Analysis of FIFA Players Dataset

Project Team Members
1. Atharva Vidyadhar Jadhav (A20548559)
   ajadhav16@hawk.iit.edu
2. Sriujan Smrutin Harihar (A20548635)
   sharihar1@hawk.iit.edu
3. Harshali Vasant Gaikwad(A20538035)
   hgaikwad@hawk.iit.edu

1. Formal Project Description and Research Goal

The objective of this project is to analyse the FIFA Official Dataset to gain insights into various aspects of international football, including player performance, team dynamics, and historical trends. The primary research goal is to provide valuable information and generate actionable insights for various stakeholders in the football industry, including teams, coaches, scouts, and enthusiasts.

2. Specific Questions to Address

The project will seek to address the following specific questions:

- What are the key performance metrics that influence player ratings in FIFA, and how have they evolved over time?
- How do team statistics, such as possession, passing accuracy, and goal-scoring, correlate with a team's success in real-world competitions?
- Are there significant differences in player performance based on their positions, age, gender, nationality, and club affiliation?
- What are the historical trends in terms of player ratings and team rankings, and can these trends predict future outcomes?

3. Proposed Methodology/Approach

To address these questions, we will follow a multi-faceted approach, combining data analysis and machine learning techniques:

Data Collection:

- Acquire the FIFA Official Dataset, including player attributes, team statistics, and historical data.
- Gather additional data from reliable sources, such as match outcomes, transfer history, and tournament results.

Data Preprocessing:

- Clean and standardise the data to handle missing values, outliers, and inconsistencies.
- Feature engineering to create relevant variables for analysis, such as player performance indices.

Exploratory Data Analysis (EDA):

Diverse set of packages, functions and graphical methods will be used to explore the FIFA dataset, methods included bar chart to statistical heavy distribution fitting

- **Player Performance Analysis:**

  Cleaning of the Variables

  Frequency Analysis

  Association Analysis

- **Team Dynamics:**

  Player Interaction

  Player's contribution

  Tactical Insights

  Team Performance

- **Country:**

  Explore the country Variables

  Exploring the national team's performance

  Country's club performance

Machine Learning Models:

- Develop predictive models to assess player ratings and team performance.
- Utilise regression models to predict player ratings and evaluate feature importance.
- Implement clustering to group players based on performance attributes.
- Time-series analysis to identify trends in player ratings and team rankings over time.

Statistical Analysis:

- Conduct hypothesis testing to determine the significance of variables in influencing player ratings and team success.
- Employ data-driven techniques to identify patterns and outliers.

Evaluation:

- Use appropriate metrics to assess model performance and evaluate the quality of predictions.
- Cross-validation to ensure model robustness and generalisation to unseen data.

4. Metrics for Measuring Analysis Results

To measure the results and the quality of the analysis, the following metrics will be used:

- **R-squared (R²)**: To measure the goodness of fit for the regression models, assessing how well they explain the variance in player ratings and team performance.
- **Mean Absolute Error (MAE)** and **Mean Squared Error (MSE)**: To evaluate the accuracy of predictive models for player ratings.
- **Hypothesis Test P-values**: To gauge the statistical significance of variables in influencing player ratings and team success.
- **Visualisation Quality**: Assess the quality of data visualisations and their effectiveness in conveying insights.

This project will provide valuable insights into the world of football by leveraging the FIFA Official Dataset. The results will be presented through reports, visualisations, and interactive dashboards to make them accessible to a wide audience. The findings will benefit football teams, scouts, fans, and other stakeholders in making data-informed decisions and predictions.

5. Related Work
a. Predict the value of the Football players.
   https://ieeexplore.ieee.org/abstract/document/9721908
b. Prediction of international soccer match results.
   https://www.degruyter.com/document/doi/10.1515/jqas-2014-0051/html
   https://link.springer.com/article/10.1007/s42452-019-1821-5
c. Effect of individual player performance on match results.
   https://link.springer.com/article/10.1007/s00521-022-07178-5
d. Player performance prediction
   https://ieeexplore.ieee.org/abstract/document/8474750

6. Software packages, applications, libraries, and associated tools, etc.
- ➢ **Libraries/Packages:** ggplot2, dplyr, plotly, plyr will be used for the project.
- ➢ **Softwares:** RStudio, R, Python.
- ➢ **Project Management and Source Control:** Kaggle Dataset/Website, Some analysis pdfs and GitHub.