Predicción del Rendimiento de Jugadores de Fútbol Utilizando la Base de Datos de FIFA y Redes Neuronales Recurrentes

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*—En este proyecto se desarrolla un modelo de predicción del rendimiento de jugadores de fútbol utilizando redes neuronales recurrentes (RNN) y datos de la base de datos FIFA. Se implementaron técnicas avanzadas de preprocesamiento de datos para crear secuencias temporales que permitieran capturar la evolución del rendimiento de los jugadores a lo largo de los años. Además, se utilizaron diferentes métricas de evaluación, incluyendo el error cuadrático medio (MSE), el error absoluto medio (MAE), y el coeficiente de determinación (R^2) para evaluar la precisión de los modelos. Los resultados obtenidos indican un MSE de 3.70, un MAE de 1.42, y un R^2 de 0.90, demostrando la efectividad del modelo RNN en la predicción del rendimiento de los jugadores. Este trabajo destaca la aplicación de redes neuronales en el análisis y la predicción de datos deportivos, sugiriendo potenciales mejoras en futuras investigaciones.**

***Index Terms*—Machine Learning, RNN, Predicción de Rendimiento, Modelos de Deportes, FIFA**

I. Introducción

El análisis y predicción del rendimiento de jugadores de fútbol han cobrado relevancia en la industria deportiva, donde el uso de herramientas de inteligencia artificial (IA) y *machine learning* (ML) ha facilitado el desarrollo de nuevos métodos para la identificación de talento y la optimización del rendimiento de los equipos. En este contexto, predecir el rendimiento futuro de un jugador no solo ofrece ventajas competitivas en la planificación táctica y estratégica, sino también en el scouting y el desarrollo de talento juvenil.

Para este proyecto, se emplearon datos históricos del videojuego FIFA debido a su disponibilidad y accesibilidad, lo cual permitió cumplir con los tiempos requeridos para el desarrollo. Aunque estos datos no son completamente representativos de jugadores en situaciones reales, los modelos desarrollados podrían adaptarse para trabajar con datos de rendimiento auténticos, ya sea de jugadores juveniles o profesionales en el entorno competitivo. Esto ampliaría las aplicaciones prácticas de los modelos y su utilidad en el ámbito profesional.

Inicialmente, se implementaron varios modelos de regresión y clasificación, como *Ridge Regression*, *Lasso Regression*, *Bayesian Ridge*, *Support Vector Regression (SVR)*, *Random Forest*, *Decision Tree*, y *Regresión Polinomial*. Sin embargo, estos modelos presentaron problemas de sobreajuste, donde el modelo se ajusta demasiado a los datos de entrenamiento, resultando en una precisión reducida en el conjunto de prueba. Algunos de los valores de error cuadrático medio (MSE) obtenidos incluyen: Ridge Regression con 3.846×10^{-12}, Lasso Regression con 1.663×10^{-4}, y SVR con 1.123×10^{-3}, entre otros. Debido a estos problemas, se optó por un modelo de redes neuronales recurrentes (RNN), el cual es más adecuado para capturar relaciones temporales en datos secuenciales, permitiendo modelar la evolución del rendimiento en el tiempo de forma más robusta.

El objetivo principal de este proyecto es desarrollar un modelo predictivo basado en RNN para anticipar el rendimiento futuro de los jugadores de fútbol, empleando datos históricos del FIFA. Para ello, se implementaron técnicas avanzadas de preprocesamiento de datos y métodos de evaluación de modelos, que permiten mejorar la precisión de las predicciones y comparar el rendimiento del modelo RNN con otros enfoques.

Las principales contribuciones de este trabajo incluyen el uso innovador de datos de FIFA en el modelado predictivo de rendimiento deportivo, el diseño de un modelo RNN ajustado a las particularidades de los datos deportivos, y la implementación de métricas de evaluación para medir la precisión y eficacia del modelo.

Este documento se estructura de la siguiente manera: la **Sección II** describe la metodología utilizada, abarcando el conjunto de datos, las técnicas de preprocesamiento y los modelos implementados; la **Sección III** presenta los resultados obtenidos, evaluando el rendimiento de los modelos con métricas específicas; y finalmente, la **Sección IV** concluye el trabajo y sugiere direcciones futuras para mejorar y ampliar este tipo de investigaciones.

II. Metodología

A. Conjunto de Datos

Para este proyecto, se empleó un conjunto de datos extraído de la plataforma Kaggle, titulado "FIFA 2021 to 2005 Complete Player Attributes" [1]. Este conjunto de datos contiene información detallada sobre los atributos de los jugadores desde el año 2005 hasta el 2021, proporcionando un historial

que permite analizar la evolución del rendimiento de los jugadores en el tiempo. Dado el objetivo de implementar un modelo predictivo, estos datos de FIFA ofrecieron una base accesible y extensa. Aunque los datos del videojuego no reflejan totalmente la realidad de los jugadores en situaciones competitivas, el modelo podría aplicarse también a datos de jugadores reales, ya sean juveniles o profesionales.

Los archivos individuales en formato .csv para cada año fueron concatenados mediante scripts en Python, específicamente con el archivo `funciones_de_limpieza.py`, el cual también se encargó de realizar una limpieza inicial de los datos y de generar un archivo consolidado, `cleaned_combined_fifa_data_filtered.csv`, con los datos filtrados y listos para el preprocesamiento.

B. Preprocesamiento de los datos

El preprocesamiento de los datos se realizó en dos etapas. En la primera etapa, los datos de múltiples años fueron concatenados y limpiados, eliminando valores nulos y duplicados, y filtrando columnas irrelevantes mediante el archivo de funciones mencionado. Además, se eliminó la información de los porteros debido a que sus métricas y características diferían significativamente de los jugadores de campo, lo cual podría afectar negativamente el desempeño del modelo.

Posteriormente, en el notebook del proyecto, se aplicaron técnicas adicionales de limpieza y estandarización, preparando los datos para el entrenamiento del modelo. La estandarización de las características fue esencial para asegurar que los modelos de machine learning pudieran operar de manera óptima y que todas las variables tuvieran una escala comparable.

C. Selección de Variables

Para seleccionar las variables más relevantes, se utilizó una matriz de correlación con la variable objetivo, *current_rating*, aplicando un umbral de 0.25 para identificar aquellas variables con una correlación significativa. Solo se incluyeron en el modelo aquellas características que superaban este umbral, lo cual permitió reducir el ruido en los datos y optimizar el rendimiento del modelo. Entre las variables seleccionadas, se incluyeron atributos técnicos (como *dribbling* y *finishing*), atributos físicos (como *stamina* y *strength*), y atributos de comportamiento en el juego (como *aggression* y *vision*), proporcionando una perspectiva completa del rendimiento de los jugadores. También se incluyó la variable *year* para indicar el año correspondiente a cada registro, permitiendo capturar mejor la evolución temporal de los jugadores.

D. Implementación de Secuencias Temporales

Para capturar la evolución en el rendimiento de los jugadores, se construyeron secuencias temporales de cinco años. Esta longitud de secuencia se eligió con el propósito de reflejar cambios a corto y mediano plazo en el rendimiento de los jugadores, como mejoras en habilidades técnicas y físicas que son comunes en jugadores jóvenes. Cada secuencia temporal fue usada como entrada para los modelos predictivos, permitiendo observar el impacto de años previos en la calificación actual.

E. Modelos Iniciales de Machine Learning

Para explorar la capacidad predictiva del conjunto de datos, inicialmente se implementaron varios modelos de machine learning orientados a la regresión, tales como *Ridge Regression*, *Lasso Regression*, *Bayesian Ridge*, *Support Vector Regression (SVR)*, *Random Forest*, *Decision Tree*, y *Regresión Polinomial (grado 2)*. Estos modelos fueron seleccionados para observar la capacidad de ajuste de modelos de regresión en la predicción de calificaciones de rendimiento.

Sin embargo, los modelos presentaron problemas de sobreajuste, mostrando altos niveles de precisión en el conjunto de entrenamiento, pero desempeños limitados en el conjunto de prueba. Algunos de los resultados obtenidos incluyeron valores de error cuadrático medio (MSE) de 3.846×10^{-12} para Ridge Regression, 1.663×10^{-4} para Lasso Regression, y 1.123×10^{-3} para SVR, indicando que no generalizaban bien. La Figura 1 ilustra visualmente el sobreajuste en uno de estos modelos iniciales.

F. Desarrollo de Modelos de RNN

Debido a los problemas de sobreajuste con los modelos iniciales, se optó por desarrollar un modelo basado en redes neuronales recurrentes (RNN) que pudiese capturar las relaciones temporales en los datos. En primer lugar, se construyó un modelo RNN simple, el cual obtuvo resultados satisfactorios para jugadores de rendimiento promedio, pero mostró limitaciones en la predicción para jugadores con calificaciones extremadamente altas o bajas.

Para mejorar la precisión del modelo en estos casos, se diseñó una arquitectura optimizada denominada *RNN_Ponderado*, en la que se introdujeron capas adicionales, como LSTM bidireccionales y GRU, además de ponderar las muestras de entrenamiento para dar mayor importancia a los jugadores con calificaciones altas. Este modelo incluyó una secuencia temporal de cinco años para capturar tendencias a mediano plazo y reflejar la evolución del rendimiento.

G. Evaluación de los Modelos

La precisión de los modelos fue evaluada mediante las métricas de error cuadrático medio (MSE), error absoluto medio (MAE) y el coeficiente de determinación (R^2). Además, se utilizó un modelo de *Gradient Boosting* para comparar los resultados obtenidos por la RNN. La Figura 1 muestra el problema de sobreajuste en los modelos iniciales de regresión, justificando el cambio hacia el modelo RNN en la siguiente etapa.

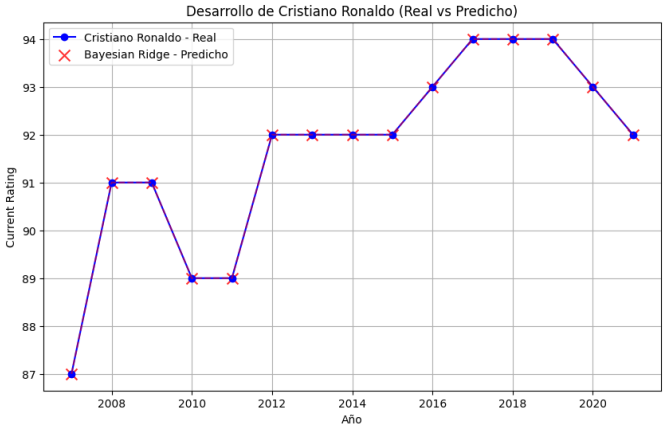

Fig. 1. Ejemplo de sobreajuste en modelos iniciales de regresión con bayesian ridge

III. Resultados

A. Predicción en Años Conocidos

Para evaluar la capacidad del modelo RNN_Ponderado en la predicción de calificaciones dentro del conjunto de validación, se realizaron predicciones para años que ya estaban cubiertos en los datos de entrenamiento. Las gráficas generadas (ver Figuras 2 y 3) muestran que el modelo captura con éxito las tendencias generales de rendimiento, aunque presenta variabilidad en precisión dependiendo del jugador y del período de predicción. Esta capacidad del modelo para capturar las tendencias se observa en la gráfica de comparación entre valores reales y predichos.

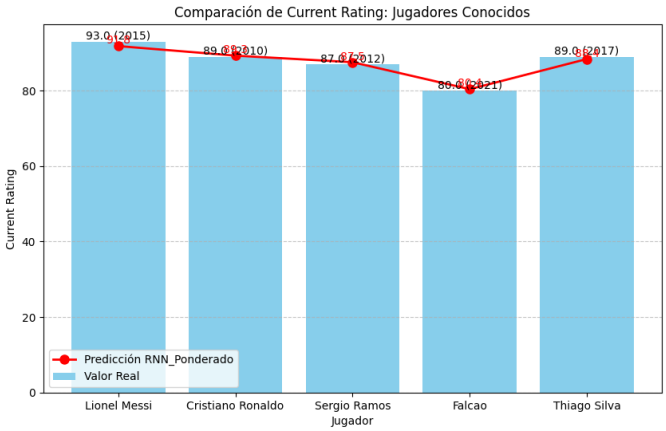

Fig. 2. Valores Reales vs Predichos de *current_rating* usando RNN_Ponderado

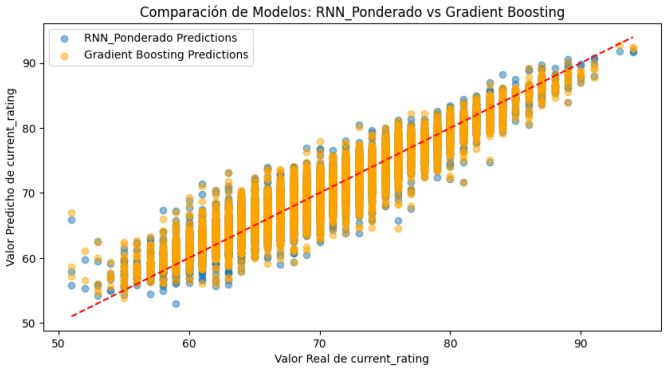

Fig. 3. Comparación de Modelos: RNN_Ponderado vs Gradient Boosting

B. Proyecciones Futuras

Para analizar la capacidad de proyección del modelo en años no observados, se realizaron predicciones para el periodo 2018-2025. Estas proyecciones, mostradas en la Figura 4, reflejan cómo el modelo interpreta los patrones de rendimiento en datos históricos y su habilidad para extender esos patrones en horizontes temporales. Aunque las predicciones son precisas en los primeros años de la proyección, se observó una disminución de precisión a medida que se extiende el horizonte de predicción, lo cual es esperable debido a la naturaleza de los datos temporales.

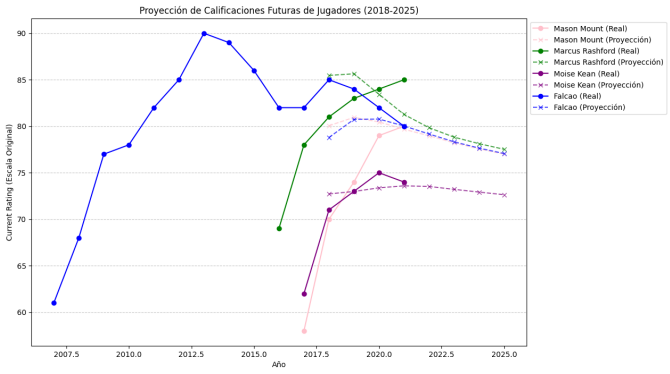

Fig. 4. Proyección de Calificaciones Futuras de Jugadores (2018-2025)

C. Evaluación de Errores

Se calcularon los errores medios absolutos (MAE) para diferentes categorías y periodos con el fin de evaluar la precisión del modelo de manera más granular. La Tabla I muestra el MAE por año, indicando una variabilidad que podría deberse a cambios en el rendimiento de los jugadores en periodos específicos.

TABLE I
MAE POR AÑO

Año	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. Visualización de Error por Posición en el Campo

La Figura 5 muestra la distribución del MAE por posición en el campo de fútbol, excluyendo a los porteros. Este gráfico ayuda a visualizar cómo la precisión del modelo varía según la posición del jugador, proporcionando una vista intuitiva del rendimiento en diferentes áreas del campo.

Fig. 5. Error Medio Absoluto (MAE) por Posición en el Campo de Fútbol

E. Análisis de Progresión en Calificación

Se analizaron los jugadores con mayor progreso en sus calificaciones a lo largo de los años, identificando a aquellos con incrementos significativos en su *current_rating*. Este análisis se realizó para evaluar cómo el modelo RNN_Ponderado reacciona ante jugadores que presentan cambios drásticos en su potencial, observando la capacidad del modelo para capturar y predecir estos incrementos sustanciales en el rendimiento. Los tres jugadores con mayor progresión en la calificación son presentados en la Figura 6, la cual muestra cómo estos jugadores experimentaron mejoras sostenidas en su rendimiento.

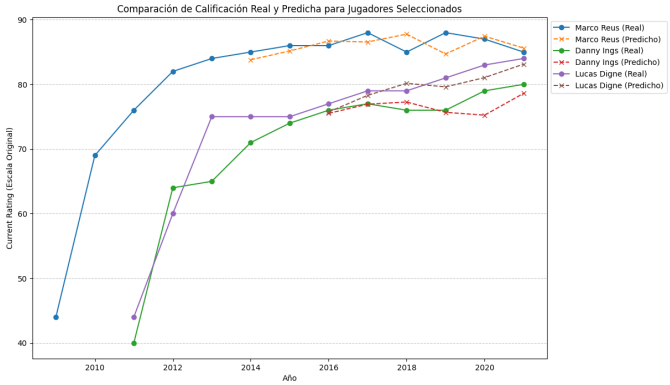

Fig. 6. Jugadores con Mayor Progresión en Calificación

F. Comparación de Jugadores Conocidos

Para evaluar la precisión del modelo en jugadores con alto perfil, se seleccionaron jugadores conocidos y se compararon sus calificaciones reales y predichas a lo largo de los años (ver Figura 7). Los resultados muestran que el modelo RNN_Ponderado predice con precisión las calificaciones de estos jugadores, reflejando tanto sus picos de rendimiento como sus declives.

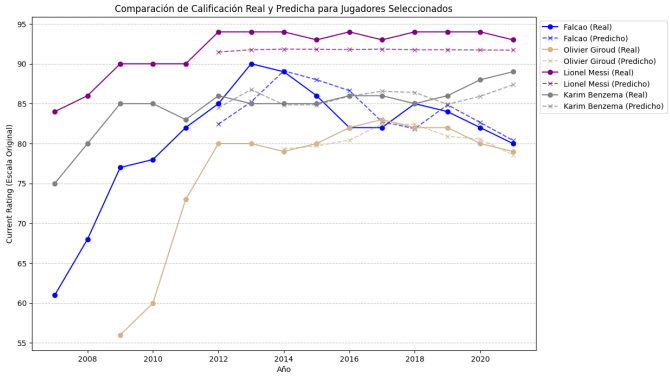

Fig. 7. Comparación de *Current Rating* entre Jugadores Conocidos (Real vs Predicho)

G. Evaluación General del Modelo RNN_Ponderado

El modelo RNN_Ponderado logró un error cuadrático medio (MSE) de 3.70, un error absoluto medio (MAE) de 1.42, y un coeficiente de determinación (R^2) de 0.90, lo cual indica un buen ajuste en el conjunto de prueba. Estos resultados reflejan que el modelo es efectivo en capturar patrones temporales en el rendimiento de los jugadores.

- **MAE para calificaciones bajas** (<= 60): 2.04
- **MAE para calificaciones altas** (>= 85): 1.34

Los resultados obtenidos muestran que el modelo tiene un mejor desempeño al predecir calificaciones altas en comparación con calificaciones bajas, lo cual puede interpretarse como una limitación en la capacidad de predicción para jugadores en las primeras etapas de desarrollo.

IV. Conclusiones y Trabajos Futuros

A. Conclusiones

En este trabajo, se desarrolló un modelo de Redes Neuronales Recurrentes (RNN) ponderado para predecir el rendimiento de jugadores de fútbol a lo largo de los años, utilizando datos históricos del videojuego FIFA. A través de las evaluaciones realizadas, se evidenció que el modelo RNN_Ponderado logra capturar las tendencias de rendimiento de los jugadores y mantiene un buen desempeño general, obteniendo un error cuadrático medio (MSE) de 3.70, un error absoluto medio (MAE) de 1.42, y un coeficiente de determinación (R^2) de 0.90.

Además, el análisis de errores demostró que el modelo es más preciso en las predicciones de jugadores con calificaciones altas, mientras que muestra mayor dificultad al predecir calificaciones de jugadores en las primeras etapas de desarrollo o con un rendimiento inestable. Asimismo, la evaluación por posición en el campo reveló diferencias en precisión dependiendo del rol del jugador, proporcionando información útil para futuras mejoras en la arquitectura del modelo.

B. Trabajos Futuros

Existen varias direcciones para mejorar y ampliar este trabajo. En futuras investigaciones, se podría considerar:

- **Integración de Datos Reales:** Utilizar datos de rendimiento real de jugadores, provenientes de partidos y estadísticas de ligas profesionales, para mejorar la capacidad de predicción y aplicabilidad del modelo en contextos reales.
- **Optimización de la Arquitectura del Modelo:** Explorar otras arquitecturas avanzadas como Transformer o modelos híbridos que combinen RNN con técnicas de atención, para mejorar la capacidad de captura de patrones de largo plazo en los datos de rendimiento.
- **Análisis por Categorías de Edad y Potencial:** Dividir los datos según categorías de edad y potencial de desarrollo para personalizar las predicciones y mejorar la precisión en jugadores jóvenes o en ascenso.
- **Proyección en Contextos Competitivos:** Implementar el modelo en simulaciones que evalúen el rendimiento de jugadores en escenarios competitivos, ayudando a los entrenadores a planificar tácticas y estrategias basadas en el rendimiento proyectado.
- **Ampliación de Variables:** Incorporar variables adicionales, como métricas físicas y psicológicas, que puedan enriquecer el modelo y permitir una comprensión más profunda del rendimiento de los jugadores.

Este trabajo proporciona una base sólida para el análisis y predicción del rendimiento de jugadores de fútbol, y las direcciones propuestas ofrecen oportunidades para desarrollar modelos más robustos y aplicables en entornos de análisis deportivo.

V. Referencias

References

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

[2] Jiménez, F. (2021). *Modelo de predicción de rendimiento en jugadores de fútbol*. Universidad del Rosario. Recuperado de https://repository.urosario.edu.co/items/de4fe326-3b59-40c3-9ec7-3e350fb24657