Volumen 11 número 3; 2026

Ciencia y Deporte

Universidad de Camagüey "Ignacio Agramonte Loynaz"

Ciencia y Deporte
ISSN 2223-1773, RNPS: 2276
Vol. 11. No. 3, septiembre-diciembre, 2026 p-e412

Artículo original

El rendimiento de jugadores de fútbol mediante big data y aprendizaje automático: un estudio de la temporada 2022–2023

[*Football player performance using big data and machine learning: a study of the 2022–2023 season*]

[*Análise do desempenho de jogadores de futebol utilizando big data e aprendizado de máquina: um estudo da temporada 2022–2023*]

Aurón Eduardo Mendoza Canales [1], *Alberto David Quesada Paz* [1], *Carlos Alberto Effio Gonzales* [1]

[1]Universidad César Vallejo, Lima, Perú.

[*]Autor para la correspondencia: amendozaca10@ucvvirtual.edu.pe

Recibido: *30/09/2026.*
Aceptado: *31/10/2026*

RESUMEN

Introducción: la toma de decisiones en el deporte ha evolucionado significativamente gracias a la incorporación de herramientas tecnológicas y analíticas. Este artículo analiza las aplicaciones, beneficios y desafíos asociados al uso de datos y apoyo en precisar la toma de decisiones deportivas, especialmente en contextos de competencia y entrenamiento.

Materiales y métodos: se realizó un estudio cuantitativo descriptivo sobre el rendimiento de jugadores de fútbol mediante big data y aprendizaje automático en la

temporada 2022–2023, con una muestra de 2,689 jugadores de las principales ligas europeas. El conjunto de datos incluyó 124 variables relacionadas con información personal, posición, minutos, goles, asistencias y acciones técnicas y defensivas. Posteriormente, se evaluó el modelo de regresión línea múltiple de, utilizando métricas para determinar la capacidad predictiva del rendimiento. Revelando la correlación entre las diferentes variables.

Resultados: el modelo de regresión lineal fue capaz de generar predicciones de goles utilizando como variables independientes los minutos jugados y las asistencias. Sin embargo, el error cuadrático medio (MSE = 2.6968) evidencia que aún existe una diferencia considerable entre los resultados predichos y los resultados reales, por lo que el modelo presenta una precisión moderada.

Conclusiones: en conjunto, el estudio demuestra que las técnicas de big data, el análisis exploratorio de datos y el aprendizaje automático constituyen herramientas valiosas para el análisis del rendimiento deportivo, permitiendo descubrir patrones y construir modelos predictivos que pueden apoyar la toma de decisiones en los eventos de fútbol

Palabras clave: toma de decisiones, análisis de datos, rendimiento deportivo, big data, aprendizaje automático.

ABSTRACT

Introduction: Decision-making in sports has evolved significantly thanks to the incorporation of technological and analytical tools. This article analyzes the applications, benefits, and challenges associated with the use of data and its support in refining sports decision-making, especially in competitive and training contexts.

Materials and methods: A descriptive quantitative study was conducted on the performance of soccer players using big data and machine learning during the 2022–2023 season, with a sample of 2,689 players from the main European leagues. The dataset included 124 variables related to personal information,

position, minutes played, goals, assists, and technical and defensive actions. Subsequently, a multiple linear regression model was evaluated using metrics such to determine its predictive capacity for performance, revealing the correlation between the different variables.

Results: The linear regression model was able to generate goal predictions using minutes played and assists as independent variables. However, the Mean Squared Error (MSE = 2.6968) indicates that there is still a considerable difference between the predicted and actual results, so the model has moderate accuracy.

Conclusions: Overall, the study demonstrates that big data techniques, exploratory data analysis, and machine learning are valuable tools for analyzing athletic performance, allowing the discovery of patterns and the construction of predictive models that can support decision-making in football events.

Keywords: Decision-making, data analysis, athletic performance, big data, machine learning.

RESUMO

Introdução: A tomada de decisões no esporte evoluiu significativamente graças à incorporação de ferramentas tecnológicas e analíticas. Este artigo analisa as aplicações, os benefícios e os desafios associados ao uso de dados e seu suporte no aprimoramento da tomada de decisões esportivas, especialmente em contextos competitivos e de treinamento.

Materiais e métodos: Foi realizado um estudo quantitativo descritivo sobre o desempenho de jogadores de futebol utilizando big data e aprendizado de máquina durante a temporada 2022–2023, com uma amostra de 2.689 jogadores das principais ligas europeias. O conjunto de dados incluiu 124 variáveis relacionadas a informações pessoais, posição, minutos jogados, gols, assistências e ações técnicas e defensivas. Posteriormente, um modelo de regressão linear múltipla foi avaliado utilizando métricas

para determinar sua capacidade preditiva de desempenho, revelando a correlação entre as diferentes variáveis.

Resultados: O modelo de regressão linear foi capaz de gerar previsões de gols utilizando minutos jogados e assistências como variáveis independentes. Contudo, o Erro Quadrático Médio (EQM = 2,6968) indica que ainda existe uma diferença considerável entre os resultados previstos e os reais, portanto o modelo apresenta uma precisão moderada.

Conclusões: De modo geral, o estudo demonstra que as técnicas de dig data, a análise exploratória de dados e o aprendizado de máquina são ferramentas valiosas para a análise do desempenho atlético, permitindo a descoberta de padrões e a construção de modelos preditivos que podem auxiliar na tomada de decisões em eventos de futebol.

Palavras-chave: tomada de decisões, análise de dados, desempenho atlético, big data, aprendizado de máquina.

INTRODUCCIÓN

La transformación digital impactó significativamente al deporte, particularmente al fútbol, al favorecer la incorporación de tecnologías orientadas a la recopilación, procesamiento y análisis de gran cantidad de información de datos generados durante los entrenamientos y las competiciones. En este contexto, los clubes profesionales emplearon sistemas de posicionamiento global (GPS), sensores biométricos, cámaras de seguimiento y plataformas de análisis de eventos para registrar información en tiempo real sobre el rendimiento físico, técnico y táctico de los futbolistas (FIFA, 2024). Estas herramientas facilitaron la obtención de datos de alta resolución, permitiendo una evaluación más objetiva del desempeño individual y colectivo.

El desarrollo del big data representó un avance significativo en la analítica deportiva, ya que permitió identificar patrones de juego, evaluar el comportamiento de los jugadores y fortalecer la planificación estratégica de los equipos. Asimismo, la incorporación de técnicas de inteligencia artificial y aprendizaje automático incrementó la precisión en la

evaluación del rendimiento deportivo al proporcionar modelos capaces de procesar grandes cantidades de información y generar predicciones con mayor exactitud (IBM, 2023).

En el fútbol, la disponibilidad de datos provenientes de múltiples fuentes impulsó la aplicación del aprendizaje automático (Machine Learning) como una herramienta para identificar relaciones complejas entre variables y construir modelos predictivos que superaron las limitaciones de los métodos estadísticos convencionales. Mediante algoritmos supervisados y no supervisados se analizaron simultáneamente indicadores físicos, técnicos y tácticos, proporcionando información relevante para entrenadores, preparadores físicos y analistas deportivos durante los procesos de establecer objetivos y definir conjunto de acciones para precisar la toma de decisiones (Rico-González *et al.*, 2023).

Las investigaciones desarrolladas en los últimos años evidenciaron que las técnicas de aprendizaje automático fueron aplicadas en diversas áreas del fútbol profesional, incluyendo la predicción del rendimiento físico, el análisis técnico-táctico, la evaluación del riesgo de lesiones, la identificación de talento y la predicción de resultados deportivos. De igual manera, la integración de modelos de inteligencia artificial con datos espacio-temporales permitió analizar con mayor precisión el comportamiento colectivo, las interacciones entre jugadores y la organización táctica de los equipos durante la competición (Teixeira *et al.,* 2025).

Paralelamente, las revisiones sistemáticas sobre inteligencia artificial aplicada al deporte demostraron un crecimiento sostenido en el uso de redes neuronales profundas, bosques aleatorios, máquinas de soporte vectorial y algoritmos de gradiente reforzado para el procesamiento de datos deportivos en tiempo real. Estas metodologías contribuyeron a optimizar la planificación del entrenamiento, mejorar las estrategias competitivas y fortalecer la toma de decisiones basada en evidencia, consolidándose como una de las principales tendencias de la analítica deportiva contemporánea (Vec *et al.,* 2024).

Asimismo, el análisis técnico y táctico basado en datos favoreció una comprensión más precisa de las demandas del juego, al integrar información procedente de sistemas de seguimiento, sensores y plataformas de análisis de eventos. En consecuencia, el big data se consolidó como un recurso estratégico para evaluar el rendimiento de los futbolistas desde una perspectiva integral, permitiendo transformar datos heterogéneos en conocimiento útil para los cuerpos técnicos y las organizaciones deportivas (Otero-Saborido *et al.,* 2021).

No obstante, la literatura también señaló la existencia de desafíos asociados con la calidad de los datos, la integración de múltiples fuentes de información, la interpretabilidad de los modelos predictivos y la validación de los algoritmos en diferentes contextos competitivos. Estas limitaciones evidenciaron la necesidad de desarrollar investigaciones que profundizaran en el potencial del big data y del aprendizaje automático para fortalecer la evaluación objetiva del rendimiento futbolístico y mejorar los procedimientos para la selección de opciones en los grupos de trabajo.

En este contexto, se analizó el rendimiento de jugadores de fútbol mediante técnicas de big data y aprendizaje automático, evaluando el potencial de diferentes algoritmos de inteligencia artificial para identificar patrones de desempeño y generar información que contribuyera a optimizar la planificación del entrenamiento y el rendimiento competitivo.

El big data hace referencia a la manipulación y estudio de enormes cantidades de información que se producen rápidamente y en distintas formas. En el fútbol, estos datos incluyen estadísticas de partidos, desplazamientos de los jugadores, frecuencia cardíaca, velocidad, aceleraciones, pases, remates y otras variables relevantes para el análisis deportivo (Stats Perform, 2024).

Las plataformas especializadas permiten procesar millones de registros para identificar patrones que serían difíciles de detectar mediante la observación tradicional. Las principales aplicaciones del big data en el fútbol incluyen la evaluación del rendimiento físico y técnico de los jugadores. Diseño de estrategias tácticas mediante análisis del rival.

Prevención de lesiones utilizando indicadores de carga física. Identificación y captación de talentos mediante análisis estadístico. Optimización de entrenamientos personalizados.

Mejora de la experiencia de los aficionados mediante análisis del comportamiento del público (FIFA, 2024).

Según IBM (2023), la inteligencia artificial permite complementar el análisis de datos para generar recomendaciones estratégicas que apoyan las decisiones del cuerpo técnico. Diversas investigaciones destacan que el uso del big data proporciona beneficios importantes: Incrementa la precisión en la toma de decisiones. Reduce el riesgo de lesiones. Mejora el rendimiento individual y colectivo. Optimiza la planificación táctica.

Favorece procesos de contratación de jugadores basados en evidencia. Incrementa la competitividad de los clubes (UEFA, 2023).

Para los desafíos. A pesar de sus ventajas, existen desafíos relacionados con la implementación del big data: Elevados costos tecnológicos. Necesidad de profesionales especializados en ciencia de datos. Protección y privacidad de la información de los deportistas. Integración adecuada entre el análisis estadístico y la experiencia del entrenador (FIFA, 2024).

Rendimiento deportivo en fútbol

El rendimiento deportivo se ha definido en la literatura como la expresión cuantificable de la eficacia de un deportista o equipo en relación con los objetivos del juego. Según Bangsbo (2014) y Reilly & Williams (2003), en el fútbol el rendimiento integra dimensiones físicas, técnicas, tácticas y psicológicas, las cuales interactúan de manera

dinámica durante la competición. Su análisis permite evaluar la contribución individual y colectiva al desempeño del equipo.

Análisis del rendimiento

El análisis del rendimiento es un proceso sistemático de observación, registro y evaluación del comportamiento deportivo con el objetivo de precisar la toma de decisiones. Hughes y Bartlett (2002) lo describen como la aplicación de métodos científicos para cuantificar acciones técnicas y tácticas en el deporte. Este análisis ha evolucionado desde la observación manual hacia sistemas automatizados basados en tecnología y datos masivos.

Big data

Conjuntos de datos de gran volumen, velocidad y variedad que requieren tecnologías avanzadas para su almacenamiento, procesamiento y análisis. Según Mayer-Schönberger y Cukier (2013), su valor no reside únicamente en la cantidad de datos, sino en la capacidad de extraer conocimiento útil a partir de ellos. En el deporte, el big data permite analizar patrones complejos de rendimiento mediante datos provenientes de sensores, GPS y sistemas de seguimiento.

Ciencia de datos en el deporte

La ciencia de datos aplicada al deporte integra estadística, informática y modelado matemático para transformar datos en conocimiento útil. Según Provost & Fawcett (2013), su objetivo es descubrir patrones, realizar predicciones y reforzar la toma de decisiones. En el fútbol, permite analizar grandes volúmenes de información táctica y física para optimizar el rendimiento.

Inteligencia artificial aplicada al deporte

La inteligencia artificial en el deporte comprende el uso de algoritmos y sistemas computacionales capaces de simular procesos de decisión humana. Según Russell &

Norvig (2021), la IA permite construir sistemas que perciben, razonan y actúan. En fútbol, se emplea para análisis táctico, scouting automatizado y predicción de resultados.

Premodelos

Los modelos predictivos permiten estimar resultados futuros a partir del análisis de datos históricos (James *et al.,* 2021).

Análisis táctico en fútbol

El análisis táctico estudia la organización y comportamiento colectivo de los jugadores durante el juego. Según Carling *et al.* (2005), permite evaluar la estructura del equipo, las transiciones y la ocupación del espacio. Actualmente, se apoya en datos de tracking y algoritmos de IA para mayor precisión.

Analítica deportiva

De acuerdo con Alamar (2013), la analítica deportiva implica el uso sistemático de datos para apoyar decisiones estratégicas dentro de equipos y organizaciones deportivas, permitiendo una evaluación objetiva del rendimiento más allá de la observación subjetiva tradicional. En esta línea, el análisis se basa en indicadores físicos, técnicos y tácticos obtenidos a partir de sistemas tecnológicos como sensores, cámaras de seguimiento y plataformas de registro de eventos.

Sistemas de seguimiento

Los sistemas de tracking permiten registrar la posición y movimiento de los jugadores en tiempo real mediante GPS. Carling *et al.* (2008) señalan que estos sistemas han revolucionado la evaluación del rendimiento al proporcionar datos espaciales precisos.

MATERIALES Y MÉTODOS

Enfoque y tipo de estudio

La presente investigación tuvo un enfoque cuantitativo debido a que se empleó datos numéricos que fueron recopilado de una base de datos deportivo para construir modelos predictivos mediante técnicas de big data y aprendizaje automático. El análisis comprenderá procedimientos estadísticos, minería de datos y algoritmos de Machine Learning con el propósito de identificar patrones asociados al rendimiento de los jugadores.

Tipo de investigación

La investigación es aplicada, ya que busca utilizar conocimientos sobre análisis de datos e inteligencia artificial para resolver un problema práctico relacionado con la evaluación y predicción del rendimiento de futbolistas profesionales.

Asimismo, el estudio será:

- Nivel predictivo, porque pretende desarrollar modelos capaces de estimar el rendimiento de los jugadores a partir de variables técnicas y físicas.
- Diseño no experimental, debido a que no se manipularán las variables, sino que se analizarán datos históricos ya existentes.
- Diseño transversal o de corte único, puesto que el análisis se realizará utilizando información correspondiente únicamente a la temporada 2022–2023.

Población y muestra

Población

La población estará conformada por todos los registros de jugadores profesionales pertenecientes a las cinco principales ligas europeas durante la temporada 2022–2023 incluidos en el dataset seleccionado.

El conjunto de datos contiene aproximadamente:

- Más de 2 500 jugadores.
- 124 variables estadísticas.
- Información correspondiente a:
    - Premier League.
    - La Liga.
    - Bundesliga.
    - Serie A.
    - Ligue 1.

Las variables incluyen indicadores ofensivos, defensivos, de pases, posesión del balón, disciplina, recuperaciones, acciones de creación de gol, entre otras.

Muestra

Debido al tamaño manejable del conjunto de datos, no será necesario realizar un muestreo, por lo que se empleará un muestreo censal, considerando la totalidad de los registros disponibles.

En consecuencia, la muestra estará conformada por los más de 2 500 jugadores incluidos en el dataset.

Instrumento de recolección de datos

Para la recolección será una ficha de registro de datos digitales, elaborada a partir del dataset descargado desde Kaggle.

La información corresponde a estadísticas oficiales recopiladas de Football Reference, posteriormente organizadas y publicadas en Kaggle para fines de análisis de datos deportivos.

La ficha permitirá registrar variables como:

- Nombre del jugador.
- Edad.
- Posición.
- Equipo.
- Liga.
- Partidos jugados.
- Minutos disputados.
- Goles.
- Asistencias.
- Remates.
- Pases completados.
- Precisión de pase.
- Recuperaciones.
- Intercepciones.
- Entradas.
- Acciones de creación de gol.
- Acciones de creación de disparo.
- Tarjetas.
- Duelos aéreos.
- Entre otras variables relevantes.

Posteriormente, la información será procesada utilizando herramientas de análisis de datos como Python, Pandas, Scikit-learn, NumPy y Power BI para el desarrollo de modelos predictivos.

Validez y fiabilidad

Validez

La validez del instrumento estará garantizada por tratarse de un conjunto de datos elaborado a partir de estadísticas oficiales obtenidas sobre el fútbol americano profesional, una de las principales plataformas internacionales de análisis estadístico del fútbol profesional. El dataset ha sido estructurado para investigaciones y proyectos de ciencia de datos, incluyendo información estandarizada sobre el desempeño de jugadores en las principales ligas europeas.

Además, la validez del proceso analítico se fortalecerá mediante:

- limpieza de datos;
- tratamiento de valores faltantes;
- normalización de variables cuando corresponda;
- selección de características (Feature Selection);
- validación cruzada (Cross Validation) durante el entrenamiento de modelos de inteligencia artificial.

Fiabilidad

La fiabilidad estará respaldada porque:

- los datos provienen de registros oficiales y estandarizados;
- las variables fueron obtenidas mediante procedimientos consistentes para todos los jugadores;
- el procesamiento será reproducible utilizando scripts desarrollados en Python;
- los modelos predictivos serán evaluados mediante métricas reconocidas en Machine Learning.

Dependiendo del algoritmo implementado, se utilizarán indicadores como:

- Exactitud (Accuracy).

- Precisión (Precision).
- Sensibilidad (Recall).
- Puntaje F1 (F1-Score).
- Área bajo la curva ROC (AUC-ROC), para problemas de clasificación.
- Error absoluto medio (MAE).
- Error Cuadrático medio (MSE).
- Raíz del error cuadrático Medio (RMSE).
- Coeficiente de determinación (R^2), para problemas de regresión.

Estas métricas permitirán verificar la consistencia, estabilidad y capacidad predictiva de los modelos desarrollados (Tabla 1).

Tabla 1.- *Resumen metodológico*

Aspecto	Descripción
Enfoque	Cuantitativo
Tipo de investigación	Aplicada
Nivel	Predictivo
Diseño	No experimental, transversal
Población	Más de 2 500 jugadores profesionales de las cinco principales ligas europeas (temporada 2022–2023)
Muestra	Muestreo censal (100 % del dataset)
Instrumento	Ficha de registro de datos digitales basada en el dataset de Kaggle
Fuente de datos	Dataset "2022-2023 Football Player Stats" (Football Reference/Kaggle)
Validez	Datos provenientes de estadísticas oficiales, limpieza de datos, selección de características y validación cruzada
Fiabilidad	Reproducibilidad del procesamiento y evaluación mediante métricas de Machine Learning (Accuracy, Precision, Recall, F1, AUC, MAE, RMSE y R^2)

RESULTADOS Y DISCUSIÓN

A continuación, se detalla los pasos que corresponden al análisis y desarrollo

Paso 1. Elección del dataset

Se buscó el dataset apropiado, para ello se analizaron tres alternativas de las cuales para el presente trabajo se eligió el siguiente dataset del siguiente link: https://www.kaggle.com/datasets/vivovinco/20222023-football-player-stats. Donde se muestra la tabla "2022–2023 Football Player Stats" debido a que cumple con el objetivo de la investigación: contiene información estadística de 2,689 jugadores pertenecientes a las principales ligas de fútbol europeas durante la temporada 2022–2023. El conjunto de datos está compuesto por 124 variables que describen aspectos como la información personal del jugador, su equipo, posición, minutos disputados, goles, asistencias, pases, tiros, acciones defensivas y otras métricas de rendimiento. Esta base de datos es adecuada para realizar análisis exploratorios, estudios estadísticos y desarrollar modelos de aprendizaje automático orientados a la evaluación y predicción del desempeño de los futbolistas (Figura 1).



Estadísticas de jugadores de fútbol 2022-2023.csv (1,33 MB)

Detalle | Compacto | Columna — 10 de 124 columnas

Rk	Jugador	Nación	Pos	Equipo	Comp	Edad
1	Brenden Aaronson	EE.UU	MFFW	Leeds United	Premier League	22
2	Yunis Abdelhamid	MAR	DF	Reims	Ligue 1	35
3	Himad Abdelli	FRA	MFFW	Enojos	Ligue 1	23
4	Salis Abdul Samed	GHA	MF	Lente	Ligue 1	22
5	Laurent Abergel	FRA	MF	Lorient	Ligue 1	30
6	Oliver Abildgaard	GUARIDA	MF	Grecia Verona	Serie A	26
7	Matthis Abline	FRA	FWMF	Auxerre	Ligue 1	19
8	Matthis Abline	FRA	FWMF	Rennes	Ligue 1	19

Figura. 1.- *Estadísticas de jugadores de fútbol 2022-2023.csv*

Paso 2.- Cargar el dataset de Kaggle

Mediante el lenguaje de Pyhton en Google Colab se subió correctamente el conjunto de datos de estadísticas de futbol (Figura 2).



Figura. 2.- *Subida de archivo dataset*

Paso 3.- Instalación biblioteca oficial de kaggle

Se instalaron las librerías oficiales de Kaggle, se descargó el dataset mediante la API de Kaggle y se procesó toda la información de la tabla (Figura 3)



Figura. 3.- *Subida de archivo dataset*

Paso 4.- Importar bibliotecas pandas

- Importa la biblioteca Pandas, utilizada para manipular datos.
- Lee el archivo players_stats.csv.
- Guarda toda la información en un DataFrame llamado df.
- df.head() muestra las primeras cinco filas para verificar que los datos se cargaron correctamente (Figura 4).

```
import pandas as pd

df = pd.read_csv("2022-2023 Football Player Stats.csv", encoding='latin1', delimiter=';
df.head()
```

	Rk	Player	Nation	Pos	Squad	Comp	Age	Born	MP	Starts
0	1	Brenden Aaronson	USA	MFFW	Leeds United	Premier League	22	2000	20	19
1	2	Yunis Abdelhamid	MAR	DF	Reims	Ligue 1	35	1987	22	22
2	3	Himad Abdelli	FRA	MFFW	Angers	Ligue 1	23	1999	14	8
3	4	Salis Abdul Samed	GHA	MF	Lens	Ligue 1	22	2000	20	20
4	5	Laurent Abergel	FRA	MF	Lorient	Ligue 1	30	1993	15	14

Figura 4.- *Importar librerías pandas y carga de datos*

Paso 5.- Limpieza de datos

- Cuenta cuántos valores faltantes (nulos) existen en cada columna.
- Elimina todas las filas que contienen al menos un valor vacío.
- Muestra el nombre de todas las columnas del conjunto de datos (Figura 5).

```
# Ver valores nulos
df.isnull().sum()

# Eliminar filas con valores vacíos
df = df.dropna()

# Ver columnas
df.columns
```

```
Index(['Rk', 'Player', 'Nation', 'Pos', 'Squad', 'Comp', 'Age', 'Born', 'MP',
       'Starts',
       ...
       'Off', 'Crs', 'TklW', 'PKwon', 'PKcon', 'OG', 'Recov', 'AerWon',
       'AerLost', 'AerWon%'],
      dtype='object', length=124)
```

Figura. 5.- *Limpieza de datos*

Paso 6.- Selección de variables

Selecciona únicamente las variables más importantes para el estudio. Con el objetivo de Reducir el conjunto de datos y trabajar únicamente con la información necesaria (Figura 6).

```
df_clean = df[[
    "Player",
    "Pos",
    "Squad",
    "Goals",
    "Assists",
    "Min"
]]
df_clean.head()
```

	Player	Pos	Squad	Goals	Assists	Min
0	Brenden Aaronson	MFFW	Leeds United	1	0.11	1596
1	Yunis Abdelhamid	DF	Reims	0	0.05	1980
2	Himad Abdelli	MFFW	Angers	0	0.00	770
3	Salis Abdul Samed	MF	Lens	1	0.00	1799
4	Laurent Abergel	MF	Lorient	0	0.08	1165

***Figura. 6**.- Limpieza de datos*

Paso 7.- Análisis Exploratorio de Datos (EDA)

El EDA (Exploratory Data Analysis) consistió e en explorar y comprender los datos antes de construir modelos predictivos. Ordena los jugadores desde el que tiene más goles hasta el que tiene menos. Después selecciona los primeros diez. Resultado Obtiene el ranking de los diez máximos goleadores (Figura 7).

```
top_goleadores = df_clean.sort_values(by="Goals", ascending=False).head(10)
top_goleadores
```

	Player	Pos	Squad	Goals	Assists	Min
1057	Erling Haaland	FW	Manchester City	25	0.16	1636
1260	Harry Kane	FW	Tottenham	17	0.04	2055
1858	Victor Osimhen	FW	Napoli	16	0.19	1401
1426	Robert Lewandowski	FW	Barcelona	14	0.27	1334
252	Wissam Ben Yedder	FW	Monaco	14	0.24	1135
2451	Ivan Toney	FW	Brentford	14	0.15	1789
184	Folarin Balogun	FW	Reims	14	0.06	1591
1358	Alexandre Lacazette	FW	Lyon	13	0.19	1923
899	Niclas Füllkrug	FW	Werder Bremen	13	0.16	1667
1609	Kylian Mbappé	FW	Paris S-G	13	0.12	1468

***Figura 7.-** Análisis exploratorio de datos*

Paso 8.- Top asistentes

Ordena los jugadores según el número de asistencias. Tuvo como resultado de los diez mejores asistentes. Se ordenó el DataFrame de mayor a menor según la columna Assists y muestra los diez primeros jugadores con el valor más alto en esa columna. Sin embargo, los resultados muestran jugadores con muy pocos minutos jugados, lo que indica que Assists probablemente representa una tasa (por ejemplo, asistencias por 90 minutos) y no el total de asistencias. Por ello, el ranking puede no reflejar a los jugadores con más asistencias totales. Para obtener un resultado más representativo, es recomendable filtrar jugadores con una cantidad mínima de minutos jugados antes de ordenar (Figura 8).

```
df_clean.sort_values(by="Assists", ascending=False).head(10)
```

	Player	Pos	Squad	Goals	Assists	Min
1876	Levin Öztunalý	MF	Union Berlin	0	10.00	9
2544	Gabriel Vidovic	FW	Bayern Munich	0	5.00	19
921	Carlos García	DF	Cádiz	0	3.33	25
902	Gianluca Gaetano	MF	Napoli	0	2.50	33
2684	Szymon ?urkowski	MF	Fiorentina	0	2.50	32
1898	Leandro Paredes	MF	Paris S-G	0	1.67	52
731	Ckene Doucouré	FWMF	Lorient	0	1.67	56
325	Badredine Bouanani	FWMF	Nice	0	1.50	177
1707	Jon Morcillo	FWMF	Athletic Club	0	1.43	63
173	Cédric Bakambu	MF	Marseille	0	1.43	59

Figura 8. - *Top de asistentes*

Paso 9.- Mostrar jugadores según la cantidad de asistencias

A continuación, se ordenó todos los jugadores según la cantidad de asistencias y muestra los diez mejores.

Sirve para identificar los jugadores con mayor capacidad para generar goles.

Permite responder preguntas como:

- ¿Qué posición anota más goles?
- ¿Los delanteros realmente marcan paso?

Se Importó la biblioteca para crear gráficos luego se seleccionó los diez mejores goleadores.
Posterior a ello se generó una figura de diez pulgadas de ancho por cinco de alto.
Se dibujó un gráfico de barras.

- Eje X → nombres de jugadores.
- Eje Y → goles.
- Gira los nombres 45° para que no se sobrepongan (Figura 9).

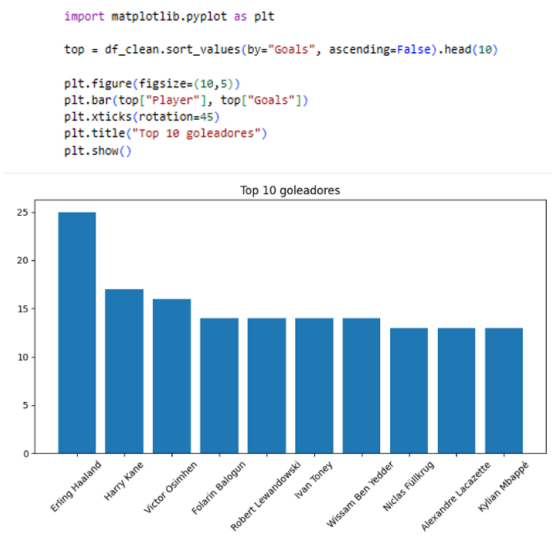


Figura 9. - *Mostrar jugadores según la cantidad de asistencias*

Paso 10. - Calcular promedio de goles por cada equipo de mayor a menor

Este código permitió identificar qué equipos tienen el mejor rendimiento ofensivo al calcular el promedio de goles por partido para cada uno. Posteriormente, ordena los resultados de mayor a menor, facilitando la comparación entre los 98 equipos presentes en el conjunto de datos. Esta técnica es usualmente utilizada en análisis deportivo para

evaluar la capacidad goleadora de los equipos y detectar cuáles son los más efectivos en ataque.

Esto significa que:

- Manchester City tiene el promedio de goles más alto del conjunto de datos, con aproximadamente 2.48 goles por partido.
- Paris S-G ocupa el segundo lugar con un promedio de 2.16 goles.
- Bayern Munich y Napoli también presentan promedios superiores a dos goles por partido.
- En el extremo inferior, Sampdoria registra el promedio más bajo, con apenas 0.32 goles por partido, lo que indica un bajo rendimiento ofensivo en comparación con el resto de los equipos (Figura 10).

```
goals_by_squad = df_clean.groupby("Squad")["Goals"].mean().sort_values(ascending=False)
display(goals_by_squad)
```

Squad	Goals
Manchester City	2.476190
Paris S-G	2.160000
Bayern Munich	2.148148
Napoli	2.125000
Monaco	2.000000
...	...
Wolves	0.483871
Cádiz	0.470588
Valladolid	0.468750
Schalke 04	0.382353
Sampdoria	0.322581

98 rows × 1 columns

dtype: float64

Figura. *10. - Calcular promedio de goles por cada equipo de mayor a menor*

Paso 11.- Relación minutos vs goles

Este código crea un gráfico de dispersión (scatter plot) para analizar la relación entre los minutos jugados y los goles anotados por los jugadores.

- Eje X (minutes): representa los minutos jugados por cada jugador.
- Eje Y (goals): representa la cantidad de goles anotados.

- Cada punto corresponde a un jugador del conjunto de datos.

A continuación, se muestra una tendencia positiva: en general, los jugadores que disputan más minutos tienen mayores posibilidades de marcar más goles. Sin embargo, la relación no es perfecta, ya que existen muchos jugadores con bastantes minutos y pocos goles, así como algunos que destacan por anotar muchos goles. Este tipo de visualización permitió identificar patrones, posibles valores atípicos (jugadores con un rendimiento excepcional) y comprender cómo influye el tiempo de juego en la cantidad de goles marcados (Figura 11).

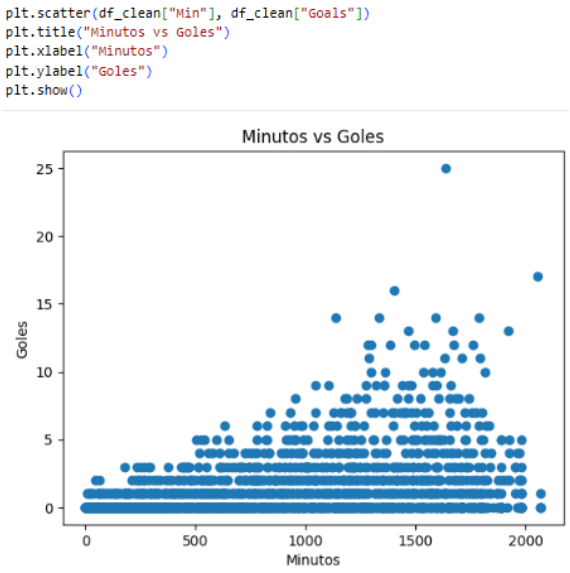


Figura 11. - *Relación minutos vs goles*

Paso 12.- Correlación big data

Este código calcula la matriz de correlación entre las variables goles (goals), asistencias (assists) y minutos jugados (minutes). La correlación toma valores entre -1 y 1 (Tabla 2):

- 1: relación positiva perfecta.
- 0: no existe relación lineal.
- -1: relación negativa perfecta.

Tabla 2.- *Correlación entre variables*

Variables	Correlación	Interpretación
Goals – Goals	**1**	Correlación perfecta consigo misma.
Goals – Assists	**0.105**	Correlación positiva muy débil; marcar más goles no implica necesariamente dar más asistencias.
Goals – Min	**0.378**	Correlación positiva moderada; los jugadores con más minutos tienden a marcar más goles.
Assists – Min	**0.007**	Correlación prácticamente nula; en este conjunto de datos, los minutos jugados no muestran una relación lineal con las asistencias.

La matriz de correlación indica que existe una relación moderada entre los minutos jugados y los goles, mientras que las asistencias presentan una relación muy débil tanto con los goles como con los minutos. Este análisis es útil en big data y análisis exploratorio de datos (EDA) para identificar patrones y relaciones entre variables antes de aplicar modelos estadísticos o de aprendizaje automático (Figura 12).

```
df_clean[["Goals", "Assists", "Min"]].corr()
```

	Goals	Assists	Min
Goals	1.000000	0.104955	0.378071
Assists	0.104955	1.000000	0.006787
Min	0.378071	0.006787	1.000000

Figura 12. - *Matriz de Correlación entre las variables goles (goals), asistencias (assists) y minutos jugados (minutes)*

Paso 13. Modelo simple de regresión lineal

Este modelo aprende a partir de otros ejemplos donde se conoce tanto las variables de entrada (minutos y asistencias) como la salida (goles), y luego utilizó ese aprendizaje para predecir los goles de nuevos jugadores. Estos valores representan la cantidad estimada de goles para los primeros cinco jugadores del conjunto de prueba. Como la regresión lineal produce valores continuos, las predicciones aparecen con decimales. Por ejemplo, un valor de 1.758 indica que el modelo estima aproximadamente 2 goles,

mientras que un valor de 0.260 sugiere que el jugador probablemente no marcará goles o anotará muy pocos. En conclusión, el modelo logró generar predicciones de goles basadas en los minutos jugados y las asistencias, aunque para determinar qué tan preciso es sería necesario comparar estas predicciones con los valores reales mediante métricas de evaluación como el MAE, el MSE o el R^2. No se puede afirmar que 1.758 o 0.962 sean las mejores predicciones solo por su valor, debidos a que son simplemente las predicciones de goles para los primeros cinco jugadores del conjunto de prueba. No indican si el modelo es bueno o malo (Figura 13).

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression

X = df_clean[["Min", "Assists"]]
y = df_clean["Goals"]

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)

model = LinearRegression()
model.fit(X_train, y_train)

pred = model.predict(X_test)
pred[:5]
```

```
array([0.25983302, 1.75800328, 0.96221523, 1.32616471, 0.88641452])
```

Figura 13.- *Modelo simple de regresión lineal para la predicción de goles a partir de los minutos jugados y las asistencias*

Paso 14.- Evaluación del modelo

Este valor indica que el Error Cuadrático Medio (MSE) del modelo es aproximadamente 2.70. El MSE calcula la media de las diferencias al cuadrado entre los diferentes valores observados y los valores estimados. Un valor de 0 representaría una predicción perfecta, por lo que, en general, cuanto menor sea el MSE, mejor será el desempeño del modelo. En este caso, un MSE de 2.6968 muestra que existe una diferencia entre las proyecciones y resultados efectivamente observados de goles. Aunque el modelo es capaz de realizar estimaciones, este resultado sugiere que aún hay margen de mejora en su precisión, por ejemplo, incorporando más variables predictoras o utilizando un modelo más adecuado para este tipo de datos (Figura 14).

```
from sklearn.metrics import mean_squared_error

error = mean_squared_error(y_test, pred)
print("Error del modelo:", error)

Error del modelo: 2.6968170281078727
```

Figura 14 - *Evaluación del modelo*

Paso 15.- Residuos vs. goles previstos

A continuación, se muestra los valores de residuos vs. goles previstos permitieron evaluar el comportamiento de los errores del modelo de predicción de goles. Se observa que:

- Los desechos se sitúan alrededor de la línea horizontal del valor cero, lo que muestra que el modelo tiene errores tanto en positivo como en negativo.
- Existe una mayor concentración de observaciones para valores predichos entre 0 y 2.5 goles, lo cual sugiere que la mayoría de los jugadores tienen un número reducido de goles estimados.
- Conforme aumentan los goles predichos, la dispersión de los residuos también aumenta. Este fenómeno indica la presencia de heterocedasticidad, es decir, la variabilidad del error no es constante.
- Se identifican algunos residuos positivos elevados (entre 8 y 12 goles aproximadamente), lo que significa que el modelo subestimó considerablemente el rendimiento goleador de ciertos jugadores.
- También aparecen residuos negativos cercanos a -3, indicando casos donde el modelo sobreestimó el número de goles.
- No se aprecia una tendencia claramente lineal en los residuos; sin embargo, la forma de abanico observada sugiere que el modelo pierde precisión para predicciones más altas (Figura 15).

```
import matplotlib.pyplot as plt
import seaborn as sns

# Calculate residuals
residuals = y_test - pred

# Create a scatter plot of predicted values vs. residuals
plt.figure(figsize=(12, 6))
sns.scatterplot(x=pred, y=residuals)
plt.axhline(y=0, color='r', linestyle='--')
plt.xlabel('Predicted Goals')
plt.ylabel('Residuals')
plt.title('Residuals vs. Predicted Goals')
plt.grid(True)
plt.show()
```

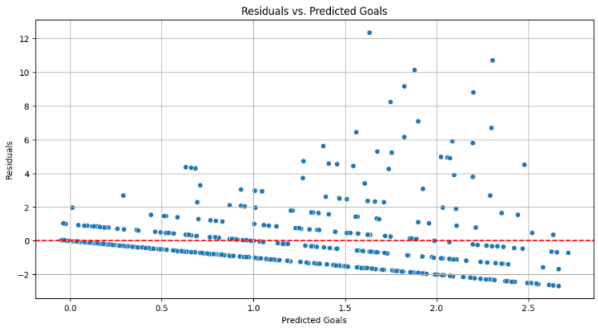


Figura 15. - *Residuos vs. Goles previstos*

Paso 16.- Distribución de los errores

Se pueden observar los siguientes aspectos:

- La mayor parte de los residuos se concentra entre -1 y 1, especialmente cerca del valor **0**. Esto indica que la mayoría de las predicciones presentan errores pequeños, lo que refleja un buen desempeño general del modelo.
- La distribución presenta una asimetría positiva (sesgo hacia la derecha**)**, ya que existe una cola larga con residuos positivos que alcanzan valores cercanos a 12. Esto significa que en algunos casos el modelo **s**ubestimó considerablemente el número real de goles.
- Los residuos negativos son menos extremos y se concentran aproximadamente entre **-**2.5 y 0, indicando que las sobreestimaciones del modelo son de menor magnitud.

- La curva de densidad (KDE) confirma que la distribución no es completamente simétrica ni sigue una distribución normal perfecta.
- Se observan algunos valores atípicos (outliers) en la parte derecha del gráfico, correspondientes a jugadores cuyo rendimiento goleador fue mucho mayor de lo que el modelo predijo (Figura 16).

```
# Create a histogram of residuals
plt.figure(figsize=(8, 5))
sns.histplot(residuals, kde=True)
plt.xlabel('Residuals')
plt.ylabel('Frequency')
plt.title('Distribution of Residuals')
plt.grid(True)
plt.show()
```

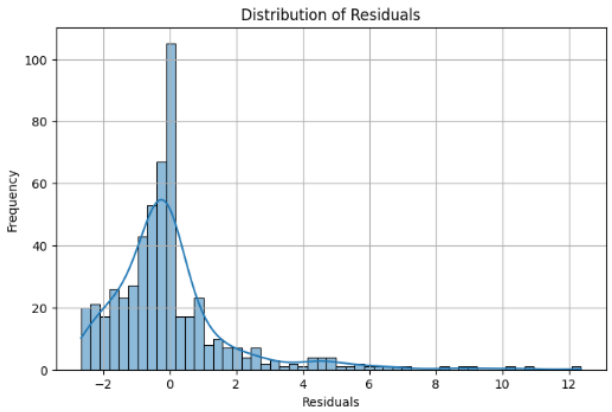


Figura 16. *Distribución de los errores*

CONCLUSIONES

El proceso de análisis de datos permitió preparar adecuadamente el conjunto de datos de estadísticas de jugadores de fútbol mediante la carga, limpieza y selección de variables relevantes, garantizando la calidad de la información utilizada para el análisis y la construcción del modelo predictivo.

El análisis exploratorio de datos (EDA) facilitó la identificación de patrones importantes en el rendimiento de los jugadores y equipos, destacando a los máximos goleadores, asistentes y los equipos con mejor promedio de goles, proporcionando una visión general del comportamiento ofensivo en la temporada 2022–2023.

La relación entre los minutos jugados y la cantidad de goles mostró una tendencia positiva moderada, evidenciando que los jugadores con mayor tiempo de

participación suelen registrar un mayor número de anotaciones, aunque esta relación no es suficiente para explicar completamente el rendimiento goleador.

La matriz de correlación reveló que los minutos jugados presentan una correlación positiva moderada con los goles (0.378), mientras que las asistencias muestran una relación muy débil con los goles (0.105) y prácticamente nula con los minutos (0.007), lo que indica que estas variables, por sí solas, poseen una capacidad limitada para explicar el número de goles anotados.

El modelo de regresión lineal fue capaz de generar predicciones de goles utilizando como variables independientes los minutos jugados y las asistencias. Sin embargo, el error cuadrático medio (MSE = 2.6968) evidencia que aún existe una diferencia considerable entre los resultados predichos y los valores reales, por lo que el modelo presenta una precisión moderada.

El análisis de residuos mostró que la mayoría de los errores se concentran alrededor de cero, indicando un desempeño aceptable para la mayoría de los jugadores. No obstante, también se identificó heterocedasticidad y la presencia de valores atípicos, lo que demuestra que el modelo pierde precisión al estimar jugadores con un alto rendimiento goleador.

La distribución de los residuos evidenció una asimetría positiva y una distribución no completamente normal, indicando que el modelo tiende a subestimar el número de goles de algunos jugadores destacados y que aún no cumple totalmente los supuestos de un modelo de regresión lineal.

En conjunto, el estudio demostró que las técnicas de big data, el análisis exploratorio de datos y el aprendizaje automático constituyen herramientas valiosas para el análisis del rendimiento deportivo, permitiendo descubrir patrones y construir modelos predictivos que pueden apoyar en mejorar la toma de decisiones en el fútbol.

Se recomienda continuar aplicando metodologías de análisis de datos y Big Data en el ámbito deportivo, ya que permiten transformar grandes volúmenes de información en conocimiento útil para optimizar el rendimiento de jugadores y equipos mediante decisiones fundamentadas en evidencia.

REFERENCIAS BIBLIOGRÁFICAS

Alamar, B. C. (2013). *Sports analytics: A guide for coaches, managers, and other decision-makers*. Columbia University Press.

Bangsbo, J. (2014). *Fitness training in football: A scientific approach*. HO+Storm.

Carling, C., Bloomfield, J., Nelsen, L., & Reilly, T. (2008). The role of motion analysis in elite soccer: Contemporary performance measurement techniques and work-rate data. *Sports Medicine, 38*(10), 839–862. https://doi.org/10.2165/00007256-200838100-00004

Carling, C., Reilly, T., & Williams, A. M. (2008). *Performance assessment for field sports*. Routledge.

Fédération Internationale de Football Association. (2024). *FIFA annual report 2024*. https://inside.fifa.com

Hughes, M., & Bartlett, R. (2002). The use of performance indicators in performance analysis. *Journal of Sports Sciences, 20*(10), 739–754. https://doi.org/10.1080/026404102320675602

IBM. (2023). *Artificial intelligence in sports analytics*. https://www.ibm.com

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R* (2nd ed.). Springer.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

Otero-Saborido, F. M., Aguado-Méndez, R. D., Torreblanca-Martínez, V. M., & González-Jurado, J. A. (2021). Technical-tactical performance from data providers: A systematic review in regular football leagues. *Sustainability, 13*(18), 10184. https://doi.org/10.3390/su131810184

Provost, F., & Fawcett, T. (2013). *Data science for business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media.

Reilly, T., & Williams, A. M. (2003). *Science and soccer* (2nd ed.). Routledge.

Rico-González, M., Pino-Ortega, J., Méndez, A., Clemente, F. M., & Baca, A. (2023). Machine learning application in soccer: A systematic review. *Biology of Sport, 40*(1), 249–263. https://doi.org/10.5114/biolsport.2023.112970

Russell, S., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson.

Stats Perform. (2024). *Sports AI and data analytics*. https://www.statsperform.com

Teixeira, J. E., Maio, E., Afonso, P., Encarnação, S., Machado, G. F., Morgans, R., Barbosa, T. M., Monteiro, A. M., Forte, P., Ferraz, R., & Branquinho, L. (2025). Mapping football tactical behavior and collective dynamics with artificial intelligence: A systematic review. *Frontiers in Sports and Active Living, 7*, Article 1569155. https://doi.org/10.3389/fspor.2025.1569155

Union of European Football Associations. (2023). *UEFA football development report 2023*. https://www.uefa.com

Vec, V., Tomažič, S., Kos, A., & Umek, A. (2024). Trends in real-time artificial intelligence methods in sports: A systematic review. *Journal of Big Data, 11*, Article 180. https://doi.org/10.1186/s40537-024-01026-0

***Conflicto de intereses*:**

Los autores declaran no tener conflictos de intereses.

***Contribución de los autores*:**

Los autores han participado en la redacción del trabajo y análisis de los documentos.

Esta obra está bajo una Licencia Creative Commons Reconocimiento-No Comercial Compartir igual 4.0 Internacional
Copyright (c) 2026 *Aurón Eduardo Mendoza Canales, Alberto David Quesada Paz, Carlos Alberto Effio Gonzales*