Optimización del Rendimiento de un Equipo de Fútbol mediante Análisis Estadístico

Objetivo:

El objetivo de esta investigación es identificar a los jugadores de fútbol más eficientes y efectivos para cada posición en el campo, utilizando un enfoque estadístico basado en métricas avanzadas de rendimiento.

1. Proceso de Webscrapping

El primer paso para realizar el **web scraping** fue identificar las páginas de interés donde se encontraban las estadísticas de los jugadores. En este caso, la página **FBREF** proporciona tablas estructuradas con estadísticas detalladas para cada jugador en varias temporadas, como se puede observar en la siguiente URL.

https://fbref.com/en/players/89ac64a6/matchlogs/2023-2024/summary/Manuel-Akanji-Match-Logs

Esta URL contiene las estadísticas de Manuel Akanji para la temporada 2023-2024. El patrón de la URL es consistente para otros jugadores, lo que permitió automatizar el proceso de extracción de datos. Por tanto, se escogió esta página para obtener los datos de los jugadores de clubes ingleses en la temporada 2023-2024.

Se utilizó la función pd.read_html() de la librería **pandas**, que es capaz de leer tablas directamente desde una página web. Para realizar un scraping eficiente de las estadísticas de múltiples jugadores, se construyó un proceso iterativo que recorriera las URLs de cada jugador. Las URLs de los jugadores estaban estructuradas de forma similar, lo que permitió generalizar el proceso. Además, se implementaron temporizadores y se utilizó una VPN para evitar que la página web bloqueara la IP, debido a su estricta política antibots.

De esta forma se pudieron obtener las siguientes métricas por partido para cada jugador:

• Fecha • Día • Competición • Ronda • Lugar • Resultado • Equipo • Oponente • Titularidad • Posición • Minutos • Goles • Asistencias • Penales Marcados • Penales Intentados	• Tiros • Tiros al Arco • Tarjetas Amarillas • Tarjetas Rojas • Toques • Tacleadas • Intercepciones • Bloqueos • Goles Esperados (xG) • Goles No Penales Esperados (npxG) • Asistencias Esperadas (xAG) • Pases Completados	• Acciones Creadoras de Gol (SCA) • Contribuciones de Gol (GCA) • Pases Intentados • Porcentaje de Pases Completados • Pases Progresivos • Conducciones • Conducciones Progresivas • Intentos de Regate • Regates Exitosos

Y para porteros se obtuvieron las siguientes métricas de la temporada:

• Goles Esperados Post-Tiro vs Goles en Contra (PSxG-GA) • Goles en Contra • Porcentaje de Paradas • Goles Esperados Post-Tiro por Tiro al Arco (PSxG/SoT) • Porcentaje de Paradas en Penales	• Toques • Porcentaje de Lanzamientos • Saques de Meta • Longitud Promedio de Saques de Meta • Porcentaje de Centros Detenidos • Acciones Defensivas Fuera del Área	• Porcentaje de Partidos con Portería a Cero • Distancia Promedio de Acciones Defensivas

2. **Preprocesamiento de Datos**

Tratamiento de Datos Nulos

Uno de los primeros pasos en el preprocesamiento fue el manejo de los valores nulos. Se revisaron las filas con datos nulos y se aplicaron varias estrategias de limpieza:

- **Filas con fecha nula**: Se determinó que las filas que contenían una columna de fecha vacía, generalmente, tenían la mayoría de sus valores nulos. Debido a esto, estas filas fueron eliminadas completamente del dataset.
- **Jugadores que no jugaron**: Otra situación común fue cuando un jugador estaba presente en la convocatoria pero no jugó el partido, resultando en estadísticas nulas. Estas filas fueron también filtradas, ya que no aportaban información relevante para el análisis.
- **Competencias fuera de la Premier League**: Al revisar el origen de las filas con valores nulos, se observó que la mayoría correspondía a competiciones que no eran la Premier League. Para garantizar la coherencia en el análisis, se decidió filtrar el dataset para incluir únicamente partidos correspondientes a la Premier League.

Este enfoque garantizó que los datos finales fueran consistentes y relevantes para el análisis.

Normalización del Estado Final del Partido
Se identificó que la columna de "estado final" contenía una mezcla de información sobre el resultado del partido, por ejemplo, valores como W 2-2 indicaban una victoria con 2 goles a favor y 2 en contra. Para facilitar el análisis, esta columna fue descompuesta en tres nuevas columnas:

- **Resultado_Final**: que indicaba si el equipo ganó (W), empató (D), o perdió (L).
- **Goles_Favor**: número de goles anotados por el equipo del jugador.
- **Goles_Contra**: número de goles recibidos.

Esta transformación permite realizar análisis más detallados de rendimiento, tanto a nivel individual como de equipo.

Creación de la Columna "Posición Normalizada"
Dado que los jugadores pueden ocupar diferentes posiciones en distintos partidos, fue necesario agrupar esas posiciones en categorías generales. Se creó la columna **Posición_Normalizada**, que asigna una posición general según el rol principal del jugador:

- **Portero**
- **Defensor**
- **Mediocampista**
- **Delantero**

Si la posición de un jugador en un partido estaba registrada como nula, se imputó utilizando la última posición en la que el jugador había sido registrado. Esto permitió consolidar las posiciones y asegurar consistencia en el análisis.

Corrección de Tipos de Datos

El siguiente paso fue la revisión y corrección de los tipos de datos en cada columna. Muchas columnas que contenían datos numéricos estaban mal clasificadas como cadenas de texto. Se realizó la conversión de las columnas relevantes para permitir el cálculo de métricas y facilitar el análisis estadístico posterior.

3. **Transformación de Datos (Creación de nuevas Métricas)**

El proceso de transformación de datos se centró en la creación de nuevas variables que permitieran obtener un análisis más completo del desempeño tanto individual de los jugadores como de los equipos rivales y aliados.

Variables Agregadas por Equipo

Se generaron variables que reflejaban el desempeño colectivo del equipo rival y del equipo aliado, basadas en la suma o el promedio de métricas clave de los jugadores:

Asistencias, Penales Marcados e Intentados, Tarjetas,Toques y Acciones Defensivas: , Goles y Asistencias Esperadas (xG, xAG), Acciones y Contribuciones de Gol (SCA, GCA), Pases y Conducciones, Intentos y Regates Exitosos.

Normalización de Datos

Para asegurar que las comparaciones entre jugadores fueran coherentes, se normalizaron las métricas individuales dividiendo cada estadística por los minutos jugados, y luego multiplicando el resultado por 90.

Las métricas normalizadas incluyeron:

- **Goles por 90 minutos**
- **Asistencias por 90 minutos**
- **Tiros por 90 minutos**
- **Tacleadas, Intercepciones, Bloqueos y Pases por 90 minutos**
- **Conducciones y Regates por 90 minutos**

Creación de Nuevas Variables

Se añadieron nuevas métricas para mejorar el análisis del desempeño individual de los jugadores:

- **xOVA**: Una métrica que mide el valor ofensivo agregado de un jugador. Se calculó sumando los goles esperados no penales (npxG) y las asistencias esperadas (xAG), restando las asistencias esperadas del equipo.
- **PPDA (Passes Allowed per Defensive Action)**: Esta métrica mide la presión defensiva de un jugador, dividiendo los pases completados del oponente entre las acciones defensivas del jugador.
- **Efectividad de Tiros**: Se calculó dividiendo los goles por los tiros realizados.

Agrupación de Estadísticas por Jugador

Se procedió a agrupar las estadísticas por jugador utilizando promedios y sumas lógicas.

Rankings

Finalmente, se crearon columnas que establecían un ranking para cada jugador en función de sus contribuciones:

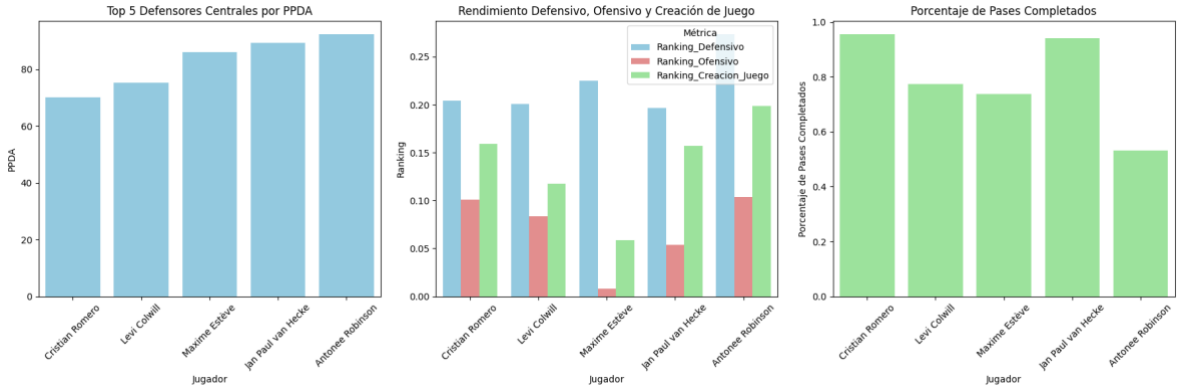
- **Ranking Ofensivo**: Calculado a partir de los goles esperados (xG) y las asistencias esperadas (xAG).
- **Ranking Defensivo**: Basado en las tacleadas, intercepciones y bloqueos por 90 minutos.
- **Ranking de Creación de Juego**: Involucraba métricas como pases progresivos, regates exitosos, y acciones creadoras de gol (SCA), además de contribuciones de gol (GCA).
- **Ranking de Pases**: Se evaluó el porcentaje de pases completados por cada jugador.

Estas métricas fueron normalizadas con el algoritmo MinMax.

4. Análisis Estadístico por Posición

A continuación, se presenta el análisis estadístico para la selección de jugadores en cada posición, con las métricas específicas que respaldan la decisión. Se explican las métricas clave utilizadas para cada posición y por qué ciertos jugadores fueron seleccionados sobre otros. Cabe aclarar que para realizar el equipo, se escogió una formación típica del fútbol: 4-3-3.

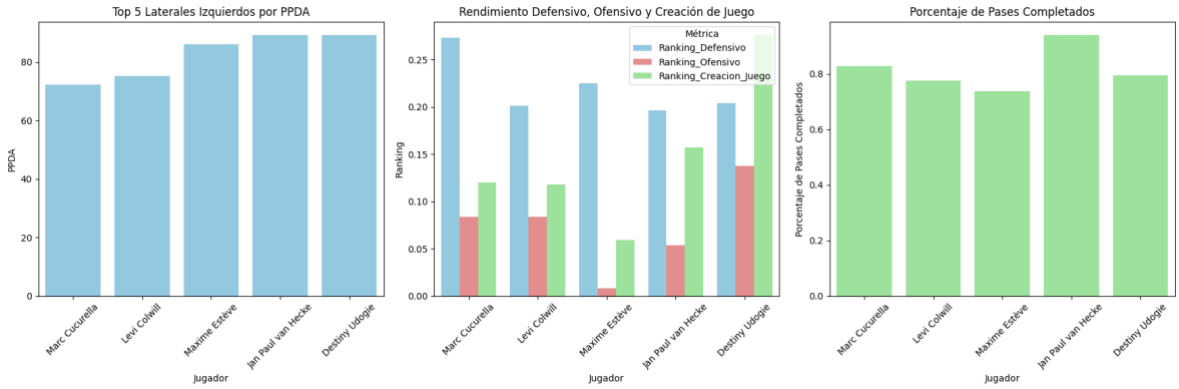
Defensas Centrales



Se parte de la métrica **PPDA** (pases permitidos por acción defensiva) como indicador clave de la presión defensiva de los centrales. **Cristian Romero** y **Levi Colwell** fueron seleccionados debido a su combinación de bajo **PPDA** y buen **Ranking Defensivo**. Ambos jugadores también tienen un alto **Porcentaje de Pases Completados**, lo cual es crucial para una defensa eficiente.

- **Cristian Romero**:
  - PPDA: 70.20
  - Ranking Defensivo: 0.204
  - Porcentaje de Pases Completados: 91.81%
- **Levi Colwill**:
  - PPDA: 75.18
  - Ranking Defensivo: 0.200
  - Porcentaje de Pases Completados: 83.81%

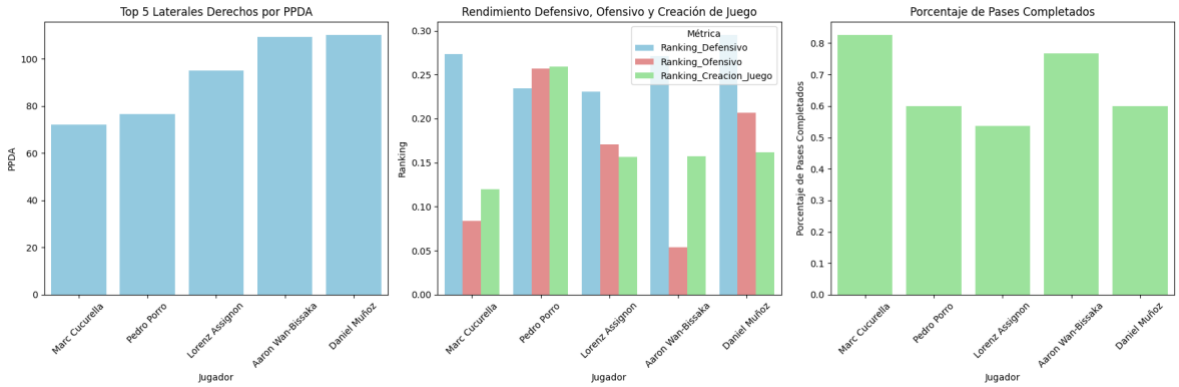
Lateral Izquierdo



Para el lateral izquierdo, se priorizó el **PPDA** junto con la capacidad de creación de juego. **Destiny Udogie** fue seleccionado por su buen **Ranking Defensivo** y notable contribución ofensiva.

- **Destiny Udogie**:
    - PPDA: 89.34
    - Ranking Defensivo: 0.204
    - Ranking Ofensivo: 0.137
    - Porcentaje de Pases Completados: 84.71%

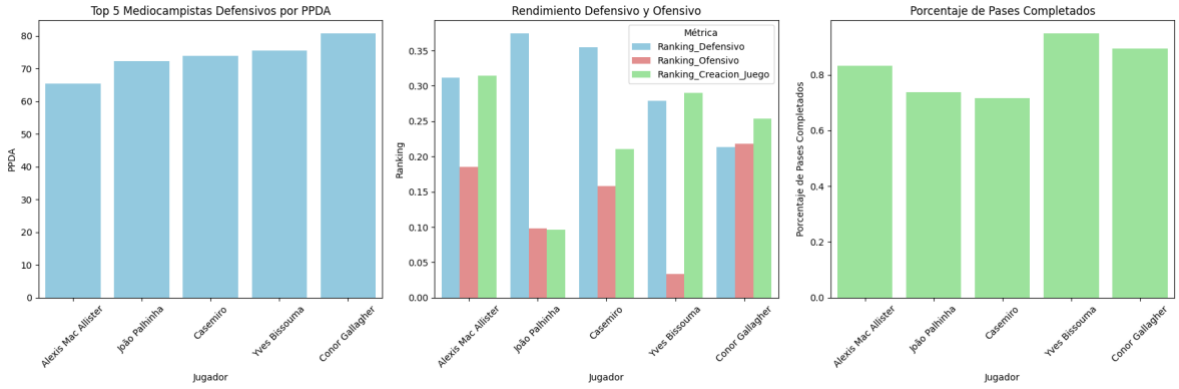
Lateral Derecho



En el lateral derecho, **Pedro Porro** fue seleccionado debido a su excelente balance entre el **Ranking Defensivo** y **Ofensivo**. Su PPDA es competitivo y tiene una gran capacidad de apoyar en ataque, con un **Ranking Ofensivo** alto.

- **Pedro Porro**:
    - PPDA: 76.74
    - Ranking Defensivo: 0.234
    - Ranking Ofensivo: 0.257
    - Porcentaje de Pases Completados: 76.11%

Mediocampista Defensivo

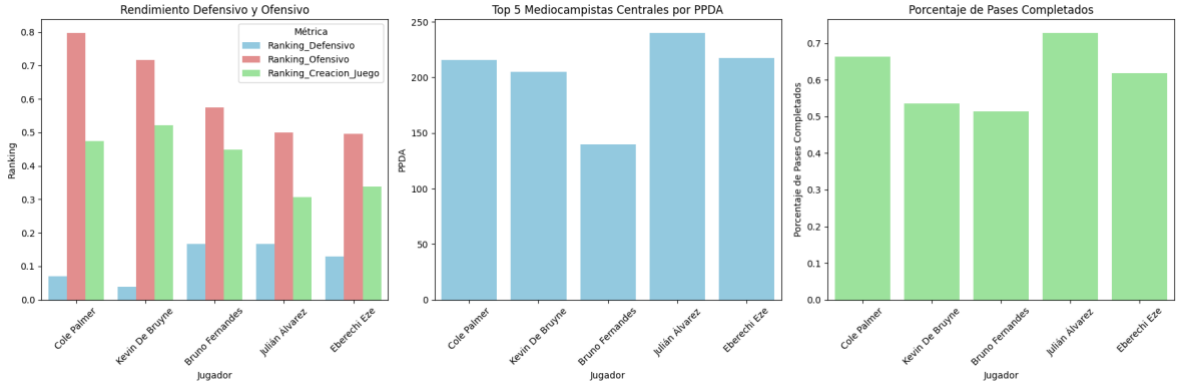


Para el mediocampista defensivo, la métrica **PPDA** fue clave para medir la capacidad de controlar el juego y bloquear al oponente. **Alexis Mac Allister** destacó por su equilibrio entre defensa y ataque, con el **Ranking Defensivo** más alto.

- **Alexis Mac Allister**:
    - PPDA: 65.44
    - Ranking Defensivo: 0.311

- Tacleadas por 90: 3.52
- Intercepciones por 90: 1.38
- Porcentaje de Pases Completados: 86.34%

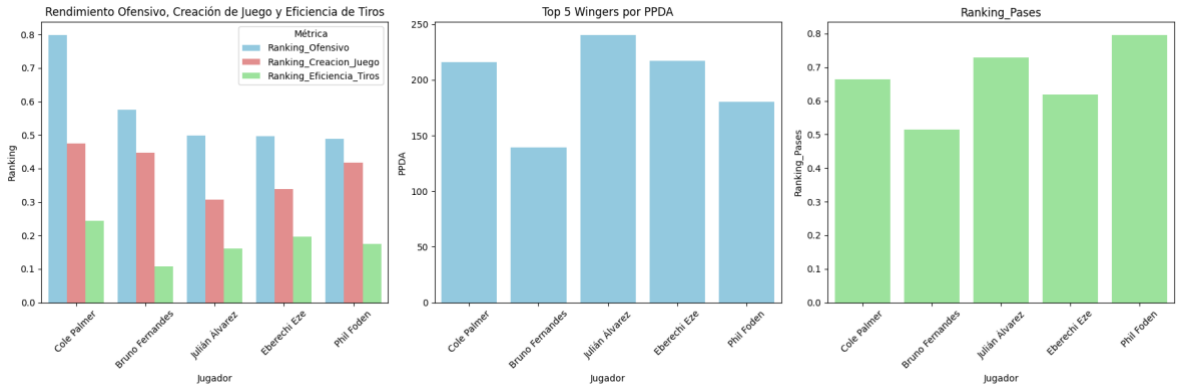
Mediocampistas Centrales



Aquí se priorizó la capacidad ofensiva con el **Ranking Ofensivo**. **Kevin De Bruyne** y **Bruno Fernández** fueron seleccionados por su habilidad en la creación de oportunidades, y ambos también tienen un buen porcentaje de pases completados.

- **Kevin De Bruyne**:
  - Ranking Ofensivo: 0.717
  - PPDA: 204.68
  - Porcentaje de Pases Completados: 73.29%
- **Bruno Fernández**:
  - Ranking Ofensivo: 0.574
  - PPDA: 139.53
  - Porcentaje de Pases Completados: 72.34%

Extremos

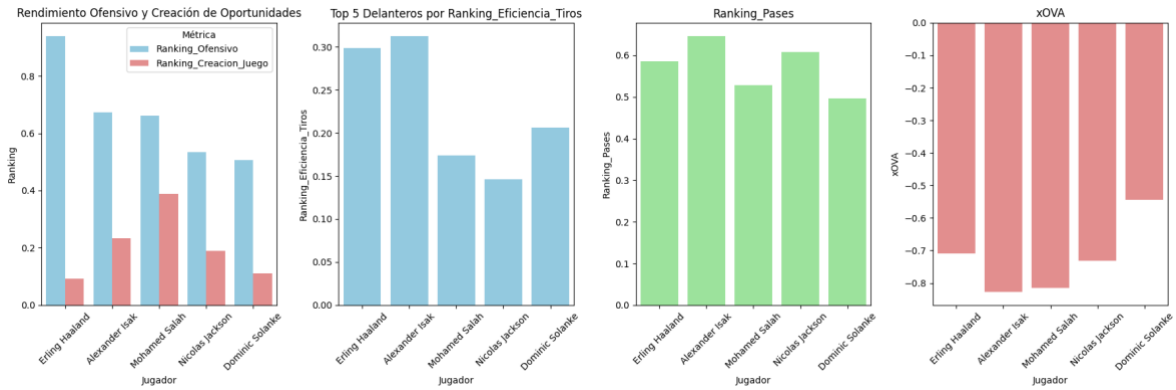


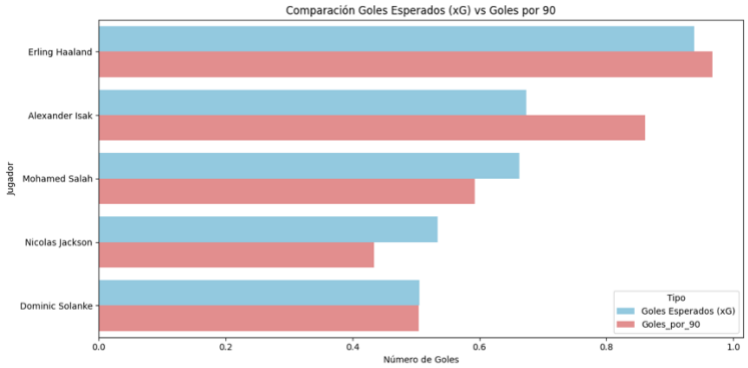
En los extremos, el **Ranking Ofensivo** fue clave, priorizando jugadores que contribuyen significativamente al ataque. **Cole Palmer** fue seleccionado por su gran capacidad ofensiva y **Phil Foden** por su equilibrio entre creación de juego y eficiencia en los tiros.

- **Cole Palmer**:
  - Ranking Ofensivo: 0.798

    - Ranking Creación de Juego: 0.474
    - Eficiencia de Tiros: 24.44%
    - Porcentaje de Pases Completados: 78.94%
- **Phil Foden**:
    - Ranking Ofensivo: 0.489
    - Ranking Creación de Juego: 0.417
    - Eficiencia de Tiros: 17.55%
    - Porcentaje de Pases Completados: 84.74%

Delantero

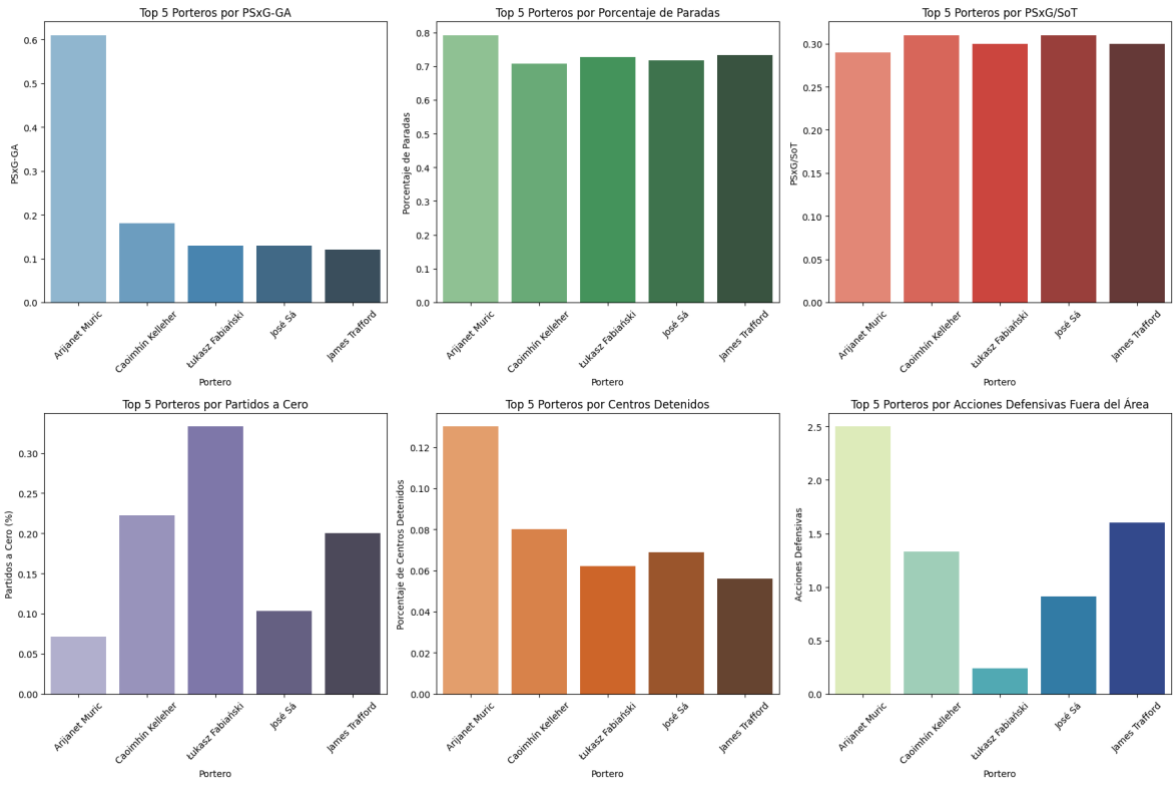




La prioridad aquí fue el **Ranking Ofensivo** y la capacidad de convertir goles. **Erling Haaland** fue seleccionado por su impresionante capacidad ofensiva, reflejada en su alto número de goles y su excelente eficiencia de tiros.

- **Erling Haaland**:
    - Ranking Ofensivo: 0.938
    - Eficiencia de Tiros: 29.85%
    - xOVA: -0.70
    - PPDA: 302.38
    - Porcentaje de Pases Completados: 75.53%

Portero



El portero fue seleccionado en base a la métrica **PSxG-GA**, que mide su capacidad para evitar goles en situaciones difíciles. **Arijanet Muric** fue el elegido por tener el valor más alto en esta métrica, además de un buen porcentaje de paradas y de centros detenidos.

- **Arijanet Muric**:
  - PSxG-GA: 0.61
  - Porcentaje de Paradas: 79.2%
  - PSxG/SoT: 0.29
  - Porcentaje de Partidos con Portería a Cero: 7.1%
  - Porcentaje de Centros Detenidos: 13.0%

5. Equipo Final

- **Portero**: Arijanet Muric (7,00 mill. €)
- **Defensa Central 1 (CD)**: Cristian Romero (65,00 mill. €)
- **Defensa Central 2 (CD)**: Levi Colwill (50,00 mill. €)
- **Lateral Izquierdo (LD)**: Destiny Udogie (45,00 mill. €)
- **Lateral Derecho (RD)**: Pedro Porro (45,00 mill. €)
- **Mediocampista Defensivo (CDM)**: Alexis Mac Allister (75,00 mill. €)
- **Mediocampista Central 1 (CM)**: Kevin De Bruyne (45,00 mill. €)
- **Mediocampista Central 2 (CM)**: Bruno Fernandes (65,00 mill €)
- **Extremo Derecho (RM)**: Cole Palmer (80,00 mill €)
- **Extremo Izquierdo (LM)**: Phil Foden (150,00 mill. €)
- **Delantero (FW)**: Erling Haaland (200,00 mill. €)

Total: 827,00 mill. €

6. Conclusión

A través de este análisis estadístico detallado, fue posible identificar a los jugadores con mejor rendimiento en cada posición, basándonos en métricas clave como PPDA, Goles Esperados (xG), Asistencias Esperadas (xAG), y otros indicadores ofensivos y defensivos. La combinación de estos análisis permitió seleccionar jugadores que no solo destacan en su capacidad individual, sino también en su contribución al desempeño del equipo, tanto en ataque como en defensa. Sin embargo, existen otras métricas y métodos de análisis que podrían proporcionar una visión aún más completa y matizada del rendimiento de los jugadores, ampliando las conclusiones obtenidas en este estudio.

7. Anexos

- **Notebooks en DataBricks:**
  - Notebook completo: **https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/1350097014768847/1712258292071631/6971521594910320/latest.html**
  - **Notebook con algunas consultas en sql: https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/1350097014768847/1712258292071580/6971521594910320/latest.html**
- **Notebook en Colab: https://colab.research.google.com/drive/1WUKSSBMgsVf2_VMgh6Vyd0gYWko8brmh?usp=sharing**