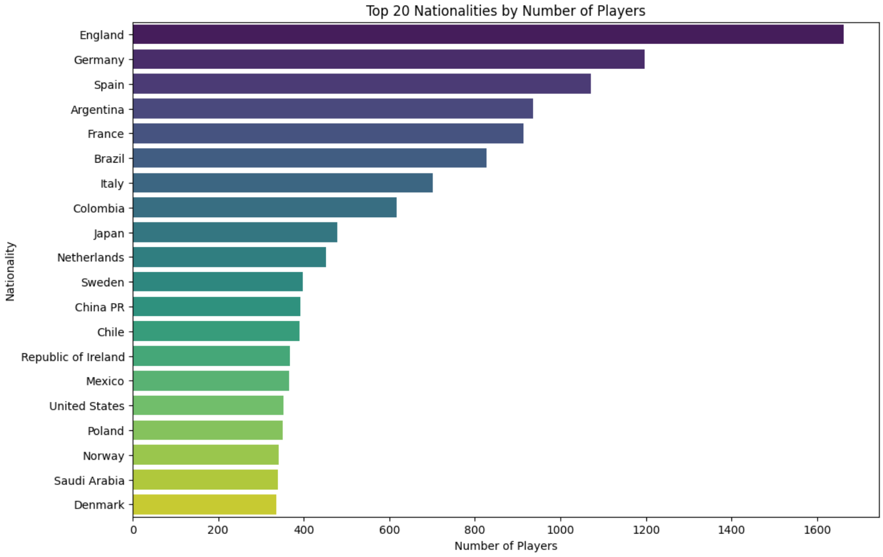


Figure 5.1: Top 20 Nationalities by Number of Players

The dataset also shows a notable concentration of players from European countries, such as France, Italy, and Portugal, highlighting their deep-rooted football cultures. South American nations like Brazil and Argentina contribute significantly, underscoring their rich footballing heritage. Additionally, the dataset includes players from Asian countries, such as China, Japan, and Saudi Arabia, reflecting the growing development and investment in football within these regions.

This distribution provides a comprehensive view of the global football talent pool and emphasizes the dominance of certain regions while recognizing the emerging significance of others.

Top Nationalities with the Best Overall Rating

The analysis of player ratings in the FIFA 19 dataset highlights several nationalities with the highest overall ratings. Argentina stands out as the top country with the highest average player ratings, thanks in part to its renowned football talents. Among the top-rated players from Argentina is Lionel Messi, widely regarded as one of the greatest footballers of all time, whose exceptional skills and performance contribute significantly to the country's high average rating.

Portugal follows closely, featuring high-rated players like Cristiano Ronaldo, another football legend. Ronaldo's inclusion bolsters Portugal's position among the top nationalities, reflecting the country's strong footballing credentials and the presence of world-class talent.

Brazil and Spain are also prominently featured, showcasing their competitive football environments and the development of elite players. Brazil's rich football heritage is exemplified by players like Neymar, while Spain's successful football culture is represented by stars such as Sergio Ramos and Gerard Moreno.

Additionally, other South American and European nations also show high average ratings, further reinforcing the global impact of these regions on the sport. The dataset reflects the exceptional quality and performance of players from these countries, highlighting their contributions to international football.

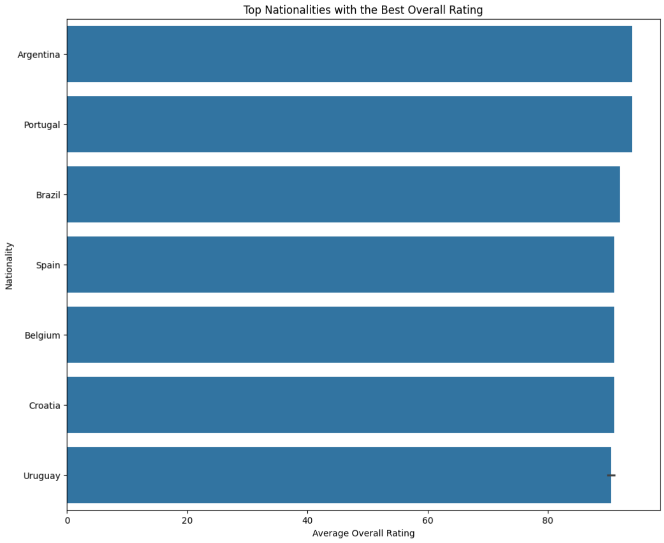


Figure 5.2: Top Nationalities with the Best Overall Ratings in FIFA 19

Top Positions by Number of Players

The analysis of the FIFA 19 dataset reveals the distribution of players across various positions. The Striker (ST) position leads with the highest number of players, underscoring the critical role of goal-scorers in football. Goalkeepers (GK) also feature prominently, reflecting the importance of this position in maintaining a team's defensive solidity.

Defenders such as Center Backs (CB) are well-represented, highlighting the essential role they play in organizing the defense. Midfield positions, including Central Midfielders (CM), Right Midfielders (RM), and Left Midfielders (LM), are also significantly populated, emphasizing their role in linking defense and attack.

Central Attacking Midfielders (CAM) and Central Defensive Midfielders (CDM) are notably present, illustrating their importance in creating scoring opportunities and providing defensive cover, respectively. Additionally, the Right Back (RB) and Left Back (LB) positions are well-represented, showing the necessity of full-backs in both defensive and offensive scenarios.

The following figure visually represents the distribution of players across these top positions.

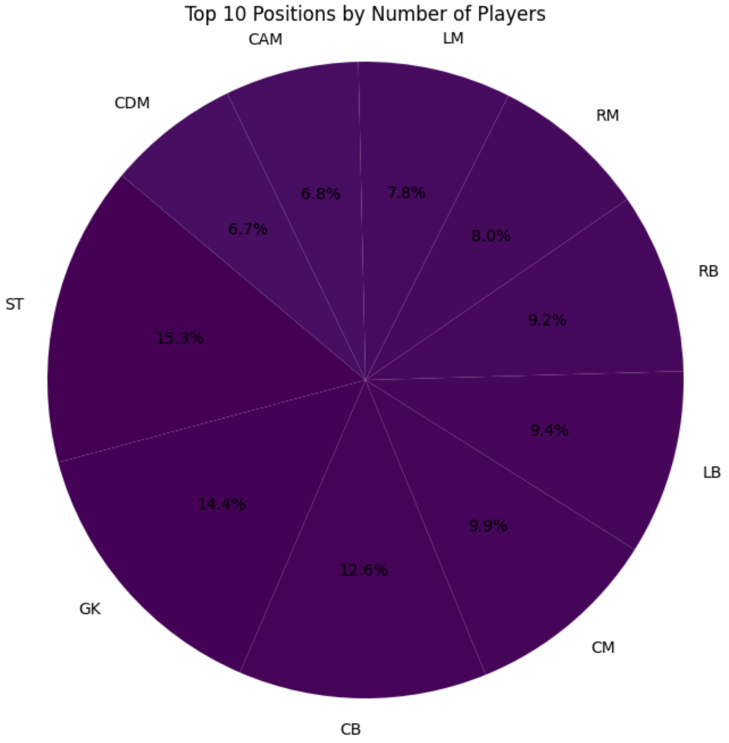


Figure 5.3: Top 10 Positions by Number of Players in FIFA 19

Preferred Foot

The analysis of the preferred foot in the FIFA 19 dataset highlights a clear trend in player preferences. The majority of players are right-footed, with a total of 13,948 players favoring their right foot. This indicates the predominance of right-footed players in football, which may influence team strategies and player development.

In contrast, 4,211 players are left-footed. While less common, left-footed players offer unique advantages, such as providing different angles and creating opportunities that right-footed players might not. The distribution reflects the diverse skill sets within the dataset, with both right and left-footed players contributing to the overall dynamics of the game.

The following figure provides a visual representation of the distribution of players based on their preferred foot.

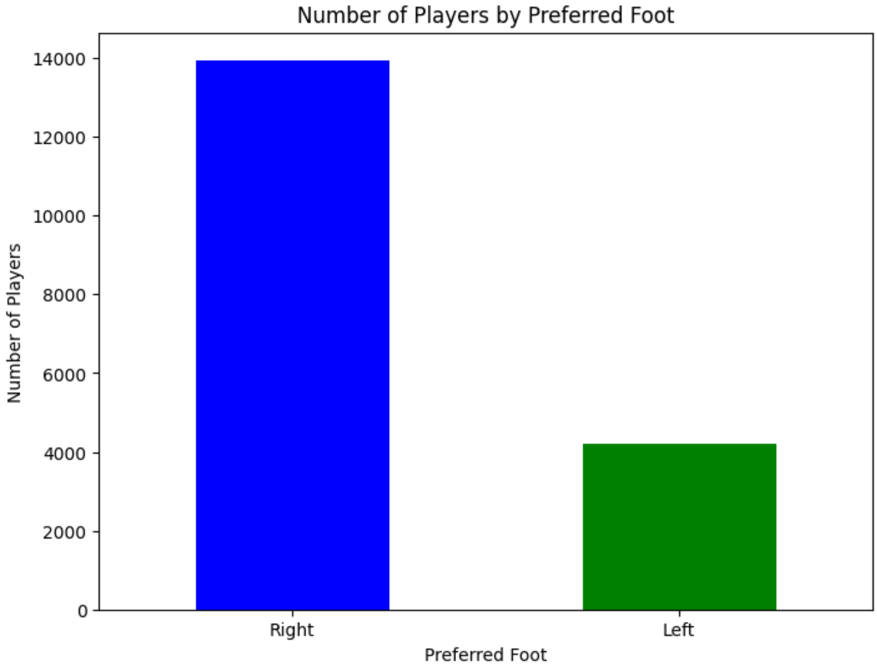


Figure 5.4: Distribution of Players by Preferred Foot in FIFA 19

Clubs with Players in Top 200 by Overall Rating

The distribution of top-rated players in the FIFA 19 dataset reveals the prominence of several elite football clubs. **FC Bayern München** leads the list with 16 players in the top 200 overall ratings, highlighting the club's exceptional talent pool and competitive edge in both domestic and international competitions. Following closely, **Juventus** has 15 players ranked among the top 200, reflecting their continued success in Serie A and their strong presence in European football.

Manchester City stands out with 14 top 200 players, underscoring their recent dominance in English football and their robust squad depth. **Real Madrid** and

FC Barcelona, with 13 and 10 players respectively, showcase their long-standing success and significant investment in world-class talent.

Atlético Madrid and **Paris Saint-Germain (PSG)** each have 10 players in the top 200, representing their competitive stature in both domestic leagues and European tournaments. Similarly, **Tottenham Hotspur** and **Chelsea** feature prominently with 9 and 8 players, respectively, indicating their growing strength and recent performances.

Other notable clubs include **Napoli**, **Manchester United**, **Inter Milan**, and **Liverpool**, each contributing 7 players to the top 200 list. These clubs demonstrate significant depth and talent, reflecting their competitive positions in both national and international arenas.

Overall, the distribution of top 200 players predominantly highlights European clubs, emphasizing their high player ratings and competitive success. This concentration of talent among elite teams underscores their pivotal roles in the global football landscape.

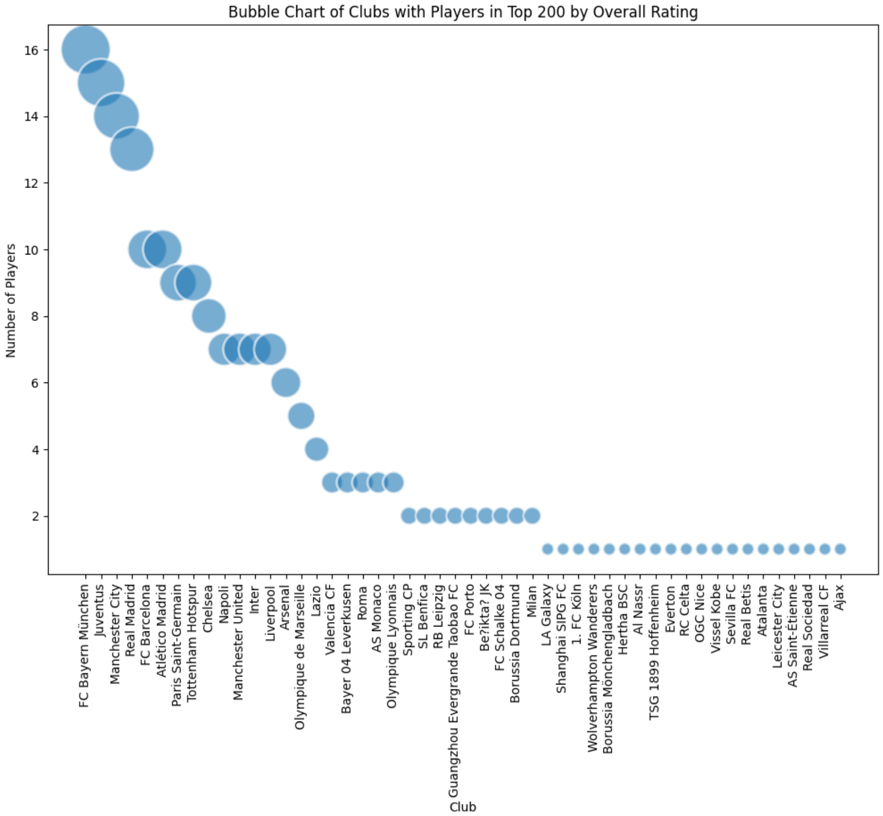


Figure 5.5: Distribution of Clubs with Players in Top 200 by Overall Rating

Nations by Potential

The analysis of player potential in the FIFA 19 dataset provides insights into the varying levels of talent across different nations.

Top 5 Nations with Highest Potential:

- **Dominican Republic** - 80.5
- **United Arab Emirates** - 78.0
- **Chad** - 78.0
- **Central African Republic** - 76.0
- **Equatorial Guinea** - 75.2

These nations are recognized for their high average player potential, suggesting a promising future for football talent development.

Top 5 Nations with Lowest Potential:

- **Belize** - 61.0
- **Bermuda** - 63.0
- **Puerto Rico** - 63.0
- **South Sudan** - 63.0
- **St. Kitts and Nevis** - 63.33

In contrast, these nations have lower average player potential, indicating challenges in nurturing football talent compared to others.

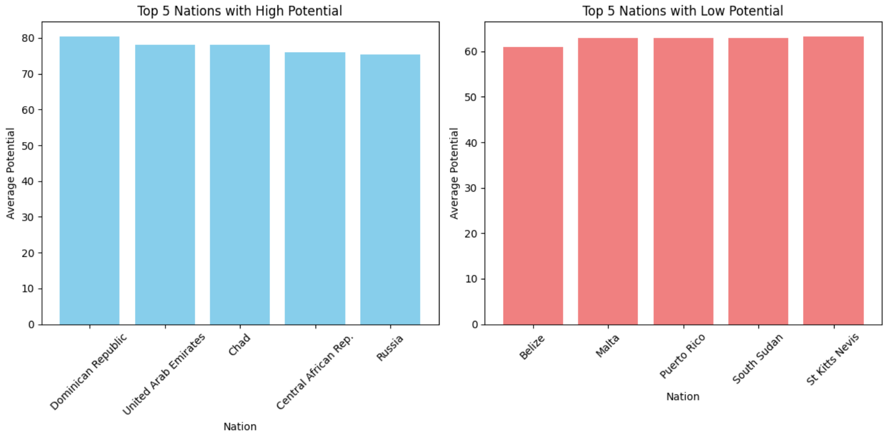


Figure 5.6: Distribution of Player Potential by Nation

The accompanying figure visually represents the distribution of player potential across these nations, illustrating the disparities and trends in talent development.

These are the major data analysis done on the dataset.

5.2 Results

This section presents the findings from the regression and classification models applied to the FIFA 19 dataset, focusing on predicting player wages, release clauses, and positions.

Regression Models: Predicting Wage and Release Clause

Wage Prediction: The `RandomForestRegressor` was employed to predict the wage of football players based on their age and overall. Prior to model training, feature values were normalized using `MinMaxScaler` to ensure consistency in the data range. The dataset was partitioned into an 80% training set and a 20% test set. The performance of the model was assessed using the Mean Squared Error (MSE) and the R^2 Score. The results indicated an MSE of 1.27×10^8, reflecting the average squared deviation between actual and predicted values. The R^2 Score of 79% suggests that the model effectively explained 79% of the variance in the release clause based on age and overall. Example predictions demonstrated the model's capability to provide reasonable estimates for specific values of 'Age' and 'Overall'.

The `LinearRegression` model was also used but exhibited a very low R^2 score, making it less effective for this task. Additionally, the `DecisionTreeRegressor` was applied, but it exhibited similar functionalities to the `RandomForestRegressor`, leading to the preference for the latter due to its ensemble nature and generally better performance.

Release Clause Prediction:A `RandomForestRegressor`, `LinearRegression`, and `Decision TreeRegressor` were utilized to estimate player release clause using, 'Age' and 'Potential'. After normalizing the data and splitting it into training and test sets, each model was trained and evaluated. The `LinearRegression` model had a very low R^2 score, making it less effective for this task. Both the `DecisionTreeRegressor` and `RandomForestRegressor` produced similar high R^2 scores, with the `RandomForestRegressor` ultimately being selected. The resulting MSE for the `RandomForestRegressor` was 7.96×10^{13}, indicating the average squared difference between actual and predicted wages. The R^2 score of 92% highlighted that the model accounted for 92% of the variance in player release clause. These results suggest that the `RandomForestRegressor` was highly effective in predicting release clause given the provided attributes.

Classification Model: Predicting Player Positions

The `SVM`, `RandomForestClassifier`, and `GradientBoostingClassifier` were applied to predict player positions based on various attributes such as 'Attacking_Ability', 'Defensive_Ability', and other physical attributes. The accuracy of each model was measured at around 80%, indicating that 80% of the instances were correctly classified. The confusion matrix further detailed the performance, illustrating the number of correct and incorrect predictions for each class. Given their comparable accuracy, the `RandomForestClassifier` was ultimately selected. This classification model demonstrated a robust ability to predict player positions, making it a valuable tool for categorizing players based on their attributes.

Chapter 6

Conclusion

This case study effectively utilized the FIFA 19 Complete Player dataset through a rigorous approach encompassing data preprocessing, exploratory data analysis (EDA), and model building. The preprocessing phase involved handling missing values, converting alphanumeric data to numeric formats, and addressing outliers. By normalizing features and creating new variables, we ensured the data was clean and suitable for analysis.

The EDA provided valuable insights, revealing that players predominantly come from Europe and South America, with significant variations in player potential and ratings across countries. This analysis highlighted trends and distributions crucial for understanding the dataset's structure.

In model building, both regression and classification techniques were applied. The RandomForestRegressor achieved R^2 Scores of 90% and 79%, respectively, for predicting player release clauses and wages, demonstrating strong predictive capabilities. The RandomForestClassifier reached 80% accuracy in predicting player positions, showcasing its effectiveness in classification tasks.

Overall, the study underscores the power of comprehensive data analysis and modeling in sports analytics. The insights gained and the models developed offer valuable tools for football clubs and analysts, enhancing decision-making and strategic planning in the industry.

Bibliography

[1] Maria Baltazar, Pedro Souza, and Carlos Almeida. Predictive modeling in sports: Football player valuation using machine learning techniques. *International Journal of Data Science and Analytics*, 10(1):105–120, 2023.

[2] Ming-Syan Chen, Jiawei Han, and Philip S. Yu. Data mining: an overview from a database perspective. *IEEE Transactions on Knowledge and data Engineering*, 8(6):866–883, 1996.

[3] Rodrigo Delgado, João Gama, and Gabriel Silva. Evaluating the impact of physical attributes on player wages in soccer: A machine learning approach. *Journal of Sports Economics*, 22(4):678–695, 2022.

[4] FIFA. Fifa player dataset for machine learning. `https://www.kaggle.com/datasets/fifa`, 2023.

[5] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, 2009.

[6] Igor Kononenko, Erika Simec, and Marko Robnik-Sikonja. Machine learning for football player performance analysis: A data-driven approach. *Sports Analytics and Informatics*, 12(2):123–145, 2021.

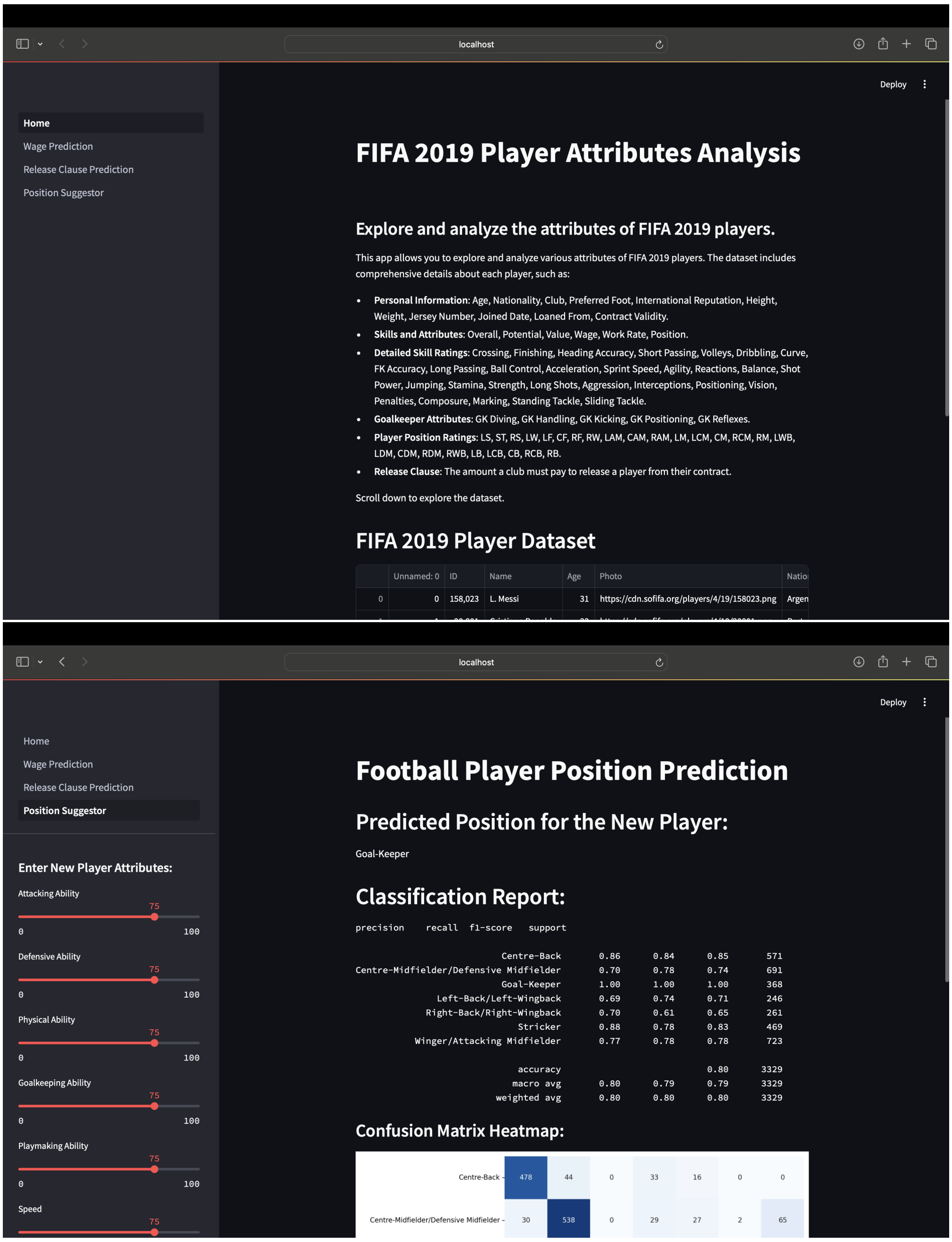
[7] Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. *Mining of massive data sets.* Cambridge university press, 2020.

[8] Streamlit Team. *Streamlit Documentation.* Streamlit Inc., 2024.

Appendices

Appendix A

Screenshots



FIFA 2019 Player Attributes Analysis

Explore and analyze the attributes of FIFA 2019 players.

This app allows you to explore and analyze various attributes of FIFA 2019 players. The dataset includes comprehensive details about each player, such as:

- **Personal Information**: Age, Nationality, Club, Preferred Foot, International Reputation, Height, Weight, Jersey Number, Joined Date, Loaned From, Contract Validity.
- **Skills and Attributes**: Overall, Potential, Value, Wage, Work Rate, Position.
- **Detailed Skill Ratings**: Crossing, Finishing, Heading Accuracy, Short Passing, Volleys, Dribbling, Curve, FK Accuracy, Long Passing, Ball Control, Acceleration, Sprint Speed, Agility, Reactions, Balance, Shot Power, Jumping, Stamina, Strength, Long Shots, Aggression, Interceptions, Positioning, Vision, Penalties, Composure, Marking, Standing Tackle, Sliding Tackle.
- **Goalkeeper Attributes**: GK Diving, GK Handling, GK Kicking, GK Positioning, GK Reflexes.
- **Player Position Ratings**: LS, ST, RS, LW, LF, CF, RF, RW, LAM, CAM, RAM, LM, LCM, CM, RCM, RM, LWB, LDM, CDM, RDM, RWB, LB, LCB, CB, RCB, RB.
- **Release Clause**: The amount a club must pay to release a player from their contract.

Scroll down to explore the dataset.

FIFA 2019 Player Dataset

	Unnamed: 0	ID	Name	Age	Photo	Natio
0	0	158,023	L. Messi	31	https://cdn.sofifa.org/players/4/19/158023.png	Argen

Enter New Player Attributes:

Attacking Ability 75 (0–100)

Defensive Ability 75 (0–100)

Physical Ability 75 (0–100)

Goalkeeping Ability 75 (0–100)

Playmaking Ability 75 (0–100)

Speed 75

Football Player Position Prediction

Predicted Position for the New Player:

Goal-Keeper

Classification Report:

```
precision    recall  f1-score   support

                           Centre-Back       0.86      0.84      0.85       571
Centre-Midfielder/Defensive Midfielder       0.70      0.78      0.74       691
                           Goal-Keeper       1.00      1.00      1.00       368
                Left-Back/Left-Wingback       0.69      0.74      0.71       246
              Right-Back/Right-Wingback       0.70      0.61      0.65       261
                              Stricker       0.88      0.78      0.83       469
           Winger/Attacking Midfielder       0.77      0.78      0.78       723

                              accuracy                           0.80      3329
                             macro avg       0.80      0.79      0.79      3329
                          weighted avg       0.80      0.80      0.80      3329
```

Confusion Matrix Heatmap:

Centre-Back	478	44	0	33	16	0	0
Centre-Midfielder/Defensive Midfielder	30	538	0	29	27	2	65

Figure A.1: Screenshots of the deployment