Predicting Football Players' Performance Using the FIFA Database and Recurrent Neural Networks

Juan Sebastian Rodríguez Salazar
Universidad del Rosario
Bogotá, Colombia
Emails: juansebasti.rodrig10@urosario.edu.co, juansebastianrsj22@gmail.com

Alejandro Rafael Vega Saavedra
Universidad del Rosario
Bogotá, Colombia
Email: alejandror.vega@urosario.edu.co

***Abstract*—This project develops a model for predicting football players' performance using recurrent neural networks (RNN) and data from the FIFA database. Advanced data preprocessing techniques were implemented to create time sequences that capture the evolution of player performance over the years. Additionally, different evaluation metrics were used, including mean squared error (MSE), mean absolute error (MAE), and the coefficient of determination (R^2) to assess model accuracy. The results indicate an MSE of 3.70, an MAE of 1.42, and an R^2 of 0.90, demonstrating the RNN model's effectiveness in predicting player performance. This work highlights the application of neural networks in the analysis and prediction of sports data, suggesting potential improvements for future research.**

***Index Terms*—Machine Learning, RNN, Performance Prediction, Sports Models, FIFA**

I. Introduction

The analysis and prediction of football players' performance have become increasingly relevant in the sports industry, where artificial intelligence (AI) and machine learning (ML) tools have facilitated the development of new methods for talent identification and team performance optimization. In this context, predicting a player's future performance not only offers competitive advantages in tactical and strategic planning but also in scouting and youth talent development.

For this project, historical data from the FIFA video game was used due to its availability and accessibility, allowing the project to meet the required development timeline. Although this data does not fully represent real players' situations, the developed models could be adapted to work with real performance data, whether from youth or professional players in competitive environments, expanding the models' practical applications.

Initially, several regression and classification models were implemented, such as *Ridge Regression*, *Lasso Regression*, *Bayesian Ridge*, *Support Vector Regression (SVR)*, *Random Forest*, *Decision Tree*, and *Polynomial Regression*. However, these models exhibited overfitting issues, where the model fits too closely to the training data, resulting in reduced accuracy on the test set. Some of the mean squared error (MSE) values obtained include: Ridge Regression with 3.846×10^{-12}, Lasso Regression with 1.663×10^{-4}, and SVR with 1.123×10^{-3}, among others. Due to these issues, a recurrent neural network (RNN) model was chosen, as it is better suited for capturing temporal relationships in sequential data, allowing a more robust modeling of performance evolution over time.

The main objective of this project is to develop a predictive model based on RNN to forecast the future performance of football players using historical FIFA data. Advanced data preprocessing techniques and model evaluation methods were implemented to improve prediction accuracy and compare the performance of the RNN model with other approaches.

The main contributions of this work include the innovative use of FIFA data in sports performance predictive modeling, the design of an RNN model tailored to the particularities of sports data, and the implementation of evaluation metrics to measure the model's accuracy and effectiveness.

This document is structured as follows: **Section II** describes the methodology used, covering the dataset, preprocessing techniques, and models implemented; **Section III** presents the results obtained, evaluating model performance with specific metrics; and finally, **Section IV** concludes the work and suggests future directions for improving and expanding this type of research.

II. Methodology

A. Dataset

For this project, a dataset titled "FIFA 2021 to 2005 Complete Player Attributes" [1] was extracted from the Kaggle platform. This dataset contains detailed information on player attributes from 2005 to 2021, providing a historical view that allows the analysis of player performance evolution over time. Given the goal of implementing a predictive model, these FIFA data offered an accessible and extensive base. Although video game data does not fully reflect players' reality in competitive environments, the model could also be applied to real players' data, whether youth or professionals.

Individual .csv files for each year were concatenated through Python scripts, specifically using the file `funciones_de_limpieza.py`, which also performed initial data cleaning and generated a consolidated file, `cleaned_combined_fifa_data_filtered.csv`, with filtered data ready for preprocessing.

B. Data Preprocessing

Data preprocessing was conducted in two stages. In the first stage, data from multiple years were concatenated and cleaned

by removing null values and duplicates and filtering irrelevant columns through the mentioned functions file. Additionally, goalkeeper data was removed since their metrics and characteristics differ significantly from field players, which could negatively affect model performance.

Subsequently, additional cleaning and standardization techniques were applied in the project notebook, preparing data for model training. Standardizing the features was essential to ensure optimal machine learning model operation and to provide a comparable scale for all variables.

C. Feature Selection

To select the most relevant variables, a correlation matrix with the target variable, *current_rating*, was used, applying a 0.25 threshold to identify those variables with significant correlation. Only features exceeding this threshold were included in the model, which helped reduce noise in the data and optimize model performance. Selected variables included technical attributes (such as *dribbling* and *finishing*), physical attributes (such as *stamina* and *strength*), and game behavior attributes (such as *aggression* and *vision*), providing a comprehensive view of players' performance. The *year* variable was also included to indicate the year corresponding to each record, allowing for better temporal evolution capture.

D. Implementation of Temporal Sequences

To capture players' performance evolution, five-year time sequences were constructed. This sequence length was chosen to reflect short- and medium-term changes in players' performance, such as improvements in technical and physical skills, which are common in young players. Each time sequence was used as input for the predictive models, allowing observation of the impact of previous years on current ratings.

E. Initial Machine Learning Models

To explore the dataset's predictive capacity, several regression-oriented machine learning models were initially implemented, such as *Ridge Regression*, *Lasso Regression*, *Bayesian Ridge*, *Support Vector Regression (SVR)*, *Random Forest*, *Decision Tree*, and *Polynomial Regression (degree 2)*. These models were selected to observe regression models' fitting capacity in predicting performance ratings.

However, the models exhibited overfitting issues, displaying high accuracy in the training set but limited performance in the test set. Some results included mean squared error (MSE) values of 3.846×10^{-12} for Ridge Regression, 1.663×10^{-4} for Lasso Regression, and 1.123×10^{-3} for SVR, indicating that they did not generalize well. Figure 1 visually illustrates the overfitting in one of these initial models.

F. RNN Model Development

Due to the overfitting issues with initial models, a model based on recurrent neural networks (RNN) was developed to capture temporal relationships in the data. Initially, a simple RNN model was built, which yielded satisfactory results for average-performing players but showed limitations in predicting players with extremely high or low ratings.

To improve the model's accuracy in these cases, an optimized architecture named *RNN_Ponderado* was designed, which included additional layers, such as bidirectional LSTM and GRU, as well as weighting the training samples to give greater importance to players with high ratings. This model included a five-year time sequence to capture medium-term trends and reflect performance evolution.

G. Model Evaluation

The models' accuracy was evaluated using mean squared error (MSE), mean absolute error (MAE), and the coefficient of determination (R^2). Additionally, a *Gradient Boosting* model was used to compare the results obtained by the RNN. Figure 1 shows the overfitting issue in the initial regression models, justifying the shift towards the RNN model in the next stage.

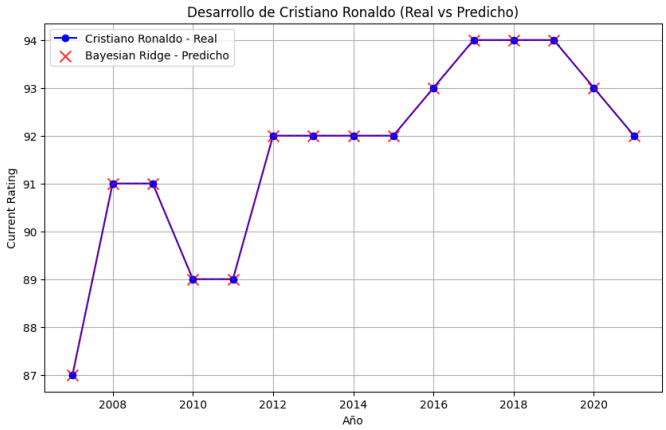

Fig. 1. Overfitting example in initial regression models with Bayesian Ridge

III. Results

A. Prediction in Known Years

To evaluate the capacity of the *RNN_Ponderado* model in predicting ratings within the validation set, predictions were made for years already covered in the training data. The generated graphs (see Figures 2 and 3) show that the model successfully captures general performance trends, though it exhibits variability in accuracy depending on the player and the prediction period. This ability to capture trends is observed in the comparison graph between actual and predicted values.

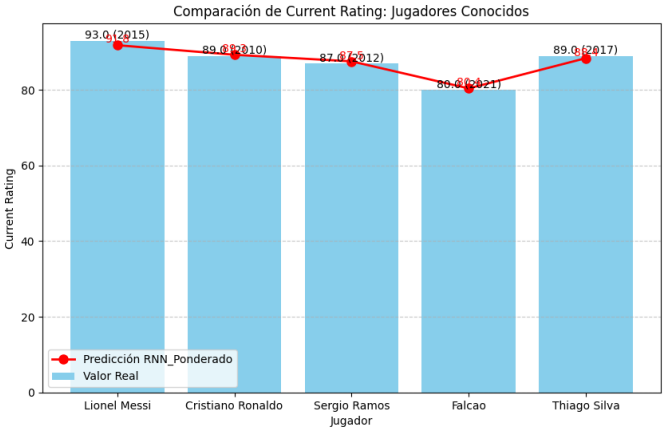

Fig. 2. Actual vs Predicted *current_rating* using *RNN_Ponderado*

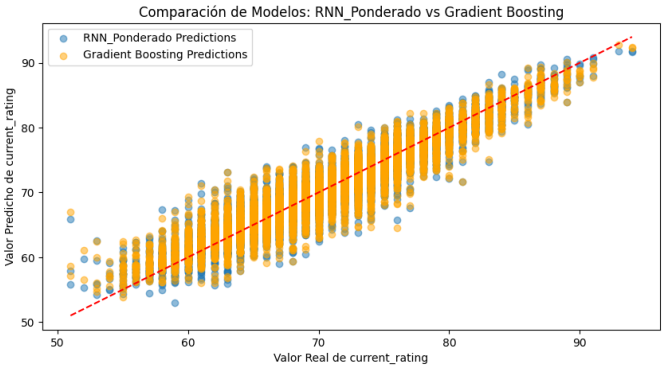

Fig. 3. Model Comparison: *RNN_Ponderado* vs Gradient Boosting

B. Future Projections

To analyze the model's projection capacity for years not observed in the data, predictions were made for the 2018-2025 period. These projections, shown in Figure 4, reflect how the model interprets performance patterns from historical data and its ability to extend those patterns into future horizons. Although predictions are accurate in the initial years of the projection, there is a decrease in accuracy as the prediction horizon extends, which is expected due to the temporal nature of the data.

C. Error Evaluation

Mean Absolute Errors (MAE) were calculated for different categories and periods to evaluate model accuracy more granularly. Table I shows the MAE by year, indicating variability that could be due to changes in player performance over specific periods.

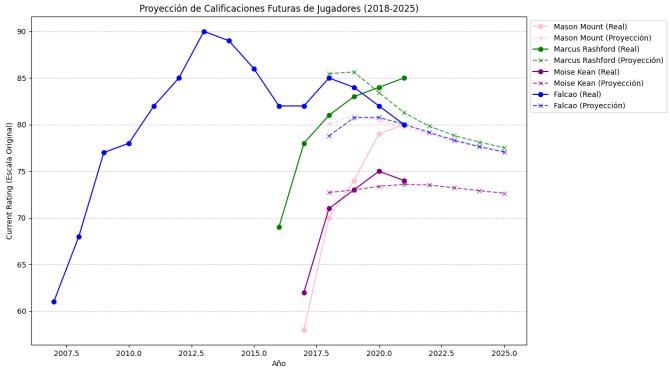

Fig. 4. Projection of Future Player Ratings (2018-2025)

TABLE I
MAE BY YEAR

Year	MAE
2005	1.38
2007	1.39
2008	1.44
2009	1.32
2010	1.39
2011	1.56
2012	1.38
2013	1.49
2014	1.27
2015	1.32
2016	1.63
2017	1.20
2018	1.20
2019	1.91
2020	1.55
2021	1.52

D. Error Visualization by Field Position

Figure 5 shows the distribution of MAE by football field position, excluding goalkeepers. This chart helps visualize how model accuracy varies according to player position, providing an intuitive view of performance across different field areas.

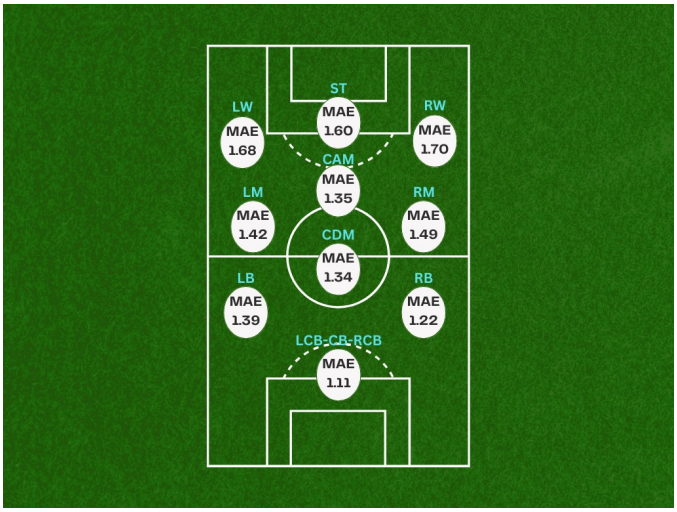

Fig. 5. Mean Absolute Error (MAE) by Football Field Position

E. Rating Progression Analysis

Players with the greatest progression in their ratings over the years were analyzed, identifying those with significant increases in their *current_rating*. This analysis evaluated how the *RNN_Ponderado* model reacts to players exhibiting drastic potential changes, observing the model's ability to capture and predict these substantial performance improvements. The three players with the most significant rating progression are shown in Figure 6, which illustrates how these players experienced sustained improvements in performance.

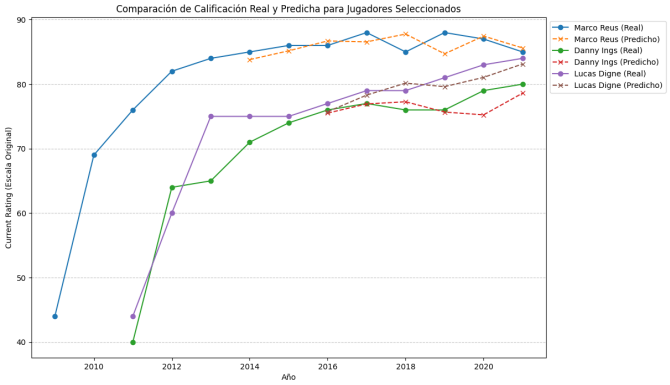

Fig. 6. Players with the Highest Rating Progression

F. Comparison of Known Players

To evaluate the model's accuracy for high-profile players, well-known players were selected, and their actual and predicted ratings were compared over the years (see Figure 7). The results show that the *RNN_Ponderado* model accurately predicts these players' ratings, reflecting both performance peaks and declines.

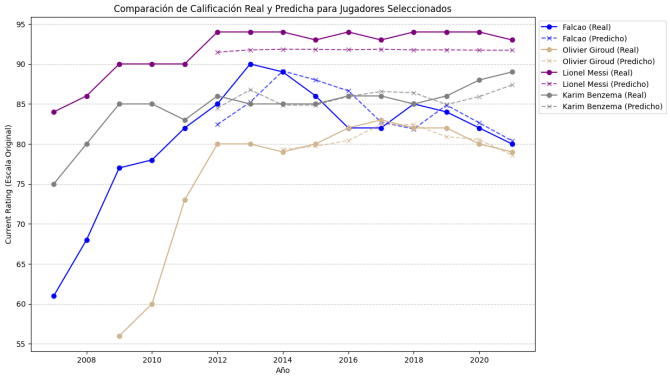

Fig. 7. Comparison of *Current Rating* between Known Players (Actual vs Predicted)

G. Overall Evaluation of the RNN_Ponderado Model

The *RNN_Ponderado* model achieved a mean squared error (MSE) of 3.70, a mean absolute error (MAE) of 1.42, and a coefficient of determination (R^2) of 0.90, indicating a good fit on the test set. These results reflect the model's effectiveness in capturing temporal patterns in players' performance.

- **MAE for Low Ratings** (<= 60): 2.04
- **MAE for High Ratings** (>= 85): 1.34

The results show that the model performs better when predicting high ratings compared to low ratings, which may be interpreted as a limitation in predictive capacity for players in early development stages.

IV. Conclusions and Future Work

A. Conclusions

This work developed a weighted Recurrent Neural Network (RNN) model to predict football players' performance over the years, using historical data from the FIFA video game. Through the evaluations conducted, it was evidenced that the *RNN_Ponderado* model captures players' performance trends and maintains a good overall performance, achieving a mean squared error (MSE) of 3.70, a mean absolute error (MAE) of 1.42, and a coefficient of determination (R^2) of 0.90.

Moreover, the error analysis showed that the model is more accurate in predicting high-rated players, while it has more difficulty in predicting ratings for players in the early development stages or with unstable performance. Additionally, the position-by-position evaluation on the field revealed precision differences depending on the player's role, providing useful information for future improvements in model architecture.

B. Future Work

There are several directions for improving and expanding this work. Future research could consider:

- **Integration of Real Data:** Using actual performance data from players, drawn from matches and professional league statistics, to improve the model's predictive capability and applicability in real contexts.
- **Model Architecture Optimization:** Exploring other advanced architectures, such as Transformers or hybrid models that combine RNN with attention techniques, to improve long-term pattern capture in performance data.
- **Analysis by Age and Potential Categories:** Segmenting data by age categories and development potential to customize predictions and improve accuracy in young or emerging players.
- **Projection in Competitive Contexts:** Implementing the model in simulations that evaluate players' performance in competitive scenarios, helping coaches plan tactics and strategies based on projected performance.
- **Variable Expansion:** Incorporating additional variables, such as physical and psychological metrics, that could enrich the model and allow a deeper understanding of players' performance.

This work provides a solid foundation for the analysis and prediction of football players' performance, and the proposed directions offer opportunities to develop more robust and applicable models in sports analysis environments.

V. References

References

[1] Daguizer. (2021). *FIFA 2005 to 2021 Complete Player Attributes*. Kaggle. Retrieved from https://www.kaggle.com/datasets/daguizer/fifa-2021-to-2005-complete-player-attributes?select=fifa05.csv

[2] Jiménez, F. (2021). *Model for predicting performance in football players*. Universidad del Rosario. Retrieved from https://repository.urosario.edu.co/items/de4fe326-3b59-40c3-9ec7-3e350fb24657