

---

# Predictive Analytics For FIFA Player Prices: An ML Approach

Sabina Anjum<sup>1</sup>, Asra Fatima<sup>2</sup>

<sup>1</sup>M. Tech Student, CSE Department, KBN University, Kalaburagi, India,  
sabinarahman169@gmail.com

<sup>2</sup>Assistant Professor, CSE Department, KBN University, Kalaburagi, India,  
fasra50@gmail.com

---

## ABSTRACT

---

Soccer extends beyond being a widely followed sport; it constitutes a flourishing industry. In the realm of player transfers, team managers face critical decisions regarding player valuation, transfer fees, and market values. Market values represent estimated prices that players can command in the football market, playing a crucial involvement in transfer negotiations. While football connoisseurs have traditionally been relied upon for market estimations, their judgments often lack accuracy and transparency. Thankfully, data analytics offer a promising alternative or supplementary methodology to expert-based player valuation, this research introduces an impartial numerical approach assessing market worth of players. Our method employs ML algorithms, specifically utilized with performance-related data extracted from [sofifa.com](http://sofifa.com). We conducted empirical investigations employing four regression models: multiple linear regression, random forests, linear regression, and decision trees, with the aim of assessing the market values of players. Additionally, our objective extends to data analysis to identify the pivotal factors impacting market value determination. Our empirical results indicate that the random forest algorithm surpasses other models in predicting market values of players. It attained the highest level of accuracy and exhibited the lowest error ratio in comparison to the baseline. These findings underscore efficacy of proposed methodology, outperforming established approaches documented in previous studies. Furthermore, we posit that our outcomes can exert a significant influence on negotiations among football players agents and clubs. Our model serves as robust foundation, streamlining the negotiation process and delivering an objective, quantitative estimation of a player's market value.

---

**Keywords**—Football, Player Value Prediction, Machine Learning, Data-Driven, Regression

---

## I. INTRODUCTION

Football, also recognized as soccer in some regions, ranks among the most universally practiced sports on a global scale. Numerous nations boast football players who contribute to local as well as international tournaments. Premier League garners unparalleled viewership, reaching 643 million households across 212 territories. According to a BBC report, the football betting market is estimated to range between \$700 billion and \$1 trillion. More recently, researchers have shifted their focus toward determining market value of players.

Forecasting systems for soccer matches carry substantial economic significance, extending beyond academic interest. The concept of players' market value has garnered considerable attention from researchers, particularly in recent times. Players' contracts are tradable commodities, allowing them to transition between teams, and their market value becomes a point of estimation in this context. While transfer fees indicate the actual transactional amounts, market values serve as estimates for these fees and play a central role in the negotiations surrounding player transfers.

Significantly, there has been a heightened focus on the assessment of players' market values in recent times. A player's transfer valuation is a computed approximation of monetary value at which a team can potentially transfer the player's contractual obligations to another team. While transfer costs directly denote transaction amounts involved, market values play a pivotal role in accelerating and influencing the course of transfer negotiations. In the realm of football expertise, encompassing team management and sports journalism, there has been a consistent recognition of the pivotal role that market values play. Crowdsourcing platforms such as Transfermarkt ([www.transfermarkt.com](http://www.transfermarkt.com)) have emerged as invaluable tools for approximating these market values in recent years.

Nevertheless, the utilization of data-driven methodologies for assessing market values has not witnessed widespread adoption within the football domain. In the context of data pertaining to electronic games, the dataset extracted from SOFIFA encompasses roughly 55 attributes associated with the proficiency of each player. This collective effort ensures that player data remains current and accurate. The SOFIFA dataset offers impartially assessed skill attributes for every player, delivering quantified proficiency information for more than 17,000 players. Its credibility unparalleled in realm of football-related data, accurately reflecting real player statistics.

Due to these distinct advantages, the SOFIFA dataset finds application in diverse research pursuits, such as predicting match outcomes, forecasting player market values, categorizing player positions, and projecting player performances.

## II. METHODOLOGY

The data flow diagram provides a clear overview of data movement, helping in understanding system functionality and communication. They are a vital tool for systems analysis, design, and documentation.

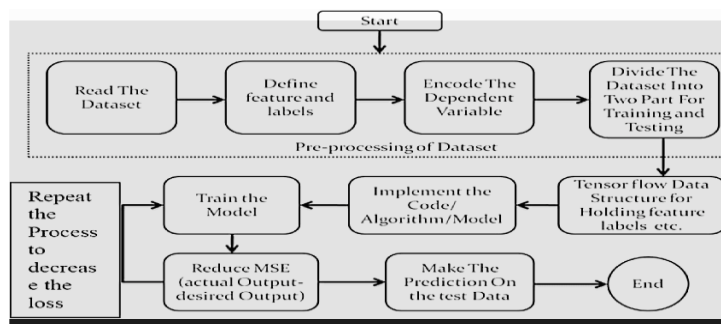


FIGURE 1. Data Flow Diagram.

The collection of training data becomes a streamlined endeavour with the aid of tools like Visual Studio Code (VSCode) and Python programming language. The approach employed in this experiment revolves around supervised ML Algorithm. In paradigm, the methods gain insights from detailed samples to construct a model, which is subsequently assessed using separate samples that weren't involved in its creation. These distinct test samples facilitate a comparison between the model's predicted outcomes and the genuine values, thus enabling an evaluation of the model's precision in forecasting real-world instances.

The approach outlined for this study encompasses the following sequential stages:

- Stage 1: Exploration of Influencing Factors on Players' Market Value:

In this initial stage, the focus is on predicting a player's market value. Concurrently, a comprehensive review of existing literature was undertaken to discern the factors influencing market value. This led to the identification of nine variables frequently recurring in the literature. Table presents the specific attributes employed within The investigation at hand.

- Stage 2: Data Preprocessing Approaches:

Data preprocessing, a pivotal facet within the domain of data mining, entails the conversion and preparation of data format for subsequent procedures. This encompasses a range of techniques including data cleansing, transformation, and reduction, all of which are essential for achieving precise machine learning outcomes. The dataset underwent a comprehensive cleansing and processing regimen, with only the most pertinent segments selected for model development.

• Stage 3: Initial Examination of Feature Subset:

To evaluate the effectiveness of the chosen subset of characteristics (shown in Step 1), an analysis upon their mutual relationships was conducted employing Pearson's correlation coefficient. The underlying rationale for employing this heuristic is as follows:

Effective feature subsets comprise attributes that exhibit strong correlation with the target class, yet display minimal correlation with each other. Moreover, the interplay between predictors (features) was also analysed. Attributes related to playing skills, such as shooting, passing, and dribbling, showed significant correlations among themselves but did not strongly correlate with player value. This situation necessitated potential attribute combination to alleviate Model intricacy or complexity. Passing exhibited the most robust correlation with player value within the realm of playing attributes. Consequently, it was selected for incorporation, while shooting and dribbling were omitted.

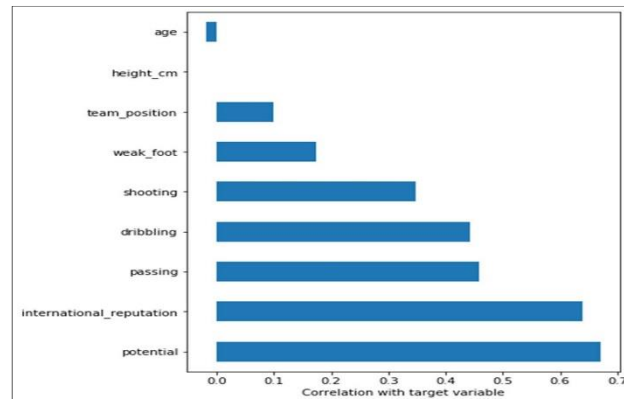


FIGURE 2. The correlation of numerical features.

• Stage 4: Comprehensive Analysis of Selected Features:

Subsequent to the determination of the refined feature subset (as shown in Step 3), the inclusion was subjected to validation through statistical significance assessment via linear regression analysis and decision trees. Ordinary Least Squares (OLS) regression models were deployed for experimental purposes. The analysis of variance provided preliminary insights into the correlation between predictors and the dependent variable. The value of P utilized. Notably, passing skills surpassed player height in importance, even though linear regression analysis indicated that passing skills lacked statistical significance.

• Stage 5: Data Segmentation:

Following data cleansing, feature subset definition, and logical validation, a random allocation of 80% of the data was designated for training and with the 20% allocated for the purposes of testing.

• Stage 6: Development of Models:

Within modelling phase, market value of the player was established for the function for 17,980 players dataset. Four regression models were employed to predict players' market values, and each model was tested with the complete set of features. Unless specified otherwise, default parameters were adhered to during model construction. Model outcomes were subsequently compared with actual values, affirming their suitability for this specific application.

• Stage 7: Model Evaluation:

To assess model performance, various metrics were computed, including Train and Test Split for estimating performance via training dataset. Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) and also Determination of Coefficient (R2) were employed to assess the performance of the regression models utilizing the testing dataset. The scikit-learn Python module was leveraged for the development of machine learning models.

### III. DESCRIPTION OF THE DATASET

The challenges associated with acquiring extensive and trustworthy data, coupled with the cost constraints highlighted in the Introduction, prompted the exploration of alternative data sources. In this context, this study turns to FIFA soccer video game data, a widely utilized resource in the literature. This dataset has shown its efficacy in predicting football match outcomes and has exhibited comparability or superiority to other football-related data sources. As such, we posit that outcomes from this video game dataset may be linked to real football player market transactions and broader analytical insights.

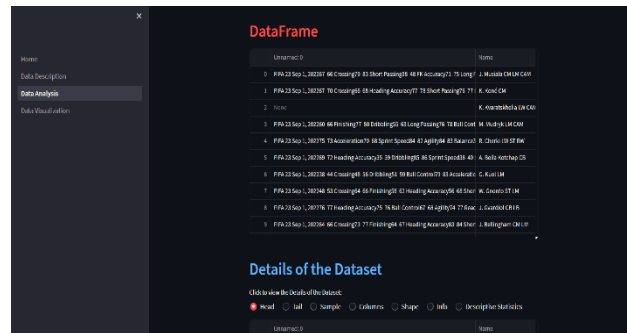


FIGURE 3(a). Data analysis and data set

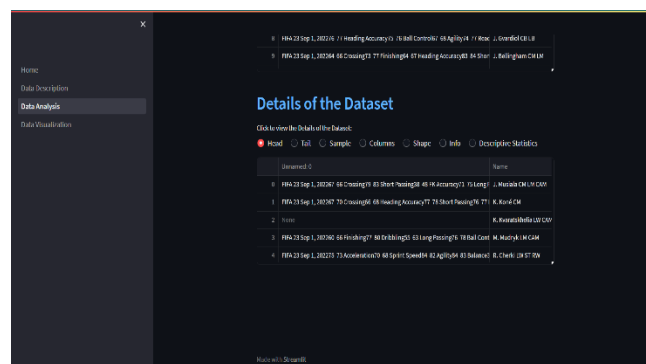


FIGURE 3(b). Data analysis and data set.

FIFA does not employ a definitive mathematical equation for determining the value of the player. Instead, scouts rely on their expertise and player ratings, introducing various biases into the market value determination process. Nevertheless, this dataset offers valuable quantitative insights into footballer performance. For the purposes of this study, the FIFA 20 dataset was utilized, comprising 17,980 instances, each representing an individual football player. These players are characterized by a comprehensive set of attributes, spanning distinct categories. These abilities are classified into three dimensions: physical, mental, and technical attributes. This data repository is readily accessible through the official game website (<http://sofifa.com/>).

### IV. DATA MODELLING

The pursuit of predicting player market values through attributes encapsulating football players' skills and traits was accomplished using four distinct supervised machine learning methods. These methodologies collectively aim to establish the optimal linkage between skill patterns and player market values. Considering that market value is a numeric entity, the chosen methodologies are tailored to predict continuous numerical results. The selected supervised ML techniques encompass a variety of paradigms, including:

**Linear Regression:** is a statistical method used for modelling the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. It is commonly used for predictive analysis and determining the extent to which independent variables affect the dependent variable.

Multiple Linear Regression: Building upon linear regression, multiple linear regression constructs a line considering multiple explanatory variables. The objective is the same: finding the line that optimally aligns with data points to minimize the RSS.

As a statistical method that aims to model the relationship between a dependent variable and multiple independent variables is generated, as depicted by the equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon \quad (1)$$

Where:

Y represents the response variable.

$X_1, X_2, \dots, X_p$  are the explanatory variables.

$\beta_0, \beta_1, \beta_2, \dots, \beta_p$  are the coefficients.

$\varepsilon$  represents the error term.

Aim is to find the line that best fits the dataset, representing an optimal the selection of a coefficient combination that minimizes the Residual Sum of Squares (RSS).

Regression Tree: As a predictive model employed within decision tree learning, this method progresses from observations (branches) to inferences about a target value (presented in the leaves). In situations where the target variable involves continuous values, these trees are specifically known as regression trees. Decision trees are widely recognized for their clarity and simplicity, making them a prevalent choice in the field of machine learning.

Random Forest Regression: Operating as an ensemble learning method, random forests create multiple decision trees during training and aggregate their outcomes. For classification, regression, and analogous tasks, these trees collectively yield a final class prediction (for classification) or an averaged prediction (for regression). While random forests often surpass individual decision trees in performance, their efficacy can be influenced by data characteristics.

These algorithms were chosen due to their wide prevalence for characterizing players and data mining realms. They represent both non-linear (e.g., decision trees) and linear methodologies (e.g., linear regression), allowing for a comprehensive performance comparison. The versatile nature of these machine learning techniques empowers this study's predictive endeavours, enhancing our understanding of the intricate relationship between player attributes and market values.

## V. DATA TRAINING

The assessment of model performance encompasses a range of metrics that provide insights into predictive accuracy. Among these metrics, The Test Split and Train technique has been utilized for the assess of Model effectiveness by employing separate data subsets. Furthermore, evaluation metrics such as MAE, RMSE, and Determination Coefficient ( $R^2$ ) are applied and appraise the regression models during the testing phase dataset. The versatility of player market value within the machine learning framework permits the exploration of various methodologies. In this context, it is treated as a regression problem, wherein market value prediction is based on performance of the players data. Within this research endeavour, we possess devised four regression techniques that leverage performance of players and skill data as characterized to establish foundational techniques for comparative analysis:

a. Test and Train Split:

A simplest yet insightful technique for assessing algorithm performance involves the division of data into distinct training and testing sets. This segregation enables training on one subset while predicting outcomes on the other, followed by an evaluation of predictions against actual outcomes.

b. Performance Metrics:

Every ML technique strives to tackle specific challenges by harnessing unique datasets. For regression problems, standard error measurements like MAE, MSE, RMSE, and  $R^2$  are employed to assess model performance. These metrics quantify the accuracy, precision, and explanatory power of the models.

Collectively, these evaluation techniques offer a comprehensive understanding of model performance. By combining established practices and specialized regression metrics, this study achieves a robust assessment of models' predictive capabilities.

A higher MSE value signifies poorer model performance. It is always non-negative, and an ideal model would yield an MSE of zero. RMSE introduced and ensure that error scales align with the scales of the target variable. RMSE retains the essence of MSE while providing results in a more interpretable unit, ensuring comparability.

Determination of Coefficient ( $R^2$ ) is an additional metric for assessment models, closely linked to both the model's MSE and the baseline's MSE. The baseline's MSE represents the simplest possible model, often involving predicting the mean of all samples. Interpreting  $R^2$ , an individual value near 1 individual value approaching zero signifies a model with minimal error, whereas a value in proximity to zero indicates that the model closely approximates the effectiveness of baseline. Each machine learning algorithm employed in this investigation was assessed using the coefficient of determination ( $R^2$ ).

## VI. RESULTS

Through the utilization of the methodologies delineated earlier, dataset outlined in Data Set Description and employing a variety of machine learning algorithms, the ensuing results are hereby presented:



FIGURE 6(a)

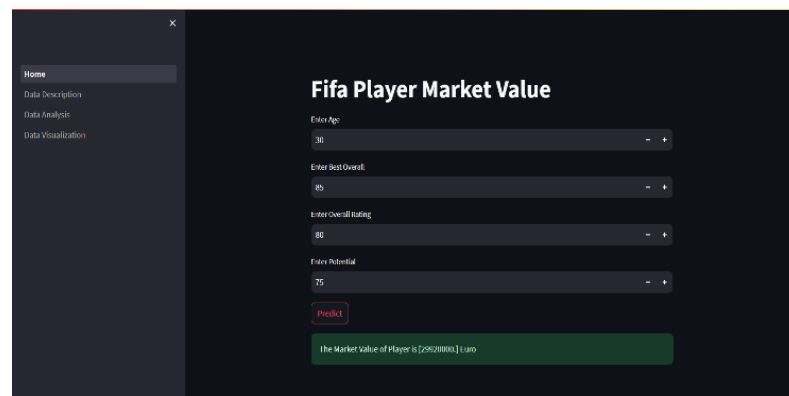


FIGURE 6(b)

FIGURE 6(a) and FIGURE 6(b) Shows the interpretation of results.

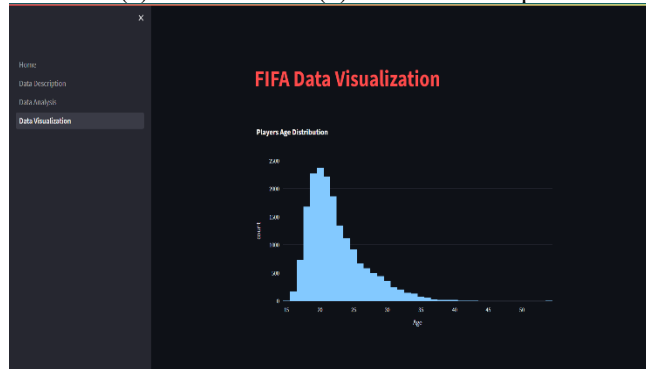


FIGURE 7(a)

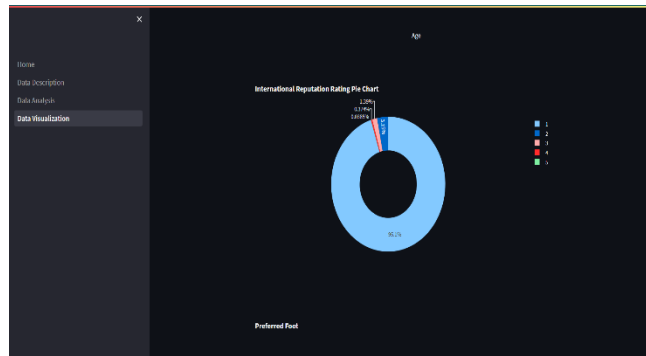


FIGURE 7(b)



FIGURE 7(c)

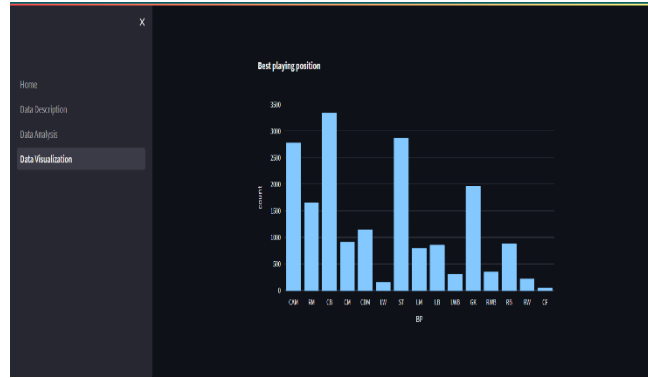
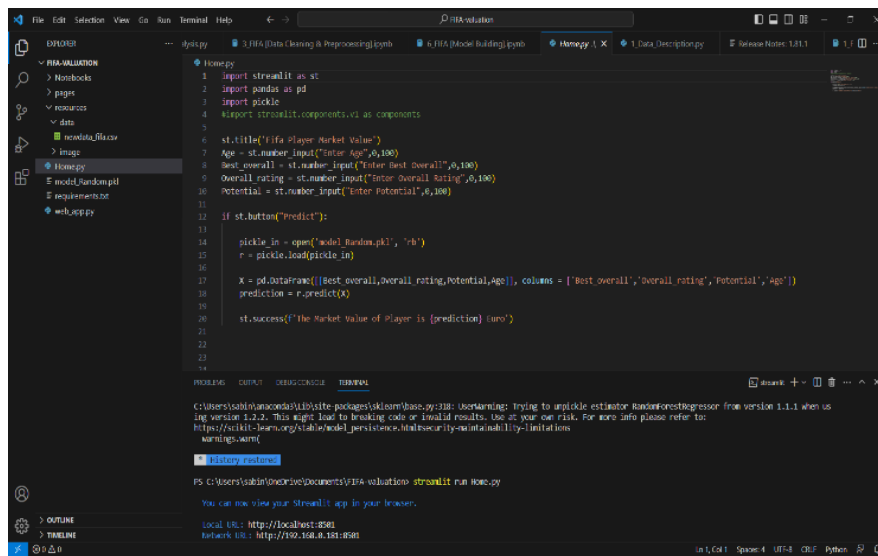


FIGURE 7(d)

FIGURE 7(a), (b), (c) and (d) Shows the data visualization.



```

1 import streamlit as st
2 import pandas as pd
3 import pickle
4 import sklearn.components as components
5
6 st.title('Fifa Player Market Value')
7 Age = st.number_input("Enter Age",0,100)
8 best_overall = st.number_input("Enter Best Overall",0,100)
9 Overall_rating = st.number_input("Enter Overall Rating",0,100)
10 Potential = st.number_input("Enter Potential",0,100)
11
12 if st.button("Predict"):
13
14     pickle_in = open('model_Random.pkl', 'rb')
15     r = pickle.load(pickle_in)
16
17     X = pd.DataFrame([Best overall,Overall rating,Potential,Age])
18     prediction = r.predict(X)
19
20     st.success("The Market Value of Player is {prediction} euro")
21
22
23
24

```

FIGURE 8. Shows the implementation

The findings elucidate the proficiency of a range of machine learning algorithms in the prediction of player market values. Notably, emphasis is placed on the evaluation of regression models, with specific attention given to the Random Forest Regression model, exhibit substantial improvements over the baseline model, underscoring their potential for robust predictive modelling in the realm of football player valuations.

## VII. CONCLUSION AND FUTURE WORK

The evolution of video game simulations to replicate football has exhibited remarkable strides over the last two decades. Moreover, substantial efforts have been dedicated to scrutinizing soccer players' skills and performances, fostering dependable game simulations that mirror the intricate dynamics of real soccer matches. Notably, FIFA datasets have showcased their efficacy in predicting match outcomes and conducting various analytical endeavours. The outcomes have consistently matched or even exceeded those obtained from alternative football data sources.

The experimental findings derived from our study unequivocally demonstrate the superiority of the suggested nonlinear approaches compared to contemporary methodologies is demonstrated. in the realm of predicting football players' market values. As a result, the contributions of this study transcend the realm of video game applications. They stand as a testament to the potency of the adopted methodology, surpassing conventional



approaches prevalent in the literature. This success is evident in tackling the same problem with identical data resources.

The ultimate objective is to construct a formidable team composed of exceptional players. Although FIFA 20 designates standard prices for in-game players, such valuations often face scepticism from the player community. Consequently, an imperative arises to ascertain player prices objectively, coupled with the foresight to anticipate price fluctuations prior to market transactions. This proactive approach facilitates the determination of the current average value at which a player is traded.

In subsequent research endeavours, the insights garnered from our study hold the potential to be harnessed for the development of a calculator integrated into the FIFA website. Such a tool would offer valuable assistance to video game players, potentially translating into financial gains. Moreover, we consider our findings to hold significant importance in the context of negotiations among football player and club's representatives. In brief, our example can serve a fundamental benchmark, optimizing the negotiation procedure and providing a quantitatively objective assessment of market rating of player.

Upon summation, our study stands as a testament to the significant advancements achievable through leveraging video game data to predict football player market values. Beyond the gaming context, the methodological superiority uncovered in our research reverberates across diverse domains. The potential for decision support tools and facilitating player-agent interactions accentuates the tangible implications of our findings, ultimately enhancing the precision, transparency, and efficacy of player valuation processes. As we conclude, these models offer a pragmatic baseline, simplifying negotiation processes and fostering an objective, quantitative estimation of a player's market value. Our research sets the stage for future exploration, where dynamic applications and innovations continue to reshape the landscape of football analytics.

## REFERENCES

1. Baboota, R., & Kaur, H. J. (2019). Predictive analysis and modeling of football results using a machine learning approach for the English Premier League. *Journal of Football Knowledge*, 35(2), 741-755.
2. Al-Asadi, M. A., & Tasdemir, S. (Year). Predicting the Value of Football Players Using FIFA Video Game Data and Machine Learning Techniques. Title of the Journal, Volume(Issue),
3. <https://ieeexplore.ieee.org/document/9721908>
4. <https://towardsdatascience.com/fifa-ultimate-team-rating-prediction-machine-learning-project-3a02767fcb38>
5. Vroonen, R. (2017). Predicting the potential of professional soccer players. In *Proceedings of the Machine Learning and Data Mining in Sports Analytics Workshop* (pp. 1-10).
6. Gacar, B. K., & Kocakoç, I. D. (2020). Regression analyses or decision trees? *Manisa Celal Bayar Üniversitesi Sosyal Bilimler Dergisi*, 18(4), 251-260.
7. Markovits, A. S., & Green, A. I. (2017). FIFA, the video game: A major vehicle for soccer's popularization in the United States. *Sport & Society*, 20(5-6), 716-734.
8. Prasetyo, D., & Harlili, D. (2016). Predicting football match results with logistic regression. In *Proceedings of the International Conference on Advanced Informatics: Concepts, Theory, and Applications* (pp. 1-5).
9. Al-Asadi, M. A., & Tasdemir, S. (2021). Empirical comparisons for combining balancing and feature selection strategies for characterizing football players using FIFA video game system. *IEEE Access*, 9, 149266-149286.
10. Siuda, P. (2021). Sports gamers practices as a form of subversiveness—the example of the FIFA ultimate team. *Critical Studies in Media Communication*, 38(1), 75-89.
11. Ho, T. K. (1995). Random decision forests. In *Proceedings of the 3rd International Conference on Document Analysis and Recognition* (pp. 278-282).
12. Prasetyo, D. (2016). Predicting football match results with logistic regression. In *Proceedings of the International Conference on Advanced Informatics: Concepts, Theory, and Applications* (pp. 1-5).