FIFA Dataset Analysis and Match Prediction

Aman Anand[1], Salil Tamboli[2], Vishwadeepak Singh Baghela[3], Prashant Johri[4]

Galgotias University, India
anandaman141@gmail. com,tambolisalil02@gmail.com, vdsbaghela@gmail.com, johri.prashant@gmail.com

Abstract::The world of football, with its vast player database and intricate gameplay, has been a subject of fascination for enthusiasts and analysts alike. The FIFA video game series, developed by EA Sports, provides an opportunity to delve into this world from a datascience perspective. In this project, we embark on a comprehensiveanalysis of FIFA player data using Python to uncover insights andpatterns that can enhance our understanding of the virtual footballworld. Data visualization is the process of representing data using common graphics, like charts, plots, infographics, and even animations, prediction of overall rating of football players using Linear Regression and Win prediction of EPL teams using Random Forest classifier.

Keywords: Linear Regression, Data visualization, Logistic Regression, Machine Learning

1. INTRODUCTION

In terms of players and spectators, football is the most popular sport in the world. $27 billion was estimated to have been made by European football clubs alone in 2017. It consequently becomes a vital component of the world economy. A growing number of football stars are in demand. significantly over the last several decades, and the worth of the Football players have e100 M+ or higher. These figures are far greater than trade data from the past, in comparison to the typical rate of inflation. Decision-makers now have access to enormous amounts of statistics in the data era. Big data are sets of data that are difficult to handle with standard tools and methods since they are not only large in size but also have great range and velocity. Owing to the rapid growth of this type of data, alternatives must be carefully considered and provided in order to manage and extract value and expertise.Able to use such diverse and suddenly changing data to gain priceless insights. Huge Records Analytics, which is the application of sophisticated analytics techniques on big data, can be used to provide such a fee. There are a variety of tools that can be used for storing and analyzing data. Some of the popular tools for storing data are:

a. Hive: Hive is a distributed Hadoop data management system. Because it provides query operations like Hive SQL for huge data access, it can be utilized for data mining.
b. Apache Hadoop: Massive volumes of data can be stored in a cluster using Apache Hadoop. It is a framework built on Java. It has the ability to process data across all nodes and can operate in parallel on a cluster. Data replication is made possible by this, increasing data availability.
c. Apache Cassandra: This database does not useSQL. It can manage massive volumes of data sinceit is scalable and features a high-performance distributed database.

Some of the popular tools for analyzing data are:

a. Tableau Public: Tableau Public is a user-friendly, straightforward tool that provides insightful data visualizations. A person can examine a theory, findthe information, and verify their conclusions.
b. Jupyter Notebook: This user-friendly tool facilitates end-to-end data science workflows, including data cleansing, statistical modelling, machine learning model construction and training,and data visualization.
c. Rapid Miner: Any number of information source types, such as Microsoft SQL, Sybase, IBM SPSS, Excel, Oracle, My SQL, Tera Data, IBM DB2, Ingress, can be included in RapidMiner.

1.1 DATA SETS

A grouping of related and comparable data or information is called a data set. It is set up to make

anentity more accessible. Because data sets offer relevant information in a unified format, they a reutilized in data analytics. It may or may not be organized. When a data set is said to be structured, it means that it is organized according to apredeter mined model or format. For example, a tabular data set is properly structured and contains information in the form of tables with rows, columns, and cells. An unstructured data set, in contrast to structured data sets, is primarily text-based and contains facts, figures, and other information. It is not organized in any redetermined format, such as tables. A data set type is selected based on the work in order to better meet the requirements. In this work, a structured data set with data organized into rows and columns is considered for future research. A structured data set in the form of rows and columns that was related to the FIFA World Cup wasused in this study. The data set includes information on players, teams, goals, matches, and other fields that are made up of columns with various data types.

1. Goals includes information about the player's name in the "Player Name" field; Id Match, which is the id assigned to the match; Team Name, which includes the names of the opposition teams, such as Portugal, the UK, and Russia; Goal Keeper, player shirt number ,and minutes.

2. The match information includes the following:

The match's ID, the home and away teams, attendance, the match day, the stage, the home and away teams' tactics, the penalty score for each team, the stadium name (which includes the name of the venue where the match was played), the temperature, humidity, wind speed ,and the winner of each match.

3. Players include each player's ID player, name, affiliated team name, birthdate, height, weight,and goals.

4. Teams have an ID team, a team name, a coach name linked to the team, and a coach country that denotes the nation of coach.

Importance of Data Analytics

It is crucial to specify the goals when analyzing data sets so thatthe next steps are obvious. We can ask questions about data thanks to analysis. It is crucial to collect data for data collectionpurposes, as this will inform future operations. Following the aforementioned actions, "Data Wrangling" becomes apparent. The process of raw data cleansing and conversion so that subsequent operations become easier to carry out and conclusions can be drawn from the results is known as data wrangling or data munging.

2. METHODOLOGY:

Linear Regression:

Making predictions based on observed data and investigating correlations between variables are made easier with the help of linear regression, a key idea in statistical modelling and machine learning. Basically, thegoal of linear regression is to create a linear relationship between a dependent variable and an independent variable, or set of independent variables. This methodology, when applied to FIFA 23, provides insights into the complex dynamics of virtual football performance by determining the relationship between individual player attributes and the overall player rating.

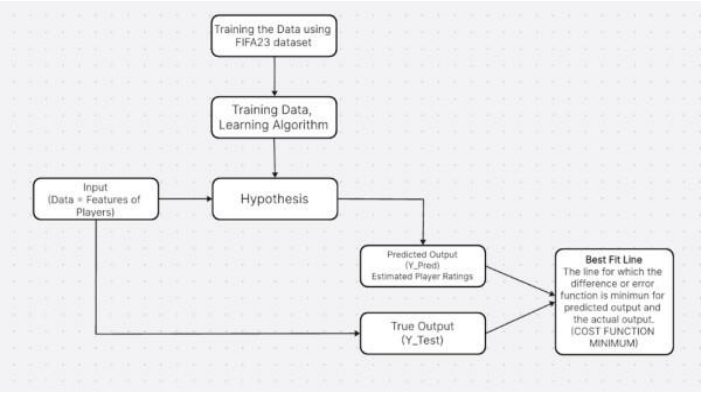

Figure 1. Flow chart Linear Regression

Linear Regression is Calculated by using best fit liney= B0 + B1* x,
And minimizing the cost function MSE (Mean Squared Error).

$$CostFunction(m, c) = \sum (Yi - (mxi + c))$$

1. Logistic Regression:
   - A statistical analysis technique called logistic regression uses previous observations of a data set to predict a binary outcome, such as yes or no.
   - A logistic regression model examines the relationship between one or more independent variables that are already present in order to predict a dependent variable in the data. To forecast whether a political candidate will win or lose an election or whether a high school student will be admitted to a specific college, for instance, one could use a logistic regression. Simple choices between two options are made possible by these binary results.
   - Logistic Regression has cost function inspired by the sigmoid function and finding the best fit line using thesame.
     The cost function for logistic regression is:

$$CostFunciton(h(x^2), y^2) = 1/m(\sum_{i=0}^{m} -yi(\log(h(xi) - (1 - yi)(\log(1 - h(xi))))))$$

2. Machine Learning:
   - A subfield of computer science and artificial intelligence (AI) called "machine learning" focuses onusing data and algorithms to simulate human learning processes and progressively increase their accuracy. An essential element of the expanding field of data science is machine learning. Algorithms are trained to make predictions or classifications and to find important information in data mining projects using statistical techniques. Subsequently, these insights inform business, which ideally influence important growth metrics.
3. Random forest classifier:
   - Popular machine learning algorithm Random Forest isa member of the supervised learning method. In machine learning, it can be applied to both classification and regression issues. It is founded on theidea of ensemble learning, which is the process of merging several classifiers to enhance the model's performance and solve a challenging issue.

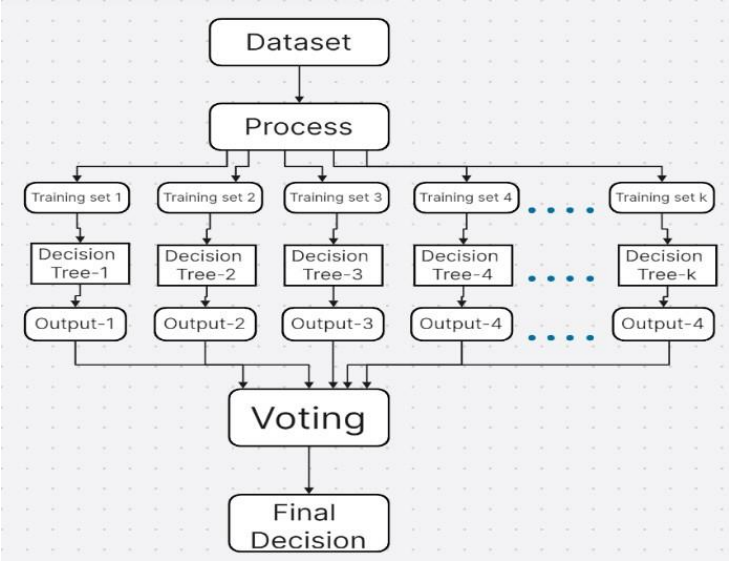

Figure 2. Flow chart of Random Forest classifier

- According to the name, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the

predictive accuracy of that dataset." Rather than depending just on a single decision tree, the random forest predicts the outcome by using the predictions from each tree and a majority vote system.

4. Objective Definition:
    - Before diving into data collection and analysis,we clearly defined our primary and secondary objectives. The overarching goal was tounderstand how FIFA ratings correlate with actual on-field performances of players.
5. Data Collection:
    - Primary Data: The primary dataset wascollected directly from the official FIFA gamedatabase for multiple years, ensuring the mostrecent and accurate player ratings.
    - Secondary Data: Secondary data sourcesincluded player performance metrics from various football databases, websites, andrepositories, offering insights into actual match performances.
6. Data Cleaning:
    - Given the vast amount of data, a significant portion of our methodology was devoted to cleaning and preprocessing. This involved:
    - Removing duplicates.
    - Handling missing values by either imputationor removal, depending on the context.
    - Converting all data into a consistent format.
7. Feature Selection:
    - Using domain knowledge and exploratory data analysis (EDA), we selected relevant features from the dataset that wouldcontribute to meaningful analysis. Variables like age, club, position, and specific skill ratings were deemed significant.
8. Statistical Analysis:
    - Basic statistical tests, including t-tests andANOVA, were conducted to identify significant differences in ratings between groups (e.g., forwards vs. defenders or premier league players vs. other leagues).
    - Correlation analysis was used to see how different player attributes relate to each otherand to their overall rating.
9. Predictive Modeling:
    - Using regression analysis and machine learning models,we tried to predict a player'sFIFA rating based on their real-world performance metrics.
    - Model performance was evaluated usingmetrics like Mean Absolute Error (MAE), R-squared value, and Root Mean Squared Error(RMSE).
10. Visualization:
    - Data visualizations played a crucial role in our methodology. Charts, graphs, and heatmaps were used extensively to represent findings and draw patterns.
11. Sensitivity Analysis:
    - We further refined our models by conducting a sensitivity analysis. This allowed us to see how small changes in input variables impacted our outcomes, ensuring our findings were not overly reliant on any single variable.
12. Peer Review:

- Before finalizing our report, we engaged in a peer review process. This involved sharing ourfindings with external experts in football analytics to gain feedback and insights, ensuring the quality and validity of our analysis.

3. RESULTS

We are making various analysis from the FIFA data set. These analysis are:

Analysis 1: Data Visualization WORK RATE OF PLAYERS:

Finding the work rate of players in FIFA: Also known asPlayer Work Rate. The pace at which a player performsboth defensive and offensive tasks on the field is knownas their player work rate. The Work Rate, which is ratedbetween low, medium, and high, describes how hard a player works to participate in attacks and defenses evenwhen they are out of position. The bar chart illustrates that the "Medium/Medium" work rate is predominant, with a staggering majority of players falling into this category. This suggests that most players in the game have a balanced offensive and defensive contribution during matches. The distribution offers a lens into the preferred in- game tactical behaviors and how they mirror real-world football dynamics, where a balanced approach to play is often the most prevalent

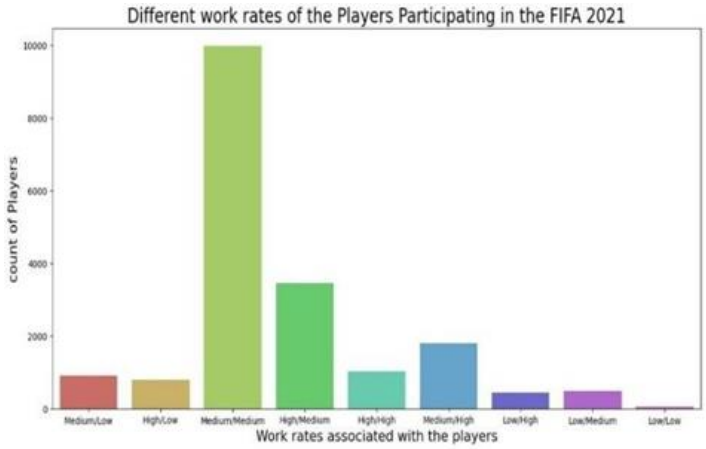

Figure 3. work rates of players

MESSI VS RONALDO SKILL SET:

The graph provides a head-to-head comparison between two football giants, Messi and Ronaldo, across various skill metrics. At first glance, both players exhibit high proficiency in numerous areas, with skill values often surpassing the 80 mark, indicating their world-class caliber.

- Both Messi and Ronaldo face a dip in one of their attributes which is defending as both the players play on a forward position.
- Messi excels remarkably in attributes such as dribbling, which is in alignment with his on-field agility and ability to navigate through tight defenses.
- Ronaldo, on the other hand, demonstrates a significant edge in the physicality domain, reflecting his robust built and athletic prowess. His strength in shot power and jumping is indicative of his aerial threat and powerful strikes, features often witnessed in his play.

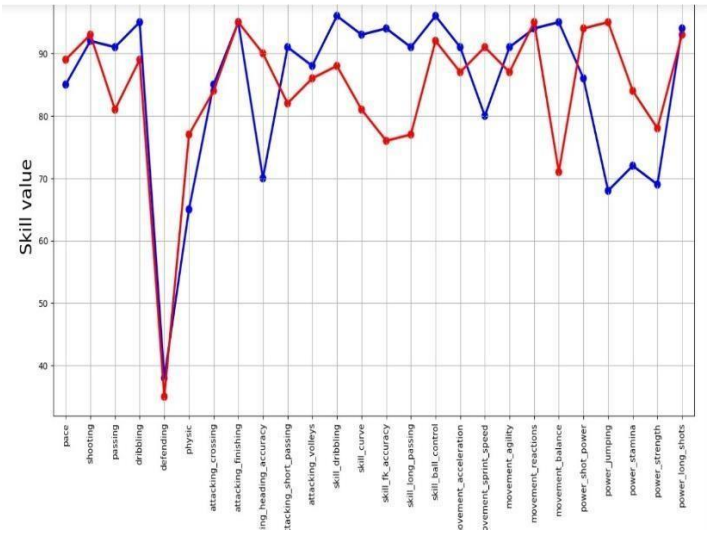

Figure 4. Skill value of Messi and Ronaldo

AGE DISTRIBUTION:

Age Distribution in top 5 clubs of Europe: Age distribution refers to how the ages of the players in the football (or soccer, in some countries) sport are distributed or arranged. This distribution can be used for a number of things, such as playerdevelopment, talent scouting, team composition, and understandingsport trends. It also offers insights into the demographics of footballplayers.

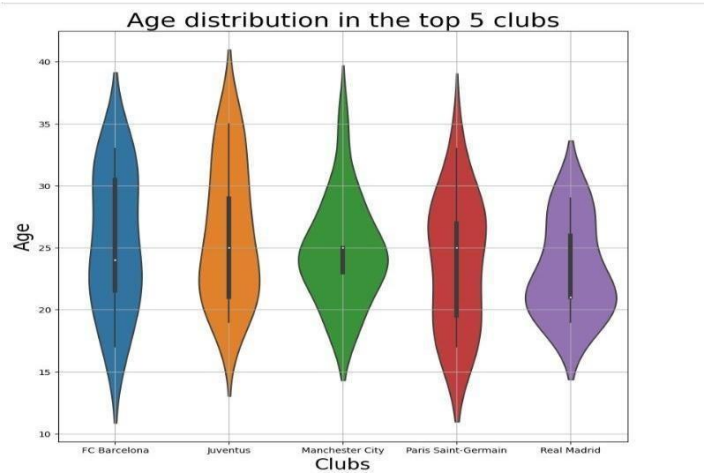

Figure 4. Age distribution of players

Analysis 2: Player Performance

Overall rating prediction of player using Linear Regression. For employing linear regression, the dataset is split into two parts one is training and the other one is testing in the ratio of 80:20.

```
features = ["pace", "shooting", "passing", "dribbling",
            "defending",  "defending"]
```

Figure 6. features used in player rating prediction

To discern the key attributes influencing a player's overall rating, we conduct a correlation analysis. Utilizing only numeric columns, we calculate the Pearson correlation coefficients between each attribute and the target variable, 'overall.' The scikit-learn library's Linear Regression class is employed to train the model on the training dataset. The trained model is then applied to the test dataset to predict overall ratings for a different subset of players. These predictions are compared with the actual ratings, and the results are meticulously analyzed.

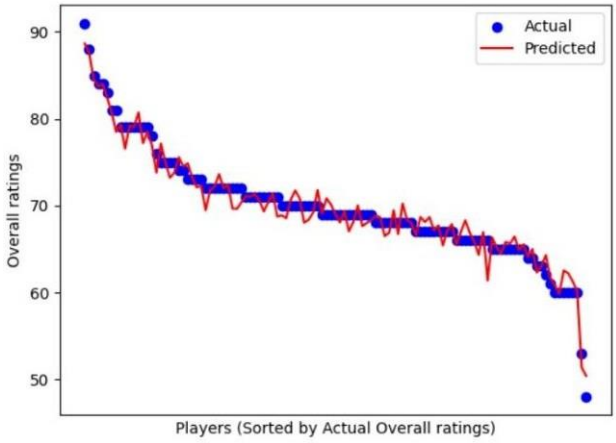

Figure 7. Players predicted rating and actual rating

Analysis 3: Match win Prediction

The model is used to predict the result of the matches and represent the accuracy in predicting the matches across all the matches in the dataset. These predictions are based on various factors such as historical data, team performance and players potential. The project's predictive model depicts very promising accuracy in predicting the match outcomes. The findings provide a foundation for ongoing improvements and insights into the factors influencing match results.

	actual_x	predicted_x	date	team_x	opponent_x	result_x	new_team_x	actual_y	predicted_y	team_y	opponent_y	result_y	new_team_y
0	0	0	2022-01-23	Arsenal	Burnley	D	Arsenal	0	0	Burnley	Arsenal	D	Burnley
1	1	0	2022-02-10	Arsenal	Wolves	W	Arsenal	0	0	Wolverhampton Wanderers	Arsenal	L	Wolves
2	1	0	2022-02-19	Arsenal	Brentford	W	Arsenal	0	0	Brentford	Arsenal	L	Brentford
3	1	1	2022-02-24	Arsenal	Wolves	W	Arsenal	0	0	Wolverhampton Wanderers	Arsenal	L	Wolves
4	1	1	2022-03-06	Arsenal	Watford	W	Arsenal	0	0	Watford	Arsenal	L	Watford
...	...	...	...	...	...	...	...	...	...	...	...	...	...
257	1	0	2022-03-13	Wolverhampton Wanderers	Everton	W	Wolves	0	0	Everton	Wolves	L	Everton
258	0	0	2022-03-18	Wolverhampton Wanderers	Leeds United	L	Wolves	1	0	Leeds United	Wolves	W	Leeds United
259	1	0	2022-04-02	Wolverhampton Wanderers	Aston Villa	W	Wolves	0	0	Aston Villa	Wolves	L	Aston Villa
260	0	0	2022-04-06	Wolverhampton Wanderers	Newcastle Utd	L	Wolves	1	0	Newcastle United	Wolves	W	Newcastle Utd
261	0	0	2022-04-24	Wolverhampton Wanderers	Burnley	L	Wolves	1	0	Burnley	Wolves	W	Burnley

262 rows × 13 columns

Table 1. Match Prediction model

4. CONCLUSION

Through our comprehensive analysis of FIFA player data, we have been able to delve deep into the intricacies and nuances that shape the world of football. Python appears to be a newly popular programming language these days and is growing quite a bit. Because of its many advantages—such as its extensive libraries, easy-to-learn syntax, enhanced readability, support for object- oriented programming, and integration—this language is becoming more and more versatile in a wide range of fields. Our investigation revealed patterns and trends, from the predominant work rates of players to the comparative skill sets of football legends. Furthermore, as the FIFA series continues to evolve, it will be intriguing to monitor shifts in player ratings and to see how emerging talents stack up against established stars. This project underscores the importance of data-driven approaches in understanding and appreciating the beautiful game.

References

[1]. M. A. Al-Asadi and S. Tasdemır, "Predict the Value of Football Players Using FIFA Video Game Data and Machine Learning Techniques," in IEEE Access, vol. 10, pp. 22631- 22645, 2022, doi: 10.1109/ACCESS.2022.3154767.Paul Z, Eaton C. Big data understanding: analytics for streaming data and enterprise class Hadoop. P. 1–166, McGraw-Hill Osborne Media, 2011.

[2]. Sharma, Mansi, Palak Mittal, Nidhi Garg, and Prateek Jain. "Analysis of FIFA World Cup Data Set." *Indian Journal of Science and Technology* 12 (2019): 39.

[3]. Karthik K. Trends in big data analytics. J Parallel Distr Com. 2014;74(7):2561–73.

[4]. D. Prasetio, ‘‘Predicting football match results with logistic regression,’’ in Proc. Int. Conf. Adv. Inform.: Concepts, Theory Appl. (ICAICTA), Apr. 2016, pp. 1–5.

[5]. R. Stanojevic and L. Gyarmati, ‘‘Towards data-driven football player assessment,’’ in Proc. IEEE 16th Int. Conf. Data Mining Workshops (ICDMW), Dec. 2016, pp. 167– 172.

[6]. X. Wu, V. Kumar, J. R. Quinlan, and J. Ghosh, ‘‘Top 10 algorithms in data mining,’’ Knowl. Inf. Syst., vol. 14, no. 1, pp. 1–37, 2008.

[7]. T. Pawlowski and C. A. Breuer Hovemann, ‘‘Top clubs’ performance and the competitive situation in European domestic football competitions,’’ J. Sports Econ., vol. 11, no. 2, pp. 186–202, 2010.

[8]. S. Majewski and U. Szczecin, ‘‘Identification of factors determining market value of the most valuable football players,’’ J. Manage. Bus. Administration. Central Eur., vol. 24, no. 3, pp. 91–104, Sep. 2016.

[9]. “A Data Science Approach to Football Team PlayerSelection”, P. Rajesh and Bharadwaj and Mansoor Alam and Mansour Tahernezhadi, {2020 IEEE International Conference on Electro Information Technology (EIT)}, year={2020}, pages={175-183} }

[10]. P. Rajesh, Bharadwaj, M. Alam and M. Tahernezhadi, "A Data Science Approach to Football Team Player Selection," 2020 IEEE International Conference on Electro Information Technology (EIT), Chicago, IL, USA, 2020, pp. 175-183, doi: 10.1109/EIT48999.2020.9208331.

[11]. ARAYA CORY, Albert Omar and AGUILAR FERNANDEZ, José Antonio. Visualization and analysis of Fifa 2021 players applying Machine Learning algorithms. *Rev. Inv. Est. I.* [online]. 2020, vol.12, n.1, pp. 13-34. ISSN 2415-2323.

[12]. Kaggle: FIFA-22 dataset retrieved from: Kaggle.com.

[13]. GFG(Geeks for Geeks) : Machine Learning. 12th February,2022. GFG:

[14]. https://www.geeksforgeeks.org/ml-linear-regression/