



Universidad de
SanAndrés

UNIVERSIDAD DE SAN ANDRÉS
BIG DATA

Final Project

“Prediccion del valor de mercado de futbolistas con Machine Learning: un análisis basado en estadísticas de rendimiento”

GASPAR HAYDUK
MARTINA HAUSVIRTH
CAROLINA DE BOECK

FEBRERO 2025

Contents

1	Introducción y Motivación	2
2	Data	4
3	Metodología y Resultados	8
4	Conclusiones y trabajo futuro	13

1 Introducción y Motivación

El uso de Machine Learning (ML) en el fútbol ha experimentado un crecimiento significativo en los últimos años, con aplicaciones que van desde el análisis táctico hasta la gestión de rendimiento y la predicción de resultados. Una de las áreas de interés más recientes es la estimación del valor de mercado de los jugadores, un aspecto crucial en la economía del fútbol profesional.

Dentro del mercado del fútbol europeo, Transfermarkt se ha consolidado como la referencia principal para la valuación de jugadores. Este portal recopila información de traspasos, desempeño y opiniones de expertos para determinar valores de mercado estimados. Sin embargo, estos valores dependen de métodos que combinan juicio subjetivo y datos cuantitativos, lo que abre la puerta a la aplicación de técnicas de ML para realizar predicciones más objetivas y replicables.

El objetivo de este trabajo es desarrollar un modelo de predicción del valor de mercado de un jugador según Transfermarkt a partir de estadísticas de rendimiento deportivo. Esta aplicación de ML permitirá evaluar hasta qué punto es posible anticipar la valuación de un jugador a partir de su desempeño en el campo y explorar qué stats deportivas generan heterogeneidad en las valuaciones de mercado. Para ello, utilizaremos las técnicas de Ridge Regression, Tree Regression y Bagging. La regresión Ridge nos servirá como marco de referencia ya que establece una relación lineal entre los predictores y la variable objetivo, por lo que podemos interpretar la influencia de cada predictor. El árbol de regresión nos permitirá explorar relaciones no lineales y entender qué características (goles, asistencias, edad, etc.) influyen más en la valoración de un jugador. Por último, el método de ensamble Bagging nos permitirá reducir el sobreajuste que suelen tener los árboles de regresión y decidir si incluir no linealidades mejora la predicción en el testing set respecto de un modelo lineal a la Ridge.

El desarrollo de modelos predictivos basados en ML para estimar el valor de mercado de los jugadores tiene implicaciones importantes para distintos actores dentro de la industria del fútbol. Para los clubes y agentes, contar con estimaciones precisas del valor de mercado de un jugador puede facilitar la toma de decisiones en traspasos, negociaciones contractuales e inversiones estratégicas. Estos modelos permiten analizar el rendimiento histórico, métricas de desempeño y tendencias de mercado para identificar oportunidades de compra y evitar sobrevaloraciones.

Además, la capacidad de prever valores de mercado contribuye a una mayor eficiencia en el mercado de transferencias, ayudando a que los clubes asignen mejor sus recursos y negocien con mayor fundamento. Al mismo tiempo, los modelos de ML pueden desempeñar un papel clave en la evaluación de riesgos asociados a la adquisición de jugadores, permitiendo analizar factores como la edad, historial de lesiones y demanda de mercado para reducir incertidumbres financieras.

Otra aplicación relevante es la identificación de talentos con alto potencial de crecimiento. Mediante el análisis de atributos individuales y patrones históricos de evolución, los clubes pueden detectar jugadores con margen de mejora y ajustar sus estrategias de formación y desarrollo en academias juveniles. También, en el ámbito de las negociaciones, el uso de modelos predictivos aporta un enfoque basado en datos que ayuda a fundamentar posturas tanto de clubes como de agentes y jugadores, promoviendo negociaciones más equitativas y transparentes.

Finalmente, la capacidad de estimar valores de mercado de manera precisa puede beneficiar a industrias relacionadas, como los deportes de fantasía y las apuestas deporti-

vas, donde el conocimiento detallado del rendimiento de los jugadores y sus expectativas de mercado puede mejorar la experiencia de los participantes y su capacidad de tomar decisiones informadas.

Además, un aspecto clave en esta investigación es la heterogeneidad en las estadísticas deportivas. Es decir, identificar cuáles son los atributos más relevantes según la posición del jugador y cómo ciertos factores pueden afectar de manera diferencial a la valuación. Por ejemplo, mientras que las acciones defensivas pueden ser críticas para un central, la participación en goles y asistencias podría ser el principal determinante del valor de un delantero.

Este análisis contribuirá tanto al campo de la economía del deporte como a la aplicación de ML en la evaluación de activos en el mercado del fútbol, proporcionando una metodología replicable y basada en datos para estimar valores de mercado en el fútbol profesional.

Revisión de Literatura

La literatura existente en economía del deporte ha explorado diversos factores que influyen en el valor de mercado de los jugadores profesionales de fútbol. Esta revisión se centra en tres líneas principales de investigación: *el efecto combinado del talento y la popularidad*, el *impacto de habilidades específicas*, y la *aplicación de nuevas tecnologías en la valoración de jugadores*.

Franck y Nüesch (2012) analizan los determinantes del valor de mercado de los jugadores en la liga alemana, estudiando la interacción entre talento y popularidad como explicaciones del fenómeno de las superestrellas. Los autores argumentan que el talento, medido por indicadores de rendimiento como goles y asistencias, incrementa la probabilidad de éxito del equipo. Además, la popularidad, cuantificada a través de menciones en la prensa, agrega un componente no relacionado directamente con el desempeño, pero igualmente relevante para el mercado de jugadores.

Por otro lado, Bryson et al. (2014) estudian el efecto de habilidades escasas, como la bipedestría, en los salarios de jugadores en las principales ligas europeas. Utilizando datos de la Bundesliga y una muestra de las cinco principales ligas europeas, encuentran una prima salarial significativa para jugadores capaces de usar ambos pies, incluso al controlar por otras medidas de rendimiento. Esta habilidad escasa resulta particularmente valiosa para posiciones ofensivas como delanteros, quienes pueden aprovecharla para anotar goles desde ángulos variados.

Un enfoque distinto es el de Morgulev et al. (2018), quienes exploran cómo el análisis predictivo ha revolucionado el ámbito deportivo, permitiendo identificar patrones de rendimiento complejos en grandes bases de datos. Utilizando métodos de aprendizaje automático, los autores demuestran cómo métricas detalladas de jugadores pueden predecir su desempeño futuro, aportando herramientas para optimizar la toma de decisiones en fichajes y transferencias.

Adicionalmente, recientes estudios han complementado estas líneas de investigación con nuevos enfoques y metodologías. Cohen y Risk (2023) proponen un marco que combina principios de las matemáticas financieras y la teoría de redes para evaluar el valor de los futbolistas en Europa. Utilizan una "matriz de pases" para capturar las interacciones entre jugadores en el campo y aplican medidas de centralidad para cuantificar la influencia individual.

Lee, Tama y Cha (2022) investigan los factores clave que afectan las tarifas de transferencia de los futbolistas de élite mundial. Emplean un modelo LightGBM optimizado con el algoritmo Tree-structured Parzen Estimator (TPE) para predecir el valor de mercado de cada jugador. Además, identifican las características más influyentes utilizando el método SHAP (SHapley Additive exPlanations). Barrio (2020) realiza un análisis económico sobre el valor de mercado de los futbolistas profesionales y su influencia en los ingresos comerciales de los clubes europeos. Utilizando datos de Transfermarkt y el informe Deloitte Football Money League, establece una relación entre el valor de mercado de los equipos y sus ingresos por taquilla, derechos televisivos y marketing.

Finalmente, Gallegos, Orrala y Paredes (2019) estudian las variables que afectan el valor de mercado de los futbolistas en los principales equipos de la CONMEBOL, encontrando que factores como el número de goles, la cantidad de equipos previos, la valoración del club actual y la cantidad de seguidores en redes sociales son determinantes significativos.

Nuestro trabajo amplía esta literatura al combinar metodologías de aprendizaje automático con el análisis del valor de mercado de los jugadores de fútbol, proporcionando un enfoque basado en datos que complementa las estimaciones tradicionales de Transfermarkt. A diferencia de estudios previos que han explorado el impacto de habilidades específicas (Bryson et al., 2014) o la influencia de la popularidad en las valuaciones (Franck y Nüesch, 2012), nuestro enfoque se centra en la capacidad predictiva de diferentes técnicas de ML—Ridge Regression, Tree Regression y Bagging—para estimar el valor de mercado a partir de estadísticas de rendimiento. Además, al analizar la heterogeneidad en la importancia de distintos atributos según la posición del jugador, contribuimos a la comprensión de los factores que influyen diferencialmente en la valuación de defensores, mediocampistas y delanteros. Finalmente, este trabajo se inserta en la creciente aplicación de nuevas tecnologías en la economía del deporte (Morgulev et al., 2018; Lee, Tama y Cha, 2022) al demostrar cómo los modelos de ML pueden mejorar la transparencia y precisión en la evaluación de jugadores, ofreciendo una herramienta replicable para clubes, analistas y agentes en la toma de decisiones estratégicas dentro del mercado de transferencias.

2 Data

Para la construcción del modelo de predicción, utilizamos dos fuentes principales de información: Sofascore y Transfermarkt. Ambas plataformas son ampliamente reconocidas en la industria del fútbol por la calidad y cantidad de datos que proporcionan sobre el rendimiento y el mercado de jugadores.

Sofascore es una plataforma web de análisis estadístico en tiempo real que recopila datos detallados de cada partido jugado en las principales ligas del mundo. Entre las métricas disponibles se incluyen pases, regates, duelos ganados, tiros a puerta, intervenciones defensivas, entre muchas otras. Estas estadísticas permiten una evaluación granular del desempeño de cada jugador en distintos aspectos del juego y proporcionan una base cuantificable para analizar su impacto en el equipo. Para cada jugador que haya disputado algún partido en las cinco grandes ligas de Europa -LaLiga Española, SerieA Italiana, Ligue 1 Francesa, Bundesliga Alemana y Premier League Inglesa - la Figura 1 muestra las estadísticas deportivas de rendimiento individual que SofaScore reporta.

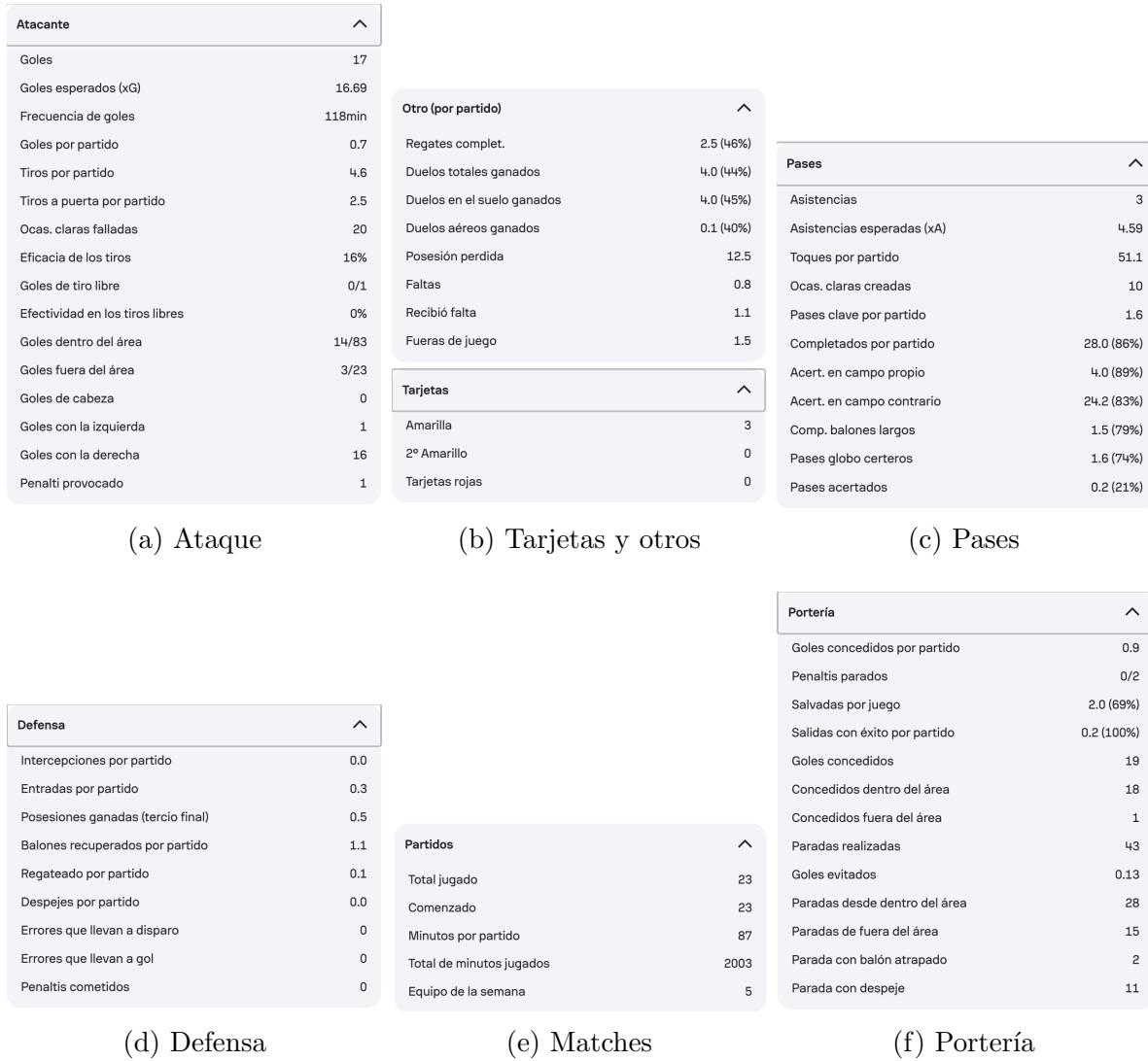


Figura 1: Estadísticas deportivas reportadas por SofaScore

Transfermarkt, por otro lado, es la referencia clave en la estimación del valor de mercado de los jugadores. El sitio web compila información sobre traspasos, contratos, edad, historial de rendimiento y factores subjetivos como la percepción del mercado y la demanda del jugador en un momento dado. Aunque estos valores no son generados exclusivamente mediante modelos analíticos, constituyen una referencia utilizada por clubes, agentes y analistas del fútbol. De este sitio web obtenemos el valor de mercado de cada jugador de las cinco grandes ligas de Europa.

Transfermarkt es particularmente influyente debido a su enfoque participativo, donde expertos y miembros de la comunidad contribuyen con información sobre el valor de los jugadores. Si bien esto introduce un componente subjetivo en la valoración, la plataforma ha demostrado ser una herramienta confiable al reflejar tendencias de mercado y el interés de los clubes por determinados jugadores. Además, sus datos son ampliamente utilizados por clubes y agentes como un punto de referencia en negociaciones de traspasos y revisiones salariales. Es importante destacar que el valor de mercado estimado por Transfermarkt no siempre coincide con las tarifas efectivamente pagadas en los traspasos de jugadores. Factores como la duración del contrato, cláusulas de rescisión, urgencia de los clubes compradores y vendedores, o incluso la existencia de múltiples ofertas en

competencia, pueden generar diferencias significativas entre la valuación de un jugador en la plataforma y el precio final de su fichaje. Además, situaciones específicas del mercado, como el interés de clubes con alto poder adquisitivo o restricciones financieras impuestas por regulaciones como el Fair Play Financiero, pueden influir en los montos pagados en transferencias sin que estos reflejen necesariamente el rendimiento deportivo del jugador.

La combinación de estos dos conjuntos de datos permite construir un modelo robusto que relacione el rendimiento deportivo de un jugador con su valor de mercado, evaluando la influencia de distintas métricas en su valoración económica. Además, el análisis de estas fuentes posibilita detectar patrones de heterogeneidad en la relación entre desempeño y valor de mercado, dependiendo de la posición en el campo y el contexto competitivo en el que se desempeñe el jugador.

Dataset

Para maximizar la ventana de partidos en las que se recolecta información acerca del rendimiento deportivo de los jugadores, el sitio web SofaScore fue scrapeado el 21 de febrero de 2025 y obtuvimos datos de 2328 jugadores. La data corresponde a 22 jornadas de la Ligue 1, 24 jornadas de LaLiga, 22 jornadas de la Bundesliga, 27 jornadas de la Premier League y 25 jornadas de la Serie A. En cuanto a Transfermarkt, el sitio web fue scrapeado el 15 de febrero y obtuvimos el valor de mercado de 2550 jugadores. Hay que tener en cuenta que en Transfermarkt puede haber jugadores dentro de un plantel que no tuvieron minutos en la liga doméstica. Por ejemplo, Claudio Echeverri figura como parte del plantel del Manchester City y Transfermarkt le asigna un valor de mercado pero todavía no disputó minutos en la Premier League; lo mismo ocurre con juveniles.

Al margen ambas bases de datos, obtenemos el valor de mercado y stats de rendimiento deportivo individual en la liga doméstica para 2286 jugadores. Por último, decidimos filtrar y considerar en el análisis a jugadores que hayan jugado al menos 7 partidos y nos quedamos con 1855 jugadores.

Para cada jugador, tenemos data de su valor de mercado reportado por Transfermarkt y los siguientes atributos: edad, el total de partidos que disputó en la liga doméstica, posición (trabajamos con dummies para cada posición), goles por partido, ocasiones claras creadas por partido, pases claves por partido, minutos por partido, asistencias por partido, pases completados por partido, pases acertados en campo rival, intercepciones por partido, entradas por partido, posesiones ganadas en el tercio final por partido, balones recuperados por partido, despejes por partido, regates completados por partido, duelos totales ganados por partido, salvadas por partido (0 para defensores, mediocampistas y delanteros), y liga (trabajamos con una dummy para cada liga).

La Tabla 1 muestra el total de jugadores según liga y posición, mientras que la Tabla 2 muestra el valor de mercado promedio según posición; como podemos ver, el puesto de delantero es en general el mejor valuado.

Tabla 2: Valor de Mercado según posición

Posición	Valor de Mercado Promedio (en millones de €)
Defensor	11.90
Delantero	17.10
Arquero	9.81
Mediocampista	15.66

Tabla 1: Cantidad de Jugadores por Posición y Liga

		Cantidad
Liga	Bundesliga	353
	LaLiga	398
	Ligue 1	312
	Premier League	389
	Serie A	403
Posición	Defensores	607
	Delanteros	377
	Arqueros	119
	Mediocampistas	752

La Figura 2 muestra la distribución del valor de mercado. Podemos observar una distribución altamente asimétrica (sesgada a la derecha). Esto significa que existen pocos jugadores con valores extremadamente altos (100M+ €), lo que crea una larga cola derecha. Para facilitar el modelado, decidimos trabajar con el valor de mercado en escala logarítmica. La Figura 3 nos muestra la distribución en dicha escala. En este caso, la distribución se vuelve mucho más simétrica y cercana a una normal.

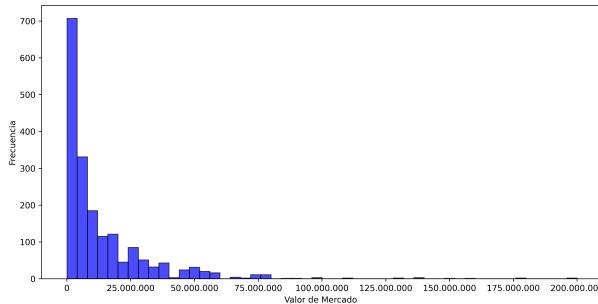


Figura 2: Distribución del valor de mercado (escala original en millones de euros)

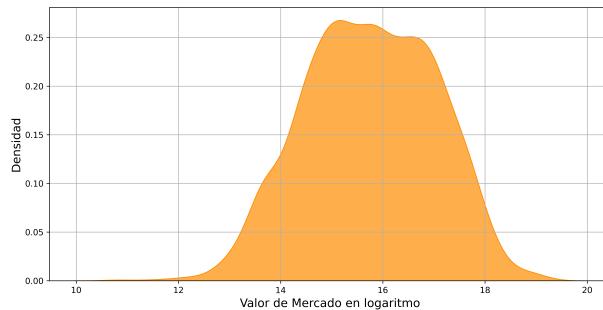


Figura 3: Distribución del valor de mercado en logaritmo

3 Metodología y Resultados

Para garantizar la validez y generalización del modelo, dividimos el conjunto de datos en dos subconjuntos; un training set para entrenar modelos y aprender patrones en los datos y otro testing set para evaluar el rendimiento de los modelos sobre datos no vistos, usando el Error Cuadrático Medio como métrica de evaluación. Destinamos el 75% (1391 observaciones) de la data a entrenar modelos y un 25% (464 observaciones) a evaluar a los mismos modelos.

Además, para optimizar los hiperparámetros de los modelos de Ridge Regression y Tree Regression (árboles de decisión), utilizaremos validación cruzada con K-Fold. Esta técnica permite dividir el conjunto de entrenamiento en K subconjuntos, entrenando el modelo varias veces con combinaciones diferentes de datos de entrenamiento y validación. Esto garantiza que el modelo no dependa excesivamente de una única partición de los datos y ayuda a seleccionar los valores óptimos de los hiperparámetros, mejorando la estabilidad y la capacidad predictiva del modelo.

Ridge Regression

Recordemos que, para un λ dado, la función a minimizar en Ridge es:

$$R_r(\beta) = \sum_{i=1}^n (y_i - x'_i \beta)^2 + \lambda \sum_{s=2}^p (\beta_s)^2 \quad (1)$$

El parámetro λ en Ridge Regression controla la penalización aplicada a los coeficientes del modelo. Su función principal es regularizar el modelo para evitar el sobreajuste. Como es práctica habitual en Ridge, estandarizamos los atributos x_i ya que si los mismos están escalas diferentes, los coeficientes también estarán en escalas distintas, lo que afectará la penalización de manera desigual.

Para determinar el valor óptimo del hiperparámetro λ en Ridge Regression, utilizamos un enfoque de validación cruzada con 10 folds (10-Fold Cross Validation) sobre una búsqueda logarítmica de 50 valores en el rango

$$\lambda \in \{10^{-4}, 10^{-3.83}, 10^{-3.67}, \dots, 10^{3.67}, 10^{3.83}, 10^4\}$$

La validación cruzada consiste en dividir el conjunto de entrenamiento en 10 subconjuntos (folds). En cada iteración, el modelo se entrena utilizando 9 folds y se evalúa en el fold restante, repitiendo este proceso 10 veces. Luego, se calcula el error cuadrático medio en todas las iteraciones para cada valor de λ seleccionando el que minimiza dicho error. El uso de una escala logarítmica en la exploración de λ permite cubrir eficientemente valores muy pequeños y muy grandes sin concentrarse excesivamente en un solo orden de magnitud. Este procedimiento garantiza que el modelo tenga una regularización óptima, reduciendo el riesgo de sobreajuste o subajuste, y mejorando su capacidad de generalización.

El proceso de validación cruzada selecciona un $\lambda = 24.4$. Con este hiperparametro, la Tabla 3 muestra los coeficientes estimados en la regularización Ridge. En dicha tabla, podemos observar que los coeficientes asociados a stats de ataque (goles por partido, Ocasioness generadas por partido, pases acertados en campo rival) son todos positivos, indicando que un mejor desempeño ofensivo aumenta el valor de mercado del jugador. Los coeficientes asociados a stats defensivas, como *intercepciones por partido* y *despejes*

por partido, son negativos; posiblemente porque los defensores suelen ser menos valorados en el mercado en comparación con mediocampistas o delanteros. Por otra parte, el coeficiente asociado a la Premier League es positivo, lo que sugiere que jugar en esta liga se asocia con valores de mercado más altos.

Tabla 3: Coeficientes Ridge

Variable	Coeficiente
Edad	-0.4617
Total jugado	0.1811
Minutos por partido	-0.0713
Pases clave por partido	-0.0199
Pases complet. por partido	0.4072
Pases acert. campo rival	0.2460
Intercepciones por partido	-0.0385
Entradas por partido	-0.1309
Pos. ganadas (tercio final)	0.0388
Balones recup. por partido	-0.0383
Despejes por partido	-0.0804
Regates complet.	0.0103
Duelos totales ganados	0.0678
Salvadas por juego	-0.0986
Goles por partido	0.2726
Asists por partido	0.0526
Ocas. creadas por partido	0.1088
Bundesliga	-0.0893
LaLiga	-0.0853
Ligue 1	-0.1369
Premier League	0.3500
Serie A	-0.0562
Defensor	-0.0405
Delantero	-0.0429
Arquero	0.2600
Mediocampista	-0.0532

Para evaluar la performance predictiva de Ridge, queda predecir el valor de mercado para el testing set y computar el Error Cuadrático medio: $MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$. La predicción arroja un MSE de 0.59.

Tree Regression

El uso de Tree Regression como alternativa a modelos lineales como Ridge Regression ofrece varias ventajas en la predicción del valor de mercado de los jugadores. A diferencia de los modelos lineales, los árboles de decisión permiten capturar relaciones no lineales y efectos de interacción entre los predictores sin necesidad de especificar una forma funcional previa. A diferencia de Ridge Regression, donde la regularización penaliza los coeficientes en función de su magnitud, los árboles de decisión no dependen de la escala de las variables. Esto se debe a que los árboles toman decisiones basadas en divisiones (splits)

en los datos, en lugar de optimizar una función de pérdida cuadrática. En un árbol de decisión, cada nodo divide el conjunto de datos con base en valores específicos de las variables, por lo que las unidades o rangos de los predictores no afectan la estructura del árbol. Esto hace que Tree Regression sea más flexible y fácil de interpretar, ya que no requiere transformación previa de los datos.

El hiperparámetro máxima profundidad controla la profundidad máxima del árbol, es decir, cuántas divisiones puede hacer el modelo antes de detenerse. Si la máxima profundidad es muy grande, el árbol puede memorizar los datos de entrenamiento, llevando a sobreajuste (overfitting). Si la máxima profundidad es muy pequeña, el modelo puede no capturar suficiente información relevante, generando subajuste (underfitting). Por esta razón, se realiza una búsqueda de hiperparámetros para determinar el valor óptimo de la máxima profundidad que minimiza el error en validación. Nosotros realizamos una búsqueda de máxima profundidad con la siguiente barrida $\{1, 2, 3 \dots, 24, 25, 26\}$. La validación cruzada con 10 folds consiste en dividir el conjunto de entrenamiento en 10 subconjuntos (folds), entrenar el modelo en 9 de ellos y evaluar su desempeño en el fold restante. Este proceso se repite 10 veces, alternando el fold de validación, y se promedian los errores para cada valor del hiperparámetro. Tras evaluar distintos valores de profundidad con validación cruzada, encontramos que la profundidad óptima del árbol es 5. La Figura 5 nos muestra el diagrama del árbol de decisión, cómo predice el valor de mercado de un jugador dados sus atributos.

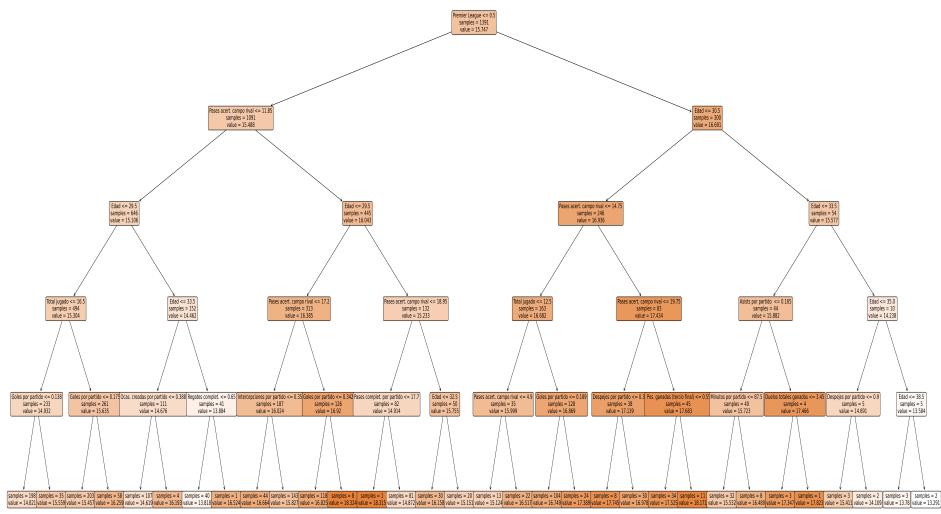


Figura 4: Diagrama del Árbol de Decisión

Por otra parte, el árbol de decisión como regresor nos permite calcular la importancia de las variables (Feature Importance). Esto nos ayuda a entender cuáles predictores tienen mayor influencia en la predicción del valor de mercado de los jugadores. En un árbol de regresión, la importancia de cada variable se calcula en función de cuánto reduce la impureza en cada división del árbol. La impureza en regresión se mide con el Error

Cuadrático Medio (MSE). Cada vez que una variable es seleccionada para dividir un nodo, contribuye a reducir la impureza total del árbol. La importancia de una variable se calcula sumando todas las reducciones de impureza en los nodos donde esa variable fue utilizada. La importancia de cada variable X_j se obtiene como:

$$\text{Importancia}(X_j) = \frac{\sum_{t \in S_j} \Delta \text{MSE}_t}{\sum_{t \in S} \Delta \text{MSE}_t}$$

donde ΔMSE es la reducción del Error Cuadrático Medio cuando la variable X_j es usada para dividir un nodo, S_j es el conjunto de nodos donde la variable X_j fue utilizada para realizar una división en el árbol y S es el conjunto de todos los nodos del árbol. La Figura 6 muestra la importancia de las variables en el modelo de Tree Regression, indicando cuáles son los predictores más influyentes en la predicción del valor de mercado de los jugadores. Pases acertados en campo rival es la variable más importante, lo que sugiere que los jugadores con mayor precisión en pases ofensivos son altamente valorados; la Edad tiene una gran importancia, confirmando que los jugadores más jóvenes suelen tener mayor valor de mercado. Jugar en la Premier League también tiene un impacto fuerte en la valuación, lo que refleja que esta liga eleva el precio de los jugadores. Los goles por partido y el total de partidos jugados también son factores clave en la valoración.

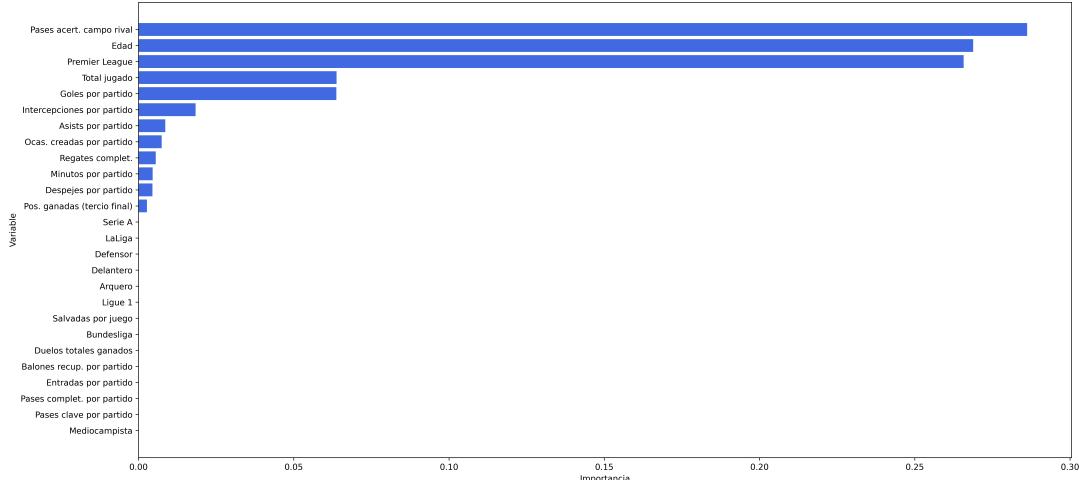


Figura 5: Importancia de cada atributo

Finalmente, evaluamos la performance predictora de este árbol de decisión sobre el testing set. El resultado es un MSE de 0.75; es decir, tenemos una peor performance que Ridge.

Bagging

Bagging es una técnica de aprendizaje de conjunto (ensemble learning) que mejora la estabilidad y precisión de los modelos de predicción al reducir la varianza. Su funcionamiento se basa en entrenar múltiples modelos sobre subconjuntos de datos generados mediante

bootstrap (muestreo con reemplazo) y luego promediar sus predicciones. Se crean B subconjuntos aleatorios a partir del conjunto de entrenamiento original, seleccionando observaciones con reemplazo (bootstrap); se entrena un árbol de decisión independiente en cada subconjunto; y la predicción final es el promedio de las B predicciones. Al promediar B árboles estamos mitigando el sobreajuste de cada árbol a los datos con los que dicho árbol fue entrenado.

Al hacer Bagging con 200 árboles de decisión entrenados con subconjuntos del training set generados mediante bootstrap, obtenemos un MSE de 0.62 cuando predecimos sobre el testing set.

Análisis de los Errores de Predicción

El Error de Predicción del modelo m para el jugador i se define como:

$$Err_i^m = y_i - \hat{y}_i^m, \quad (2)$$

donde y_i es el valor de mercado (en logs) del jugador i y \hat{y}_i^m es el valor de mercado (en logs) predicho para el jugador i según el algoritmo $m \in \{\text{Ridge, Tree, Bagging}\}$.

Dado que estamos trabajando en logaritmos, convertimos el error de pronóstico como porcentaje del valor de mercado como:

$$ErrRelativo_i^m = e^{Err_i^m} - 1 \quad (3)$$

Así tenemos una medida del porcentaje en que el modelo m subestima o sobreestima el valor de mercado en logs para el jugador j .

La Figura 6 muestra el error relativo de los modelos Ridge, CART y Bagging en función del valor de mercado real (en logaritmo). Cada punto representa una observación del testing set; el eje y indica cuánto subestima o sobreestima cada modelo el valor de mercado de un jugador. Podemos observar que a medida que el valor de mercado real aumenta, el error relativo crece significativamente. Esto indica que los modelos tienden a subestimar a los jugadores más caros, ya que muchos errores están por encima de la línea horizontal (cero). Los modelos tienen mejor desempeño en jugadores promedio y que la dificultad principal está en predecir a los jugadores más valiosos. Esta mala predicción en los jugadores más valiosos puede deberse a la falta de variables que reflejen rendimiento en torneos más importantes que las ligas domésticas como la Champions League, Eurocopa, Copa América o Mundial de Fútbol; otra interpretación posible es que las valoraciones de jugadores como Mbappé, Vinicius Júnior y Lamine Yamal (160 millones de €, 200 millones de £ y 180 millones de €, respectivamente) reflejen la intransferibilidad del jugador, son jugadores que no están disponible en el mercado y por ello una valoración estratosférica es una forma de decir que no tienen precio.

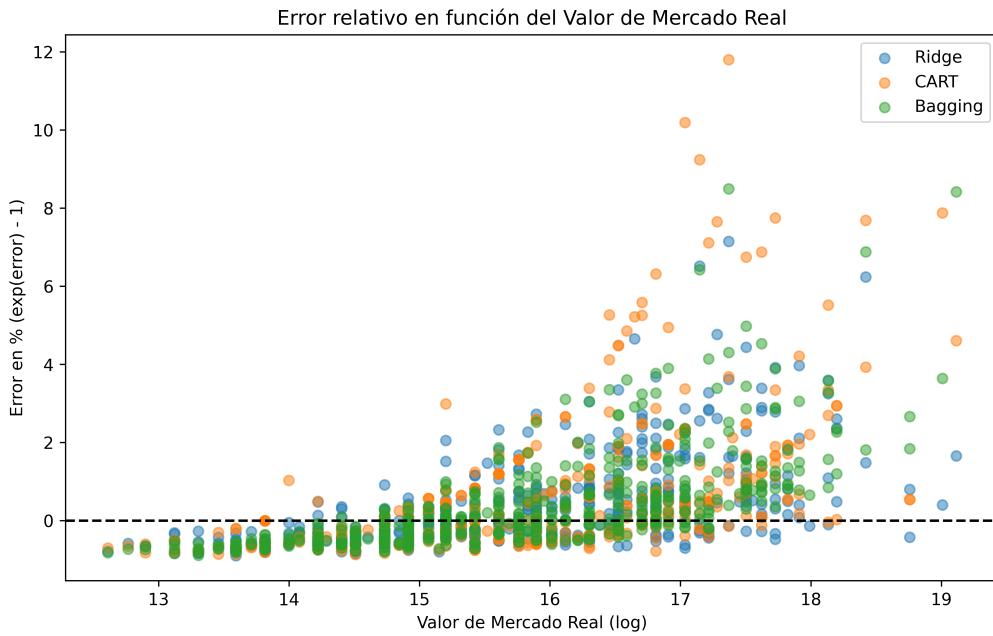


Figura 6: Error Relativo de Predicción

4 Conclusiones y trabajo futuro

Este estudio ha demostrado cómo las técnicas de Machine Learning pueden ser utilizadas para estimar el valor de mercado de los futbolistas a partir de sus estadísticas de rendimiento en las cinco principales ligas europeas. A través de la aplicación de modelos de regresión Ridge, árboles de decisión y Bagging, se logró identificar qué factores son determinantes en la valuación de los jugadores y comparar el desempeño predictivo de distintos enfoques.

Uno de los hallazgos más relevantes es que la regresión Ridge mostró el mejor desempeño en términos de error cuadrático medio, lo cual sugiere que la mejor forma de predecir el valor de mercado de un futbolista a partir de su rendimiento deportivo es mediante una estructura lineal. Sin embargo, el análisis mediante árboles de decisión permitió capturar relaciones no lineales y resaltar la importancia de predictores como los pases acertados en campo rival, la edad, jugar en la Premier League y el total de partidos jugados.

A pesar del buen desempeño predictivo de los modelos, se observó una tendencia sistemática a subestimar el valor de los jugadores más caros. Esto podría deberse a la falta de información sobre el rendimiento en torneos de mayor prestigio, como la Champions League, la Eurocopa, la Copa América o el Mundial de Fútbol. Dado que estos torneos representan el más alto nivel de competencia y suelen influir en la percepción del valor de los jugadores, su incorporación podría mejorar significativamente la capacidad predictiva del modelo, especialmente en la estimación de los futbolistas con mayores valoraciones.

De cara al futuro, una posible mejora en la metodología consistiría en integrar datos de rendimiento en estas competiciones internacionales. Esto permitiría evaluar si el desempeño en contextos de máxima exigencia tiene un peso diferencial en la valorización de los jugadores. Además, el uso de datos longitudinales sobre la evolución del valor

de mercado a lo largo del tiempo podría ayudar a modelar mejor la trayectoria de los futbolistas y predecir cambios en su valuación.

En conclusión, este trabajo ha proporcionado un marco metodológico sólido para la estimación del valor de mercado de los jugadores utilizando técnicas de Machine Learning. Aunque existen desafíos y limitaciones, la incorporación de información adicional sobre el rendimiento en torneos de élite podría contribuir a una mejor comprensión de los factores que determinan la valuación de los futbolistas en el mercado actual.

References

- [1] Müller, O., Simons, A., Weinmann, M. (2017). "Beyond crowd judgments: Data-driven estimation of market value in association football." *European Journal of Operational Research*, 263(2), 611-624.
- [2] Ribeiro, H. V., Mukherjee, S., Zeng, X. H. T. (2020). "Talent vs luck: The role of randomness in success and failure." *Advances in Complex Systems*, 23(2), 2050004.
- [3] García, F., Gutiérrez, F., Fernández, C. (2021). "Predicting football players' market value using machine learning techniques." *Journal of Sports Analytics*, 7(3), 153-167.
- [4] Carmichael, F., Thomas, D., Ward, R. (2001). "Production and efficiency in association football." *Journal of Sports Economics*, 2(3), 228-243.
- [5] Wager, S., Athey, S. (2018). "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests." *Journal of the American Statistical Association*, 113(523), 1228-1242.
- [6] Murphy, K. P. (2022). *Probabilistic Machine Learning: An Introduction*. MIT Press.
- [7] Varian, H. R. (2014). "Big data: New tricks for econometrics." *The Journal of Economic Perspectives*, 28(2), 3-27.
- [8] Sosa Escudero, W. (2019). *Big Data*. 9a edición, Siglo XXI Editores, Buenos Aires.
- [9] Breiman, L. (2001). "Random forests." *Machine Learning*, 45(1), 5-32
- [10] Transfermarkt. (2024). "Player market values and transfer history." Retrieved from www.transfermarkt.com.
- [11] Sofascore. (2024). "Football statistics and live scores." Retrieved from www.sofascore.com.