***FIFA 18 Player Performance Prediction and Anomaly Detection***

**1.Abstract**

Player performance analysis in football is essential for clubs, analysts, and the gaming industry to evaluate talent, optimize team strategies, and make informed financial decisions. The FIFA 18 dataset provides extensive player attributes, enabling advanced machine learning techniques to predict player performance and detect anomalies in wages or market value. This study develops a machine learning pipeline to forecast a player's future performance based on key attributes such as dribbling, passing, and agility. Additionally, clustering techniques are employed to identify players whose performance deviates significantly from market expectations, potentially revealing undervalued or overpaid players.

A comprehensive data preprocessing approach, including data cleaning, feature engineering, normalization, and dimensionality reduction, is implemented to enhance model efficiency. A Random Forest Regression model is used for performance prediction, while K-Means clustering is applied for anomaly detection. The models are evaluated using performance metrics such as Root Mean Squared Error (RMSE) and silhouette score to ensure accuracy and reliability. The results highlight the effectiveness of data-driven approaches in football analytics, aiding in talent scouting, player valuation, and financial management. This study demonstrates how machine learning can revolutionize decision-making in football by providing objective and actionable insights for clubs, managers, and analysts

**2.Introduction and Problem Formulation**

Football clubs, analysts, and gaming industries increasingly rely on data-driven insights to assess player performance and market value. The growing availability of player data from sources such as FIFA player databases enables the use of machine learning techniques to predict future performance and detect anomalies in player valuation. Traditional scouting methods rely on subjective assessments by coaches and analysts, which can be inconsistent and inefficient when analysing thousands of players. Big data techniques are used to process vast climate datasets for real-time decision-making, machine learning models can analyze extensive football player data to predict performance trends and detect market anomalies, optimizing recruitment strategies and financial decision-making in sports management.

(Al-Jarrah, O. Y., 2015)

**Problem Statement**

**1. Predicting Player Performance**

A player's overall performance is influenced by various attributes, including technical skills (dribbling, passing, finishing), physical characteristics (speed, stamina, strength), and mental aspects (composure, reactions, vision). Given a player's current attributes, this study aims to answer the following questions:

* Can a player’s overall performance be predicted based on attributes such as dribbling, passing, and agility?
* How do player characteristics correlate with future performance trends?  
  By leveraging machine learning models, we can identify the most important features that contribute to a player's overall rating and predict future performance levels with greater accuracy.

**2. Anomaly Detection in Player Market Value**

Player wages and market values are often influenced by multiple factors, including on-field performance, club reputation, league competitiveness, and transfer history. However, some players may be overpaid or undervalued due to subjective negotiations, media influence, or transfer market dynamics. This study seeks to answer:

* *Are there players whose wages or market value do not align with their actual performance?*
* Can machine learning help detect undervalued or overpaid players in the market?  
  To address these questions, we apply K-Means clustering to group players based on performance attributes and financial metrics. Players in small or distinct clusters may be potential anomalies, indicating market inefficiencies.

Machine learning is well-suited for these challenges due to its ability to process large-scale data, detect hidden patterns, and generate data-driven insights. In this study, we employ Random Forest Regression for player performance prediction and K-Means clustering for anomaly detection. These techniques allow us to make accurate predictions about future player performance and identify players who are overvalued or undervalued, aiding clubs, scouts, and analysts in making better financial and strategic decisions.

**3. Implementation**

This section outlines the end-to-end implementation process, from data preprocessing and feature engineering to model training and evaluation for player performance prediction and anomaly detection.

**1. Dataset Overview**

The FIFA 18 dataset consists of over 17,000 football players, each with 70+ attributes, covering personal information, skill ratings, and market value. The dataset was obtained from EA Sports’ FIFA 18 video game. *"A dataset is often depicted as a table, with each row representing an instance. Each column contains the value of one of the variables (attributes) for each of the instances."(Max Bramer.,2007).*  
The FIFA 18 dataset follows this structure, where each row represents an individual player, and columns include various attributes such as skill ratings, club affiliation, wages, and market value. By standardizing this data, we ensure compatibility with machine learning models.

The key attributes include:

* Personal Information: Name, Age, Nationality, Club
* Skill Ratings: Dribbling, Passing, Agility, Strength, Reactions, Vision, Stamina
* Market Data: Player Wages, Market Value
* Position Data: Preferred Playing Positions

These attributes are used for predicting player performance and identifying anomalies in wages and market value.

**2. Data Preprocessing**

Handling Missing Values

* Numerical attributes with missing values were replaced using the column mean.
* Categorical attributes (e.g., Club, Nationality) were filled using the mode (most frequent value).
* Columns with more than 50% missing values were dropped to prevent bias.

Feature Selection

From the 70+ features, only the most relevant attributes were selected for analysis:

*For Player Performance Prediction*

* Dribbling, Long Passing, Short Passing, Agility, Balance, Sprint Speed, Ball Control, Reactions, Stamina, Vision, Composure, Strength, Age, Potential
* These attributes were chosen based on their influence on overall performance.

*For Anomaly Detection*

* Overall Rating, Market Value, Wage, Strength, Sprint Speed, Dribbling, Positioning
* These were selected to identify players who may be undervalued or overpaid.

Data Normalization & Standardization

* Min-Max Scaling was applied to normalize values between 0 and 1.
* Standard Scaling was used to ensure that all attributes have equal variance.
* This helps improve model accuracy by preventing any single attribute from dominating.

Dimensionality Reduction using PCA

Principal component analysis (PCA) is a multivariate technique that analyses a data table in which observations are described by several inter-correlated quantitative dependent variables. (Abdi, H. and Williams, L. J., 2010).

* Principal Component Analysis (PCA) was applied to reduce the number of features while preserving the most important information.
* The best number of principal components (K) was determined using cross-validation

**3. Player Performance Prediction Model**

Model Selection:

A random forest is a classifier consisting of a collection of decision trees, where each tree is constructed by applying an algorithm on the training set with an additional random vector.( . Shalev-Shwartz, Shai., 2014)

* Random Forest Regression was selected due to its ability to handle non-linear relationships between player attributes and performance.
* It also provides feature importance scores, helping interpret the most influential attributes.

Feature Engineering:

The removal of these irrelevant and redundant features reduces the computational and storage costs without significant loss of information or negative degradation of the learning performance. (Li, J. and Liu, H., 2017).

* Players with similar skill sets were grouped to create new composite features.
* Attributes such as ball control, vision, and composure were weighted higher for performance prediction.

Training the Model:

* The dataset was split into 80% training and 20% testing.
* The Random Forest Regressor was trained on the selected features.
* Model performance was evaluated using Root Mean Squared Error (RMSE).

Hyperparameter Tuning:

This is relevant to your hyperparameter tuning and model evaluation process, where cross-validation (such as k-fold CV) is used to optimize the Random Forest and K-Means models, ensuring robust player performance predictions and anomaly detection.

* Grid Search Cross-Validation was used to optimize:
  + Number of trees (n\_estimators)
  + Maximum depth of trees
* The best configuration reduced RMSE significantly.

4**. Anomaly Detection using Clustering**

*Clustering Algorithm:*

*FIFA 18* dataset consists of various attributes of football players, which can be high-dimensional and complex, making it important to optimize the clustering process to detect meaningful patterns in performance vs. market value. ( Abiodun,M., Ikotun., 2023)

* K-Means Clustering was applied to group players based on performance and wage data.
* The optimal number of clusters (K) was determined using Silhouette Score evaluation.

*Hyperparameter Tuning*

Grid Search for Optimal K

* Cross-validation was performed by testing different values of K.
* The model was evaluated using the Silhouette Score to determine the best number of clusters.
* The range tested: K = [3, 4, 5, 6, 7].

*Evaluating Clusters:*

* The Silhouette Score was used to measure cluster quality.
* The highest score was obtained when K=5, meaning five distinct player groups.

*Identifying Anomalies:*

* The smallest cluster was flagged as anomalies.
* These players had unexpected wage-to-performance ratios:
  + Overpaid but underperforming players.
  + Undervalued but high-performing players.

**4.Evaluation**

The evaluation of this study focuses on assessing the performance of the machine learning models used for predicting player performance and detecting anomalies in market value. The models were evaluated using various metrics to determine their effectiveness in addressing the research questions.

**Player Performance Prediction**

The Random Forest Regression model was selected to predict a player's overall performance based on attributes such as dribbling, passing, agility, and stamina. The model was trained using 80% of the dataset and evaluated on 20% of the test data.

The performance of the model was assessed using Root Mean Squared Error (RMSE), a commonly used metric for regression tasks. Before hyperparameter tuning, the model achieved an RMSE of 3.0470, indicating a moderate error in predictions. After tuning the hyperparameters, including adjusting the number of trees and depth of the model, the RMSE improved significantly to 2.0744, demonstrating an enhanced ability to predict player performance accurately.

A graph of blue dots

AI-generated content may be incorrect.

Figure 1 : player performance prediction.

The scatter plot (Figure 1) of actual vs. predicted performance values further confirms the effectiveness of the model, as most data points align closely with the expected trend. However, some deviations indicate potential areas for further refinement, such as incorporating additional contextual data like player form, injuries, and team influence.

The scatter plot helps visualize how well the model predicts player performance, with points aligning along the diagonal indicating accurate predictions. By training on historical player data, the model can track performance trends over multiple seasons, identifying rising stars (above the line) or declining players (below the line). This approach allows teams to forecast future player ratings based on skill evolution, training impact, and aging effects

**Anomaly Detection in Market Value**

For anomaly detection, K-Means clustering was used to identify players whose market value and wages did not align with their performance. The clustering model grouped players into clusters based on attributes like overall rating, value, and wage.

To determine the optimal number of clusters (K), the Silhouette Score was used, which measures the quality of clustering. The best result was achieved with 4 clusters, yielding a Silhouette Score of 0.7144, indicating well-separated and meaningful player groups.

A graph with red dots

AI-generated content may be incorrect.

Figure 2: Anomalies in player performance.

A visualization of the anomalies (Figure 2) shows players flagged as outliers—those with either exceptionally high or low market values relative to their performance. These players could represent undervalued bargains or overpaid transfers, which is valuable information for club scouts and financial analysts.

Overall, the evaluation results demonstrate the effectiveness of machine learning in predicting player performance and detecting anomalies. The models provide valuable insights into player valuation, helping clubs make data-driven decisions in recruitment and contract negotiations.

**5.Discussion and Future Work**

**Discussion**

The results of this study highlight the effectiveness of machine learning models in analysing football player performance and market value. The Random Forest Regression model provided a reliable means to predict a player's future performance based on technical and physical attributes. The reduction in RMSE from 3.0470 to 2.0744 after tuning suggests that hyperparameter optimization plays a crucial role in improving predictive accuracy.

The anomaly detection component using K-Means clustering successfully identified outliers in market valuation, revealing players who may be undervalued or overpaid. A Silhouette Score of 0.7144 indicates a well-defined clustering structure, supporting the hypothesis that player wages and market value are not always aligned with actual performance.

However, some limitations were observed:

1. Static Data: The FIFA 18 dataset represents a single season, lacking temporal aspects such as player form fluctuations, injuries, and tactical role changes.
2. Feature Selection Constraints: The attributes considered for prediction may not capture the full scope of player performance, as team dynamics, coaching strategies, and playing style are not included.
3. Market Value Estimation Complexity: Factors such as player popularity, endorsements, and club financial health significantly influence market value but were not considered in the clustering model.

**Future Work**

To enhance the accuracy and applicability of this research, the following areas could be explored:

1. Time-Series Analysis: Incorporating historical performance data to capture player development trends over multiple seasons.
2. Advanced Deep Learning Models: Utilizing Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks to analyze sequential player data.
3. Integration of Real-World Factors: Expanding the dataset to include injury history, playing time, and tactical adjustments to refine predictions.
4. Hybrid Clustering Approaches: Combining unsupervised clustering with supervised models to improve anomaly detection accuracy.
5. Comparison with Real Market Trends: Validating the predicted anomalies against real-world transfers to assess the effectiveness of the model in detecting undervaluation or overpayment.

By addressing these aspects, future research could develop a more robust and dynamic system for player performance analysis and financial evaluation.

**6.Conclusion and Summary**

This study applied machine learning techniques to predict football player performance and detect anomalies in market value using the FIFA 18 dataset. The Random Forest Regression model was used for performance prediction, reducing RMSE from 3.0470 to 2.0744 after tuning, demonstrating improved accuracy.

For anomaly detection, K-Means clustering successfully identified outliers in player valuation, achieving a Silhouette Score of 0.7144, indicating well-defined clusters. These insights can help football clubs, analysts, and scouts make data-driven decisions regarding player recruitment and salary structuring.

Despite its success, the study faced limitations related to static data, limited feature selection, and real-world market complexities. Future work should focus on time-series analysis, deep learning approaches, and integrating real-world financial factors to enhance model accuracy.

Overall, this research demonstrates the power of machine learning in football analytics, providing valuable insights for performance forecasting and financial decision-making in professional football.

**References:**

1.[Al-Jarrah, O. Y., Yoo, P. D., Muhaidat, S., Karagiannidis, G. K., & Taha, K. (2015). Efficient Machine Learning for Big Data: A Review. Big Data Research, 2, 87-93.](https://canvas.mdu.se/courses/15546/files/2073163/download?wrap=1)

2. Max Bramer (2007). [Chapter 2: Data for Data Mining in the *Principles of Data Mining*](https://canvas.mdu.se/courses/15546/files/2073190/download?wrap=1)[Download Chapter 2: Data for Data Mining in the Principles of Data Mining](https://canvas.mdu.se/courses/15546/files/2073190/download?download_frd=1), 3rdEd. ISBN: 9781447173076 (online) 9781447173069 (print).

3. Li, J. and Liu, H. [Challenges of Feature Selection for Big Data Analytics](https://canvas.mdu.se/courses/15546/files/2073188/download?wrap=1)[Download Challenges of Feature Selection for Big Data Analytics](https://canvas.mdu.se/courses/15546/files/2073188/download?download_frd=1). IEEE Intelligent Systems, 2017, vol. 32, no. 2, pp. 9-15. doi:10.1109/MIS.2017.38.

4. Abdi, H. and Williams, L. J. (2010), [Principal component analysis](https://canvas.mdu.se/courses/15546/files/2073179/download?wrap=1). WIREs Comp Stat, 2: 433-459. doi:10.1002/wics.101.

5. James, Gareth., Witten, Daniela., Hastie, Trevor., Tibshirani, Robert.  [*An Introduction to Statistical Learning with Applications in R Links to an external site.*](https://static1.squarespace.com/static/5ff2adbe3fe4fe33db902812/t/6009dd9fa7bc363aa822d2c7/1611259312432/ISLR+Seventh+Printing.pdf). New York, NY: Springer New York , 2013. ISBN: 978-1-4614-7138-7. pg. number 176-181.

6. Shalev-Shwartz, Shai; Ben-David, Shai. [*Understanding machine learning : from theory to algorithms*Links to an external site.](https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/understanding-machine-learning-theory-algorithms.pdf). New York, NY: Cambridge University Press, 2014 ISBN: 978-1-107-05713-5.pg number 255-256.

7.Abiodun,M., Ikotun., Absalom,E. Ezugwu.,Laith, Abualigah ., Belal ,Abuhaija., Jia, Heming. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. [Information Sciences](https://www.sciencedirect.com/journal/information-sciences) [Volume 622](https://www.sciencedirect.com/journal/information-sciences/vol/622/suppl/C), April 2023, Pages 178-210.