



Pontificia Universidad Católica de Valparaíso
Instituto de Estadística
Magíster en Estadística

Informe Técnico

Análisis predictivo mediante métodos de Machine Learning

Nombre: Gabriel Ávila C.

Asignatura: EST852-1 Machine Learning

Profesor: Alejandro Veloz

Resumen ejecutivo

El análisis de datos ha adquirido un rol fundamental en el fútbol moderno. Clubes profesionales, medios especializados y casas de apuestas utilizan modelos predictivos para evaluar rendimientos y anticipar resultados. En este contexto, el presente trabajo constituye un primer acercamiento al uso de técnicas de Machine Learning aplicadas al fútbol, evaluando la pertinencia y capacidad predictiva de dos métodos ampliamente utilizados: la regresión lineal con regularización LASSO y el modelo de clasificación Random Forest.

Para ello se emplea una base de datos pública compuesta por 9.380 partidos de la liga inglesa entre 2000/01 y 2024/25, con 22 variables originales. A partir de esta información, se construyeron nuevos atributos con el objetivo de enriquecer el set de predictores y capturar dinámicas temporales relevantes para el modelamiento.

El primer enfoque utiliza LASSO para predecir la diferencia de goles, un objetivo particularmente desafiante debido a que la mayoría de los encuentros se define por márgenes estrechos y la serie presenta una alta incidencia de valores atípicos. El segundo enfoque aplica Random Forest para clasificar si el equipo local gana, buscando complementar la aproximación regresiva y evaluar el potencial de un modelo no lineal basado en árboles.

Los resultados evidencian un desempeño predictivo limitado en ambos casos. El modelo LASSO trabajando con 18 variables obtuvo un MAE de 1,273 y un R^2 de 0.20. Por su lado, el modelo de clasificación alcanzó una precisión (accuracy) de 63,3% utilizando 29 variables relevantes, mostrando una fuerte inclinación a predecir no triunfos del equipo local aunque con una sobreinterpretación sobre estos.

En conjunto, estos hallazgos sugieren, que la predicción de resultados futbolísticos presenta restricciones inherentes, tanto por la naturaleza altamente aleatoria del deporte, así como también por el nivel de granularidad y la calidad de las variables disponibles. Aun utilizando técnicas robustas, la capacidad predictiva es acotada, lo que destaca la complejidad del fenómeno y la necesidad de datos de mayor detalle para mejorar los análisis y resultados futuros.

Introducción

El análisis de datos se ha convertido en una herramienta central para la toma de decisiones en múltiples áreas, incluyendo el deporte. En el fútbol, su incorporación ha permitido optimizar el rendimiento y apoyar la planificación táctica. Un ejemplo de ello es el reportaje de The New York Times (2019)¹ sobre el Liverpool FC., donde se describe cómo el análisis cuantitativo pasó a integrarse en el trabajo técnico del club. Asimismo, trabajos de estadística también plantea interés, por ejemplo, respecto de propiedades de la distribución del número de goles anotados en un partido (Baio & Blangiardo, 2010).²

Paralelamente, el crecimiento del negocio deportivo ha impulsado aún más el uso de modelos predictivos, especialmente en la industria de las apuestas, que capitaliza la disponibilidad de grandes volúmenes de datos para estimar probabilidades y orientar decisiones de los usuarios.

En este contexto surgen las preguntas que guían este informe: ¿es posible predecir el resultado de un partido?, ¿qué tan factible es estimar la victoria del equipo local a partir de variables históricas?, y ¿cuál es el aporte real de las técnicas predictivas tanto para clubes como para la industria asociada?

Para abordarlas, se analiza una base de datos de Kaggle (original de football-data.co.uk) correspondiente a la Premier League 2020/21-2024/25, que incluye información sobre goles, tiros, corners, tarjetas y otras variables del desempeño de los equipos durante el encuentro.

Con estos datos se implementan dos modelos: regresión LASSO y Random Forest. El primero busca predecir la diferencia de goles; el segundo, la probabilidad de triunfo del equipo local. Finalmente, se comparan ambos métodos según su capacidad predictiva y según los criterios de selección o importancia de variables que ofrece cada enfoque incluso cuando los objetivos de predicción no son idénticos.

1. Fundamentos de los modelos utilizados

El informe considera dos enfoques propios del Machine Learning (ML): un modelo de regresión lineal regularizado mediante LASSO y un modelo de clasificación basado en Random Forest (RF). Ambos corresponden a métodos supervisados, aunque difieren en su estructura y en el tipo de variable objetivo. A continuación, se presentan de forma resumida los fundamentos teóricos de cada modelo.

1.1. Regresión lineal y regularización LASSO

La regresión lineal es uno de los modelos predictivos fundamentales y constituye la base de métodos más complejos. Su objetivo es aproximar el valor esperado de una variable respuesta continua a partir de un conjunto de predictores. Bajo el supuesto de linealidad, el modelo general se expresa como:

$$\hat{y}_i = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \dots + \theta_p x_{ip}, \quad (1)$$

donde θ_0 es el intercepto, θ_j los coeficientes asociados a cada predictor y p el número de variables explicativas.

¹<https://www.nytimes.com/es/2019/05/29/espanol/liverpool-champions.html>.

²<https://www.tandfonline.com/doi/abs/10.1080/02664760802684177>.

La estimación tradicional mediante mínimos cuadrados ordinarios (OLS) minimiza la suma de errores cuadráticos, pero puede volverse inestable en presencia de multicolinealidad, un gran número de predictores o riesgo de sobreajuste. Para abordar estos problemas se emplean técnicas de regularización. Una de las más utilizadas es LASSO (*Least Absolute Shrinkage and Selection Operator*), que incorpora una penalización basada en la norma L_1 , resolviendo:

$$\min_{\theta} \left\{ \frac{1}{2n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^p |\theta_j| \right\}, \quad (2)$$

La penalización L1 tiende a contraer algunos coeficientes a cero (valores mayores generan una mayor contracción de los coeficientes), permitiendo actuar simultáneamente como método de regularización y selección de variables. Esto resulta especialmente útil en contextos con numerosos predictores potenciales —como el deportivo— donde solo algunas características contribuyen de forma relevante a explicar la diferencia de goles.

1.2. Árboles de decisión o Random Forest

Random Forest es un método de ensamble que combina múltiples árboles de decisión para mejorar la capacidad predictiva. Cada árbol se construye a partir de una muestra Bootstrap del conjunto de datos y, en cada nodo, se evalúa solo un subconjunto aleatorio de predictores, lo que introduce variabilidad controlada y reduce la correlación entre árboles.

En problemas de clasificación, la predicción final se obtiene mediante votación mayoritaria. Entre sus principales ventajas destacan: (i) robustez frente al sobreajuste, (ii) capacidad para capturar relaciones no lineales e interacciones, y (iii) disponibilidad de medidas de importancia de variables basadas en reducción de impureza o permutaciones.

2. Descripción del dataset: Aspectos generales

La base de datos contiene **9.380 observaciones** y **22 variables**. No presenta valores perdidos (“NaN”) ni registros duplicados. En cuanto a su composición, incluye 16 variables numéricas (enteras), 5 categóricas (tipo string) y 1 variable de fecha. El detalle de cada una se presenta en el **Cuadro 1**.

Cuadro 1: Resumen del dataset original

Característica	Nro
Número de observaciones	9.380
Número de variables originales	22
Variables numéricas	16
VARIABLES CATEGÓRICAS	5
VARIABLES DE FECHA	1
VARIABLES SELECCIONADAS PARA EL MODELO	18

En cuanto al número de partidos, cada temporada incluye 380 encuentros, correspondientes a las dos rondas disputadas por los 20 equipos participantes. Sin embargo, se identificó que las temporadas 2003 y 2004 registran únicamente 335 partidos. Dado que no se cuenta con información adicional que esclarezca esta discrepancia —presumiblemente asociada a registros

incompletos o inconsistencias en la fuente— dichas temporadas igualmente se incluirán en el análisis.

3.2. Creación de nuevas variables para el modelo

Dado que el objetivo es predecir el resultado antes de que el partido ocurra, es necesario construir variables basadas exclusivamente en información previa. A partir de las variables originales se generaron nuevos atributos que resumen el rendimiento reciente y acumulado de cada equipo. Este proceso de *feature engineering* permite incorporar información dinámica sobre la forma previa, un enfoque ampliamente utilizado en modelos predictivos aplicados al análisis del fútbol.

Temporadas

A partir de la fecha del partido (`matchdate`) se ajusta la variable `season` permitiendo identificar y ordenar las temporadas coherentemente. Esta corrección evita mezclar encuentros de campañas distintas y previene inconsistencias en la construcción de atributos con dependencia temporal que podrían distorsionarse por pausas entre temporadas.

- `season`: variable ordinal que identifica la temporada a la que pertenece cada partido.

Puntos obtenidos por partido

Estas variables son un puente para la creación de otras que miden el rendimiento tanto del equipo local como el visitante. La construcción de esta sigue el patrón reglamentario: partido ganado: 3ptos, empate: 1pto y derrota 0.

- `home/away_points`: contabiliza los puntos obtenidos en base al resultado del equipo local y visitante respectivamente.

Puntos obtenidos en los últimos cinco partidos

Para cada equipo y condición (local o visita) se calcularon los puntos obtenidos en sus cinco partidos previos, con el fin de capturar su rendimiento reciente antes del encuentro a predecir. Todas las series fueron desplazadas mediante un *lag* de una unidad, evitando el uso de información del propio partido.

- `home/away_form_last5`: promedio de puntos obtenidos por el equipo local/visitante en sus últimos cinco partidos como local/visitante.

Rendimiento ofensivo y defensivo reciente

Para caracterizar el comportamiento reciente de los equipos, se calcularon los promedios de goles anotados y recibidos en los cinco encuentros previos, permitiendo capturar la mayor incertidumbre propia del inicio de temporada:

- `home/away_attack_last5`: promedio de goles anotados por el equipo local y visitante, respectivamente.
- `home/away_defense_last5`: promedio de goles recibidos por el equipo local y visitante, respectivamente.

Puntos acumulados previo al partido (trayectoria de la temporada)

Para capturar el desempeño acumulado de cada equipo dentro de una temporada, se transformó la base de datos a un formato largo (*long format*) en el que cada registro correspondía a un equipo en un partido, independiente de si actuaba como local o visitante. A partir de esta estructura se calculó la suma acumulada de puntos obtenidos por cada equipo, aplicando un desplazamiento temporal (*lag*) para evitar la incorporación de información del propio encuentro.

- `home/away_points_before`: puntos acumulados por cada equipo local/visitante antes de disputar el partido.

Eficacia de cara a portería

Con el objetivo de sintetizar el rendimiento ofensivo previo, se construyeron indicadores de eficacia basados en la proporción entre goles anotados y tiros al arco en partidos anteriores. Se consideraron dos ventanas temporales —3 y 10 encuentros— con el fin de capturar tanto la forma reciente como una tendencia de mayor estabilidad.

- `eff_on_target_home/away_last3`: eficacia ofensiva en los últimos tres partidos del equipo local y visitante respectivamente.
- `eff_on_target_hom/away_last10`: eficacia ofensiva en los últimos diez partidos del equipo local y visitante respectivamente.

Elo de rendimiento histórico entre equipos

Con el objetivo de capturar la fortaleza relativa de cada equipo a lo largo del tiempo, se incorporó un indicador basado en el sistema Elo, ampliamente utilizado en deportes y modelos predictivos futbolísticos. Este sistema actualiza la “calificación” de un equipo según su desempeño acumulado, asignando un puntaje más alto a aquellos que obtienen resultados consistentes por encima de lo esperado.

En este trabajo, el Elo se construyó a partir del historial de enfrentamientos directos entre los equipos, considerando tanto victorias como diferencias de goles. De esta manera, se incorpora un componente de rendimiento estructural que complementa las métricas recientes basadas en forma o promedios móviles.

- `elo_home/away_before`: calificación Elo del equipo local/visitante antes del partido.
- `elo_diff_before`: diferencia entre ambos valores, que resume la brecha de fortaleza entre los equipos previo al encuentro.

Resumen de variables construidas

El proceso de *feature engineering* generó un conjunto amplio de nuevas variables orientadas a capturar distintos aspectos del rendimiento reciente y acumulado de los equipos. Muchas de las variables que se muestran (ejemplo las que considerando los últimos 5 partidos) se incluyeron luego de un pre proceso que considero distintos números de partidos, siendo la mejor opción. Asimismo, cabe destacar que en el caso de datos acumulados en las ultimas observaciones, para evitar valores nulos, se consideran partidos de la temporada pasada. En síntesis, las variables construidas fueron:

Cuadro 2: Resumen de variables construidas para el análisis

Categoría	Variables
Temporada y puntos	season; home_points, away_points; home_points_before, away_points_before.
Forma reciente	home_form_last5, away_form_last5.
Rendimiento ofensivo y defensivo	home_attack_last5, away_attack_last5; home_defense_last5, away_defense_last5.
Precisión y eficacia ofensiva	shot_acc_home, shot_acc_away; eff_on_target_home_last3, eff_on_target_away_last3; eff_on_target_home_last10, eff_on_target_away_last10.
Elo de rendimiento histórico	elo_home_before, elo_away_before, elo_diff_before

Con el fin de ofrecer una visión clara del dataset final utilizado para el análisis, se resume a continuación (**Cuadro 3**) su estructura y composición tras el proceso de *feature engineering*.

Cuadro 3: Resumen estructural del dataset

Característica	Cantidad	Porcentaje
Número total de variables	50	100 %
Variables numéricas (int64, int32, float64)	43	86 %
Variables categóricas (object)	6	12 %
Variables tipo fecha	1	2 %

Nota: Se incluyen las variables target creadas.

3. Análisis descriptivo

Este análisis busca proporcionar una visión general del comportamiento de las variables que serán utilizadas en los modelos predictivos. El objetivo es caracterizar la distribución de los atributos relevantes y evaluar la presencia de asimetrías, variabilidad y patrones asociados en las variables utilizadas en los modelos.

3.1. Descripción general del dataset

El conjunto de datos utilizado reúne información detallada de partidos de la liga inglesa de fútbol, incorporando estadísticas de rendimiento ofensivo y defensivo, además de variables fácticas asociadas a cada encuentro. En conjunto, estos elementos proporcionan un contexto adecuado para la construcción del modelo predictivo. El **Cuadro 4** presenta un resumen de las principales características estructurales del dataset.

Cuadro 4: Resumen estructural del dataset

Característica	Cantidad	Porcentaje
Número total de variables	50	100 %
Variables originales	22	44 %
Variables derivadas (feature engineering)	28	56 %
Variables utilizadas en los modelos	18	36 %
Variables excluidas	32	64 %

Adicionalmente, se generaron boxplots para cada variable con el objetivo de analizar su distribución y detectar la presencia de datos atípicos. Para facilitar la comparación entre variables con escalas distintas, todas ellas fueron previamente estandarizadas, permitiendo así una visualización más coherente y homogénea.

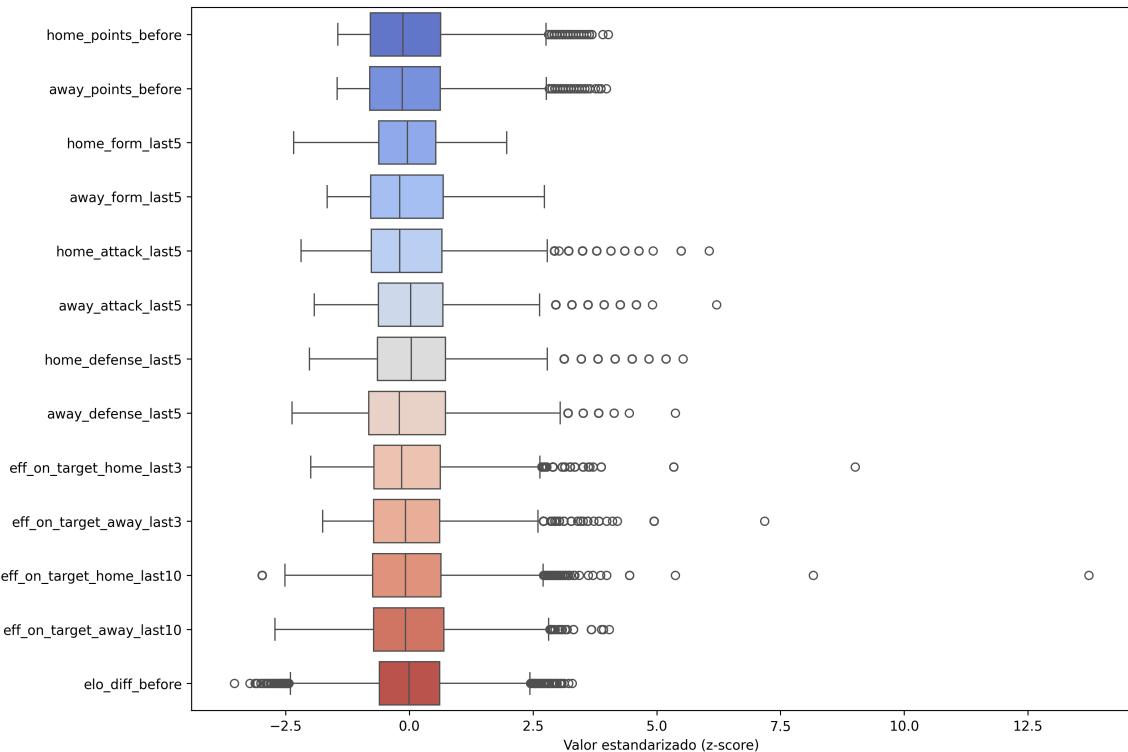


Figura 1: Distribución de las variables utilizadas en los modelos.

El análisis exploratorio revela una alta concentración de valores atípicos en variables asociadas a rendimiento acumulado y eficiencia ofensiva. Estos outliers reflejan, en términos prácticos, situaciones de equipos con rendimiento muy superior o inferior al promedio (por ejemplo, clubes dominantes en la parte alta de la tabla, rachas goleadoras o marcadas ineficiencias en la definición), más que errores de registro.

En consecuencia, los valores extremos se interpretan como expresión de la heterogeneidad propia de la competición y de la naturaleza de las rachas del rendimiento futbolístico, lo cual plantea un desafío adicional para los modelos lineales y refuerza la conveniencia de contrastarlos con métodos no lineales como Random Forest.

3.2. Distribución de las variables objetivo (target)

Previo al modelamiento es fundamental examinar el comportamiento de las variables objetivo. La **Figura 2** muestra la distribución de la diferencia de goles (`goal_diff`) y de la variable binaria que indica si el equipo local ganó el encuentro (`home_win_bin`).

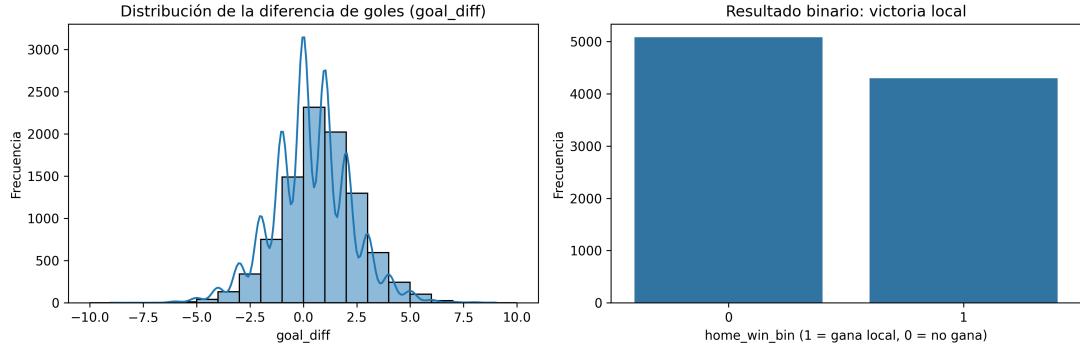


Figura 2: Distribución de las variables objetivo: diferencia de goles y victoria del equipo local.

La distribución de `goal_diff` es aproximadamente simétrica y se concentra en torno al cero, lo que refleja que la mayoría de los partidos se decide por diferencias estrechas. Solo una fracción reducida de encuentros presenta resultados extremos. Esta estructura, típicamente asociada al fútbol profesional, anticipa la dificultad del problema de regresión.

Por su parte, `home_win_bin` presenta clases relativamente balanceadas, con una ligera mayor frecuencia de partidos en que el equipo local no gana. Esta propiedad es favorable para el modelo de clasificación, pues evita problemas de desbalance severo.

3.3. Análisis de correlación

Para evaluar la asociación entre las variables predictoras y los objetivos, se calculó una matriz de correlaciones de Spearman que incluye todas las variables empleadas en los modelos. La elección de Spearman resulta adecuada, ya que este coeficiente no impone supuestos de linealidad y es robusto frente a valores extremos, algo especialmente relevante considerando la presencia de múltiples outliers observados anteriormente.

Adicionalmente, se generó un gráfico específico para examinar la correlación entre cada predictor y la variable objetivo del modelo lineal (`goal_diff`). Esto permite identificar no solo las asociaciones internas entre predictores, sino también su aporte individual potencial al rendimiento del modelo.

Los resultados muestran que las correlaciones entre las variables creadas y la diferencia de goles son, en general, débiles a moderadas, con valores que rara vez superan 0.20 o 0.25. Este patrón es coherente con la alta variabilidad e imprevisibilidad del resultado exacto de un partido de fútbol, donde factores no incluidos en la base—como lesiones, contexto táctico, decisiones arbitrales o dinámicas psicológicas—pueden influir de forma significativa y que no están en este modelo. La baja correlación directa también anticipa la limitada capacidad de los modelos lineales para capturar relaciones en este dominio, algo que se confirma posteriormente con el rendimiento del modelo LASSO.

Aun así, algunas variables destacan por su mayor asociación con la diferencia de goles, en particular la diferencia de Elo previa (`elo_diff_before`), la forma reciente del equipo local (`home_form_last5`) y el rendimiento ofensivo reciente (`home_attack_last5`).

Estas relaciones, aunque moderadas, reflejan dinámicas futbolísticas esperables: equipos con mejor

forma, historial más fuerte o mayor capacidad goleadora tienden a tener mayor probabilidad de anotar diferencias de goles positivas. Sin embargo, la magnitud reducida del coeficiente evidencia que estas ventajas estructurales no garantizan un resultado claro.

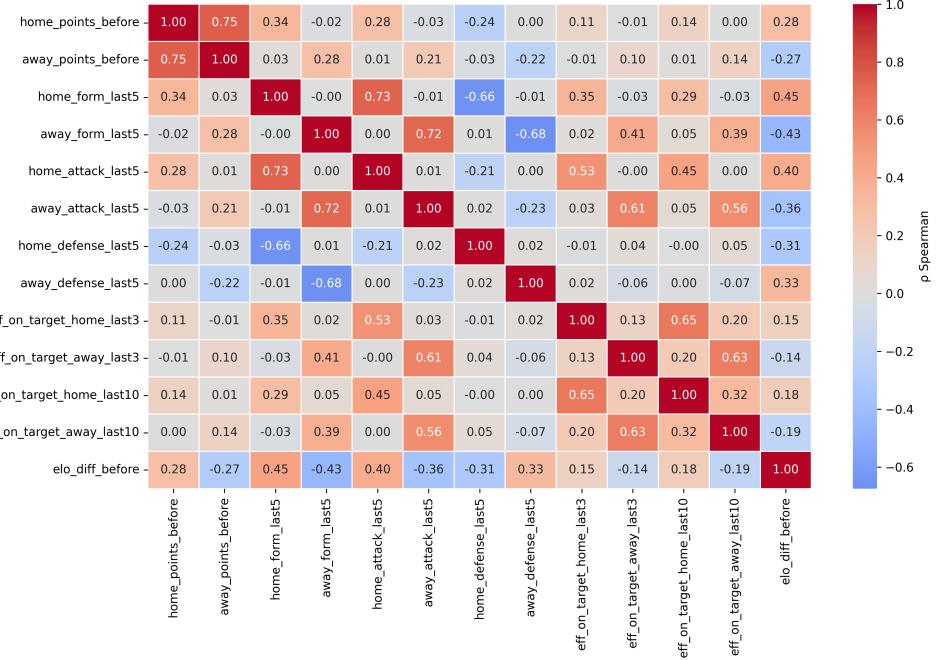


Figura 3: Matriz de correlaciones de Spearman entre las variables utilizadas en los modelos.

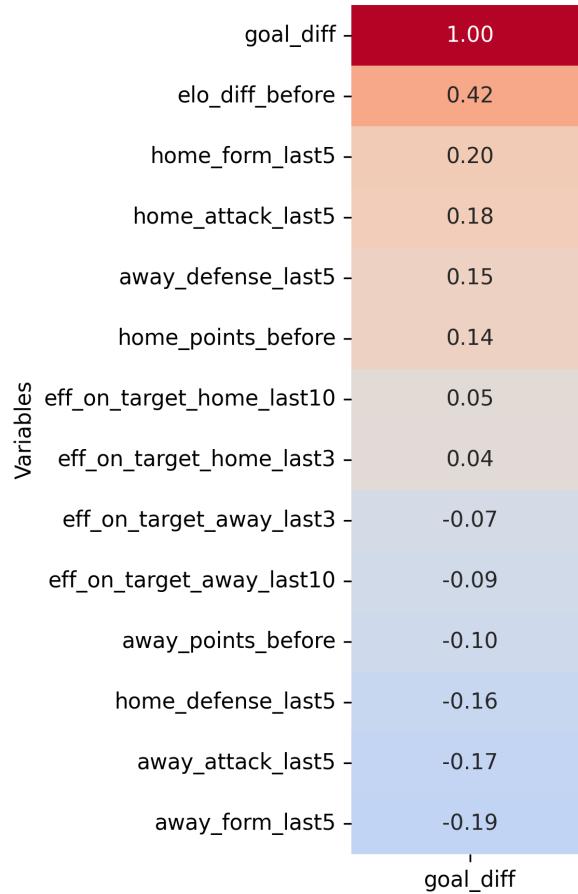


Figura 4: Correlación de Spearman entre los predictores y las variables objetivo.

3.4. Resumen estadístico de las variables utilizadas en los modelos

El **Cuadro 5** viene a ser un complemento de cierre de los boxplot, sintetizando los valores generales de las variables.

Cuadro 5: Estadísticas descriptivas de las variables utilizadas en el modelo

Variable	count	mean	std	min	25 %	50 %	75 %	max
home_points_before	9.380	26,41	18,32	0,00	12,00	24,00	38,00	100,00
away_points_before	9.380	26,58	18,20	0,00	12,00	24,00	38,00	99,00
home_form_last5	9.334	1,63	0,70	0,00	1,20	1,60	2,00	3,00
away_form_last5	9.334	1,13	0,68	0,00	0,20	1,00	1,60	3,00
home_attack_last5	9.380	1,54	0,70	0,00	1,00	1,40	2,00	5,80
away_attack_last5	9.380	1,18	0,61	0,00	0,75	1,00	1,20	5,00
home_defense_last5	9.380	1,18	0,58	0,00	0,80	1,20	1,60	4,40
away_defense_last5	9.380	1,53	0,65	0,00	1,00	1,40	2,00	5,00
eff_on_target_home_last3	9.334	0,27	0,14	0,00	0,17	0,25	0,36	1,50
eff_on_target_away_last3	9.332	0,26	0,15	0,00	0,15	0,25	0,35	1,33
eff_on_target_home_last10	9.334	0,27	0,09	0,00	0,20	0,26	0,32	1,50
eff_on_target_away_last10	9.332	0,26	0,09	0,00	0,19	0,25	0,32	0,64
elo_diff_before	9.380	0,77	152,06	-537,06	-91,26	0,00	94,02	500,58

4. Aplicación del método LASSO y principales resultados

Variable objetivo (target)

El objetivo del modelo es predecir la diferencia final de goles del partido. Para ello se construyó la variable `goal_diff`, definida como:

$$\text{goal_diff} = \text{fulltimehomegoals} - \text{fulltimeawaygoals}.$$

Un valor positivo indica victoria del equipo local, un valor negativo representa victoria del visitante y un valor cercano a cero refleja un partido equilibrado.

Especificación del modelo

Como variables explicativas se utilizaron los atributos construidos a partir de información previa al partido: puntos acumulados antes del encuentro (`home_points_before`, `away_points_before`), forma reciente en los últimos cinco partidos (`home_form_last5`, `away_form_last5`), promedios de goles a favor y en contra en los cinco encuentros previos (`home_attack_last5`, `away_attack_last5`, `home_defense_last5`, `away_defense_last5`), medidas de eficacia ofensiva (`eff_on_target_home_last3`, `eff_on_target_away_last3`, `eff_on_target_home_last10`, `eff_on_target_away_last10`) y el diferencial de *rating* histórico tipo Elo previo al partido (`elo_diff_before`).

Además, se incluyeron variables categóricas para la temporada y los equipos involucrados (`season`, `hometeam`, `awayteam`), codificadas mediante *one-hot encoding*, lo que genera un conjunto extendido de variables binarias para cada categoría.. El modelo estimado puede escribirse de forma general como:

$$\widehat{\text{goal_diff}}_i = \beta_0 + \sum_{j=1}^p \beta_j x_{ij} + \varepsilon_i,$$

donde x_{ij} corresponde a las variables explicativas previamente descritas y los coeficientes β_j se estiman mediante regresión lineal con regularización LASSO.

El parámetro de penalización α se seleccionó mediante validación cruzada (*cross-validation*) con tres, cinco y diez particiones, obteniéndose valores óptimos cercanos a $\hat{\alpha} \approx 0,005$ en 3, $\hat{\alpha} \approx 0,0015$ en el caso de 5 y 10 particiones. Esta penalización induce esparsidad en los coeficientes, de modo que sólo un subconjunto de variables mantiene coeficientes distintos de cero, lo que permite identificar los predictores con mayor capacidad explicativa sobre la diferencia de goles.

Desempeño del modelo

En el conjunto de entrenamiento el modelo alcanzó un MAE de aproximadamente 1,24 goles y un R^2 de 0,213, mientras que en el conjunto de prueba obtuvo un MAE cercano a 1,27 goles y un R^2 de 0,206. Esto indica que, en promedio, el error absoluto al predecir la diferencia final de goles se sitúa en torno a un gol y que el modelo sólo logra explicar una fracción acotada de la variabilidad total del resultado.

Entre las variables con coeficientes distintos de cero destacan el diferencial de *rating* histórico previo al partido (`elo_diff_before`), la diferencia reciente de capacidad ofensiva entre local y visita y la forma reciente del equipo local. En conjunto, estos resultados son coherentes con la idea de que los equipos más fuertes y en mejor racha tienden a obtener una mayor diferencia de goles, aunque el alto grado de aleatoriedad inherente al juego limita el poder predictivo del modelo.

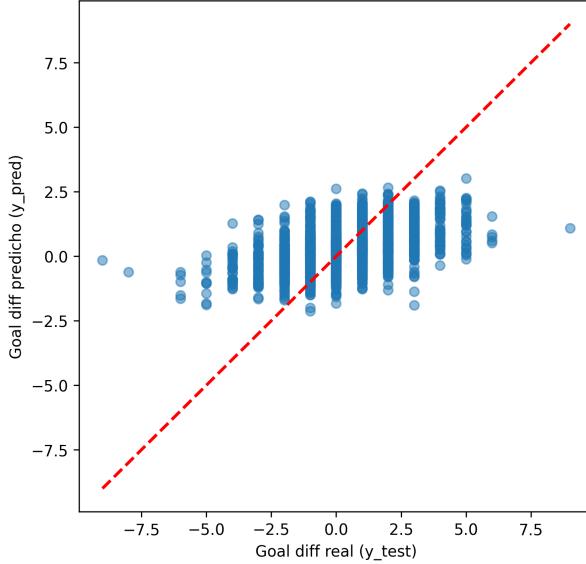


Figura 5: Valores reales vs. predichos de la diferencia de goles (modelo LASSO).

4.1. Análisis de los valores alfa

Como se explicó anteriormente, el parámetro de penalización α controla la intensidad con que el modelo LASSO contrae los coeficientes hacia cero. En esta sección se evalúa empíricamente la sensibilidad del modelo frente a cambios moderados en dicho parámetro. Para ello, a partir del valor óptimo α^* obtenido mediante validación cruzada, se estimaron tres versiones del modelo fijando α en $\alpha^*/5$, α^* y $5\alpha^*$.

El **Cuadro 6** resume las métricas de desempeño para cada uno de estos valores.

Cuadro 6: Evaluación de valores de α en el modelo LASSO

Nro	alpha	MSE	RMSE	MAE	R ²
alpha 1	0,000341	2,747158	1,657455	1,279688	0,204041
alpha 2	0,001704	2,740812	1,655540	1,273225	0,205879
alpha 3	0,008519	2,756045	1,660134	1,273123	0,201466

Como se observa, las diferencias en MSE, RMSE y MAE son mínimas y el coeficiente de determinación R^2 oscila en un rango muy acotado en torno a 0,20. Esto indica que el modelo es relativamente estable frente a variaciones razonables de α en este intervalo y que la elección de α^* mediante validación cruzada ofrece un compromiso adecuado entre ajuste y regularización, sin pérdidas relevantes de capacidad predictiva.

4.2. Relevancia de las variables en nuestro modelo

El análisis de importancia de variables permite identificar qué atributos tienen mayor influencia en la predicción de la diferencia de goles bajo el modelo LASSO. La **Figura 6** muestra los coeficientes absolutos de cada predictor, ordenados de mayor a menor relevancia, mostrando el top 20 superior.

Los resultados indican que el diferencial de rating histórico (`elo_diff_before`) es, con diferencia, la variable más influyente, lo que coincide con la idea de que los equipos más fuertes estructu-

ralmente tienden a obtener mejores resultados. Le siguen algunas dummies asociadas a equipos específicos, reflejando diferencias persistentes de desempeño entre clubes, así como variables relacionadas con la forma reciente y temporadas particulares. En conjunto, estas evidencias sugieren que la combinación entre fortaleza histórica y rendimiento reciente desempeña un rol clave en la predicción, mientras que otros factores incorporados en el modelo parecen tener un aporte menor, coherentemente con la naturaleza altamente estocástica del fútbol.

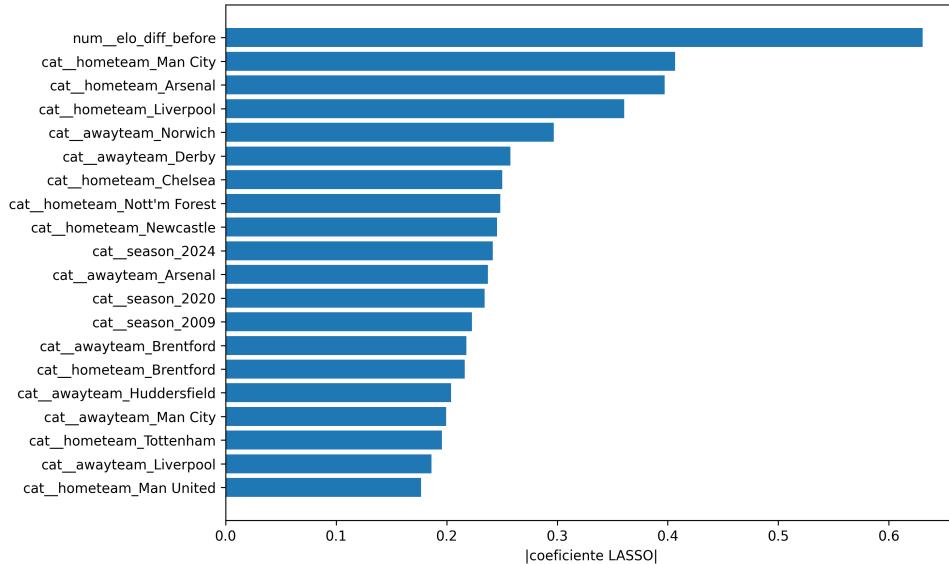


Figura 6: Distribución del ratio de victorias por equipo.

5. Random Forest

Variable objetivo (target)

Para la tarea de clasificación se definió la variable binaria `home_win_bin`, que toma el valor 1 si el equipo local gana el partido y 0 en caso contrario (empate o derrota). De este modo, el problema se formula como una clasificación supervisada entre “gana local” y “no gana local”.

Desempeño del modelo

El modelo Random Forest obtuvo un *accuracy* cercano al 65 % sobre el conjunto de prueba, con un rendimiento relativamente equilibrado entre ambas clases. De acuerdo al *classification report*, la clase 0 (no gana local) presenta un *recall* de 0.75, mientras que la clase 1 (gana local) alcanza un *recall* de 0.53. Esto indica que el modelo identifica con mayor facilidad los partidos en que el equipo local no gana, mientras que tiene más dificultades para reconocer correctamente las victorias del local.

La matriz de confusión confirma esta asimetría: el modelo acierta 754 veces al predecir que el local no gana, pero solo 453 veces al predecir una victoria local, cometiendo 399 falsos negativos para esta última clase. Este comportamiento sugiere que el Random Forest capta parte importante de la estructura del problema, pero su capacidad para clasificar correctamente las victorias locales sigue siendo limitada.

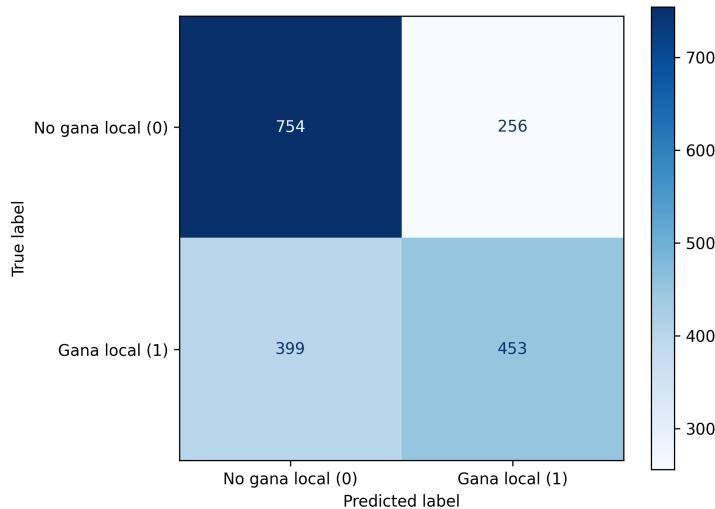


Figura 7: Distribución del ratio de victorias por equipo.

En conjunto, los resultados muestran que el modelo supera a una clasificación aleatoria y logra extraer patrones relevantes del desempeño de los equipos. Sin embargo, su capacidad predictiva es moderada y podría mejorarse mediante un ajuste más fino de hiperparámetros, la incorporación de nuevas variables explicativas o técnicas avanzadas como *boosting*. Aun así, el modelo ofrece una aproximación inicial útil para entender los determinantes de la victoria local en la Premier League.

6. Discusión de los resultados

Tras corregir los problemas de fuga de información y restringir los predictores a variables observables antes del partido, el desempeño de los modelos disminuye respecto de versiones preliminares. No obstante, estas métricas reflejan de manera más honesta la dificultad intrínseca de anticipar resultados futbolísticos a partir de estadísticas históricas.

El modelo LASSO explica en torno al 17 % de la variabilidad la diferencia de goles. Si bien este valor es modesto, es coherente con la alta incertidumbre propia de los marcadores en fútbol y confirma que, incluso disponiendo de información detallada del rendimiento previo, la capacidad para predecir la magnitud exacta del resultado es necesariamente limitada.

En la tarea de clasificación, el modelo Random Forest alcanza un *accuracy* del 63 % en el conjunto de prueba, mejorando alrededor de 9 puntos porcentuales respecto de un clasificador ingenuo que siempre predice que el equipo local no gana. Las métricas de precisión y *recall* son relativamente equilibradas entre ambas clases, lo que sugiere que el modelo captura patrones útiles en los datos, aun cuando su capacidad predictiva sigue siendo moderada.

En conjunto, los resultados muestran que es posible extraer cierta señal predictiva a partir del desempeño histórico de los equipos. Sin embargo, también evidencian que el fútbol continúa siendo un fenómeno con alta variabilidad aleatoria, difícil de anticipar mediante modelos que utilizan únicamente estadísticas de partidos previos y con ello, la necesidad de contar con datos de mayor calidad.

7. Conclusión

Un resultado central del trabajo no es únicamente el valor específico de las métricas obtenidas, sino la constatación de que pequeñas decisiones en la construcción de variables pueden inflar de manera artificial el desempeño de los modelos. La corrección de estas decisiones, aunque reduce los indicadores numéricos, mejora la validez externa de las predicciones y entrega una visión más realista de lo que puede —y no puede— lograrse con métodos de *Machine Learning* en este contexto.

Asimismo, sería valioso incorporar en futuros análisis variables adicionales ampliamente utilizadas en la literatura, tales como historial reciente entre los equipos, métricas avanzadas de rendimiento (pases, posesión, presión), factores contextuales (descanso, lesiones, cambios de alineación) y condiciones del partido. Por razones de tiempo, estas variables quedan fuera del presente informe; sin embargo, representan una línea natural para mejorar los modelos actuales y avanzar hacia predicciones más precisas.

Anexo

A. Variante de los modelos presentados

Como ejercicio complementario, se evaluó una variante de ambos modelos incorporando dos tipos adicionales de información: (i) una medida histórica del rendimiento entre los mismos equipos (diferencia de goles del último enfrentamiento directo), y (ii) los goles anotados por el equipo local y visitante al término del primer tiempo.

La inclusión de esta información, especialmente la correspondiente al marcador parcial del partido, produce una mejora sustantiva en las métricas de desempeño. Este resultado es consistente con la teoría y con la intuición: a medida que se observa información ocurrida durante el propio encuentro, la incertidumbre se reduce significativamente y aumenta la capacidad predictiva del modelo.

No obstante, estas variables no se incorporaron en el análisis principal del informe, pues el objetivo central fue evaluar modelos predictivos basados únicamente en información disponible antes del inicio del partido. En consecuencia, los resultados aquí expuestos deben interpretarse como una referencia ilustrativa del potencial del modelo bajo escenarios con información adicional.

Cuadro 7: Comparación del desempeño de los modelos (versión original y extendida)

Modelo	MSE	RMSE	MAE	R ²
LASSO (sin goles HT)	2,741	1,656	1,273	0,206
LASSO (con goles HT)	1,552	1,246	0,957	0,551
Random Forest (sin goles HT)	—	—	—	—
Accuracy: 0,648				
Random Forest (con goles HT)	—	—	—	—
Accuracy: 0,764				

B. Código utilizado

El código completo está disponible en: https://github.com/GaboAvila90/machine_learning_PL