Universidad del Desarrollo
Facultad de Ingeniería

GENERACIÓN DE RANKING DE FUTBOLISTAS EN BASE A ESTADÍSTICAS HISTÓRICAS Y UN ENFOQUE DE CIENCIA DE DATOS

POR: MAURICIO ANDRÉS BERNIER MORA

Proyecto de grado presentado a la Facultad de Ingeniería de la Universidad del Desarrollo para optar al grado académico de Magíster en Data Science

PROFESORA GUÍA: LORETO BRAVO

Diciembre 2024

SANTIAGO

Contenido

INTRODUCCIÓN

En los últimos años, el análisis del rendimiento de jugadores de fútbol ha experimentado un cambio significativo gracias al auge de las técnicas de análisis de datos y la creciente disponibilidad de datos estadísticos detallados. El fútbol, un deporte tradicionalmente evaluado mediante observación subjetiva, ha comenzado a adoptar herramientas avanzadas de ciencia de datos para evaluar el desempeño de los jugadores y optimizar decisiones tanto dentro como fuera del campo. Esto ha permitido a clubes deportivos y analistas utilizar un enfoque más cuantitativo y objetivo para medir el rendimiento de los jugadores, facilitando la identificación de patrones y la predicción del desempeño futuro.

Actualmente, las estadísticas de los jugadores no sólo incluyen métricas clásicas como goles, asistencias o minutos jugados, sino también un abanico de datos más específicos como pases completados, duelos ganados, kilómetros recorridos y contribuciones defensivas. Estos datos pueden ser aprovechados para generar modelos predictivos que ayuden a estimar el rendimiento futuro de los jugadores o incluso a identificar el impacto potencial de ciertos atributos en el éxito colectivo del equipo. Además, la dimensionalidad de los datos y las relaciones complejas entre los atributos presentan un desafío que requiere la aplicación de técnicas avanzadas como el Análisis de Componentes Principales (PCA) y modelos basados en Machine Learning.

El análisis de ranking de jugadores en deportes ha demostrado ser una herramienta valiosa, tanto para la gestión de plantillas como para la planificación estratégica. Equipos de alto nivel alrededor del mundo han comenzado a aplicar estos modelos para predecir el desarrollo de jóvenes promesas, optimizar fichajes y ajustar tácticas de juego en función de las habilidades de sus jugadores. A medida que la ciencia de datos sigue transformando la forma en que se aborda el fútbol, la creación de modelos que logren plasmar en un ranking el rendimiento de los jugadores con mayor precisión y fiabilidad se vuelve cada vez más relevante, sobre todo para efectos de scouting y así apoyar las decisiones de contratación en los clubes de fútbol y del deporte en general.

Por ejemplo, el Brentford Football Club, bajo la dirección de su propietario Matthew Benham, ha revolucionado la gestión deportiva mediante el uso intensivo de análisis de datos y Big Data. Benham, un físico graduado de la Universidad de Oxford aplicó su experiencia en modelos estadísticos y análisis predictivo, adquirida en su empresa de apuestas deportivas Smartodds, para transformar al Brentford en un club competitivo y financieramente sostenible.

La estrategia del Brentford se centra en identificar y fichar a jugadores subvalorados o con potencial no desarrollado, utilizando algoritmos y estudios de mercado para evaluar su rendimiento y adecuación al equipo. Este enfoque ha permitido al club adquirir talento a bajo costo y venderlo posteriormente con ganancias significativas. Por ejemplo, Neal Maupay fue fichado por 1,6 millones de libras y vendido al Brighton por 19,8 millones, mientras que Ollie Watkins llegó por 150.000 euros y fue transferido al Aston Villa por 34 millones.

Además, el Brentford ha implementado un equipo B en lugar de una academia juvenil tradicional, enfocándose en jugadores de 17 a 20 años rechazados por otros clubes o provenientes de mercados menos explorados. Esta estructura permite al club desarrollar

talento de manera eficiente y adaptada a su modelo de juego, reduciendo costos y aumentando la eficacia en la formación de jugadores.

La aplicación de Big Data también se extiende al análisis de partidos y al desarrollo de estrategias de juego, optimizando el rendimiento del equipo en el campo. Este enfoque integral ha llevado al Brentford a ascender a la Premier League en 2021, después de 74 años de ausencia, demostrando el éxito de una gestión basada en datos y análisis estadístico. De hecho, en los últimos 3 años, el Brentford se ubica en lugar 18° en cuanto a magnitud de gastos en fichajes, pero deportivamente, está en el 9° lugar de la tabla de posiciones de la Premier League (al 13/12/2024).

Por otro lado, está la historia de un club de béisbol, la cual se muestra en la película Moneyball, que cuenta la historia real de Billy Beane, el Gerente General de los Oakland Athletics, un equipo de las Grandes Ligas que enfrenta serias limitaciones presupuestarias. La trama se desarrolla durante la temporada de 2002, cuando Beane debe reconstruir su equipo tras perder a varias de sus estrellas, que son fichadas por equipos con mayores recursos financieros. Desesperado por competir en un sistema donde el dinero parece determinar el éxito, Beane encuentra una solución poco convencional al asociarse con Peter Brand, un joven economista que propone una nueva forma de evaluar jugadores basada en estadísticas avanzadas.

Brand introduce a Beane en un enfoque analítico que ignora las métricas tradicionales del béisbol y se centra en datos menos valorados, pero altamente efectivos para predecir el rendimiento de los jugadores. Juntos, desafían las prácticas tradicionales de los scouts y fichan a jugadores subestimados por el resto de la liga, pero que poseen habilidades clave para contribuir al éxito del equipo.

A lo largo de la película, Beane enfrenta la resistencia de sus colegas, los medios y los fanáticos, que consideran su método como una amenaza al espíritu del deporte. Sin embargo, su enfoque comienza a dar frutos, culminando en una histórica racha de 20 victorias consecutivas, un logro sin precedentes que demuestra el potencial del análisis de datos en el béisbol, y que puede ser aplicado al deporte en general.

El presente informe se estructura de la siguiente manera:

- **Objetivos generales y específicos**, donde se detallarán las metas del proyecto.
- **Carta Gantt**, que ilustrará la planificación temporal del proyecto.
- **Metodología**, que describirá los pasos a seguir para alcanzar los objetivos propuestos, incluyendo las técnicas y modelos a emplear.
- **Revisión bibliográfica**, en la que se expondrán estudios previos y trabajos relevantes sobre el análisis y predicción del rendimiento de jugadores en el fútbol.

OBJETIVO GENERAL

Desarrollar un modelo para calcular y analizar un ranking de jugadores de fútbol, basado en métricas de rendimiento derivadas de estadísticas de partidos y su evolución a lo largo del tiempo.

OBJETIVOS ESPECÍFICOS

- Realizar un preprocesamiento y análisis exploratorio de las estadísticas históricas de jugadores y clubes.
- Identificar las variables clave que puedan explicar el resultado de un partido (victoria o derrota) a nivel de equipo.
- Obtener ponderadores de importancia para cada variable en relación con su contribución al éxito de los equipos (victorias).
- Proponer y calcular una métrica de rendimiento de jugadores, a través de una función o fórmula basada en las variables clave y los ponderadores obtenidos.
- Generar un ranking de jugadores y evaluar las diferencias entre ellos con base en su rendimiento y evolución.
- Visualizar y comunicar los resultados del ranking mediante gráficos y análisis interpretativos.

REVISIÓN BIBLIOGRÁFICA

El análisis del rendimiento de jugadores en el fútbol ha demostrado ser una de las áreas donde la ciencia de datos tiene un impacto significativo. Con el acceso a grandes volúmenes de datos estadísticos y técnicas avanzadas de análisis, se ha logrado evaluar el desempeño individual y colectivo de manera más precisa. Esto ha permitido optimizar decisiones relacionadas con la gestión de jugadores y equipos, así como explorar nuevas metodologías para comparar a los jugadores y crear rankings que reflejen su contribución al juego.

Un ejemplo destacado y ampliamente referenciado en la evaluación del rendimiento de jugadores es el trabajo de Luca Pappalardo et al. (2019), quienes desarrollaron PlayeRank, un marco metodológico diseñado para analizar el desempeño individual en el fútbol. Este sistema, basado en datos de eventos capturados durante los partidos, se distingue por su enfoque multidimensional que incluye métricas ofensivas, defensivas y de contribución a la posesión del balón. A diferencia de metodologías tradicionales que se centran en una o dos dimensiones del rendimiento, PlayeRank utiliza un enfoque integral que permite evaluar a los jugadores desde diferentes perspectivas.

Un aspecto innovador de PlayeRank es su capacidad para asignar ponderadores específicos a cada dimensión del juego, dependiendo de su importancia relativa en el contexto del partido o

del rol del jugador. Esto asegura que los rankings generados sean justos y comparables entre jugadores de roles similares, como delanteros, mediocampistas o defensores. Además, el sistema no solo mide el rendimiento en un partido aislado, sino que integra la evolución de las métricas a lo largo del tiempo, permitiendo identificar tendencias en el desempeño de los jugadores. Esta integración temporal es especialmente relevante para este proyecto, ya que se busca construir rankings que reflejen tanto el rendimiento actual como el desarrollo a lo largo de múltiples temporadas.

En términos metodológicos, PlayeRank combina estadísticas individuales con datos colectivos del equipo, lo que proporciona un análisis más robusto del impacto de cada jugador en los resultados del partido. Por ejemplo, un jugador podría destacar en métricas individuales como intercepciones o asistencias, pero su valor relativo puede aumentar o disminuir dependiendo del contexto del equipo y del partido. Esta capacidad de contextualizar las métricas individuales en un marco colectivo es un componente clave que este proyecto podría adaptar, especialmente para ponderar las variables que influyen en los rankings.

La validación de PlayeRank se llevó a cabo utilizando datos de competiciones de alto nivel, lo que demuestra la robustez y versatilidad del modelo en diferentes contextos. Los resultados mostraron que el sistema podía distinguir eficazmente a jugadores destacados dentro de roles específicos, como mediocampistas ofensivos, laterales defensivos o delanteros, y producir rankings confiables y consistentes. Esta capacidad para segmentar jugadores por posición es fundamental para este proyecto, ya que permitirá generar rankings más precisos y significativos.

Además, PlayeRank introduce la posibilidad de realizar análisis de "qué pasaría si", simulando el impacto de cambios en las métricas ponderadas o en el rendimiento de los jugadores. Esto abre la puerta a exploraciones más detalladas sobre cómo diferentes variables influyen en los rankings y cómo podrían ajustarse para reflejar mejor la evolución del rendimiento.

El uso de datos estadísticos para evaluar a los jugadores no se limita a PlayeRank. Hansoo Lee et al. (2020) emplearon el algoritmo LightGBM para modelar el valor de mercado de los jugadores de fútbol. Aunque el estudio se centra en predicciones financieras, su metodología de identificar variables clave a partir de datos de rendimiento proporciona un marco sólido para calcular métricas que alimenten rankings objetivos. De manera similar, Abdessatar Ati et al. (2022) combinan algoritmos de Machine Learning con enfoques multicriterio para evaluar múltiples dimensiones del rendimiento, lo cual inspira este proyecto en su objetivo de integrar aspectos individuales y colectivos en la generación de rankings.

Otro enfoque metodológico relevante es el análisis de redes aplicado al fútbol. Los trabajos de Korte et al. (2021) y Buldú et al. (2018) destacan la importancia de analizar las relaciones entre los jugadores para comprender su impacto en el rendimiento del equipo. Estos estudios identifican jugadores clave dentro de las dinámicas colectivas, lo que complementa las métricas tradicionales y puede enriquecer los rankings propuestos al incluir dimensiones como la conectividad y el liderazgo dentro del campo.

La evolución del rendimiento también ha sido explorada desde diferentes perspectivas. Seife Dendir (2016) analizó las edades óptimas de rendimiento de los jugadores, concluyendo que

estos alcanzan su máximo entre los 25 y 27 años, dependiendo de la posición. Esta información podría ser valiosa para contextualizar los rankings según la etapa de carrera de cada jugador. Por otro lado, Merzah et al. (2024) desarrollaron un modelo para clasificar el rendimiento en categorías como "activo", "normal" o "débil", utilizando múltiples algoritmos de Machine Learning. Este enfoque destaca la importancia de la clasificación y agrupación de jugadores, lo cual puede ser incorporado en los rankings mediante la segmentación por niveles de rendimiento.

En términos de innovación metodológica, Stephan Wolf et al. (2020) adaptaron el algoritmo Elo para evaluar el rendimiento de jugadores individuales en fútbol. Este sistema, ya utilizado en otros deportes, ofrece la ventaja de facilitar comparaciones entre jugadores de diferentes ligas y temporadas. Su aplicabilidad a este proyecto radica en su capacidad para refinar las ponderaciones asignadas a los atributos de los jugadores, asegurando que los rankings reflejen fielmente su contribución al juego.

Por último, el enfoque presentado por Kati et al. (2022) sobre la evolución del rendimiento de los jugadores y su valor de mercado es especialmente relevante para este proyecto. Utilizando modelos de aprendizaje automático como Multi-Layer Perceptron, el autor destaca cómo las estadísticas históricas pueden traducirse en métricas consistentes que reflejen tanto el rendimiento actual como su proyección futura. Esta idea puede ser adaptada para calcular y ajustar los scores que determinarán los rankings finales.

La combinación de enfoques en estos estudios proporciona una base sólida para este proyecto, que busca construir un ranking de jugadores de fútbol chileno basado en métricas objetivas de rendimiento y su evolución temporal. Desde la identificación de variables clave hasta la integración de técnicas como ponderaciones dinámicas, esta metodología permitirá ofrecer una herramienta robusta y justa para evaluar y comparar a los jugadores en diferentes contextos.

METODOLOGÍA

El presente proyecto se desarrollará en varias fases interrelacionadas que permitirán abordar de manera sistemática el análisis y la predicción del rendimiento de jugadores de fútbol. A continuación, se describe cada fase, y en la figura 2 se muestra un diagrama resumido.

Análisis Exploratorio de Datos

Esta fase inicial tiene como objetivo entender la estructura de los datos disponibles y preparar el conjunto de datos para su posterior análisis. Se utilizarán estadísticas históricas de jugadores y clubes recopiladas durante un período de cinco años, que incluyen variables como goles, asistencias, duelos ganados, pases completados, entre otras.

- Preprocesamiento de datos: Se revisará existencia de duplicados y se manejarán valores faltantes (si hubiera) a través de métodos de imputación o eliminación de registros incompletos.

- Exploración descriptiva: Se realizarán análisis estadísticos descriptivos para identificar distribuciones y posibles outliers en las variables seleccionadas.
- Se identificará cuáles variables o estadísticas están presentes tanto en el dataset de equipos como en el de jugadores, para así poder calcular una métrica de rendimiento de jugadores.

Reducción de Dimensionalidad mediante Análisis de Componentes Principales (PCA)

Una vez que los datos han sido limpiados y normalizados (de ser necesario), se realizará un Análisis de Componentes Principales (PCA) con el objetivo de analizar la factibilidad de reducir la dimensionalidad del conjunto de datos. El PCA permite identificar las combinaciones lineales de los atributos originales que explican la mayor parte de la variabilidad.

- Aplicación del PCA: Se calcularán los componentes principales utilizando el paquete scikit-learn de Python. El número de componentes retenidos se determinará de acuerdo con el criterio de captura de al menos el 80-90% de la varianza total.
- Interpretación de componentes: Los componentes principales serán analizados para interpretar qué atributos originales tienen mayor peso en cada componente. Además, en esta etapa se debe concluir si estas componentes pueden ser interpretadas y si serán útiles para los análisis posteriores.

Obtención de Ponderadores y Cálculo del Score de Rendimiento

Se obtendrán coeficientes que reflejen la importancia relativa de cada variable o estadística en los resultados de los partidos (victorias, empates o derrotas). Los coeficientes se pueden calcular mediante modelos de Random Forest o Regresión Logística, por ejemplo. Los coeficientes obtenidos de estos modelos servirán como ponderadores, reflejando la importancia relativa de cada variable en el éxito del equipo.

Usando los ponderadores obtenidos, se calculará un score o métrica de rendimiento para cada jugador y partido, aplicando los pesos a los atributos individuales de cada jugador. Esta métrica será una función o fórmula (por ejemplo, combinación lineal) que estará basada en los ponderadores o coeficientes resultantes y las variables o estadísticas clave. Además, la idea es que esta métrica se calcule a través del tiempo (temporada, partido) para los jugadores, para analizar la evolución y se pueda construir un ranking a partir de esto.

Generación de Ranking y Visualización de Resultados

Para la generación del ranking, se definirán criterios que permitan evaluar el rendimiento de los jugadores de manera integral. En primer lugar, se calculará el score promedio de cada jugador en base a los datos obtenidos en los partidos analizados. Este promedio proporcionará una medida general del rendimiento del jugador durante el período de estudio. Además, se

considerará la consistencia del rendimiento, evaluando la variabilidad o volatilidad de los scores mediante métricas como la desviación estándar o el coeficiente de variación. Los jugadores más consistentes tendrán menor variabilidad en sus scores, lo cual puede ser una característica deseable dependiendo del análisis. También se incorporará la evolución del rendimiento a lo largo del tiempo, calculando la pendiente de una regresión lineal ajustada a los scores de cada jugador. Esto permitirá identificar tendencias positivas o negativas en su desempeño.

El ranking final se generará combinando estas métricas de manera ponderada, considerando el score promedio, la consistencia y, finalmente, la evolución. Los pesos asignados a cada criterio dependerán del objetivo del análisis, asegurando un balance entre el rendimiento histórico y el potencial de mejora. Asimismo, se generarán rankings separados por posición para evitar comparaciones directas entre jugadores con roles significativamente distintos. Adicionalmente, se evaluará la factibilidad de construir rankings globales, utilizando scores normalizados que equilibren las diferencias entre posiciones.

La visualización de los resultados será un componente crucial para interpretar y comunicar los hallazgos. Se presentarán gráficos para mostrar los jugadores mejor rankeados por posición, desglosando las contribuciones de cada criterio al score total. Se utilizarán gráficos de líneas para ilustrar la evolución de los scores a lo largo del tiempo, destacando las tendencias de los jugadores más relevantes. Además, gráficos radiales pueden ser útiles para comparar múltiples métricas de rendimiento de un jugador en particular, proporcionando una vista integral de sus fortalezas y áreas de mejora. También se evaluará un análisis de clústeres para agrupar jugadores según sus características de rendimiento, categorizándolos en niveles como élite, promedio y en desarrollo. Los jugadores con scores significativamente altos o bajos respecto a su grupo serán identificados como valores atípicos, y su rendimiento será analizado con mayor detalle para comprender las posibles causas.

El producto final incluirá una tabla resumida que presente los rankings de los jugadores con información relevante como el score final, la posición, el promedio de scores, la consistencia y la evolución. Esta tabla permitirá una interpretación clara y rápida de los resultados, facilitando la identificación de los jugadores con mejor rendimiento y potencial.

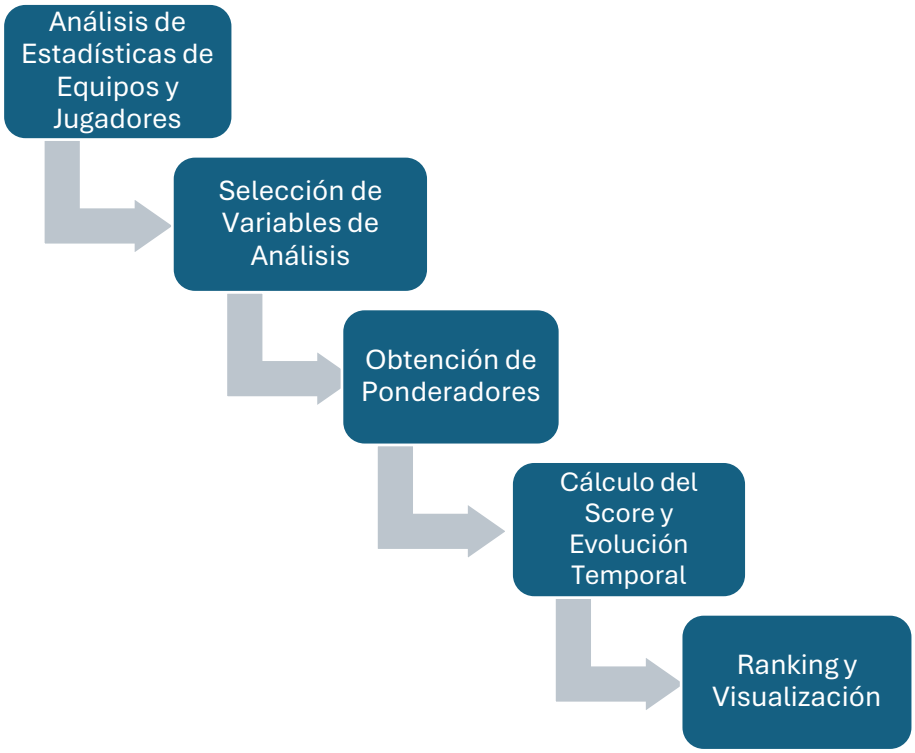


Figura 1. Diagrama Resumen de Metodología

DESARROLLO

Preprocesamiento y Análisis Exploratorio

Los datos disponibles contienen estadísticas del fútbol chileno de 5 años hasta el año 2023, que están compuestas por estadísticas a nivel de equipos, y estadísticas a nivel de jugadores, como se muestra en la figura 3. Estas bases contienen varias estadísticas asociadas a partidos, y no todas coinciden necesariamente en ambas.

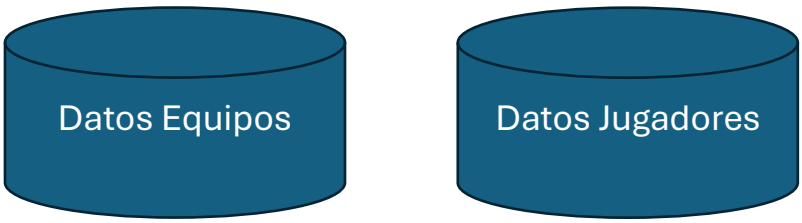


Figura 2. Datos Disponibles

Las estadísticas de equipos vienen en archivos Excel por separado, un archivo por equipo y por año. Por otro lado, las estadísticas de jugadores vienen también en archivos Excel por separado, uno por cada jugador. Entonces, el primer preproceso consiste en construir un dataset consolidado para equipos, y otro para jugadores.

Luego de consolidar, el dataset de estadísticas de equipos posee las siguientes características principales:

Número de Columnas con Datos Numéricos	104
Número de Registros	2.854
Cantidad de Registros sin algún Dato	322
Variables Explicativas	101
Variables Dependientes	1

Tabla 1. Características Originales de Datos de Equipos

Hay que considerar que las primeras 3 columnas numéricas de los datos de equipos contienen información más bien relacionada con el resultado, o que no representan variables explicativas, tales como: Duración, Goles. En el caso de la variable Goles, en realidad está muy relacionada con el resultado de los partidos, y es más bien una consecuencia de las estadísticas del juego. Por lo tanto, estas primeras 2 columnas no serán tomadas en cuenta como atributos o variables que puedan explicar el resultado final, así basándonos más en la construcción del juego. Además, se incorpora una columna adicional a la base, que tendrá como valor 1 si es que el resultado es una victoria, y 0 si no es así. Esta representaría la variable a explicar o dependiente, y se denominará variable "Resultado", y que se construye a partir de una columna que menciona el resultado en formato texto. Por otro lado, las 322 filas sin datos, las cuales contienen alguna celda vacía, no se considerará en el análisis, por lo que se eliminan.

Además, las estadísticas de equipos consideran algunos registros que no corresponden al torneo de Primera A del fútbol chileno, por ejemplo: Primera B, Copa Chile, o Copas Internacionales. Dado que el foco del análisis será en la liga local del fútbol chileno, se considerarán sólo aquellos registros. Es relevante notar que la importancia de las variables en la victoria de un equipo puede ser distinta según el nivel de la liga a considerar, por lo que se acotará el análisis sólo a la liga chilena, y dado que ahí está concentrada la mayor cantidad de registros disponibles.

Con esto, las características de estadísticas de equipo quedan de la siguiente forma:

Número de Columnas con Datos Numéricos	105
Número de Registros	2.014
Cantidad de Registros sin algún Dato	0
Variables Explicativas	101
Variables Dependientes	1

Tabla 2. Características Finales de Datos de Equipos

Por otro lado, el dataset consolidado de estadísticas por jugador tienen las siguientes características generales:

Número de Columnas con Datos Numéricos	68
Número de Registros	10.143
Cantidad de Registros sin algún Dato	0
Variables Explicativas	67

Tabla 3. Características de Datos de Jugadores

Las variables explicativas son 67, dado que no se considerará como tal la variable numérica de Minutos Jugados. Además, estos datos consideran registros asociados a 60 jugadores.

De todas formas, y para ser consecuente con el hecho de enfocarse sólo en las estadísticas de la liga chilena, se van a considerar los registros de jugadores sólo en partidos del torneo nacional, dejando de lado por ejemplo los correspondientes a Copas Internacionales, Ligas Extranjeras, o Partidos por Selección Nacional, dado que la exigencia y nivel son distintos. Además, se eliminan algunos registros repetidos que se pudieron detectar.

Dado esto, la base de estadísticas de jugadores queda de la siguiente forma:

Número de Columnas con Datos Numéricos	68
Número de Registros	4.624
Cantidad de Registros sin algún Dato	0
Variables Explicativas	67

Tabla 4. Características Finales de Datos de Jugadores

Ahora bien, para poder continuar con el análisis, es necesario saber qué variables o atributos son comunes en los datos a nivel de equipos y a nivel de jugadores. Eso se debe a que se determinarán coeficientes para medir la importancia de cada variable en el resultado de los equipos, para después aplicarlo a las estadísticas de los jugadores, de tal manera de calcular un score o métrica de rendimiento, que a la vez sea consistente con el aporte que se realiza al resultado. Al realizar el cruce, se encuentra que existen 31 variables que son comunes, las cuales se analizan en histogramas que se muestran en la figura 4. En general se puede apreciar que las variables no están dispersas, y que se pueden percibir algunos leves sesgos, más que todo positivos.

Por otro lado, se construye una matriz de correlación de las variables, la cual se muestra en la figura 5. Se pueden ver algunas relaciones positivas fuertes (rojo), las cuales están principalmente concentradas entre las variables asociadas a Pases, y en menor medida aquellas asociadas a Duelos. En cuanto a las relaciones negativas (azul), se pueden destacar las existentes entre Despejes y Pases, o Saques de meta y Pases.

Además, en la figura 6, se muestran los diagramas de cajas (boxplots) de las variables comunes que se van a considerar. En general se pueden apreciar algunos valores atípicos, lo cual se puede deber a que:

- El fútbol es un deporte con eventos poco frecuentes pero significativos (centros, tiros, tarjetas, etc.).
- Muchas estadísticas del fútbol tienen una distribución sesgada. Ejemplo: La mayoría de los jugadores realiza un número moderado de pases, pero unos pocos (como los mediocampistas centrales) pueden registrar números excepcionalmente altos en ciertos partidos.
- Las posiciones tienen responsabilidades muy distintas, lo que genera grandes variaciones en ciertas métricas.
- Dentro de una posición, también puede haber valores extremos debido a jugadores que destacan en roles específicos.

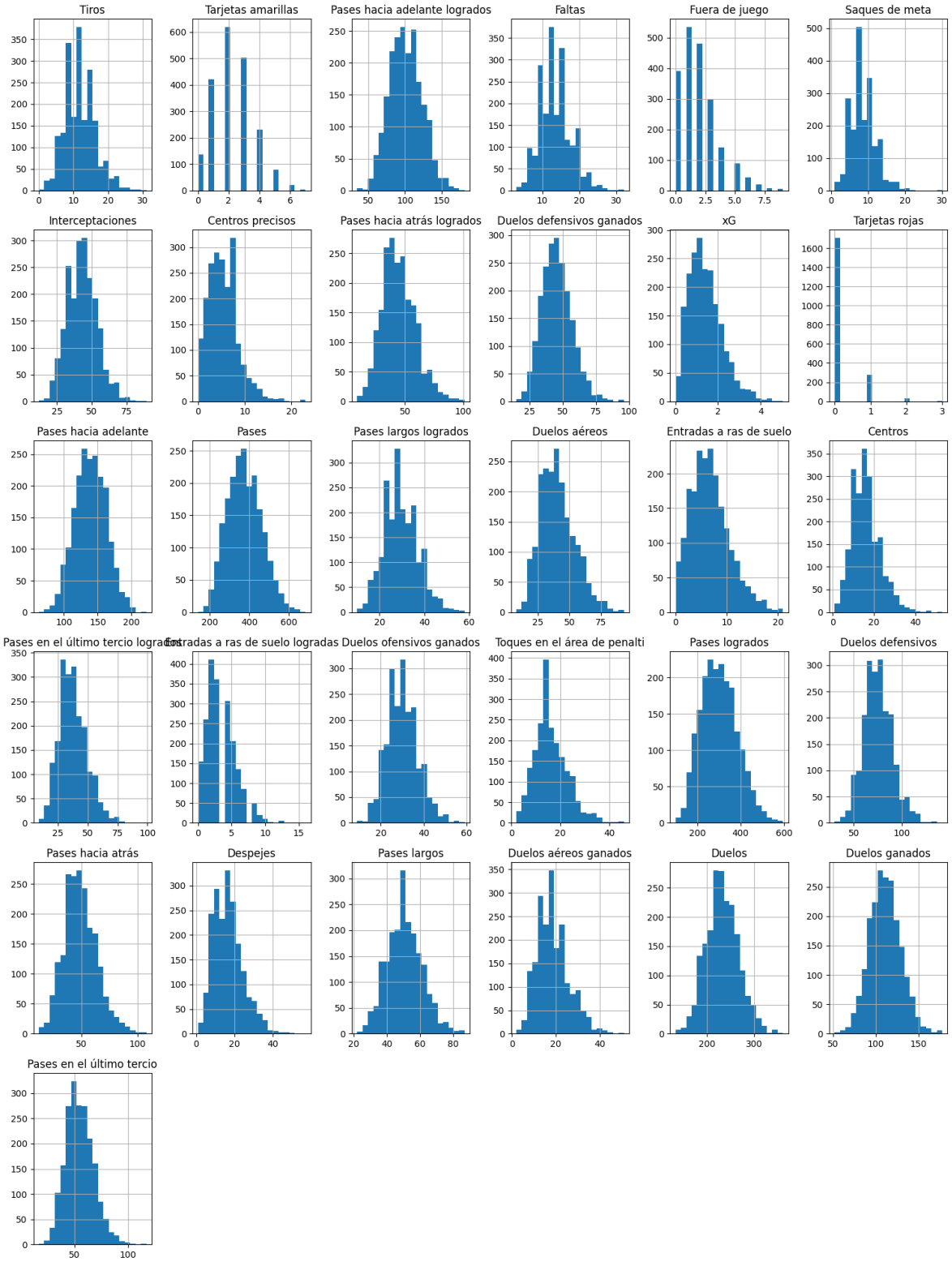


Figura 3. Histogramas de Variables Comunes

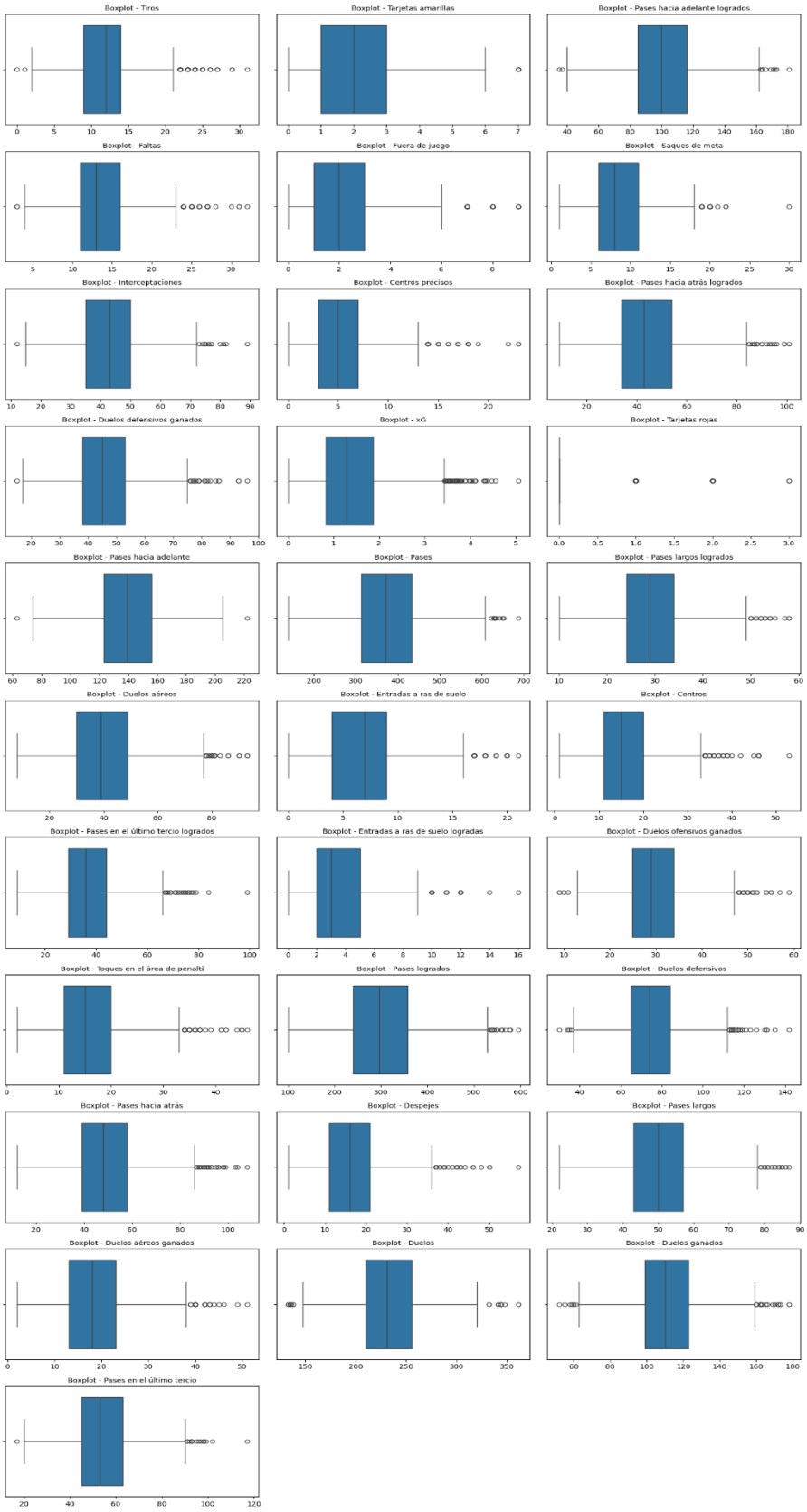


Figura 4. Boxplots de Variables Comunes

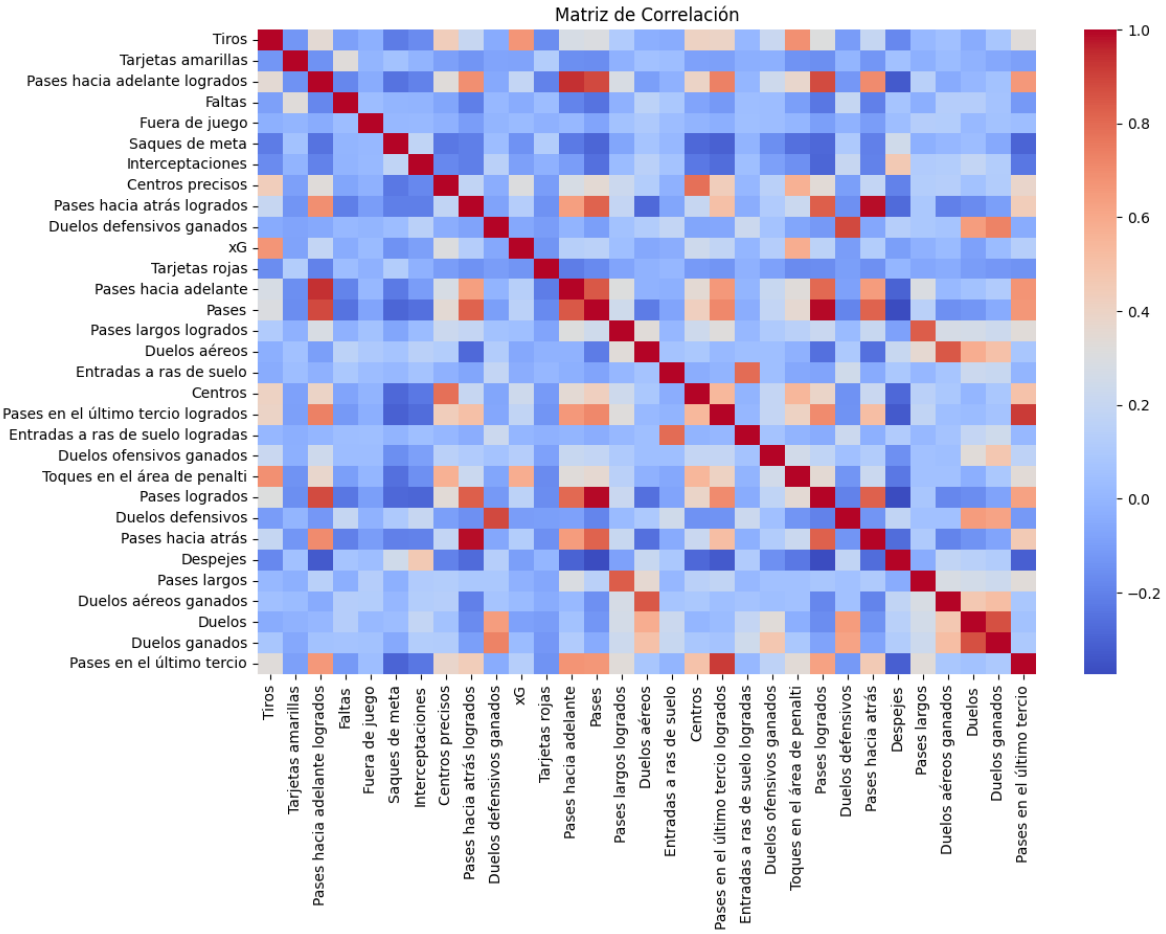


Figura 5. Matriz de Correlación Variables Comunes

Análisis de Componentes Principales (PCA)

A continuación, se realiza un análisis PCA, para ver si existen oportunidades de reducir la dimensionalidad. En primer lugar, se analiza la cantidad de componentes que explican la varianza, lo cual se presenta en la figura 6. Se puede apreciar que el 80% de la varianza acumulada se puede explicar con las primeras 11 componentes principales.

El paso siguiente es ver si las componentes principales son interpretables, y así asociarlas a conceptos asociados a las variables que más influyen en dichas componentes. Para eso, se realiza un análisis para ver qué variables tienen mayor "carga" en cada componente. Las cargas son los coeficientes que relacionan las variables originales (por ejemplo, "pases", "tiros", etc.) con cada componente principal, e indican el peso o la influencia de cada variable en la construcción del componente principal. Una carga grande (en valor absoluto) indica que la variable tiene una fuerte influencia en el componente principal, y una carga pequeña (cercana a 0) indica que la variable tiene poca influencia en ese componente. Una carga positiva significa que la variable y el componente principal están directamente relacionados: a medida que aumenta la variable, aumenta el valor del componente. Una carga negativa significa que

están inversamente relacionados: a medida que aumenta la variable, el valor del componente disminuye. En la figura 7 se presentan estas cargas para las primeras 11 componentes principales, en que las barras azules indican carga positiva, y las rojas representan cargas negativas.

Se puede ver que no es posible una interpretabilidad clara, dado que no hay una claridad de variables dominantes en los componentes, y además, existen variables que se repiten en los componentes principales, como los pases. En este sentido, parece no ser conveniente usar estas componentes para continuar con los análisis, pero de todas formas, da un indicio de qué variables son las más relevantes y ayudará a validar los análisis posteriores. Este análisis PCA nos muestra que las variables asociadas a Pases y Duelos son (en principio) muy importantes para explicar los resultados y rendimientos. Por esta razón, se continuará con un modelo de Random Forest que permita identificar variables importantes para estudio de rendimientos.

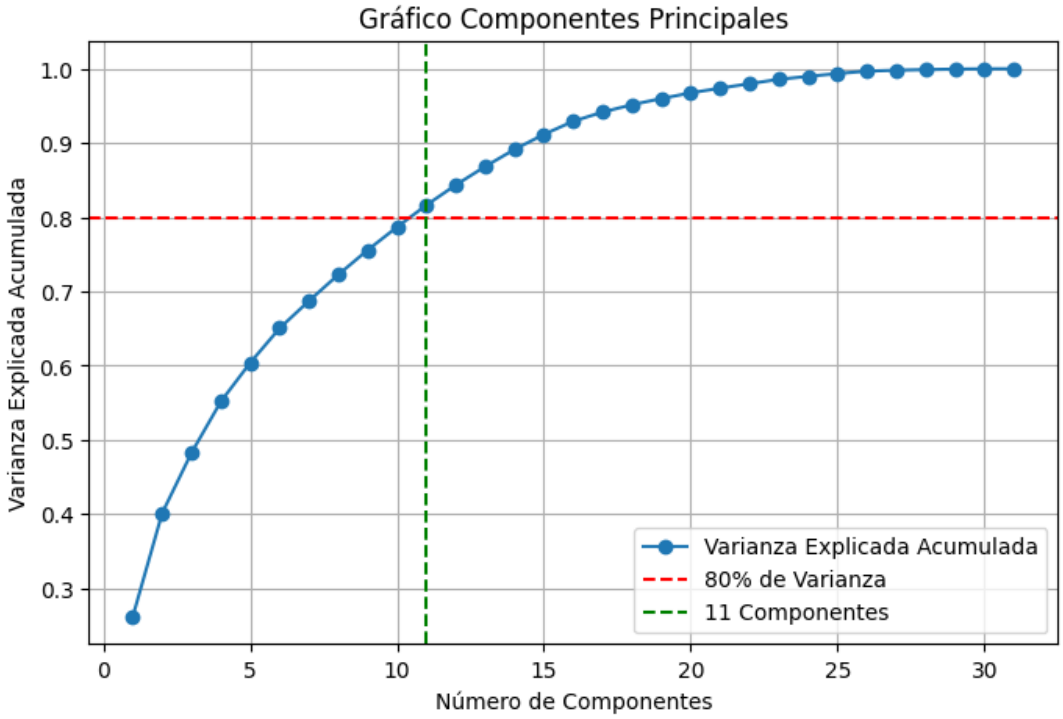


Figura 6. Gráfico Componentes Principales

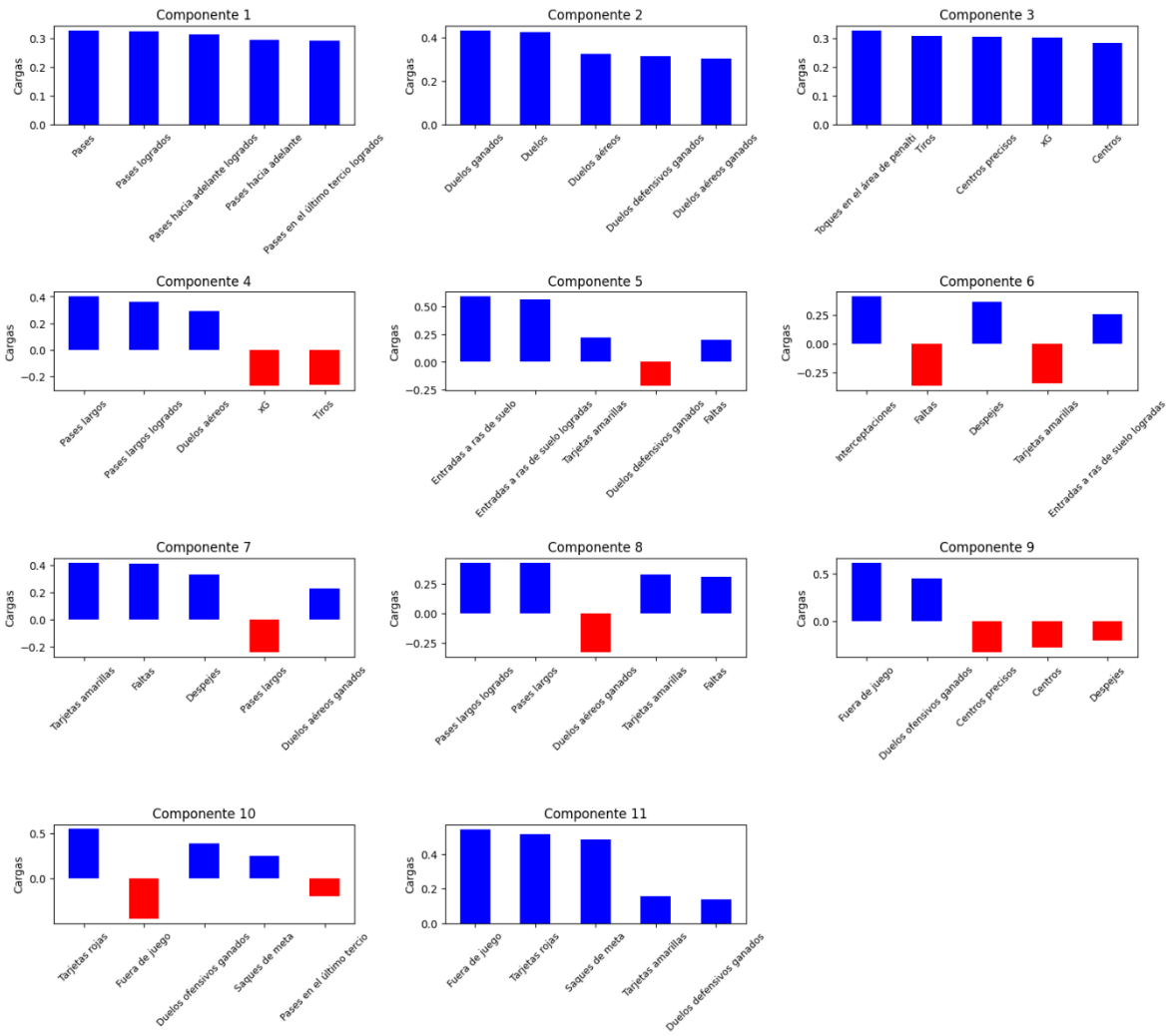


Figura 7. Variables más Influyentes por Componente

Obtención de Ponderadores

Para la estimación de coeficientes (ponderadores) se recurrirá a Random Forest. Este modelo es capaz de modelar relaciones complejas y no lineales entre las variables predictoras y la variable objetivo. Esto es especialmente útil si se cree que el impacto de ciertas estadísticas en el resultado del partido no es lineal. Por ejemplo, en fútbol, la relación entre "tiros" y "probabilidad de ganar" podría no ser lineal: después de cierto número de tiros, el beneficio adicional podría ser menor. Random Forest puede capturar este tipo de relaciones.

Además, Random Forest puede capturar interacciones entre variables sin necesidad de especificarlas de antemano. Si, por ejemplo, la "posesión" combinada con la "precisión de pases" tiene un efecto distinto en la probabilidad de ganar que cada variable por separado, Random Forest puede identificar esta interacción de forma automática. Esto es ventajoso en datasets de estadísticas de equipos donde pueden existir múltiples combinaciones de variables que afectan el resultado de un partido.

Por otro lado, los árboles de decisión en Random Forest dividen los datos basándose en umbrales, por lo que son menos sensibles a valores atípicos (outliers) y pueden manejar datos desbalanceados mejor que otros modelos lineales. Esto puede ser beneficioso en deportes, donde ciertos valores extremos en estadísticas (como un número inusualmente alto de tiros o pases en un solo partido) no afectan significativamente la importancia general de las variables.

En el trabajo de Luca Pappalardo et al. (2019) se utiliza Linear Support Vector Classifier, pero de todas formas se prefiere Random Forest por sus ventajas ya mencionadas, sobre todo en términos de considerar las relaciones no lineales, correlaciones altas, requiere menos ajuste de hiperparámetros, y funciona bien incluso si hay ruido o datos faltantes, gracias al ensamblado de múltiples árboles.

Aunque Random Forest no proporciona coeficientes con signo, ofrece una medida de importancia de variables basada en la reducción de impureza (como la reducción de entropía o del índice Gini) para cada variable. Esto permite identificar las variables más relevantes para predecir la victoria. La importancia de las variables en Random Forest es útil para obtener una visión general de qué factores son los más significativos, sin preocuparse de la dirección (positiva o negativa) de la relación.

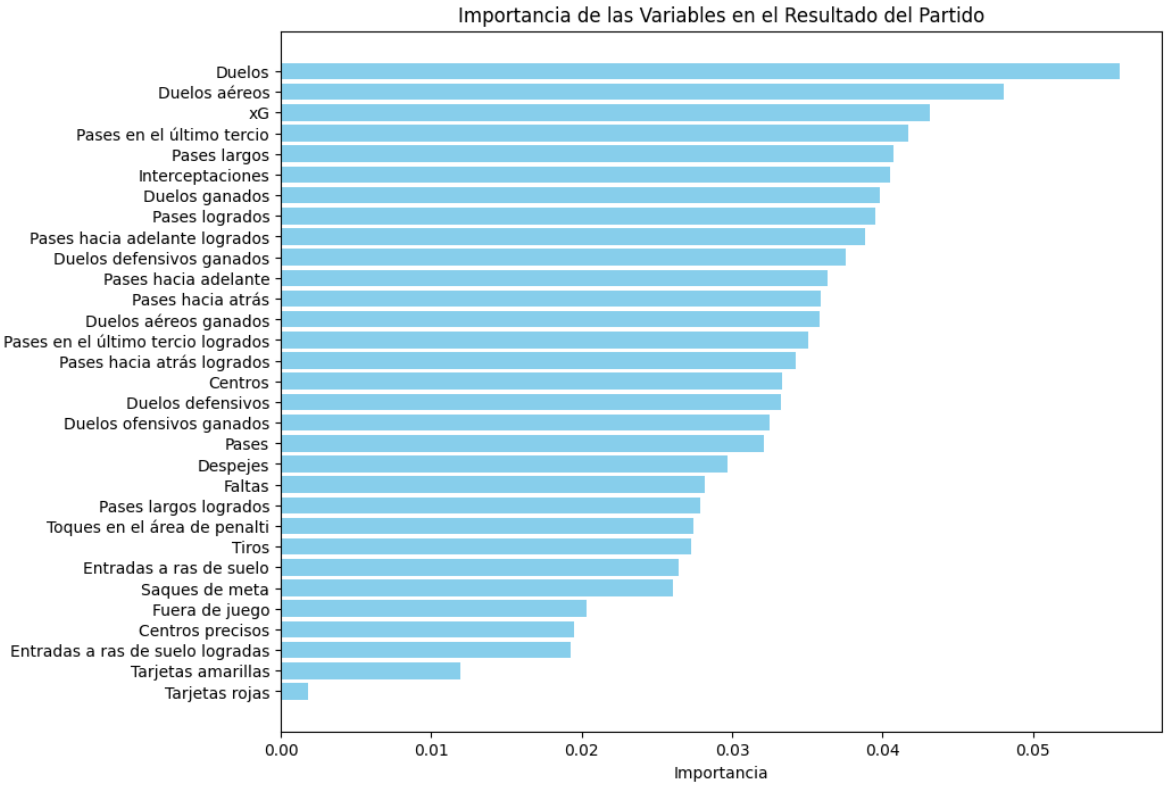


Figura 8. Ponderadores de Importancia de cada Variable en el Resultado

En la figura 8 se puede ver qué variables contribuyen más al éxito (victoria) de los equipos. Las cinco variables más importantes en este sentido serían: Duelos, Duelos aéreos, xG, Pases en el último tercio, Pases largos. Por otro lado, las cinco variables que menos influyen son:

Tarjetas rojas, Tarjetas amarillas, Entradas a ras de suelo logradas, Centros precisos, Fuera de juego.

Cabe señalar que en el modelo aplicado se incorpora Grid Search, que es una técnica de optimización que busca encontrar la combinación de hiperparámetros que maximiza el desempeño del modelo. En el caso de Random Forest, los hiperparámetros controlan aspectos clave del modelo, como el número de árboles, la profundidad máxima de los árboles, y otros factores que afectan su capacidad predictiva.

En Random Forest, los valores predeterminados de los hiperparámetros pueden no ser los mejores para un dataset específico. Optimizar estos valores ayuda a:

Mejorar el Desempeño del Modelo:

- Encontrar una configuración que maximice métricas como accuracy, recall, F1-score, etc.

Evitar Overfitting:

- Ajustar parámetros como la profundidad máxima de los árboles (max_depth) o el tamaño mínimo de las hojas (min_samples_leaf) reduce la complejidad del modelo y lo hace más generalizable.

Evitar Underfitting:

- Parámetros como el número de árboles (n_estimators) aseguran que el modelo capture patrones relevantes sin ser demasiado simple.

En cuanto a la performance de este modelo, a continuación se muestra una matriz de confusión:

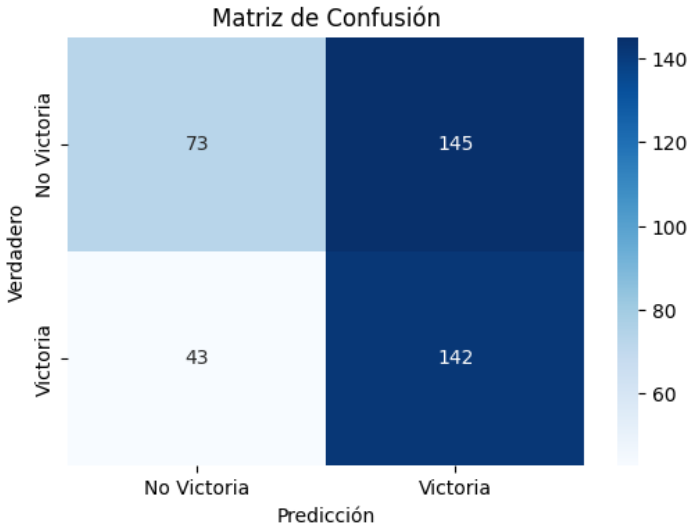


Figura 9. Matriz de Confusión Modelo Predictivo Victorias

Se puede apreciar que el modelo, en el conjunto de prueba, predice correctamente un 77% de resultados que sí terminan en victoria. En efecto, el valor de recall da un 77%, lo cual indica

que tiene un buen desempeño al identificar correctamente las victorias, lo cual es importante para el objetivo de minimizar los falsos negativos (victorias que el modelo predice como derrotas o empates). Además, las métricas de Macro Average y Weighted Average Accuracy son equilibradas, con valores cercanos a 0.55, lo que indica un mejor balance global entre ambas clases.

Con estas variables y coeficientes, es posible definir y calcular una métrica o score para el rendimiento de los jugadores. Esta debiera ser una función del tipo:

$$Score = f(variables, coeficientes)$$

Por ejemplo, podría ser una función lineal o producto escalar entre los vectores de coeficientes y valores de las variables, del tipo:

$$Score_i = \sum_j Wj * Xij$$

En que,

Wj = coeficiente (ponderador) asociado a la variable i

Xij = valor de la variable j para el jugador i.

Un modelo como el propuesto ofrece varias ventajas, como por ejemplo:

- Fácil de interpretar: un aumento en una variable (como "pases completados") tiene un impacto directamente proporcional al ponderador Wj.
- Transparente: cada variable contribuye de manera explícita al score final.
- Útil para modelos iniciales, para luego analizar si es necesario recurrir a un modelo más complejo o no lineal.

Luego, teniendo tanto los coeficientes y los valores de las estadísticas de los jugadores en sus partidos, se puede aplicar la función de Score en cada partido. En la tabla 5 se pueden apreciar una descripción general de estos resultados obtenidos para los Score.

Promedio	5,8
Desv. Estándar	2,9
Valor Mínimo	0
Valor Máximo	16,7
1er Cuartil	3,7
2do Cuartil	5,9
3er Cuartil	7,9

Tabla 5. Descripción General de los Valores de Score

De todas maneras, necesario realizar una normalización para apreciar de mejor forma las comparaciones. En la tabla 6 se muestra la descripción general de los valores normalizados.

Promedio	0,35
Desv. Estándar	0,17
Valor Mínimo	0
Valor Máximo	1
1er Cuartil	0,22
2do Cuartil	0,35
3er Cuartil	0,47

Tabla 6. Descripción General de los Valores de Score Normalizados

De lo anterior se puede deducir que el coeficiente de variación es de 0,49, lo cual es un valor alto y muestra que en general el rendimiento de los jugadores es bien variable, tanto a nivel de jugadores como a nivel de partidos. Esto es consistente con un punto relevante que se menciona en el estudio de Luca Pappalardo et al. (2019), en que se señala que, incluso en los jugadores de elite, se ve alta variabilidad de rendimiento. Algo que diferencia un jugador de elite versus uno más bien normal, es que aquellos de elite tienen peaks más frecuentes de rendimiento. Respecto a la distribución de los Scores Normalizados, que se muestra en la figura 10, se puede apreciar que hay una leve tendencia hacia la izquierda, lo cual muestra en general que los rendimientos no son muy altos en la Liga Chilena, algo que puede reflejar el bajo nivel de esta liga versus otras de Sudamérica o Europa.

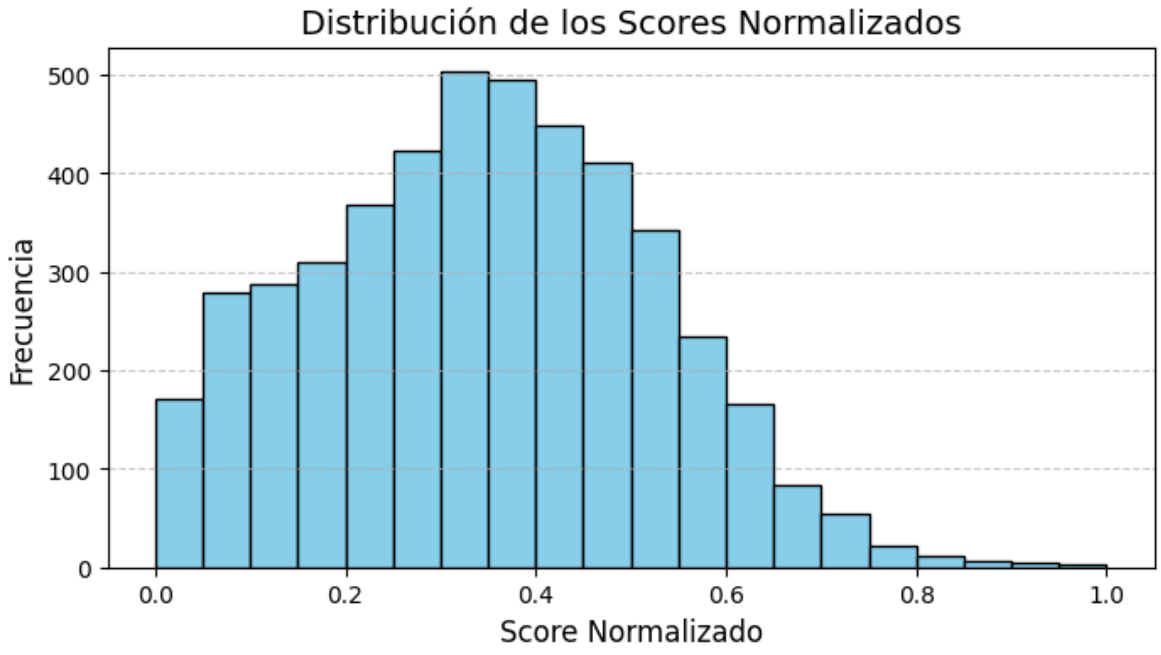


Figura 10. Histograma de Scores Normalizados

Además, se genera un boxplot de estos Scores Normalizados, lo cual se presenta en la figura 11. Se pueden visualizar ciertos outliers, que corresponden a ciertos jugadores que han obtenido muy buenos rendimientos en los partidos. Dentro de estos jugadores con altos rendimientos, se puede mencionar a: César Cortés, Erick Wiemberg, Gabriel Suazo, Ignacio Vásquez, Luciano Aued, Leonardo Gil, Matías Zaldivia, y Tomás Aránguiz. Sin duda se trata de jugadores reconocidos a nivel nacional, e incluso de selección como G. Suazo y M. Zaldivia.

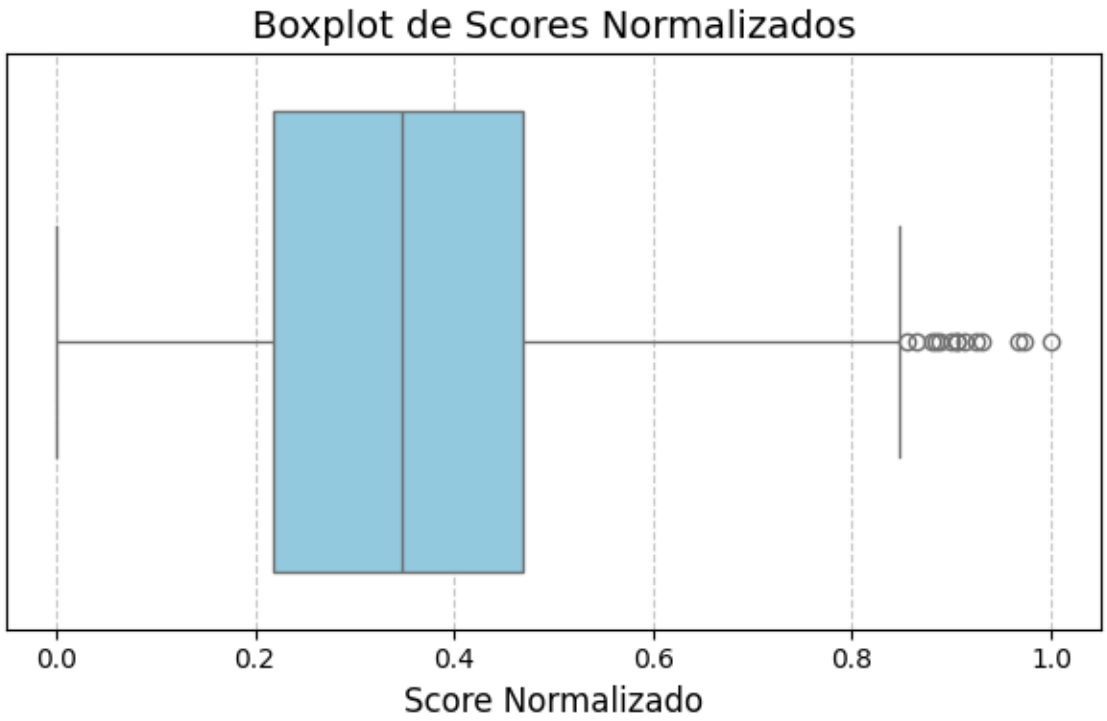


Figura 11. Boxplot de Scores Normalizados

Hasta el momento se han mezclado en los análisis jugadores de diversas posiciones en el campo de juego, lo cual no justo por el hecho de cada posición cumple una función distinta en la cancha. Por ejemplo, un delantero probablemente realice menos pases que un volante, lo cual genera distintos valores de rendimiento. Por eso, se realizó la división en las siguientes posiciones: Defensa, Volante, Delantero. No se considera la posición de arquero, dado que es una función muy distinta al resto, y en el dataset existen muy pocos registros de jugadores en dicha posición. En la figura 12 se muestra la distribución de jugadores por posición en el dataset. Se puede apreciar la poca participación de los arqueros dentro de los registros, y que además existe una predominancia de defensas y volantes, por sobre los delanteros.

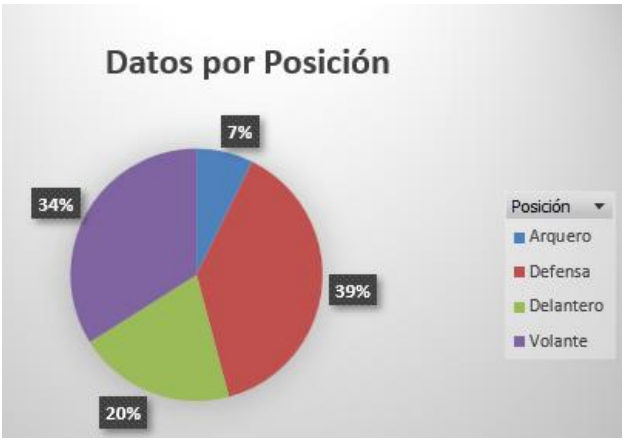


Figura 12. Distribución de Jugadores por Posición

Al realizar la separación entre Defensas, Volantes y Delanteros en el análisis, se muestra en la figura 13 los valores promedio de los Score Normalizados por posición. Llama la atención que los Defensas y Volantes tengan mayor Score promedio que los Delanteros. En los Defensas, el Score promedio da un valor de 0.4, en los Volantes un 0.37, mientras que en los Delanteros un 0.21. Defensas y volantes pueden tener valores más altos en variables con mayores pesos, como Pases o Duelos, que son más comunes en estas posiciones. Los delanteros suelen participar en menos acciones que los volantes o defensas, ya que su contribución se mide más por eventos clave (Por ejemplo goles, que no fueron considerados en el análisis). Los volantes y defensas, al intervenir en más jugadas, pueden acumular valores más altos en estadísticas generales. El modelo de Random Forest asigna pesos a las variables en función de su capacidad para predecir el resultado del partido (victoria o no). Si las variables más relevantes están asociadas a roles defensivos o de construcción de juego, es razonable que los defensas y volantes tengan scores más altos. El cálculo del score se basa en estadísticas generales, no necesariamente ajustadas al contexto de la posición: un defensa que sobresale en intercepciones puede acumular un score alto, aunque no se espera que anote goles. Un delantero puede tener valores bajos en variables defensivas, lo que podría afectar su score total.

El modelo probablemente favorece variables defensivas o de construcción de juego debido a su relevancia para predecir victorias. Esto no significa que los delanteros no tengan un impacto, pero su rol específico puede estar subrepresentado en las variables utilizadas para calcular el score. Si se hubieran considerado los goles, probablemente los Delanteros hubieran tenido mayor Score que las otras posiciones. Por lo mismo, esto refuerza el hecho de que es justo realizar las comparaciones por posición.

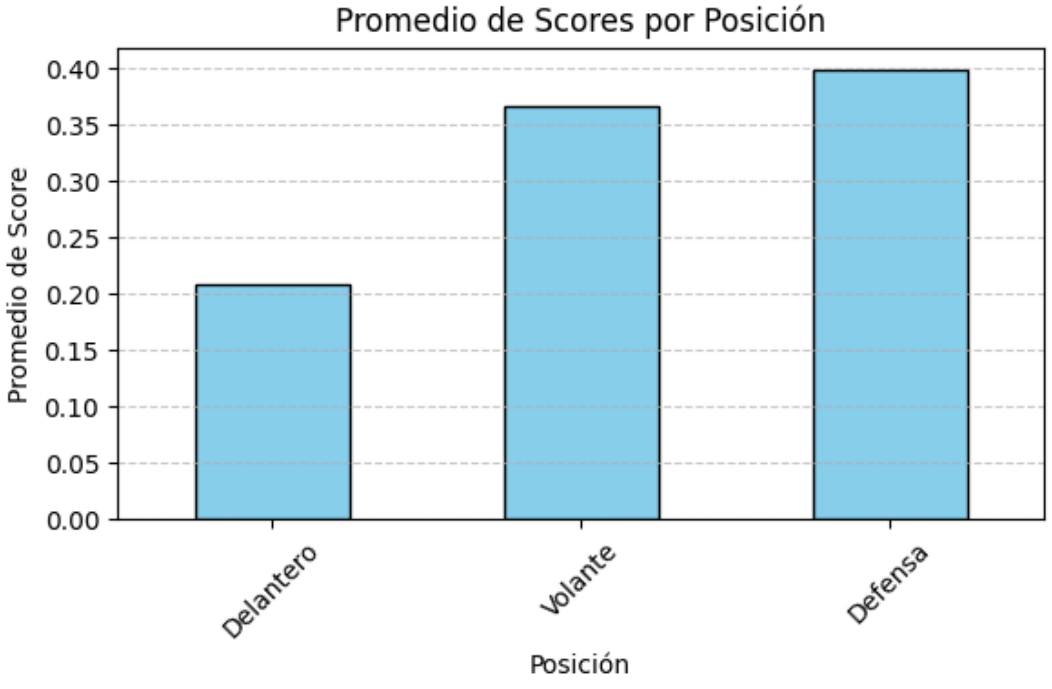


Figura 13. Promedios de Scores Normalizados por Posición

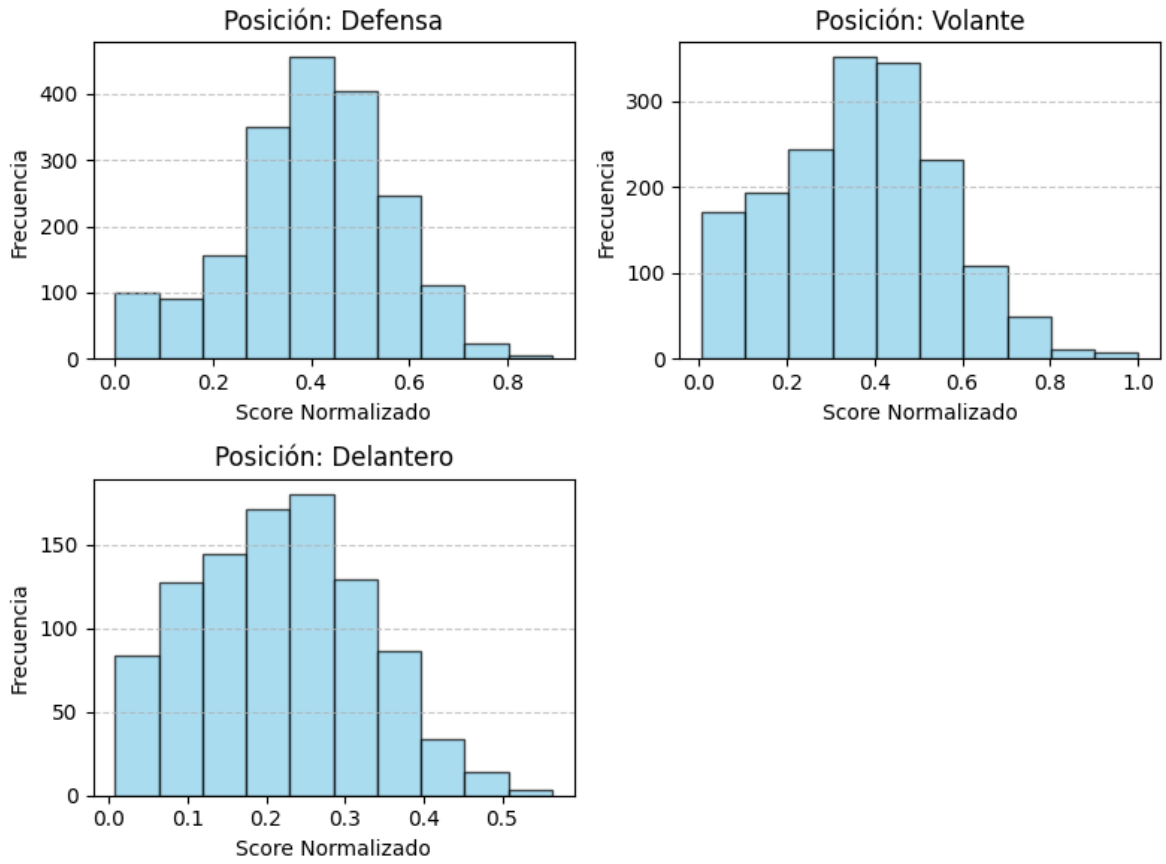


Figura 14. Histograma de Scores Normalizados por Posición

En la figura 14 se muestra la distribución de Scores promedios normalizados por posición. Si bien la posición de delantero tiene menores valores comparados con los defensas y volantes, se puede apreciar una dispersión un poco menor pero un sesgo mayor hacia la izquierda.

Con esto, ya se puede pensar en elaborar un ranking de jugadores por posición, en base a este Score Normalizado promedio. Pero como se mencionó anteriormente, es más justo realizar este ranking por posición, dadas las distintas funciones que se cumplen dentro del campo de juego. En la figura 15 se muestran los 5 mejores jugadores por posición. En general se puede ver que se trata de jugadores reconocidos en al ámbito futbolístico, y que han tenido logros importantes en el último tiempo. Por ejemplo, en los defensas, M. Zaldivia es actualmente referente de la U. de Chile, y ha sido nominado a la selección. Por otro lado, E. Wiemberg y E. Amor han hecho una gran campaña en Colo Colo en la pasada Copa Libertadores, y G. Suazo hace algunas temporadas fue transferido al fútbol francés, y ha tenido buenas actuaciones en la selección chilena. En cuanto a los volantes, L. Aued ha hecho grandes campañas en la U. Católica, lo que llevó a que fuera transferido al fútbol argentino. Además, L. Gil ha hecho grandes campañas en Colo Colo, E. Pavez es referente en el mismo equipo y seleccionado nacional, y C. Cortés ha tenido grandes temporadas en diversos clubes. Por último, en relación con los delanteros, P. Solari fue transferido a River Plate de Argentina por sus buenas actuaciones en Colo Colo, J. Alfaro está siendo contratado actualmente por la U. de Chile, y M. Bolados ha tenido varias nominaciones a la selección nacional. Aparecen otros jugadores de

no tanto renombre, como M. Moya, T. Aránguiz, F. Espinoza, y Y. Zapata, pero que sin duda, según estos análisis, tienen un buen potencial. En el Anexos 1 se pueden ver los rankings completos por posición.

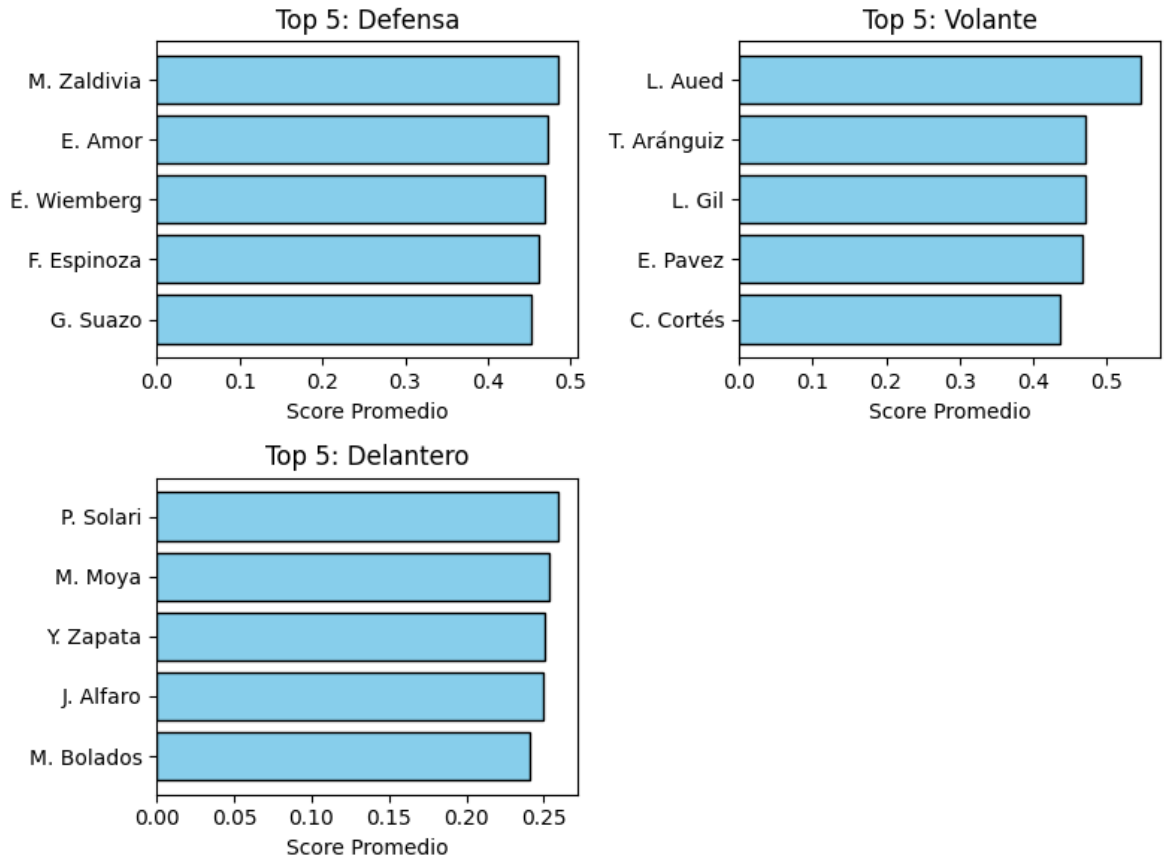


Figura 15. Ranking de Jugadores por Posición, según Score Promedio Normalizado

Por otro lado, ya teniendo los Scores, es posible realizar un análisis de cluster, con el objetivo de clasificar a jugadores según rendimiento, y así acotar la búsqueda de talentos por parte de los clubes a través de las personas que están relacionadas con el scouting. Para esto, se propone clasificar a los jugadores en tres agrupaciones según rendimiento: Alto, Medio, Bajo. Luego, se recurre a K Means estableciendo estos 3 cluster, y se realiza por posición para ser consistente con las comparaciones anteriores. En las figuras 16, 17 y 18 se presentan los resultados de este análisis de cluster, en gráficos de burbujas y mostrando en color rojo los rendimientos bajos, en amarillo los rendimientos medios, y en verde los rendimientos altos. En cuanto a la calidad de los cluster, se recurre al coeficiente de Silhouette, en cual da mayor o igual a 0,5 para cada posición (rango 0,52 – 0,65). Un coeficiente de Silhouette igual o mayor a 0.5 indica que los clústeres están razonablemente bien definidos y separados. Esto significa que el clustering es sólido y confiable para interpretar los resultados. Además, implica que los jugadores dentro de un mismo clúster tienen scores promedio similares. Las diferencias entre clústeres (bajo, medio, alto rendimiento) son significativas, y la elección de k =3 (bajo, medio, alto rendimiento) parece adecuada para los datos. No parece haber superposición significativa entre los clústeres.

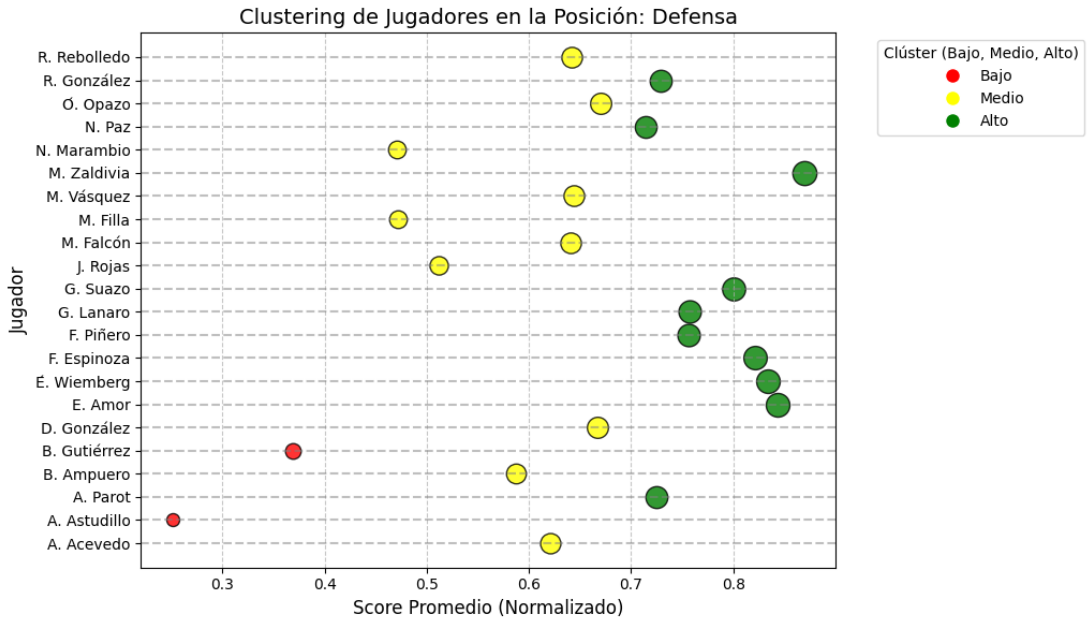


Figura 16. Análisis de Cluster para los Defensas

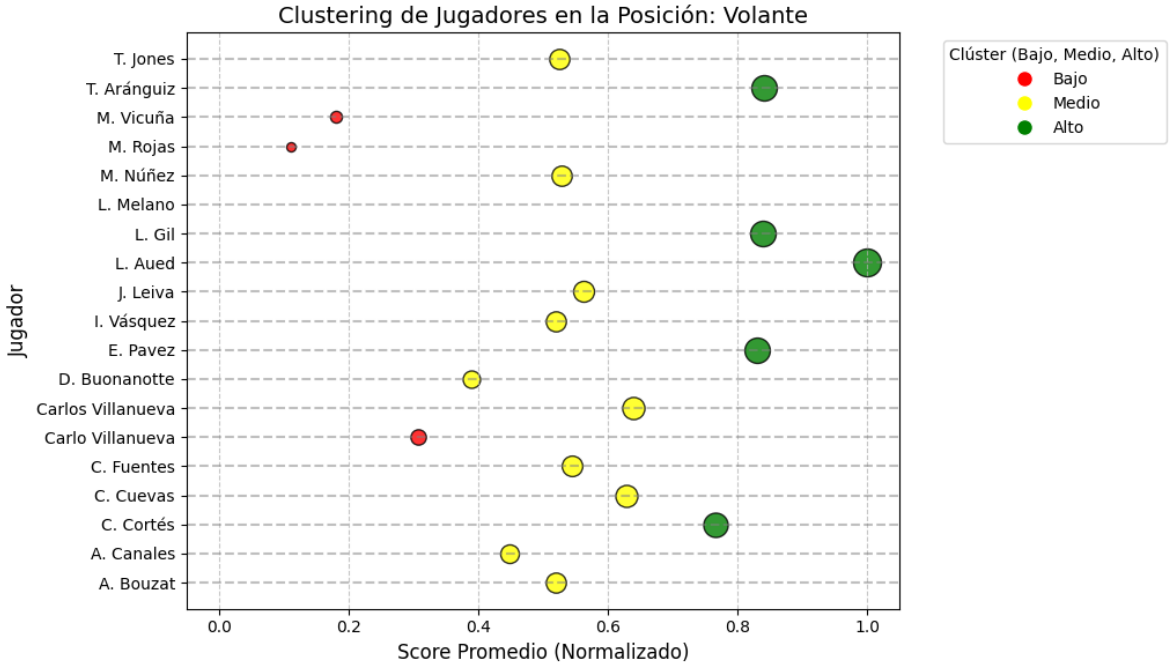


Figura 17. Análisis de Cluster para los Volantes

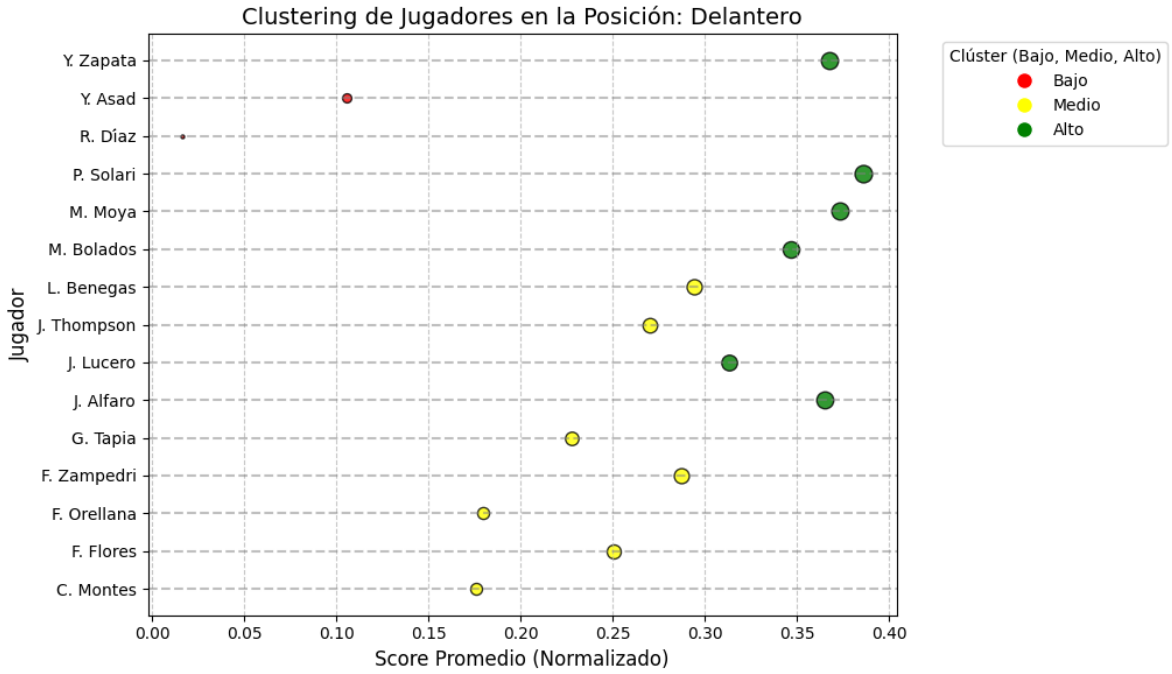


Figura 18. Análisis de Cluster para los Delanteros

El ranking de jugadores y la clasificación en clústeres según su rendimiento ofrecen una herramienta valiosa para mejorar la toma de decisiones en el scouting y la gestión deportiva de los clubes de fútbol. Para el scouting, esta información permite identificar de manera eficiente a los jugadores con mejor desempeño en términos objetivos, lo que puede ser particularmente útil para encontrar talentos en posiciones específicas o para detectar jugadores con un alto rendimiento sostenido. Además, la clasificación en clústeres facilita la comparación entre jugadores con perfiles similares, permitiendo evaluar el potencial de un jugador en un contexto competitivo y adaptado a las necesidades del equipo.

Desde una perspectiva de gestión deportiva, los rankings y clústeres pueden ser útiles para optimizar la conformación de plantillas. Los datos permiten a los clubes identificar fortalezas y debilidades dentro de su equipo, priorizando las posiciones que requieren refuerzos o ajustes. La segmentación de jugadores por rendimiento también puede apoyar la personalización de estrategias de entrenamiento, enfocándose en mejorar aspectos específicos de jugadores con menor rendimiento dentro de su grupo, o potenciando a quienes están cerca de dar un salto significativo en su nivel de juego. Por último, estos resultados también ofrecen una base sólida para justificar inversiones deportivas ante accionistas o propietarios, al basarse en análisis objetivos y consistentes del rendimiento de los jugadores, lo que puede ser clave para planificar estrategias a largo plazo.

Un análisis interesante sería contrastar los rankings y Scores obtenidos en este análisis con las calificaciones o valoraciones de los jugadores realizadas por plataformas especializadas durante las últimas temporadas. Este contraste podría validar la efectividad de las métricas utilizadas y ofrecer una perspectiva más enriquecedora, al incorporar opiniones expertas y subjetivas en la evaluación del rendimiento. Este enfoque ayudaría a entender mejor las

fortalezas y limitaciones de las metodologías aplicadas, destacando su aplicabilidad en contextos reales y prácticos.

Una plataforma relacionada es FotMob, la cual proporciona resultados en vivo, estadísticas detalladas, calificaciones y noticias actualizadas de una amplia variedad de competiciones de fútbol a nivel mundial. Con una base de datos que abarca más de 375 ligas y torneos, FotMob permite a los aficionados seguir de cerca a sus equipos y jugadores favoritos. La plataforma ofrece alertas personalizadas, alineaciones, tablas de posiciones, máximos goleadores y programación televisiva, todo en una interfaz intuitiva y fácil de usar. A continuación, se va a comparar la evaluación realizada por el modelo del presente proyecto, versus la valoración efectuada por la plataforma, para aquellos jugadores que están dentro del top 5 presentado anteriormente.

En las tablas 7, 8 y 9 se muestra la comparación por posición. En FotMob, las categorías asociadas con las calificaciones de los jugadores representan un rango de puntuaciones basado en su rendimiento en un partido:

- Calificaciones muy altas (0.8 o superior). Representa un rendimiento destacado, como marcar goles, realizar asistencias, o tener una gran influencia en el juego.
- Calificaciones altas (0.7 - 0.79). Indica un buen rendimiento, con contribuciones positivas al equipo.
- Calificaciones medias (0.6 - 0.69). Representa un desempeño adecuado, pero sin destacarse significativamente.
- Calificaciones bajas (menos de 0.6). Esto sugiere un rendimiento por debajo del promedio, errores importantes, o poca influencia en el juego.

Jugador	Score Modelo	Cluster Modelo	Score Plataforma	Categoría Plataforma
M. Zaldivia	0,88	Alto	0,72	Alto
E. Amor	0,85	Alto	0,73	Alto
E. Wiemberg	0,84	Alto	0,71	Alto
F. Espinoza	0,82	Alto	0,69	Medio
G. Suazo	0,81	Alto	0,72	Alto

Tabla 7. Comparación Scores con FotMob para Defensas

Jugador	Score Modelo	Cluster Modelo	Score Plataforma	Categoría Plataforma
L. Aued	1,0	Alto	0,70	Alto
T. Aránguiz	0,84	Alto	0,71	Alto
L. Gil	0,84	Alto	0,74	Alto
E. Pavez	0,83	Alto	0,70	Alto
C. Cortés	0,77	Alto	0,70	Alto

Tabla 8. Comparación Scores con FotMob para Volantes

Jugador	Score Modelo	Cluster Modelo	Score Plataforma	Categoría Plataforma
P. Solari	0,39	Alto	0,71	Alto
M. Moya	0,37	Alto	0,67	Medio
Y. Zapata	0,37	Alto	0,67	Medio
J. Alfaro	0,37	Alto	0,70	Alto
M. Bolados	0,35	Alto	0,69	Medio

Tabla 9. Comparación Scores con FotMob para Delanteros

Cabe señalar que para la plataforma FotMob muy pocos jugadores han tenido puntuaciones sobre 0,8. Algunos ejemplos son L. Messi, K. Mbappé, R. Lewandowski, y C. Ronaldo, pero sólo en algunas temporadas. Por lo tanto, un Score en torno a 0,7 es bastante valorable. En general, se puede ver en las comparaciones que hay una consistencia en la categorización de los jugadores. Se ven diferencias en los Score, principalmente en delanteros, dado que el enfoque de este estudio no considera los goles y más bien la construcción de juego. En todo caso, se puede apreciar que los jugadores que están en el top 5 del ranking de este estudio, tienen buenas calificaciones en la plataforma. Existen algunos casos no tan coincidentes, como F. Espinoza, M. Moya, Y. Zapata y M. Bolados, pero de todas formas son jugadores que debieran ser tomados en cuenta, dado que contribuyen de manera positiva al juego y a los resultados de los partidos, sobre todo en aspectos de la elaboración de juego.

Sin embargo, podría ser discutible el hecho de considerar sólo los Scores promedio para estos análisis, dado que pueden existir otros factores que pueden influir en el potencial de un jugador a futuro. Por ejemplo, un jugador puede tener un rendimiento promedio no muy alto en los últimos años, pero quizás puede tener una tendencia al alza, lo que puede transformarlo en atractivo en un mercado de pases, y además, la inversión podría ser menor que en el caso de un jugador que ya tiene un rendimiento promedio alto. Existen clubes en etapa de crecimiento y que tienen recursos limitados para invertir, y que por lo mismo puede ser rentable para ellos tener una política más austera en las contrataciones, pero mirando hacia el futuro con la idea de convertirse en el club poderoso. Desde este punto de vista, puede ser interesante analizar un poco la evolución que han tenido los jugadores en cuanto a sus rendimientos partido a partido, e idealmente incorporarlo de alguna forma en el análisis.

Se parte por un análisis gráfico, por posición, y considerando jugadores en distintos cluster en las figuras 19, 20 y 21. En estos gráficos, aparte de la evolución, se grafican las líneas de tendencia y se muestran los coeficientes de variación en cada caso. Hay casos interesantes que se pueden tomar en consideración, antes de tomar una decisión. Por ejemplo, en los defensas, E. Amor está clasificado como nivel alto, pero se ve una tendencia hacia la baja. Por otro lado, hay que defensa del cluster de nivel medio (J. Rojas) que muestra una tendencia al alza en su rendimiento, y que podría ser de todas maneras considerado en un mercado de fichajes. En cuanto a los volantes, hay jugadores que pueden ser interesantes para considerar, que están clasificados en nivel medio y que pueden tener buen potencial en base a su evolución positiva: D. Buonanotte y Carlo Villanueva. Por último, en relación con los delanteros, hay un par de jugadores en nivel medio que muestran una tendencia positiva: J. Alfaro (en negociaciones con la U. de Chile), y F. Flores. Llama de todas maneras la atención, que un jugador como F. Zampedri, quien es goleador histórico de la U. Católica, esté

clasificado como nivel medio y con tendencia a la baja. Seguramente si consideráramos los goles anotados como una de las variables, este jugador tendría un Score altísimo, pero en general, los centro delanteros son más directos de evaluar, dado que las decisiones de contratación se basan más que todo en los goles anotados. Además, las fortalezas de un goleador de área como Zampedri, no se basan mucho en las variables relacionadas con la construcción de juego.

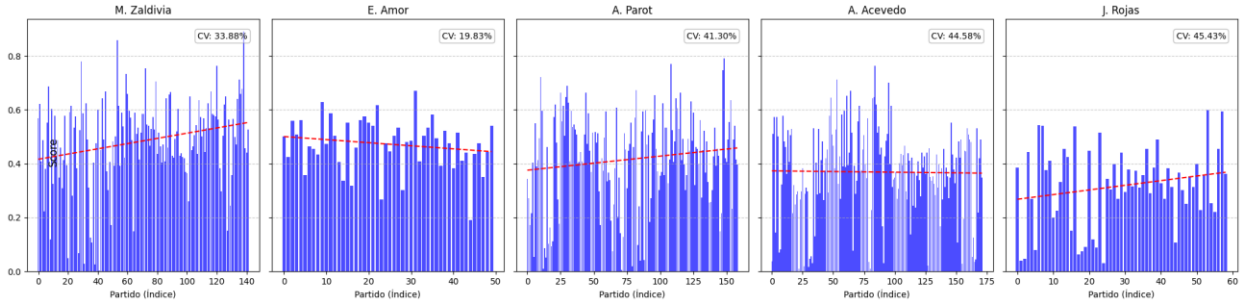


Figura 19. Evolución de Scores de Defensas

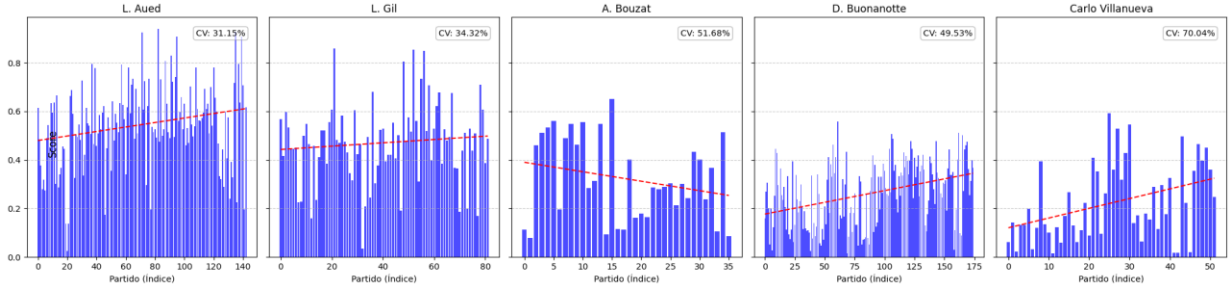


Figura 20. Evolución de Scores de Volantes

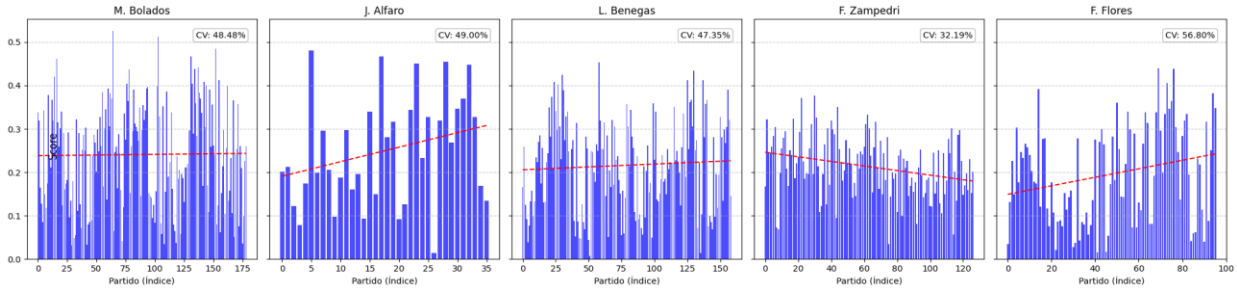


Figura 21. Evolución de Scores de Delanteros

Dado esto, resulta interesante analizar la posibilidad de calcular un nuevo Score, el cual sería ajustado por la tendencia. Este enfoque permitirá reflejar tanto el rendimiento acumulado (promedio) como la dirección de cambio en el tiempo (tendencia), para ver cómo influyen en los resultados.

El cálculo puede estructurarse como:

$$\text{Score Ajustado} = \alpha \cdot \text{Promedio del Score} + \beta \cdot \text{Tendencia}$$

Donde:

- α y β son pesos que reflejan la importancia relativa del promedio y la tendencia.
- La tendencia puede calcularse como la pendiente de la regresión lineal sobre los scores partido a partido.

Para estos cálculos, se considerará un valor de α = 0.8 y β=0.2, lo cual es perfectamente modificable de acuerdo con la importancia que se le quiera dar a cada variable. En la figura 21 se puede ver cómo quedarían los rankings de jugadores por posición de acuerdo con este nuevo Score ajustado. Cae señalar que este nuevo Score también es normalizado. Se aprecia que, por lo menos en los top 5 de jugadores por posición, son los mismos nombres, pero en varios casos la posición relativa en el ranking cambia. Están los ejemplos de J. Alfaro, F. Espinoza, o L. Gil, que suben de ranking. Por otro lado, hay jugadores como P. Solari, M. Zaldivia, o T. Aránguiz que bajan de ranking, pero de todas maneras se mantienen dentro de los top 5 por posición. En al Anexo 2 se pueden ver los rankings completos por posición.

A continuación, veremos si se ven modificaciones de los resultados en relación con los análisis de cluster. En las figuras 22, 23, y 24 se muestran los resultados obtenidos. En general, no se ven cambios considerables, pero sí algunas modificaciones en las composiciones de los cluster. Por ejemplo, en los defensas, A.Parot pasa de nivel medio a alto, mientras que M. Filla pasa de nivel medio a nivel bajo. En cuanto a los volantes, se ven más que todo descensos en las categorías de los clusters: por ejemplo, A. Bouzat, D. Buonanotte, o A. Canales. Por último, en relación con los delanteros, se ven algunos ascensos de categoría, como los casos de L. Benegas y J. Thompson, del nivel medio a alto. El coeficiente de Silhouette, en este caso, da un poco menor, en el rango 0.47 – 0.52.

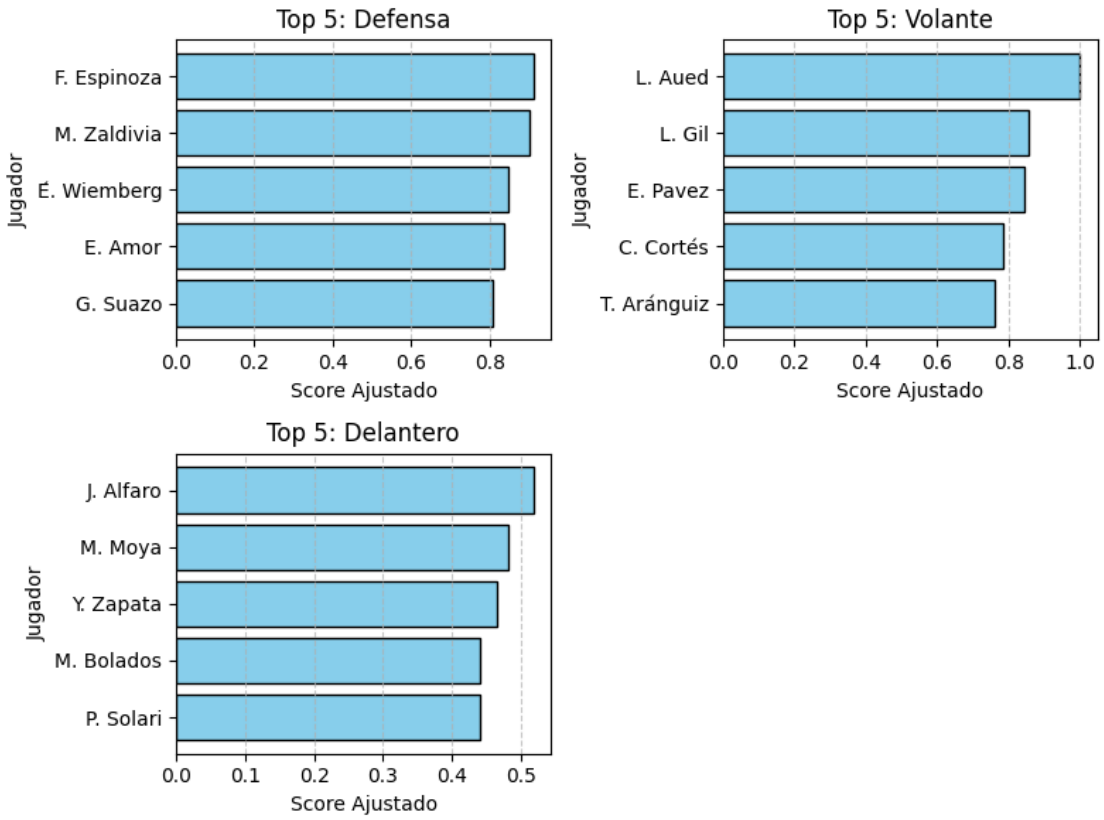


Figura 21. Ranking de Jugadores por Posición, según Score Ajustado

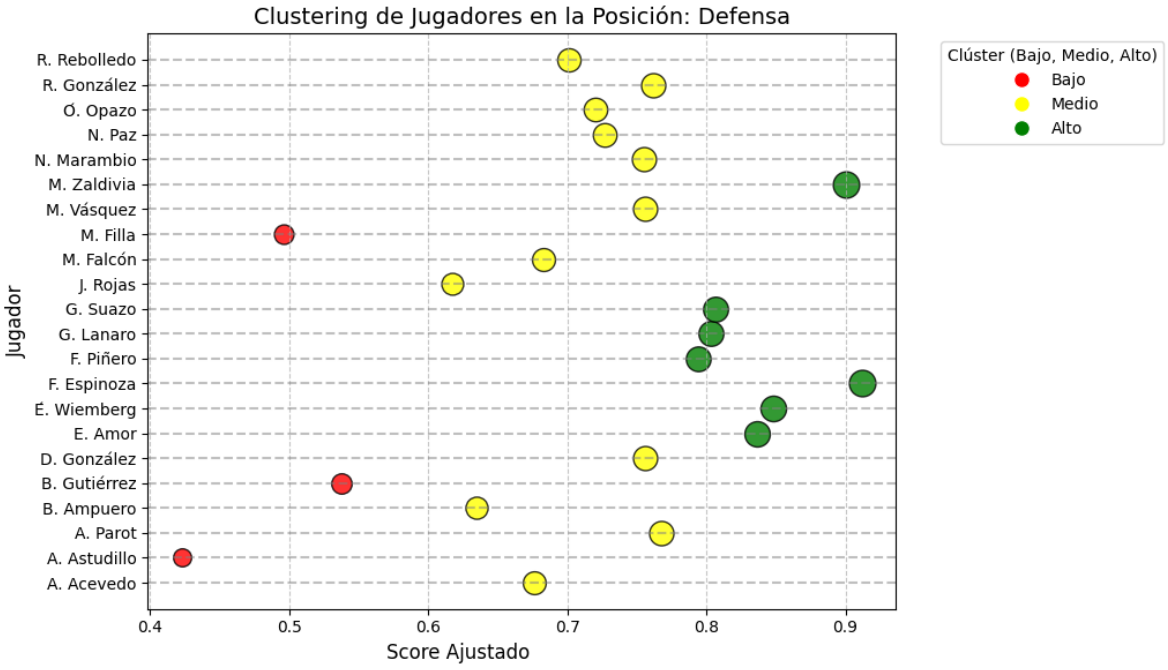


Figura 22. Análisis de Cluster para los Defensas, según Score Ajustado

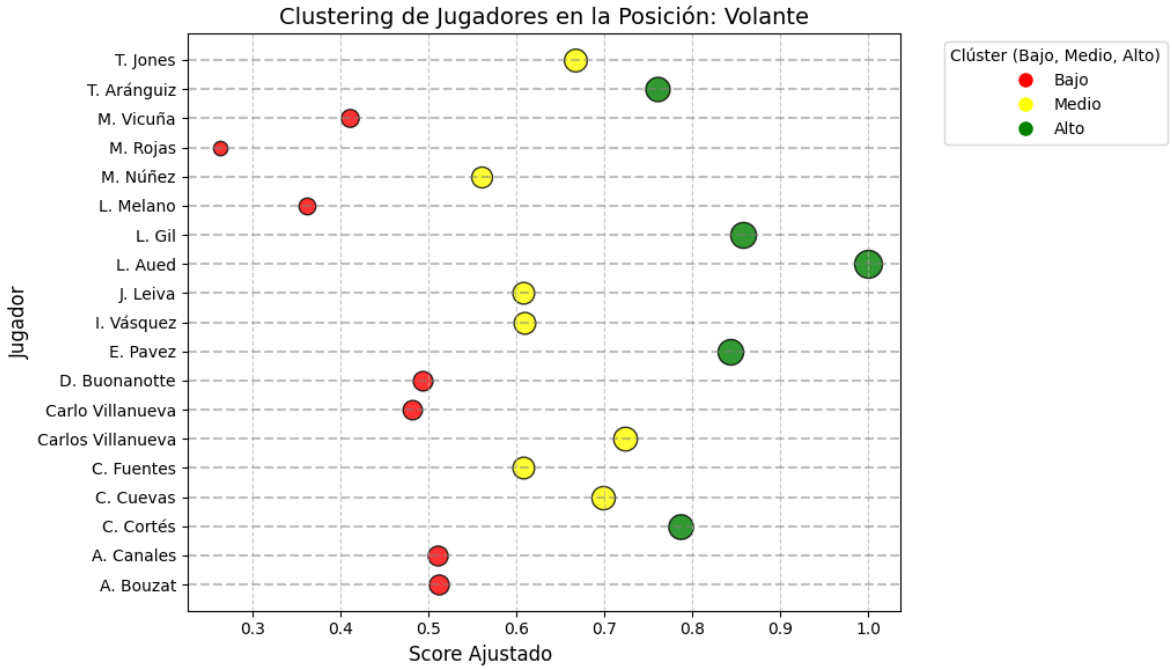


Figura 23. Análisis de Cluster para los Volantes, según Score Ajustado

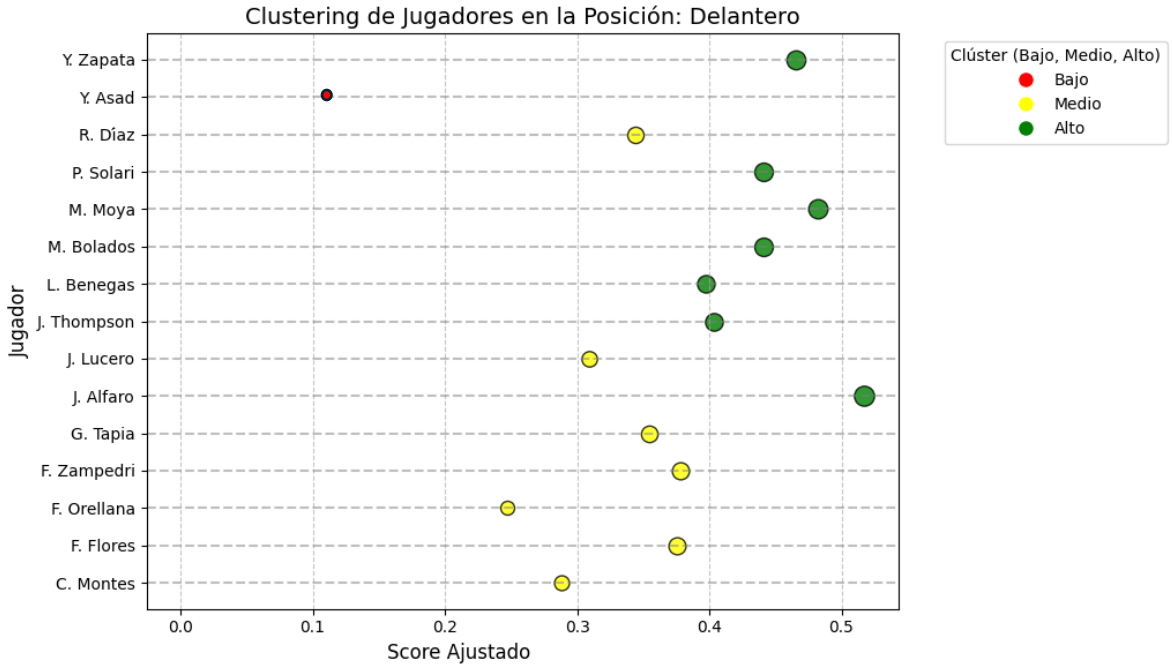


Figura 24. Análisis de Cluster para los Delanteros, según Score Ajustado

En resumen, el análisis comparativo entre el Score promedio normalizado y el score ajustado refleja diferencias sutiles en los resultados obtenidos, principalmente en la calidad del clustering y los rankings de algunos jugadores. El Score promedio normalizado muestra un desempeño levemente mejor en términos del coeficiente de Silhouette, indicando una coherencia algo mayor en la separación de los clústeres. Esto sugiere que, para fines de

clasificación y ranking de jugadores, el score promedio podría ser más efectivo al proporcionar una representación más clara de las diferencias de rendimiento entre ellos.

Por otro lado, el Score ajustado, que incorpora la tendencia, introduce una perspectiva adicional sobre la evolución del rendimiento de los jugadores a lo largo del tiempo. Aunque su impacto en los rankings es limitado debido al peso predominante del promedio, puede ser valioso para identificar jugadores con mejoras consistentes o aquellos que mantienen un rendimiento estable. Sin embargo, su uso podría ser más relevante en análisis enfocados en proyecciones futuras o en contextos donde la consistencia a lo largo del tiempo sea un factor crítico.

CONCLUSIONES

Las conclusiones de este proyecto destacan la relevancia y potencial de la ciencia de datos en la evaluación del rendimiento de futbolistas, a través de la construcción de un ranking objetivo basado en métricas de rendimiento ajustadas por tendencias y contextos posicionales. El análisis permitió identificar jugadores destacados, evaluar su evolución y clasificar su rendimiento en clústeres bien definidos, proporcionando una base relevante para apoyar la toma de decisiones en scouting y gestión deportiva, con el propósito de concentrar los esfuerzos.

El uso de métodos como Random Forest (incluyendo Grid Search) para obtener ponderadores y la incorporación de una técnica de clustering resalta cómo las metodologías avanzadas pueden superar enfoques subjetivos tradicionales. A pesar de que el coeficiente de Silhouette mostró ligeras diferencias entre los modelos de scores promedio y ajustados por tendencia, ambos enfoques demostraron ser consistentes en la identificación de jugadores clave, como los casos analizados en los rankings y clústeres. La inclusión de la tendencia como componente del score ajustado aporta una perspectiva evolutiva, útil para evaluar el potencial de los jugadores en mercados de fichajes o proyectos a largo plazo. Además. esta inclusión propuesta de las tendencias es perfectamente personalizable, en relación con asignarle más o menos importancia según los objetivos del scouting (Por ejemplo, si se quisiera privilegiar jugadores de proyección o con potencial, en algún mercado de fichajes).

Una de las decisiones metodológicas más relevantes fue centrarse en estadísticas asociadas a la construcción de juego, como pases, duelos ganados o centros precisos, en lugar de incluir métricas más obvias como los goles. Este enfoque permitió valorar el rendimiento desde una perspectiva más integral, destacando el impacto de las contribuciones indirectas al éxito del equipo. Si bien los goles son fundamentales en el resultado de los partidos, priorizar las métricas asociadas al proceso y la construcción del juego permitió identificar jugadores que, aunque no sean goleadores, tienen un rol crucial en la generación de oportunidades y el control del juego. Además, las estadísticas de goles (o asistencias) son masivamente publicadas y fáciles de obtener, y claramente son los principales factores para decidir en contratar a un delantero. El enfoque presentado en este estudio pretende ser complementario en ese sentido, ya que se concentra en el rendimiento “no obviamente visible” y que generalmente pasa más desapercibido y es difícilmente captado por periodistas y comentaristas deportivos.

Aunque inicialmente se consideró el uso de análisis PCA para reducir la dimensionalidad del conjunto de variables, finalmente no se aplicó directamente en este proyecto. Sin embargo, la exploración de su estructura proporcionó indicios importantes sobre qué tipos de variables tienen un mayor impacto en los rendimientos, permitiendo centrar el análisis en métricas clave. Este enfoque reafirma que la selección adecuada de variables, incluso sin reducción dimensional explícita, puede contribuir significativamente a comprender los factores que influyen en el desempeño deportivo.

Estos resultados son consistentes con trabajos previos como el de Pappalardo et al. (2019), donde el enfoque multidimensional y temporal de evaluación resultó crucial para rankings más justos. Asimismo, el análisis de redes y dinámicas colectivas propuesto por Buldú et al. (2018) refuerza la importancia de considerar el contexto posicional y las interacciones dentro del campo. Por otro lado, estudios como el de Wolf et al. (2020) subrayan la validez de combinar métricas objetivas con métodos estadísticos robustos para producir evaluaciones confiables, una aproximación también aplicada en este proyecto. También es relevante para el análisis el estudio de Merzah et al. (2024), en el sentido de reforzar la importancia de clasificar a los jugadores en relación con su rendimiento.

En términos de objetivos, el proyecto logró cumplir con los planteamientos iniciales al proponer un modelo que combina estadísticas individuales y contextuales para estimar el rendimiento de los jugadores. Los rankings y clústeres obtenidos muestran coherencia con las valoraciones observadas en plataformas externas, validando así la utilidad del enfoque adoptado. No obstante, algunos ajustes, como la integración de métricas específicas por posición (por ejemplo, goles y asistencias para delanteros), podrían aumentar la precisión y aplicabilidad del modelo, aunque de todas formas eso implicaría que los delanteros tengan un score mucho mejor que el resto, y sería injusto para los defensas y volantes. Además, esta inclusión no favorecería a todos los delanteros, sino que sólo a los centro delanteros, ya que por ejemplo los extremos derecho e izquierdo no suelen destacar en ese ámbito.

Dentro de las limitaciones, aparte del plazo establecido para este proyecto, se puede mencionar el dataset, en que se contaba sólo con registros asociados a 60 jugadores, siendo que en una temporada generalmente juegan más de 200. Si bien está la posibilidad de conseguir más datos, estos generalmente son pagados ya que están en plataformas especializadas en estadísticas (Ejemplos: Wyscout, KPI Football). Por esta razón, en los rankings no aparecen ciertos jugadores destacados del medio nacional (Ejemplos: Alexander Aravena, Víctor Méndez).

En cuanto a trabajos futuros, sería interesante explorar nuevas dimensiones del análisis, como el uso de datos de pases entre jugadores para aplicar enfoques de ciencia de redes. Este tipo de análisis permitiría evaluar la importancia de los jugadores no sólo desde su rendimiento individual, sino también desde su conectividad, centralidad y contribución colectiva en el campo de juego. Además, el análisis de redes permitiría identificar comunidades de jugadores dentro de los partidos, un concepto muy relacionado con las “pequeñas sociedades en el fútbol” que fue constantemente utilizado por el ex entrenador argentino César Luis Menotti, campeón del mundo en 1978. Del mismo modo, la integración de análisis de imágenes para capturar patrones de movimiento o posicionamiento táctico podría enriquecer la evaluación

del desempeño. Otra dirección prometedora sería analizar datos de jugadores de divisiones inferiores o segundas categorías para identificar talentos emergentes, priorizando aquellos con una evolución positiva y un costo potencialmente bajo, algo que podría ser de gran utilidad para equipos de primera división. Por último, es perfectamente posible pensar en un modelo predictivo basado en Machine Learning del rendimiento de jugadores, por ejemplo, usando como datos de entrenamiento los rendimientos de años pasados, y ocupar como datos de test aquellos rendimientos del presente o último año.

Es importante destacar que este modelo, si bien es una herramienta poderosa, no reemplaza por completo el juicio humano en la toma de decisiones. Existen factores difíciles de cuantificar, como la influencia de los representantes, las preferencias tácticas o el paladar futbolístico de los entrenadores, y los antecedentes extra futbolísticos de los jugadores. Estos aspectos cualitativos y subjetivos desempeñan un papel importante en el proceso de selección y gestión de jugadores. En este contexto, el modelo desarrollado actúa como una guía objetiva que permite acotar las búsquedas y reducir la incertidumbre en las decisiones, proporcionando un punto de partida robusto para el análisis.

Por otro lado, un tema relevante a analizar es la factibilidad de generalizar estos resultados. La importancia o los coeficientes obtenidos para las variables que contribuyen a la victoria en los partidos de fútbol no son necesariamente universales, ya que pueden variar significativamente según el nivel de la liga, la competición y las características de los equipos que componen el dataset utilizado como base. Estas métricas dependen en gran medida del contexto en el que se recopilan los datos, ya que las dinámicas de juego y las estrategias tácticas pueden diferir ampliamente entre ligas con distintos niveles de calidad y competitividad. Por ejemplo, en ligas de menor nivel técnico, puede que variables como los duelos físicos o los despejes sean más determinantes debido al estilo de juego más directo y menos elaborado. En cambio, en competiciones de alto nivel, como la Champions League o las principales ligas europeas, las estadísticas relacionadas con la posesión, los pases en el último tercio y las transiciones rápidas podrían tener un peso más significativo, reflejando un estilo de juego más técnico y estratégico. La generalización de estos coeficientes a otros contextos requiere precaución, ya que cada competición tiene sus propias particularidades, influenciadas por factores como la calidad técnica de los jugadores, las tácticas predominantes, las condiciones del terreno de juego y la intensidad física. Si bien los resultados obtenidos en este proyecto proporcionan una base valiosa para evaluar la importancia de las variables en un contexto específico, es probable que los pesos asignados a cada variable necesiten ser recalibrados al analizar datos de otras ligas o competiciones. Esto asegura que las conclusiones reflejen adecuadamente las dinámicas particulares de cada entorno futbolístico. Luego, se recomienda que estos resultados sean utilizados para ligas de similar nivel y exigencia que la chilena.

En términos prácticos, este modelo tiene aplicaciones claras en el ámbito del scouting y el desarrollo de talentos. Para los clubes, permite identificar talentos emergentes, optimizar decisiones de contratación y maximizar la eficiencia de los recursos asignados a fichajes. Además, los rankings y clústeres generados ofrecen una base para diseñar entrenamientos específicos orientados a mejorar aspectos críticos del rendimiento, como precisión en pases, capacidad defensiva o eficiencia ofensiva. Así, se posiciona como una herramienta estratégica

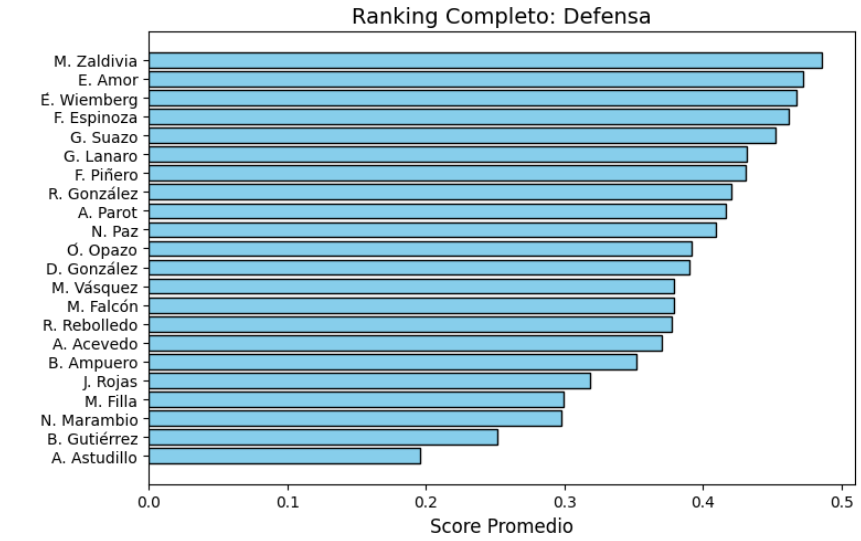
para alinear decisiones técnicas y deportivas con los objetivos organizacionales de los clubes, y perfectamente aplicable a otros deportes.

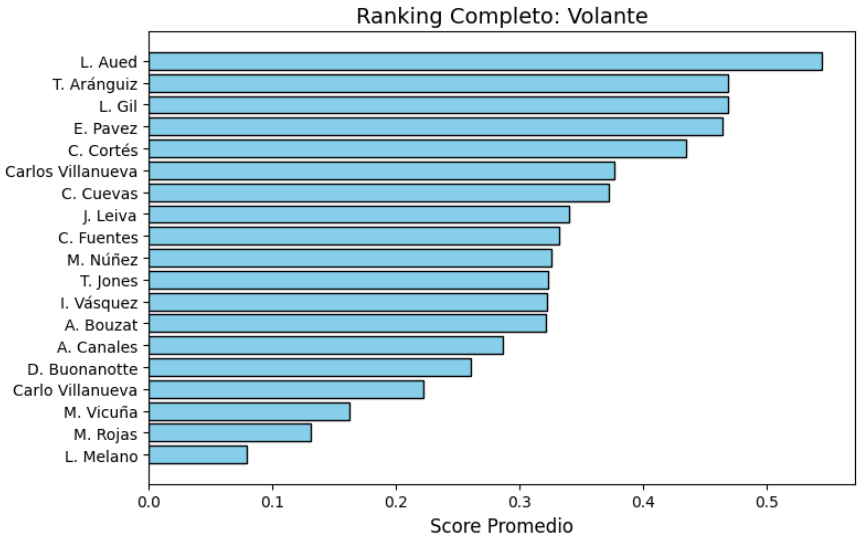
BIBLIOGRAFÍA

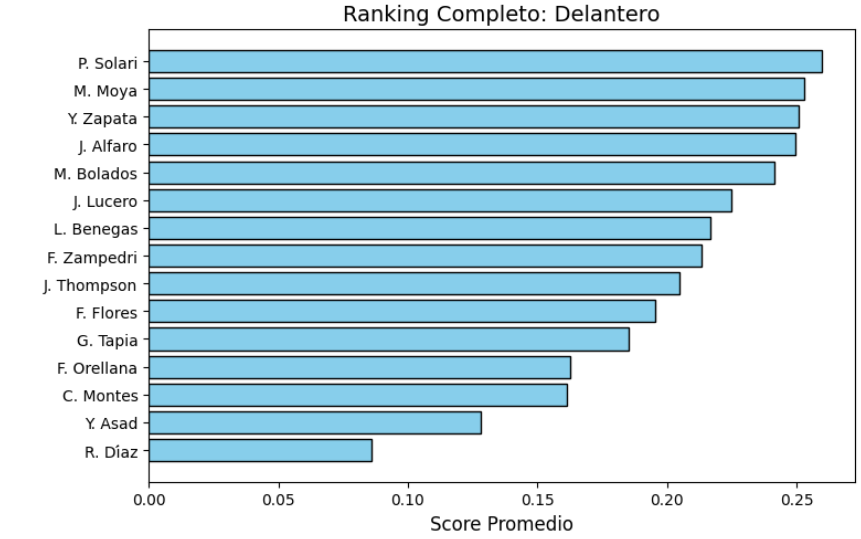
- Apostolou, K., & Tjortjis, C. (2019). Sports analytics for football league table and player performance prediction. *International Journal of Computer Science in Sport, 18*(1), 1-10. https://doi.org/10.1515/ijcss-2019-0001
- Buldú, J. M., Busquets, J., Martínez, J. H., Herrera-Diestra, J. L., Echegoyen, I., Galeano, J., & Luque, J. (2018). Using network science to analyse football passing networks: Dynamics, space, time, and the multilayer nature of the game. *Frontiers in Psychology, 9*, 1900. https://doi.org/10.3389/fpsyg.2018.01900
- Chavan, A. (2022). Recruitment of suitable football player using machine learning techniques. *International Journal of Engineering Research and Technology, 10*(8), 1321-1325. https://doi.org/10.37591/IJERT
- Dendir, S. (2016). When do soccer players peak? A note. *Journal of Sports Analytics, 2*(2), 89-105. https://doi.org/10.3233/JSA-160016
- Kati, Ö. F. (2022). Predicting the slope of a football player's performance, market value and wage using machine learning on a FIFA dataset. Tilburg University, Cognitive Science & Artificial Intelligence Department.
- Korte, F., Unkelbach, J., & Wolf, T. (2021). Play-by-play network analysis in football: A new framework to understand team and player interaction. *Journal of Sports Science & Medicine, 20*(1), 78-88. https://doi.org/10.1123/jssm.2020-0221
- Lee, H., Cho, S., & Kim, J. (2020). Prediction of football player value using Bayesian ensemble approach. *Journal of Artificial Intelligence Research, 67*(1), 101-115. https://doi.org/10.1613/jair.2020-0221
- Li, C., Zhu, H., & Ma, Y. (2022). Prediction of football player salary using machine learning techniques. *Journal of Financial Analysis, 29*(1), 233-248. https://doi.org/10.1016/j.jfa.2022.07.013
- Merzah, B. M., Croock, M. S., & Rashid, A. N. (2024). Intelligent classifiers for football player performance based on machine learning models. International Journal of Electrical and Computer Engineering Systems, 15(2), 173-182. https://doi.org/10.31247/ijeces.v15i2.2996
- Pappalardo, L., Cintia, P., Rossi, A., Massucco, E., Ferragina, P., Pedreschi, D., & Giannotti, F. (2019). PlayeRank: Data-driven performance evaluation and player ranking in soccer via a machine learning approach. *ACM Transactions on Intelligent Systems and Technology, 10*(5), 1-27. https://doi.org/10.1145/3343172
- Ati, A., & El Haddad, J. (2022). Using multi-criteria decision-making and machine learning for football player selection and performance prediction. *Procedia Computer Science, 205*, 728-734. https://doi.org/10.1016/j.procs.2022.07.100
- Wolf, S., Gudmundsson, J., & Horton, M. (2020). A football player rating system based on the Elo algorithm. *Journal of Sports Analytics, 6*(1), 1-12. https://doi.org/10.3233/JSA-200393

- Jana, A., Das, S., & Saha, D. (2021). Football player performance analysis using particle swarm optimization. *Journal of Physics: Conference Series, 1911*, 012011. https://doi.org/10.1088/1742-6596/1911/1/012011
- Li, C., Ma, Y., & Zhu, H. (2022). Evaluating football player market value using machine learning. *Sports Analytics, 15*(1), 113-127. https://doi.org/10.1016/j.sporan.2022.02.010
- FotMob. (n.d.). Live football scores, fixtures, stats, and news. Retrieved December 8, 2024, from https://www.fotmob.com

ANEXO 1: Rankings Completos en base a Score Promedio







ANEXO 2: Rankings Completos en base a Score Ajustado