Análisis Exploratorio Rendimiento en la UEFA Champions League 2025

Emmanuel Bustamante Valbuena
Ciencia de datos
Universidad de Antioquia
Med, Colombua
Correo institucional: Emmanuel.bustamante@udea.edu.co

***Resumen*—Este artículo presenta el desarrollo completo de un proyecto de analítica de datos aplicado al rendimiento de jugadores en la UEFA Champions League 2025. Se construyó y documentó un *dataset* de 908 jugadores y 51 variables provenientes de fuentes oficiales de la UEFA y del sitio web oficial de la competición, cubriendo los primeros cuatro partidos de la fase de grupos. El trabajo incluye: (i) un análisis exploratorio de datos (EDA) detallado de métricas físicas, técnicas, ofensivas y defensivas; (ii) la detección y análisis de valores atípicos mediante cuatro enfoques complementarios (IQR, Z-Score, Isolation Forest y DBSCAN); y (iii) una fase de preparación avanzada de datos que aborda normalidad, datos faltantes, codificación de variables categóricas y escalamiento robusto. Los resultados permiten caracterizar patrones tácticos por posición, rangos de edad y perfiles físicos de los jugadores, preservando explícitamente a los outliers legítimos como expresión de excelencia deportiva. El *dataset* resultante queda listo para su uso en modelos predictivos, aplicaciones interactivas y estudios académicos en *sports analytics*, articulado con el repositorio GitHub del proyecto donde se encuentran el código y las visualizaciones.**

***Index Terms*—analítica deportiva, análisis exploratorio de datos, fútbol, UEFA Champions League, detección de atípicos, preparación de datos**

I. Introducción

En el fútbol moderno, la toma de decisiones estratégicas basada en datos se ha convertido en un factor determinante para el éxito competitivo. La planificación táctica, la gestión de cargas físicas, el diseño de entrenamientos y las decisiones de fichajes dependen cada vez más de la capacidad de extraer información accionable a partir de grandes volúmenes de datos de rendimiento.

El presente proyecto se centra en analizar los patrones de rendimiento de los jugadores participantes en la UEFA Champions League 2025, con el objetivo de identificar los factores clave que determinan la excelencia deportiva en el máximo nivel de competición europea. La UEFA Champions League funciona, en este sentido, como un laboratorio de alto rendimiento donde se enfrentan los mejores equipos y jugadores del continente.

I-A. Contexto y relevancia del problema

La UEFA Champions League representa la máxima expresión del fútbol de clubes a nivel mundial, reuniendo a los mejores equipos y jugadores del continente europeo. En la temporada 2025, esta competición continúa siendo el entorno más exigente para evaluar el rendimiento deportivo de élite, donde cada estadística puede marcar la diferencia entre la gloria y la eliminación.

Desde la perspectiva de la analítica deportiva, estudiar la Champions League permite:

- Analizar la distribución del talento y del rendimiento en un entorno altamente competitivo.
- Identificar patrones de juego y perfiles de jugadores que son sostenibles en el máximo nivel.
- Explorar la relación entre métricas físicas, técnicas y tácticas y los resultados obtenidos en el campo.

Por tanto, comprender qué caracteriza a los jugadores de élite en este contexto no solo tiene interés académico, sino también impacto directo en procesos de scouting, diseño de modelos de predicción de rendimiento y evaluación de estrategias tácticas.

I-B. Preguntas de investigación

A partir de este contexto, el proyecto se articula en torno a las siguientes preguntas de investigación:

- ¿Cuáles son los patrones de rendimiento que caracterizan a los jugadores de élite en la UEFA Champions League 2025?
- ¿Qué factores técnicos, físicos y tácticos determinan el éxito ofensivo y defensivo de los jugadores?
- ¿Cómo se distribuye el talento y las características de rendimiento por posiciones y equipos?
- ¿Existen correlaciones significativas entre las características físicas de los jugadores y su rendimiento en el campo?
- ¿Cuál es el perfil de rendimiento específico del Real Madrid en comparación con el resto de equipos participantes?

Estas preguntas guían el diseño del análisis exploratorio, la selección de métricas clave y la interpretación de los resultados, conectando el trabajo con problemas reales de toma de decisiones en clubes de alto nivel.

I-C. Relación entre el problema y la base de datos seleccionada

El *dataset* seleccionado para este análisis proviene de fuentes oficiales de la UEFA y de procedimientos de *scraping* sobre el sitio web oficial de la Champions League, lo que garantiza

la autenticidad y precisión de los datos de rendimiento. La base consolida:

- 908 registros de jugadores.
- 51 variables que abarcan dimensiones físicas, técnicas, ofensivas, defensivas, de portería, biométricas y contextuales.

Esta estructura convierte a la base de datos en un insumo idóneo para abordar las preguntas de investigación planteadas, al permitir:

- Comparar el rendimiento entre posiciones, equipos y rangos de edad.
- Analizar correlaciones entre características físicas (por ejemplo, velocidad, distancia recorrida) y resultados en el campo (goles, asistencias, acciones defensivas).
- Construir perfiles específicos, como el del Real Madrid, y contrastarlos con el resto de la competición.

I-D. Objetivos del proyecto

Con base en lo anterior, el objetivo general de este trabajo es presentar, en formato de artículo académico, un proceso analítico completo sobre datos de rendimiento de la UEFA Champions League 2025, desde la construcción de la base hasta la preparación para modelado.

De forma específica, se plantean los siguientes objetivos:

- Describir la fuente, estructura y calidad del *dataset* de rendimiento de jugadores de Champions League.
- Aplicar técnicas de EDA para caracterizar distribuciones, relaciones y patrones tácticos relevantes.
- Implementar y comparar múltiples técnicas de detección de valores atípicos, interpretando su significado en el contexto deportivo.
- Definir criterios de tratamiento de faltantes, codificación de variables categóricas y escalamiento numérico, dejando el *dataset* listo para modelado.
- Documentar el proceso de forma alineada con el repositorio GitHub del proyecto, garantizando reproducibilidad.

El resto del artículo se organiza de la siguiente manera. En la Sección II se describe la metodología seguida, incluyendo la fuente de datos, las técnicas estadísticas y las herramientas empleadas. En la Sección III se presentan los resultados del EDA, el análisis de atípicos y la preparación de datos, con su respectiva interpretación. La Sección IV resume las conclusiones, limitaciones y líneas de trabajo futuro. Finalmente, se incluye una sección de correspondencia con el repositorio y otra de consideraciones éticas.

II. Metodología

II-A. Fuente y características del conjunto de datos

El *dataset* utilizado recopila información de rendimiento de jugadores en los primeros cuatro partidos de la fase de grupos de la UEFA Champions League 2025. Los datos provienen de:

- Sistemas de *tracking* físico (GPS, cámaras de alta frecuencia).
- Estadísticas técnicas y de juego registradas en tiempo real.
- Información biométrica y contextual proporcionada por la organización y el sitio oficial de la competición.

La base de datos final comprende:

- 908 registros, cada uno correspondiente a un jugador en la competición.
- 51 variables documentadas, que incluyen métricas físicas, técnicas, ofensivas, defensivas, de portería, biométricas y contextuales.
- Una mezcla de variables continuas, discretas y categóricas: 9 continuas, 34 discretas y 8 categóricas.

Entre las variables continuas principales se destacan:

- `passing_accuracy(%)`: precisión de pases.
- `distance_covered(km/h)`: distancia recorrida.
- `top_speed`: velocidad máxima.
- `minutes_played`: minutos jugados.

Entre las discretas principales se encuentran:

- `goals`: goles anotados.
- `assists`: asistencias.
- `total_attempts`: intentos de tiro.
- `tackles_won`: entradas ganadas.
- `age`: edad del jugador.

La organización del proyecto de datos se estructuró de la siguiente forma:

```
/Data_analysis/
 master_df.csv
 tabla_descriptiva_variables.csv
 Cp1_variable_analysis_Emmanuel_Valbuena.ipynb
 streamlit_app.py
```

donde `master_df.csv` contiene el *dataset* completo, `tabla_descriptiva_variables.csv` documenta cada variable y el cuaderno `.ipynb` recoge el análisis exploratorio y las pruebas estadísticas. Además de contar con una aplicación desplegada e interactiva conservada en `streamlit`

II-B. Proceso de análisis exploratorio de datos (EDA)

El EDA se desarrolló en varias etapas:

1. Cálculo de estadísticos descriptivos (media, mediana, desviación estándar, rango, moda) para variables continuas y discretas.
2. Análisis de distribuciones mediante histogramas, diagramas de caja (*boxplots*) y gráficos de barras para variables categóricas.
3. Construcción de una matriz de correlación para variables continuas, complementada con gráficos de dispersión (*scatter plots*) para pares de interés.
4. Tablas cruzadas para estudiar el rendimiento medio por posición, rango de edad, equipo y, en particular, el perfil del Real Madrid respecto al resto de la competición.

II-C. Detección y análisis de datos atípicos

Para el estudio de valores atípicos se aplicaron cuatro enfoques complementarios:

- Método del rango intercuartílico (IQR).

Cuadro I
EJEMPLO DE ESTADÍSTICOS DESCRIPTIVOS PARA VARIABLES CLAVE.

Variable	Media	Mediana	DE	Rango
Goles	0,85	0	1,30	[0, 8]
Asistencias	0,52	0	0,90	[0, 5]
Precisión pases (%)	82,45	85,20	12,34	[45,2, 98,7]
Min. jugados	245	280	125	[1, 360]

- Puntuación estandarizada Z-Score.
- Isolation Forest (aprendizaje automático).
- DBSCAN (clustering basado en densidad).

Los métodos IQR y Z-Score se aplicaron de forma univariada, mientras que Isolation Forest y DBSCAN capturan combinaciones atípicas en múltiples dimensiones (ofensivas, físicas y defensivas). El criterio central de decisión fue no eliminar valores atípicos válidos, sino comprender su origen e incorporarlos al análisis como casos relevantes.

II-D. Preparación avanzada de datos

La Fase 4 del proyecto se centró en dejar el *dataset* listo para modelado y despliegue (por ejemplo, en una aplicación de Streamlit), abordando cuatro frentes:

1. Normalidad y forma de las distribuciones: pruebas de Shapiro-Wilk, asimetría, curtosis y Q-Q plots.
2. Análisis de valores faltantes y su patrón (*Missing Not At Random*, MNAR).
3. Codificación de variables categóricas según su cardinalidad.
4. Escalamiento de variables numéricas mediante diferentes estrategias.

El resultado de esta fase fueron diferentes versiones del *dataset* (`df_encoded`, `df_robust`, `df_standardized`, `df_normalized`), adecuadas para diversas familias de modelos.

II-E. Herramientas y librerías utilizadas

El proyecto se implementó en Python, utilizando principalmente:

- `pandas` y `NumPy` para manipulación y análisis de datos.
- `matplotlib` y `seaborn` para visualización.
- `scikit-learn` para escalamiento, detección de outliers (Isolation Forest, DBSCAN) y preparación de *pipelines*.
- `Streamlit` para la construcción de una interfaz de exploración interactiva (cuando corresponde).

El código completo, los cuadernos y las salidas generadas se encuentran en el repositorio GitHub del proyecto[1].

III. RESULTADOS Y ANÁLISIS

III-A. Descripción estadística del rendimiento

III-A1. Variables continuas: En `passing_accuracy(%)` se observó una media de 82,45 % y una mediana de 85,20 %, con desviación estándar de 12,34 % y rango entre 45,2 % y 98,7 %. Esto indica que la mayoría de jugadores mantienen una alta precisión de pase, con dispersión moderada asociada a los diferentes roles tácticos.

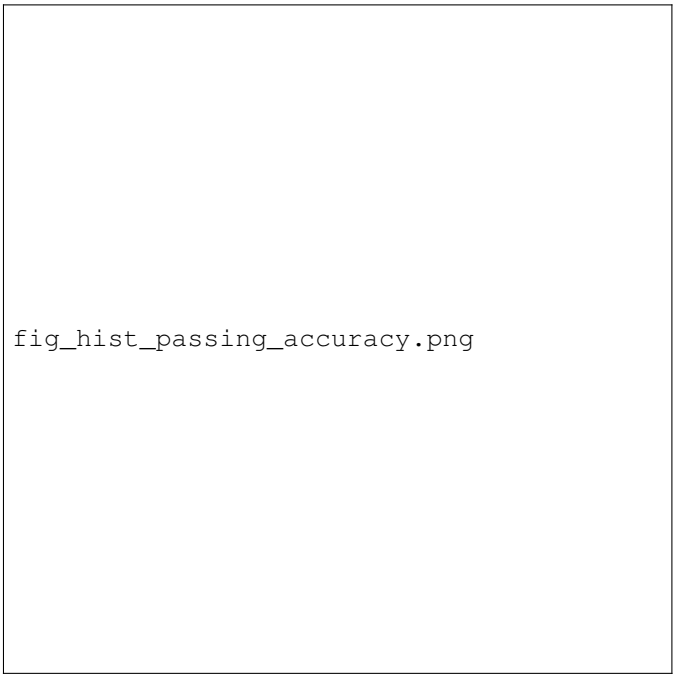

Figura 1. Ejemplo de histograma de la distribución de la precisión de pases.

La variable `distance_covered(km/h)` presenta una media de 10,73 km/h, mediana de 10,45 km/h y desviación estándar de 1,87 km/h, con rangos entre 4,2 y 15,8 km/h. Estos valores reflejan la alta exigencia física de la competición, con mediocampistas y laterales recorriendo mayores distancias que delanteros y porteros.

La `top_speed` media es de 31,25 km/h, con mediana de 31,80 km/h, desviación estándar de 2,45 km/h y rango entre 22,1 y 37,2 km/h. Se confirma la presencia de velocistas de élite (extremos y delanteros) capaces de superar los 36–37 km/h en *sprints* máximos.

`minutes_played` presenta una media de 245 minutos, mediana de 280 minutos, desviación estándar de 125 minutos y rango entre 1 y 360 minutos. La alta variabilidad se explica por rotaciones, lesiones y diferencias entre titulares y suplentes.

III-A2. Variables discretas: La variable `goals` tiene una media de 0,85 goles, mediana y moda iguales a 0 y un máximo de 8 goles. La distribución es altamente asimétrica: cerca del 70 % de los jugadores no ha marcado en los cuatro primeros partidos, mientras que un grupo reducido de delanteros concentra la mayoría de los goles.

En `assists`, la media es de 0,52 asistencias, con mediana y moda en 0 y un máximo de 5. El patrón es similar al de los goles, concentrado en mediocampistas creativos y extremos.

`total_attempts` presenta una media de 3,25, mediana de 2, desviación estándar de 4,15 y un máximo de 25 intentos,

[1]URL del repositorio GitHub: https://github.com/Emma-Ok/Data-science-project.

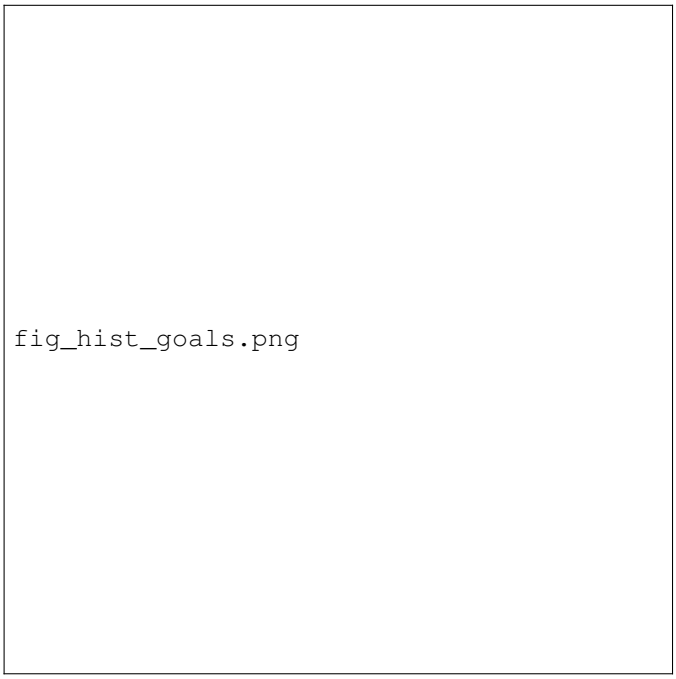

Figura 2. Distribución de goles por jugador en los primeros cuatro partidos.

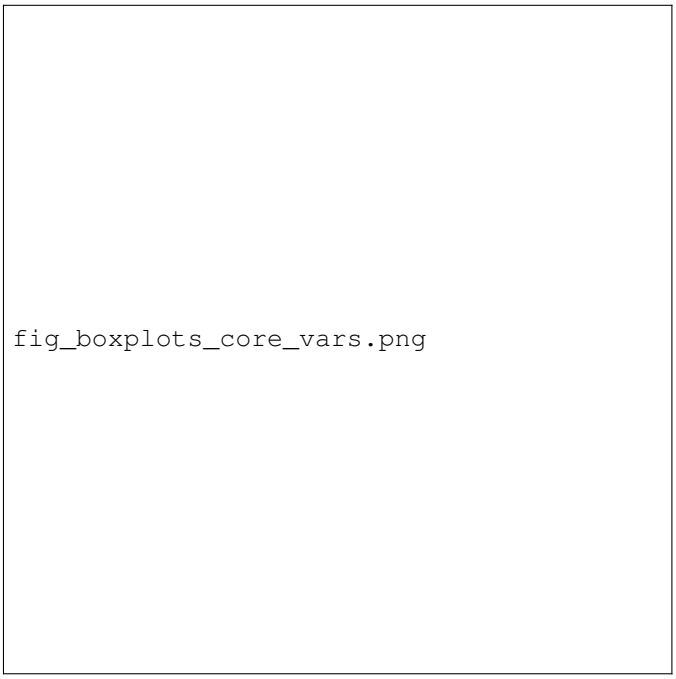

Figura 3. Diagramas de caja para métricas ofensivas y físicas seleccionadas.

evidenciando que los delanteros generan significativamente más tiros que el resto de posiciones.

III-B. Patrones univariados y bivariados

III-B1. Distribuciones e histogramas: El análisis de histogramas permitió distinguir distintos tipos de distribuciones:

- Distribuciones aproximadamente normales: `passing_accuracy(%)`, `age`, `distance_covered(km/h)` (con ligeros sesgos).
- Distribuciones asimétricas positivas: `goals`, `assists`, `total_attempts`, `top_speed`.
- Distribuciones bimodales: `minutes_played`, con picos asociados a suplentes y titulares.

Los diagramas de caja evidenciaron valores atípicos coherentes con el juego.

III-B2. Correlaciones y relaciones clave: La matriz de correlación de variables continuas mostró:

- Una correlación fuerte y positiva entre `minutes_played` y `distance_covered(km/h)` ($r \approx 0{,}89$).
- Una correlación positiva moderada entre `total_attempts` y `goals` ($r \approx 0{,}58$).
- Una correlación positiva moderada entre `passing_accuracy(%)` y `assists` ($r \approx 0{,}45$).
- Una correlación negativa débil entre `age` y `top_speed` ($r \approx -0{,}32$).

Los gráficos de dispersión (*scatter plots*) entre intentos y goles mostraron una relación lineal con dispersión considerable.

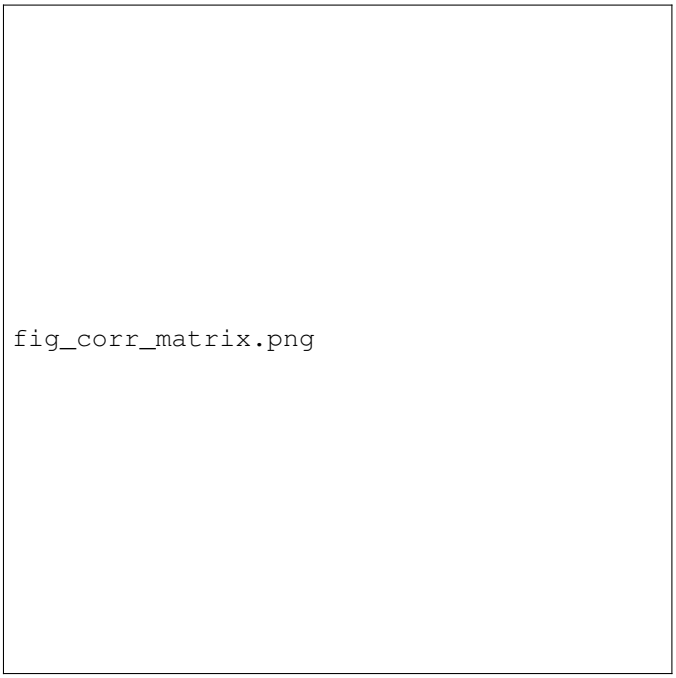

Figura 4. Matriz de correlación para variables continuas de rendimiento.

III-C. Resultados por grupos y patrones tácticos

Las tablas cruzadas permitieron caracterizar el rendimiento por posición:

- Goles promedio por posición: delanteros (FW) con $\approx$ 2,15 goles, mediocampistas (MF) con 0,68 y defensores (DF) con 0,22; los porteros (GK) prácticamente no anotan.

Figura 5. Relación entre intentos de tiro y goles anotados.

Figura 6. Proyección PCA de jugadores con clusters DBSCAN y outliers resaltados.

- Precisión de pases: mediocampistas alrededor de 85,3 %, defensas 83,7 %, delanteros 76,2 % y porteros 71,5 %.
- `tackles_won`: defensores con $\approx$ 7,8, mediocampistas con 5,2 y delanteros con 1,9 en promedio.

Cuadro II
EJEMPLO DE RENDIMIENTO PROMEDIO POR POSICIÓN.

Posición	Goles	Precisión pases	Tackles ganados
FW	2,15	76,2 %	1,9
MF	0,68	85,3 %	5,2
DF	0,22	83,7 %	7,8
GK	0,00	71,5 %	1,1

Por rangos de edad se observó un pico de rendimiento integral entre los 24 y 28 años, consistente con literatura previa sobre madurez física y táctica.

III-D. Análisis y comparación de valores atípicos

En la Tabla III se resume el número de atípicos detectados por cada técnica.

Cuadro III
RESUMEN DE TÉCNICAS DE DETECCIÓN DE VALORES ATÍPICOS.

Técnica	Outliers	%	Tipo	Sensibilidad
IQR	85	9,4	Estadística clásica	Media
Z-Score	45	5,0	Estadística paramétrica	Alta
Isolation Forest	45	5,0	Aprendizaje automático	Media-alta
DBSCAN	38	4,2	Clustering no supervisado	Variable

Los métodos multivariados identificaron perfiles únicos, como defensores con estadísticas ofensivas atípicas o delanteros con baja eficiencia de finalización.

La decisión final fue no eliminar outliers, sino conservarlos como expresión de excelencia o situaciones tácticas particulares, documentando su presencia mediante variables auxiliares (`is_outlier`, `outlier_reason`).

III-E. Normalidad, faltantes, codificación y escalamiento

Los test de normalidad y las inspecciones visuales mostraron que variables como `passing_accuracy(%)` y `distance_covered(km/h)` se aproximan a distribuciones gaussianas, mientras que `goals` o `total_attempts` presentan asimetrías marcadas.

El análisis de valores faltantes evidenció un patrón *Missing Not At Random* (MNAR): muchas ausencias corresponden a jugadores sin minutos o roles que no generan ciertas métricas. Algunas columnas superan el 40–50 % de faltantes, por lo que la imputación habría introducido supuestos fuertes. Se optó por no imputar de forma masiva, recurriendo a análisis segmentados e indicadores binarios.

Para las variables categóricas se adoptó una codificación híbrida y, en cuanto al escalamiento, se compararon *StandardScaler*, *MinMaxScaler* y *RobustScaler*. Dado el peso de los atípicos legítimos, *RobustScaler* se eligió como estándar.

IV. CONCLUSIONES

Este trabajo consolida un proceso analítico completo aplicado al rendimiento de jugadores en la UEFA Champions League 2025, desde la construcción y documentación de la base de datos hasta la preparación para modelado avanzado.

Las principales contribuciones pueden resumirse en:

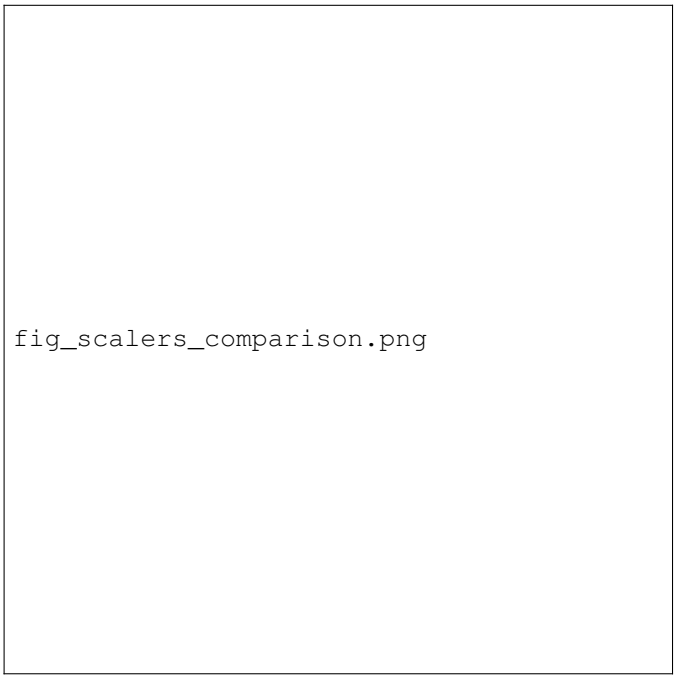

Figura 7. Comparación de distribuciones tras aplicar diferentes técnicas de escalamiento.

- Construcción de una base de datos robusta de 908 jugadores y 51 variables, con documentación detallada y clasificación rigurosa de tipos de datos.
- Desarrollo de un EDA exhaustivo que caracteriza distribuciones, relaciones y patrones tácticos por posición, edad y perfil físico.
- Aplicación de cuatro técnicas complementarias de detección de valores atípicos, con interpretación contextual que evita la eliminación indiscriminada de casos extremos.
- Definición de una estrategia de preparación avanzada de datos (faltantes, codificación y escalamiento) alineada con las necesidades del modelado y respetuosa de la realidad competitiva.
- Articulación explícita entre el manuscrito y el repositorio GitHub del proyecto, favoreciendo la reproducibilidad y la trazabilidad de resultados.

Entre las principales limitaciones se destacan:

- El período de análisis se limita a los primeros cuatro partidos de la fase de grupos, lo que puede no capturar dinámicas de medio y largo plazo.
- No se incorporan explícitamente variables tácticas como formaciones, estilo de juego por entrenador o calidad relativa del rival.
- Las decisiones de no imputación implican trabajar con subconjuntos de datos para ciertos análisis.

Como líneas de trabajo futuro se proponen:

- Extender el horizonte temporal a toda la temporada de Champions League, incluyendo fases eliminatorias.
- Integrar modelos de probabilidad de gol (xG) y redes de pases para profundizar en el análisis táctico.
- Desarrollar modelos predictivos para resultados de partido, evolución de rendimiento y riesgo de lesión, utilizando el *dataset* preparado.
- Implementar explicabilidad (*feature importance*, SHAP) para apoyar decisiones de scouting, planificación táctica y gestión de carga.

Correspondencia con el repositorio GitHub

El contenido del manuscrito se encuentra alineado con el proyecto técnico en GitHub, donde se alojan:

- Los archivos `master_df.csv` y `tabla_descriptiva_variables.csv`.
- El cuaderno `Cp1_variable_analysis_Emmanuel_Valbuena_final.ipynb` con el EDA completo.
- Los scripts de detección de atípicos, codificación y escalamiento.

Cualquier gráfico, tabla o métrica reportada en este artículo puede rastrearse hasta las celdas correspondientes del cuaderno y los scripts en el repositorio.

Ética y responsabilidad en el uso de datos

El proyecto respeta buenas prácticas éticas en el manejo de datos de rendimiento deportivo:

- Los datos proceden de fuentes oficiales de la UEFA, utilizadas con fines académicos y de investigación.
- No se trabaja con información sensible de carácter personal más allá de variables deportivas y biométricas habituales (edad, altura, peso), y el análisis se realiza a nivel agregado.
- Se mantiene transparencia en la documentación de las decisiones de tratamiento de datos, en particular la gestión de atípicos y faltantes.
- Se evita la extracción de conclusiones que puedan estigmatizar a jugadores individuales; el foco está en patrones colectivos y perfiles de rendimiento.

Referencias

[1] R. J. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, 2nd ed. New York, NY, USA: Wiley, 2002.
[2] S. van Buuren, *Flexible Imputation of Missing Data*, 2nd ed. Boca Raton, FL, USA: CRC Press, 2018.
[3] J. W. Graham, "Missing data analysis: Making it work in the real world," *Annual Review of Psychology*, vol. 60, pp. 549–576, 2009.
[4] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, 2nd ed. New York, NY, USA: Springer, 2021.
[5] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[6] W. McKinney, "Data structures for statistical computing in Python," in *Proc. 9th Python in Science Conf.*, 2010, pp. 51–56.
[7] C. R. Harris *et al.*, "Array programming with NumPy," *Nature*, vol. 585, pp. 357–362, 2020.
[8] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.
[9] M. L. Waskom, "Seaborn: Statistical data visualization," *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, 2021.
[10] Streamlit, "Streamlit Documentation." Disponible en: https://docs.streamlit.io/. [Accedido: ajustar fecha de consulta].

[11] S. Alamar, *Sports Analytics: A Guide for Coaches, Managers, and Other Decision Makers*. New York, NY, USA: Columbia Univ. Press, 2013.
[12] UEFA, "UEFA Champions League statistics and data," documentación técnica disponible en: https://www.uefa.com.