Volume 11 issue 3; 2026

Ciencia y Deporte

Universidad de Camagüey "Ignacio Agramonte Loynaz"

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

Original article

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

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

[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] César Vallejo University, Lima, Peru.

[*] Corresponding author: amendozaca10@ucvvirtual.edu.pe

Received: *30/09/2026.*
Accepted: *31/10/2026*

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 support in refining sports decision-making, especially in competition 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 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, enabling the discovery of patterns and the construction of predictive models that can support decision-making in football events.

Keywords: decision making, data analysis, sports performance, big data, machine learning.

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.

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.

INTRODUCTION

Digital transformation has significantly impacted sports, particularly football, by fostering the adoption of technologies for collecting, processing, and analyzing the vast amounts of data generated during training and competitions. In this context, professional clubs have employed global positioning systems (GPS), biometric sensors, tracking cameras, and event analytics platforms to record real-time information on the physical, technical, and tactical performance of players (FIFA, 2024). These tools have facilitated the acquisition of high-resolution data, enabling a more objective evaluation of individual and collective performance.

The development of big data represented a significant advance in sports analytics, as it allowed for the identification of playing patterns, the evaluation of player behavior, and the strengthening of team strategic planning. Likewise, the incorporation of artificial intelligence and machine learning techniques increased the accuracy of sports performance assessment by providing models capable of processing large amounts of information and generating more accurate predictions (IBM, 2023).

In football, the availability of data from multiple sources has driven the application of machine learning as a tool to identify complex relationships between variables and build predictive models that overcome the limitations of conventional statistical methods.

Supervised and unsupervised algorithms have been used to simultaneously analyze physical, technical, and tactical indicators, providing relevant information for coaches, physical trainers, and sports analysts during the processes of setting objectives and defining action plans to refine decision-making (Rico-González *et al.,* 2023).

Research conducted in recent years has shown that machine learning techniques have been applied in various areas of professional football, including predicting physical performance, technical-tactical analysis, injury risk assessment, talent identification, and predicting sporting results. Similarly, the integration of artificial intelligence models with spatio-temporal data has allowed for more precise analysis of collective behavior, player interactions, and the tactical organization of teams during competition (Teixeira *et al.,* 2025).

Simultaneously, systematic reviews on artificial intelligence applied to sports have demonstrated sustained growth in the use of deep neural networks, random forests, support vector machines, and reinforced gradient algorithms for real-time sports data processing. These methodologies have contributed to optimizing training planning, improving competitive strategies, and strengthening evidence-based decision-making, establishing themselves as one of the main trends in contemporary sports analytics (Vec *et al.,* 2024).

Furthermore, data-driven technical and tactical analysis facilitated a more precise understanding of the demands of the game by integrating information from tracking systems, sensors, and event analysis platforms. Consequently, big data became established as a strategic resource for evaluating footballers' performance from a holistic perspective, enabling the transformation of heterogeneous data into useful knowledge for coaching staff and sports organizations (Otero-Saborido *et al.,* 2021).

However, the literature also highlighted challenges related to data quality, the integration of multiple information sources, the interpretability of predictive models, and the validation of algorithms in different competitive contexts. These limitations underscored the need for further research exploring the potential of big data and

machine learning to strengthen the objective evaluation of football performance and improve the procedures for selecting options within working groups.

In this context, the performance of soccer players was analyzed using big data and machine learning techniques, evaluating the potential of different artificial intelligence algorithms to identify performance patterns and generate information that would contribute to optimizing training planning and competitive performance.

Big data refers to the manipulation and study of enormous amounts of information that are produced rapidly and in various forms. In football, this data includes match statistics, player movements, heart rate, speed, accelerations, passes, shots, and other variables relevant to sports analysis (Stats Perform, 2024).

Specialized platforms allow for the processing of millions of records to identify patterns that would be difficult to detect through traditional observation. The main applications of big data in football include evaluating players' physical and technical performance and designing tactical strategies through opponent analysis.

Injury prevention using physical workload indicators. Talent identification and recruitment through statistical analysis. Optimization of personalized training programs.

Improving the fan experience through audience behavior analysis (FIFA, 2024).

According to IBM (2023), artificial intelligence allows for the enhancement of data analysis to generate strategic recommendations that support the coaching staff's decisions. Various studies highlight that the use of big data provides significant benefits: It increases the accuracy of decision-making. It reduces the risk of injuries. It improves individual and team performance. It optimizes tactical planning.

It promotes evidence-based player recruitment processes. It increases the competitiveness of clubs (UEFA, 2023).

Regarding the challenges. Despite its advantages, there are challenges related to the implementation of big data: High technological costs. Need for professionals specialized in data science. Protection and privacy of athletes' information. Proper integration between statistical analysis and the coach's experience (FIFA, 2024).

Sports performance in football

Sports performance has been defined in the literature as the quantifiable expression of an athlete's or team's effectiveness in relation to the objectives of the game. According to Bangsbo (2014) and Reilly & Williams (2003), in soccer, performance integrates physical, technical, tactical, and psychological dimensions, which interact dynamically during competition. Analyzing these dimensions allows for the evaluation of individual and collective contributions to team performance.

Performance analysis

Performance analysis is a systematic process of observing, recording, and evaluating athletic behavior to refine decision-making. Hughes and Bartlett (2002) describe it as the application of scientific methods to quantify technical and tactical actions in sports. This analysis has evolved from manual observation to automated systems based on technology and big data.

Big data

Large, high-volume, high-volume, and varied datasets require advanced technologies for storage, processing, and analysis. According to Mayer -Schönberger and Cukier (2013), their value lies not only in the quantity of data but also in the ability to extract useful knowledge from it. In sports, big data allows for the analysis of complex performance patterns using data from sensors, GPS, and tracking systems.

Data science in sports

Data science applied to sports integrates statistics, computer science, and mathematical modeling to transform data into useful knowledge. According to Provost & Fawcett (2013), its goal is to discover patterns, make predictions, and strengthen decision-making. In soccer, it allows for the analysis of large volumes of tactical and physical information to optimize performance.

Artificial intelligence applied to sports

Artificial intelligence in sports encompasses the use of algorithms and computer systems capable of simulating human decision-making processes. According to Russell & Norvig (2021), AI allows for the construction of systems that perceive, reason, and act. In football, it is used for tactical analysis, automated scouting, and predicting results.

Pre-models

Predictive models allow us to estimate future results based on the analysis of historical data (James *et al*., 2021).

Tactical analysis in football

Tactical analysis studies the organization and collective behavior of players during a game. According to Carling *et al.* (2005), it allows for the evaluation of team structure, transitions, and space occupation. Currently, it relies on tracking data and AI algorithms for greater accuracy.

Sports analytics

According to Alamar (2013), sports analytics involves the systematic use of data to support strategic decisions within sports teams and organizations, allowing for an objective evaluation of performance beyond traditional subjective observation. In this vein, the analysis is based on physical, technical, and tactical indicators obtained from technological systems such as sensors, tracking cameras, and event recording platforms.

Tracking systems

Tracking systems allow the recording of players' position and movement in real time using GPS. Carling *et al.* (2008) point out that these systems have revolutionized performance evaluation by providing precise spatial data.

MATERIALS AND METHODS

Focus and type of study

This research employed a quantitative approach, utilizing numerical data collected from a sports database to construct predictive models using big data and machine learning techniques. The analysis will encompass statistical procedures, data mining, and machine learning algorithms to identify patterns associated with player performance.

Type of research

The research is applied, as it seeks to use knowledge about data analysis and artificial intelligence to solve a practical problem related to the evaluation and prediction of the performance of professional soccer players.

Likewise, the study It will be:

- Predictive level, because it aims to develop models capable of estimating player performance based on technical and physical variables.
- Non-experimental design, because the variables will not be manipulated, but existing historical data will be analyzed.
- Cross-sectional or single-cut design, since the analysis will be performed using information corresponding only to the 2022–2023 season.

Population and sample

Population

The population will consist of all records of professional players belonging to the five major European leagues during the 2022–2023 season included in the selected dataset.

The dataset contains approximately:

- More than 2,500 players.
- 124 statistical variables.
- Information pertaining to:
    - Premier League.
    - The League.
    - Bundesliga.
    - Serie A.
    - Ligue 1.

The variables include offensive indicators, defensive indicators, passing indicators, ball possession indicators, discipline indicators, recoveries, goal-making actions, among others.

Sample

Due to the manageable size of the data set, sampling will not be necessary, so a census sampling will be used, considering all available records.

Consequently, the sample will consist of the more than 2,500 players included in the dataset.

Data collection instrument

For data collection, a digital data registration form will be used, created from the dataset downloaded from Kaggle.

The information corresponds to official statistics compiled from Football Reference, subsequently organized and published on Kaggle for the purpose of sports data analysis.

The form will allow you to record variables such as:

- Player's name.
- Age.
- Position.
- Equipment.
- League.
- Matches played.
- Minutes played.
- Goals.
- Attendance.
- Auctions.
- Passes completed.
- Passing accuracy.
- Recoveries.
- Interceptions.
- Tickets.
- Goal creation actions.
- Shot creation actions.
- Cards.
- Aerial duels.
- Among other relevant variables.

Subsequently, the information will be processed using data analysis tools such as Python, Pandas, Scikit- learn, NumPy and Power BI for the development of predictive models.

Validity and reliability

Validity

The validity of the instrument is guaranteed because it is a dataset compiled from official statistics obtained on professional American football, one of the leading international platforms for statistical analysis of professional football. The dataset has been structured for research and data science projects, including standardized information on player performance in the major European leagues.

Furthermore, the validity of the analytical process will be strengthened by:

- data cleansing;
- treatment of missing values;
- normalization of variables where appropriate;
- feature selection (Feature Selection);
- Cross validation during the training of artificial intelligence models.

Reliability

Reliability will be supported because:

- The data comes from official and standardized records;
- The variables were obtained using consistent procedures for all players;
- The processing will be reproducible using scripts developed in Python;
- The predictive models will be evaluated using metrics recognized in Machine Learning.

Depending on the algorithm implemented, indicators such as the following will be used:

- Accuracy.
- Precision.
- Sensitivity (Recall).
- F1 Score (F1-Score).
- Area under the ROC curve (AUC-ROC), for classification problems.
- Mean absolute error (MAE).
- Mean Squared Error (MSE).
- Root Mean Squared Error (RMSE).
- Coefficient of determination (R^2), for regression problems.

These metrics will allow verification of the consistency, stability and predictive capacity of the developed models (Table 1).

Table 1 *- Methodological Summary*

Aspect	Description
Approach	Quantitative
Type of research	Applied
Level	Predictive
Design	Non-experimental, cross-sectional
Population	More than 2,500 professional players from the top five European leagues (2022–2023 season)
Sample	Census sampling (100% of the dataset)
Instrument	Digital data registration form based on the Kaggle dataset
Data source	Dataset "2022-2023 Football Player Stats" (Football Reference/Kaggle)
Validity	Data from official statistics, data cleaning, feature selection, and cross-validation
Reliability	Reproducibility of processing and evaluation using Machine Learning metrics (Accuracy, Precision, Recall, F1, AUC, MAE, RMSE and R^2)

RESULTS AND DISCUSSION

The following details the steps involved in the analysis and development.

Step 1. Choosing the dataset

dataset was sought, and three alternatives were analyzed. For this study, the following dataset was chosen from the link: https://www.kaggle.com/datasets/vivovinco/20222023-football-player-stats. The table "2022–2023 Football Player Stats" is shown here because it meets the research objective: it contains statistical information on 2,689 players from the main European football leagues during the 2022–2023 season. The dataset consists of 124 variables that describe aspects such as the player's personal information, team, position, minutes played, goals, assists, passes, shots, defensive actions, and other performance metrics. This database is suitable for conducting exploratory analyses, statistical studies, and developing machine learning models aimed at evaluating and predicting footballer performance (Figure 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

Figure 1.- *Football player statistics 2022-2023.csv*

Step 2.- Load the Kaggle dataset

Python language in Google Colab , the football statistics dataset was successfully uploaded (Figure 2).



Figure 2.- *Uploading a dataset file*

Step 3.- Installing the official Kaggle library

Kaggle libraries were installed, the dataset was downloaded using the Kaggle API, and all the information in the table was processed (Figure 3).

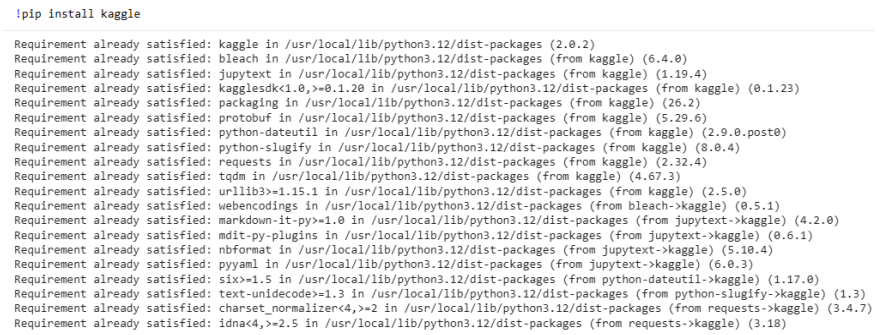


Figure 3.- *Uploading a dataset file*

Step 4.- Import pandas' libraries

- Import the Pandas library, used to manipulate data.
- Read the players_stats.csv file.
- Store all the information in a DataFrame called df .
- df.head () displays the first five rows to verify that the data was loaded correctly (Figure 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

Figure 4. *- Importing pandas' libraries and loading data*

Step 5.- Data Cleaning

- Count how many missing (null) values exist in each column.
- Remove all rows that contain at least one empty value.
- Displays the names of all columns in the dataset (Figure 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)
```

Figure 5 *- Data Cleaning*

Step 6.- Selection of variables

Select only the most important variables for the study. The goal is to reduce the dataset and work only with the necessary information (Figure 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

Figure 6. *- Data Cleaning*

Step 7.- Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) consisted of exploring and understanding the data before building predictive models. It ordered the players from the one with the most goals to the one with the fewest. Then it selected the top ten. Result: It obtained the ranking of the ten top scorers (Figure 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

Figure 7.- *Exploratory data analysis*

Step 8.- Top attendees

The players were ordered by the number of assists. The result was the top ten assist providers. The DataFrame was sorted from highest to lowest by the Assists column, displaying the top ten players with the highest value in that column. However, the results show players with very few minutes played, indicating that Assists likely represents a rate (e.g., assists per 90 minutes) and not the total number of assists. Therefore, the ranking may not reflect the players with the most total assists. For a more representative result, it is recommended to filter players with a minimum number of minutes played before sorting (Figure 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

Figure 8. - *Top attendees*

Step 9.- Show players according to the number of assists

Next, all players were ordered according to the number of assists and the top ten were shown.

It helps to identify the players with the greatest ability to generate goals.

It allows you to answer questions such as:

- Which position scores the most goals?
- Do forwards really set the pace?

The library for creating charts was imported, then the top ten goalscorers were selected.
Afterward, a figure ten inches wide by five inches high was generated.
A bar chart was then drawn.

- X-axis → player names.
- Y-axis → goals.
- Rotate the names 45° so that they do not overlap (Figure 9).

```
import matplotlib.pyplot as plt

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

plt.figure(figsize=(10,5))
plt.bar(top["Player"], top["Goals"])
plt.xticks(rotation=45)
plt.title("Top 10 goleadores")
plt.show()
```

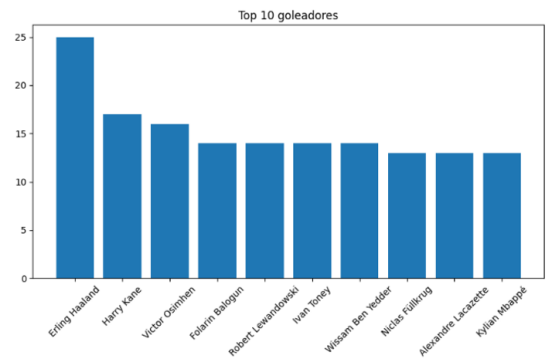


Figure 9. - *Show players according to the number of assists*

Step 10. - Calculate the average number of goals for each team from highest to lowest

This code made it possible to identify which teams have the best offensive performance by calculating the average goals per game for each. It then sorts the results from highest to lowest, facilitating comparison between the 98 teams. present in the dataset. This technique is commonly used in sports analysis to evaluate teams' scoring ability and identify the most effective attacking teams.

This means that:

- Manchester City has the highest average number of goals in the data set, with approximately 2.48 goals per game.
- Paris SG is in second place with an average of 2.16 goals.
- Bayern Munich and Napoli also have averages of more than two goals per game.
- At the bottom end, Sampdoria records the lowest average, with just 0.32 goals per game, indicating a low offensive performance compared to the rest of the teams (Figure 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

***Figure** 10. - Calculate the average number of goals for each team from highest to lowest*

Step 11.- Minutes vs. Goals Ratio

This code creates a scatter plot. plot) to analyze the relationship between minutes played and goals scored by players.

- X-axis (minutes): represents the minutes played by each player.
- Y-axis (goals): represents the number of goals scored.
- Each point corresponds to one player in the data set.

The following shows a positive trend: in general, players who play more minutes are more likely to score more goals. However, the relationship is not absolute. This is a perfect visualization, as there are many players with significant playing time but few goals, as well as some who stand out for scoring many goals. This type of visualization allowed us to identify patterns, potential outliers (players with exceptional performance), and understand how playing time influences the number of goals scored (Figure 11).

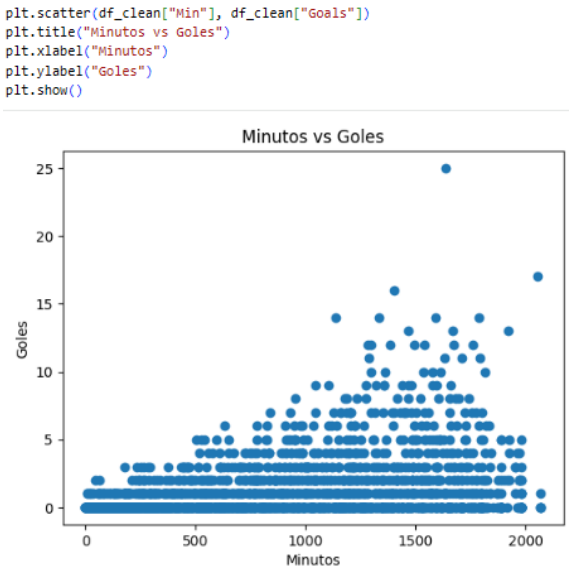


Figure 11. - *Minutes vs. Goals Relationship*

Step 12.- Big data correlation

This code calculates the correlation matrix between the variables goals (goals), assists, and minutes played. The correlation takes values between -1 and 1 (Table 2):

- 1: perfect positive relationship.
- 0: there is no linear relationship.
- -1: perfect negative relationship.

Table 2.- *Correlation between variables*

Variables	Correlation	Interpretation
Goals – Goals	**1**	Perfect correlation with itself.
Goals – Assists	**0.105**	Very weak positive correlation; scoring more goals does not necessarily imply giving more assists.
Goals – Min	**0.378**	Moderate positive correlation; players with more minutes tend to score more goals.
Assists – Min	**0.007**	There is virtually no correlation; in this data set, minutes played do not show a linear relationship with assists.

The correlation matrix indicates that there is a moderate relationship among the minutes played Goals and assists show a very weak relationship with both goals and minutes played. This analysis is useful in big data. and exploratory data analysis (EDA) to identify patterns and relationships between variables before applying statistical or machine learning models (Figure 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

Figure *12. -* *Correlation Matrix between the variable's goals, assists and minutes played*

Step 13. Simple linear regression model

This model learns from other examples where both the input variables (minutes played and assists) and the output (goals) are known, and then uses that learning to predict goals for new players. These values represent the estimated number of goals for the first five players in the test group. Because linear regression produces continuous values, the predictions appear with decimals. For example, a value of 1.758 indicates that the model estimates approximately 2 goals, while a value of 0.260 suggests that the player is unlikely to score any goals or will score very few. In conclusion, the model successfully

generated goal predictions based on minutes played and assists, although to determine its accuracy, it would be necessary to compare these predictions with the actual values using evaluation metrics such as MAE, MSE, or R^2. It cannot be concluded that 1.758 or 0.962 are the best predictions based solely on their values, as they are simply the goal predictions for the first five players in the test group. They do not indicate whether the model is good or bad (Figure 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])
```

Figure 13.- *Simple linear regression model for predicting goals based on minutes played and assists*

Step 14.- Model evaluation

This value indicates that the Mean Squared Error (MSE) of the model is approximately 2.70. The MSE calculates the mean of the squared differences between the observed and estimated values. A value of 0 would represent a perfect prediction, so, in general, the lower the MSE, the better the model's performance. In this case, an MSE of 2.6968 shows that there is a difference between the projected and actually observed goal results. Although the model is capable of making estimates, this result suggests that there is still room for improvement in its accuracy, for example, by incorporating more predictor variables or using a model better suited to this type of data (Figure 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
```

Figure 14 - *Model evaluation*

Step 15.- Waste vs. expected goals

The following table shows the residual values versus predicted goals, which allowed for the evaluation of the error behavior of the goal prediction model. It can be observed that:

- The waste is located around the horizontal line of the zero value, which shows that the model has errors in both positive and negative values.
- There is a higher concentration of observations for predicted values between 0 and 2.5 goals, which suggests that most players have a low number of estimated goals.
- As predicted goals increase, the dispersion of the residuals also increases. This phenomenon indicates the presence of heteroscedasticity, meaning that the variability of the error is not constant.
- Some high positive residuals were identified (between 8 and 12 goals approximately), which means that the model considerably underestimated the goal-scoring performance of certain players.
- Negative residuals close to -3 also appear, indicating cases where the model overestimated the number of goals.
- There is no clearly linear trend in the residuals; however, the observed fan shape suggests that the model loses accuracy for higher predictions (Figure 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()
```

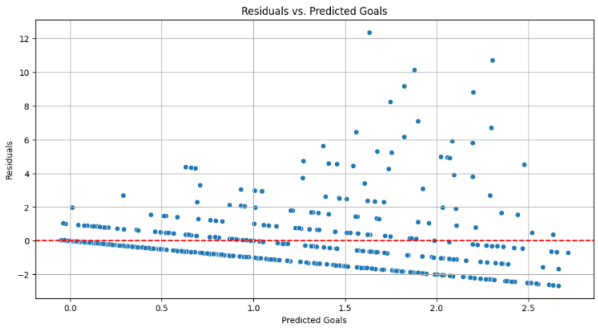


Figure 15. - *Waste vs. Expected Goals*

Step 16.- Distribution of the mistakes

The following aspects can be observed:

- Most of the residuals are concentrated between -1 and 1, especially near the value **0.** This indicates that most predictions have small errors, reflecting good overall model performance.
- The distribution exhibits positive skewness (skewed to the right)**,** as there is a long tail with positive residuals reaching values close to 12. This means that in some cases the model significantly underestimated the actual number of goals.
- The negative residuals are less extreme and are concentrated approximately between **-2.5** and 0, indicating that the model overestimations are of a smaller magnitude.
- The density curve (KDE) confirms that the distribution is not completely symmetrical nor does it follow a perfect normal distribution.

- outliers are observed on the right side of the graph, corresponding to players whose goal-scoring performance was much higher than the model predicted (Figure 16).

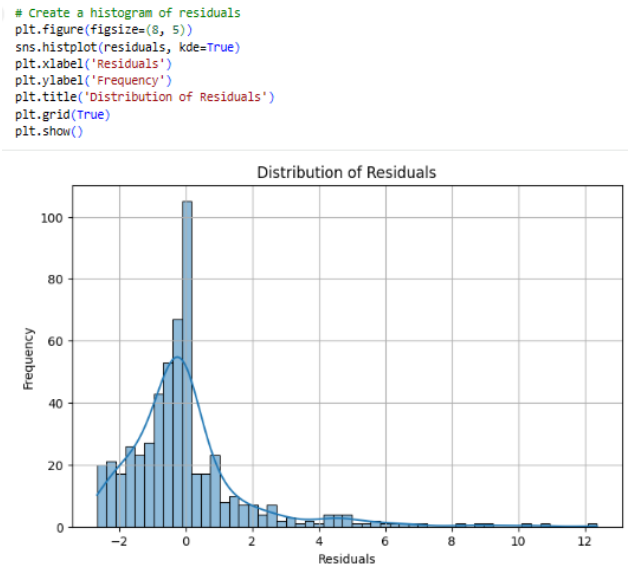


Figure 16. - *Error distribution*

CONCLUSIONS

The data analysis process allowed for the proper preparation of the football player statistics dataset by loading, cleaning, and selecting relevant variables, ensuring the quality of the information used for analysis and the construction of the predictive model.

Exploratory data analysis (EDA) facilitated the identification of important patterns in player and team performance, highlighting top scorers, assist providers, and teams with the best goal average, providing an overview of offensive behavior in the 2022–2023 season.

The relationship between minutes played and the number of goals showed a moderate positive trend, indicating that players with more playing time tend to

register a greater number of goals, although this relationship is not sufficient to fully explain goal-scoring performance.

The correlation matrix revealed that minutes played have a moderate positive correlation with goals (0.378), while assists show a very weak relationship with goals (0.105) and practically zero with minutes (0.007), indicating that these variables, on their own, have a limited capacity to explain the number of goals scored.

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.

Residual analysis showed that most errors are concentrated around zero, indicating acceptable performance for most players. However, heteroscedasticity and outliers were also identified, demonstrating that the model loses accuracy when estimating players with high goal-scoring performance.

The distribution of the residuals showed a positive skewness and a not completely normal distribution, indicating that the model tends to underestimate the number of goals scored by some outstanding players and that it does not yet fully meet the assumptions of a linear regression model.

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

It is recommended to continue applying data analysis and Big Data methodologies in the sports field, as they allow large volumes of information to be transformed into useful knowledge to optimize the performance of players and teams through evidence-based decisions.

BIBLIOGRAPHIC REFERENCES

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

Conflict of interest:
The authors declare no conflicts of interest.

Authors' contribution:
The authors have participated in the writing of the work and analysis of the documents.



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.
Copyright (c) 2026 *Aurón Eduardo Mendoza Canales, Alberto David Quesada Paz, Carlos Alberto Effio Gonzales*