Trabajo Fin de Máster

máster universitario en modelización y análisis de datos económicos

Título: Creación de un modelo de probabilidad de victoria para partidos de balonmano mediante técnicas de aprendizaje automático

Título (en inglés): Creating a win probability model of handball matches based on machine learning techniques.

N.º de palabras (excluidos bibliografía y anexos): 8000

Curso académico: 2022/23

Autor(a): Jose Ramón Cortés Alcaide

Correo electrónico: joseramon.cortes@alu.uclm.es

Tutor(a) 1: Francisco Pascual Romero Chicharro

Tutor(a) 2: Eusebio Angulo Sánchez-Herrera

Fecha: Lunes, 2 de Octubre de 2023

Firma autor(a)

Firma tutor (de solo 1, si hay varios)

Firmado por ROMERO CHICHARRO FRANCISCO PASCUAL - ***8051**
El día 03/10/2023 con un certificado emitido por AC

Facultad de Derecho y Ciencias Sociales

Ronda de Toledo, s/n • 13071-Ciudad Real (España) • Tlf. +34 926 295 300

Resumen / Abstract

En este trabajo tiene como finalidad aumentar el conocimiento de qué variables son más decisivas en un deporte escasamente estudiado como balonmano, el cual es muy interesante a la hora de ser explorado debido a su dinamicidad y complejidad. Se harán uso de técnicas de aprendizaje automático, tanto clásicas como más novedosas para afrontar el objetivo de poder explicar qué variables contribuyen a la victoria de un equipo en un encuentro, dependiendo de factores como la naturaleza o condiciones específicas de este. Los datos utilizados pertenecen a la federación internacional y europea de balonmano, tanto masculinos como femeninos y tiene aplicaciones diversas, desde otorgar a entrenadores o aficionados el poder de analizar cuáles fueron las claves globales de un partido, poder establecer métodos de valoración de jugadores o contribuir al entendimiento de este deporte para su evolución.

This study aims to deepen the understanding of which variables are most crucial in a sport as under-researched as handball, which is particularly captivating to delve into due to its dynamism and complexity. Machine learning techniques, both traditional and more contemporary, will be utilized to address the goal of explaining which variables contribute to a team's victory in a match, based on factors such as its nature or specific conditions. The data used are sourced from the international and European handball federations, covering both male and female categories. The applications of this study are manifold, from granting coaches or enthusiasts the ability to pinpoint the overarching keys to a match, establishing player evaluation methods, to contributing to the comprehension and evolution of this sport.

Facultad de Derecho y Ciencias Sociales

Ronda de Toledo, s/n • 13071-Ciudad Real (España) • Tlf. +34 926 295 300

Agradecimientos

En el momento de culminación de este Trabajo de Fin de Máster, quisiera hacer un inciso para transmitir mi más sincero y profundo agradecimiento a todas aquellas personas que han sido esenciales en esta travesía académica. En especial, a los docentes del curso, cuya sabiduría y dedicación han despejado mi camino y cuyo apoyo me ha permitido avanzar con solidez. Una mención particularmente afectuosa a mis tutores del trabajo, Francisco Pascual y Eusebio Angulo, quienes con su paciencia, orientación y pasión por el conocimiento me han inspirado y guiado hacia el logro de este proyecto. A mis seres queridos, que siempre han creído en mí y que, con sus palabras de aliento y motivación, me animaron a perseguir lo que verdaderamente me entusiasma: la ciencia de datos. A todos ellos, mi eterno reconocimiento y gratitud. En cada página de este trabajo, en cada análisis y reflexión, hay una parte de vosotros. Gracias.

Facultad de Derecho y Ciencias Sociales

Ronda de Toledo, s/n • 13071-Ciudad Real (España) • Tlf. +34 926 295 300

Índice de contenidos

1. Introducción

1.1. Justificación del tema.

Los datos y la estadística son fundamentales para comprender el mundo conforme se conoce hoy en día. Como dijo Peter Sondergaard: "La información es el aceite del Siglo XXI, y la analítica es el motor de combustión." Tanto es así, que el análisis de datos se ha convertido en una solución que afecta de forma transversal en todos los sectores, incluso en el de deporte.

En el balonmano, concretamente, a pesar de ser menos popular en España que el fútbol o el baloncesto, ofrece oportunidades únicas para la investigación (Schwenkreis, 2020). El gran auge de la tecnología y la facilidad de accesibilidad a los datos han beneficiado enormemente la ciencia del deporte, permitiendo evaluar rendimientos a modo de ponderación y poder predecir resultados de los partidos (Anguera & Hernández Mendo, 2015). (Romero et al., 2020) utilizan una metodología *Fuzzy* para evaluar objetivamente el rendimiento de los jugadores de balonmano en base a su actuación en el partido y (Angulo et al., 2022 y López-Gómez et al., 2022) aplican técnicas *softcomputing* y algoritmos *metaheurísticos* para evaluar el rendimiento de los porteros que es un puesto específico determinante en balonmano.

Los analistas de datos han ganado relevancia en el deporte de alto rendimiento, como se observa en deportes americanos y especialmente en la Premier League. En balonmano, equipos y selecciones ya incorporan analistas para examinar datos antes, durante y tras los partidos. Investigaciones, como la de Skejo et al., (2020) y Oytun et al., (2020), han aplicado técnicas de ciencia de datos en balonmano. Wagner et al., (2022) destacan sistemas de valoración de jugadores basados en estadísticas, pero falta un modelo que precise qué hace ganar a equipos y con qué probabilidad.

Los modelos probabilísticos de victoria, populares en la analítica deportiva, apoyan la toma de decisiones, análisis del rendimiento y predicciones. Estos también son valiosos para apuestas, ayudando a apostadores y casas (Hill, 2022; Robberechts et al., 2019). El desarrollo de tales modelos en balonmano es valioso para entrenadores, jugadores y aficionados. Herramientas como PIVOT (Muller et al., desconocido) pueden reforzar a estos modelos. No obstante, predecir resultados deportivos es desafiante debido a la aleatoriedad. Aoki, Assunção y Vaz De Melo (2017) cuantifican esta dificultad y

proponen un modelo gráfico de probabilidad para estimar la habilidad de un equipo y ponderar el factor suerte. Sin embargo, la precisión en la predicción deportiva sigue siendo un reto por factores aleatorios en los enfrentamientos.

1.2. Estructura del documento

El estudio será dividido, para su mejor lectura e interpretación, en las siguientes partes:

1. Introducción: El esfuerzo de esta parte estará centrado en justificar y motivar el estudio, poniendo de manifiesto los objetivos que se persiguen y las herramientas utilizadas para alcanzar dichos propósitos.
2. Marco teórico: En la siguiente parte se establecerá un despliegue de los conceptos que se deben entender para seguir el estudio satisfactoriamente debido a que serán utilizados o referenciados.
3. Marco práctico: Después se crearán los modelos que mejor se ajusten a los objetivos y restricciones mencionadas, se evaluarán y se interpretarán. También se realizará el proceso de revisión y mejora para cumplir los objetivos de la mejor forma posible.
4. Conclusiones: Se plasmarán las conclusiones que se pueden sacar del estudio realizado, así como sus limitaciones y posibles líneas de investigación futuras.

1.3. Objetivos.

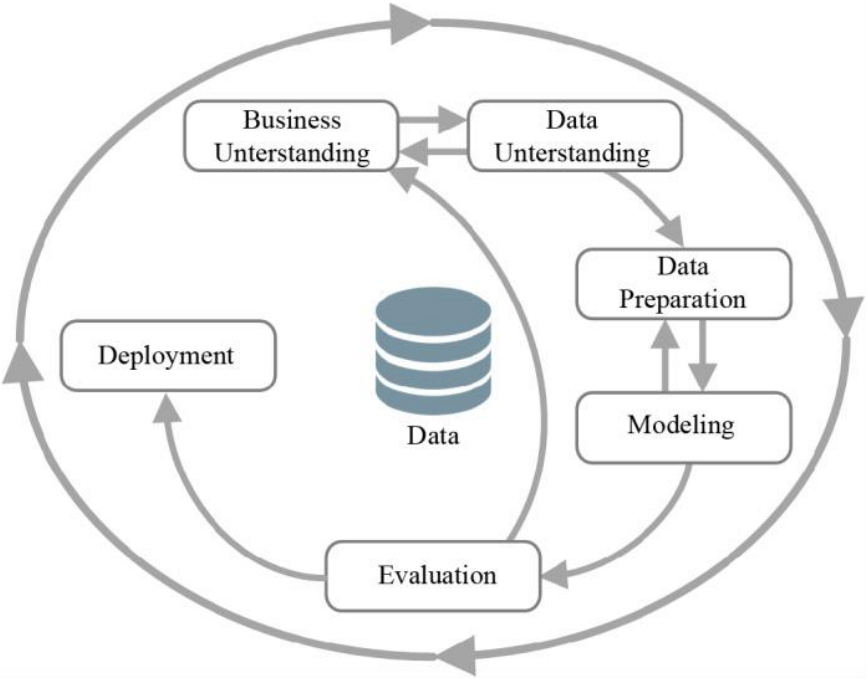
Los propósitos centrales de este estudio abarcan: predecir las probabilidades de triunfo de un equipo mediante un modelo de clasificación, establecer un modelo de probabilidad de victoria "*in-game"* basado en evidencia y técnicas de ciencia de datos y aprendizaje automático, y generar un modelo de valoración de jugadores derivado de este. Se aspira a impulsar el balonmano en la academia española. Asimismo, se persiguen objetivos específicos para alcanzar con éxito el propósito principal:

1. Recopilar datos esenciales sobre equipos, jugadores y eventos deportivos.
2. Distinguir variables cruciales que determinan el éxito de un equipo y la influencia de jugadores en dicho éxito.
3. Desarrollar análisis exploratorios para identificar tendencias, patrones y relaciones entre variables.
4. Explorar y escoger técnicas de modelado adecuadas, como regresión logística o redes neuronales.
5. Entrenar y corroborar el modelo con conjuntos de datos de entrenamiento y prueba.
6. Ajustar parámetros para optimizar precisión y adaptabilidad.
7. Aplicar procesos de extracción y transformación de datos para su modelización.
8. Usar métodos estadísticos para filtrar variables no esenciales.
9. Determinar la probabilidad de victoria "*in-game*" y desglosar momentos clave.
10. Contrastar nuestro modelo de valoración con otros para garantizar su solidez.

1.4. Metodología

Se ha utilizado para llevar a cabo el trabajo una metodología estándar para la realización de proyectos basados en minería de datos. La metodología CRISP-DM (Wirth, 2000), *Cross-Industry Standard Process for Data Mining* especifica las fases necesarias de un proyecto de minería de datos, así como las tareas necesarias en cada fase. Además, ofrece un resumen del ciclo vital de minería de datos tal y como se observa en la imagen. Una descripción más detallada de la metodología puede observarse en el Anexo 1.

FIGURA 1.1. CICLO DE VIDA DE MINERÍA DE DATOS. FUENTE: HUBER ET AL., 2018.



1.5. Marco tecnológico

Para buscar la mejor eficiencia y ejecución de este estudio, así como sus modificaciones y revisiones se han utilizado diferentes recursos de software y soluciones en línea.

- Github[1]: Es un portal creado para alojar el código y los archivos necesarios para el desarrollo. Se puede colaborar o leer el código con la descarga de la aplicación. Así pues, Git permite comparar el código de un archivo para analizar las diferentes versiones, restaurar antiguas y fusionar cambios de diferentes versiones.
- Google Collaboratory[2]: Mayormente conocido como *Colab*, es un servicio gratuito en la nube alojado por Google para fomentar la investigación sobre *Machine Learning* e Inteligencia Artificial. No es necesario realizar ninguna instalación o configuración debido a que los recursos se almacenan en una máquina de Google (Santos, 2020).
- Python[3]: Es un lenguaje de programación conocido por tener una sintaxis legible, intuitiva y limpia. Además, es posible usarlo en diferentes ámbitos y contextos debido a su licencia de código abierto, lo que hace que esté presente en plataformas como Google, Youtube o Facebook. Las librerías que se han utilizado utilizadas para la manipulación, modelado y visualización de datos se pueden ver en el Anexo 2.
- R: Es un lenguaje de programación en la que hay implementadas multitud de librerías que facilitan la implantación de técnicas estadísticas, enmarcas dentro de la plataforma GNU. Uno de sus principales atractivos es que incluye un lenguaje de programación simple y efectivo que permite condiciones, ciclos, funciones, etc. (Sosa et al., 2010). Las librerías utilizadas para este trabajo han ido enfocadas al análisis exploratorio de datos, otro de los fuertes del programa y quedan listados en el Anexo 2.

[1] https://es.wikipedia.org/wiki/GitHub
[2] Google Colab: ¿qué es y cómo usarlo? | Alura Cursos Online
[3] https://docs.python.org/es/3/tutorial/

2. Marco teórico

Para entender la estructura del documento y las técnicas aplicadas, es crucial familiarizarse con términos y técnicas específicas que se utilizarán en el estudio, recurriendo a repositorios como Google Scholar o Scopus para respaldar la investigación y ofrecer un marco teórico. Se llevará a cabo una revisión bibliográfica específica de las técnicas y modelos empleados para cumplir los objetivos propuestos.

2.1. Modelos de clasificación

Un modelo de clasificación, basado en estadísticas, categoriza observaciones en lugar de predecir valores numéricos. En este trabajo, empleamos tales algoritmos para calcular probabilidades de victoria, decidir si un evento sucede o no (como ganar un partido). La variable de respuesta es generalmente cualitativa, por ejemplo, el diagnóstico de una enfermedad o, en nuestro caso, el resultado de un partido. Estos modelos predicen la probabilidad de pertenencia a una categoría, formando la base para clasificaciones subsiguientes. Su selección dependerá de factores como tipo de datos y la capacidad de interpretar el modelo posteriormente (James et al., 2021).

2.2. Regresión logística

La regresión logística es un modelo estadístico comúnmente utilizado en problemas de clasificación binaria, donde el objetivo es predecir si la observación pertenecerá a una de dos posibles categorías. Esta se diferencia de la conocida regresión lineal en que, en lugar de predecir el valor de la variable respuesta, modela la probabilidad de que esta pertenezca a una categoría en particular y se garantiza que las predicciones se encuentren entre 0 y 1 (0 suele ser la no ocurrencia de un evento y 1 su antónimo).

La expresión matemática de este modelo es:

$$P(X)= \frac{e^{(\beta_0+\beta_1)X}}{1+e^{(\beta_0+\beta_1)X}}$$

Este modelo se ajusta a los datos usando la “máxima verosimilitud”. Las predicciones resultantes siempre se encuentran entre 0 y 1, independientemente del valor de X (James et al., 2021).

2.3. Naive Bayes

Este algoritmo es un clasificador que utiliza el teorema de Bayes para clasificar nuevas observaciones basándose en las evidencias de los datos de entrenamiento. Este algoritmo se llama “ingenuo” porque asume que todas las características son independientes entre sí, lo cual es una suposición que no siempre se cumple en la práctica. Pero sí que simplifica mucho el modelado de los datos de entrenamiento cuando no es lo suficientemente grande (James et al., 2021).

2.4. Árboles de clasificación

Los árboles de clasificación son técnicas que clasifican variables mediante reglas de decisión basadas en datos de entrada. Su estructura gráfica simula el proceso humano de toma de decisiones, siendo fácilmente interpretables. Para valorar su rendimiento, es preferible estimar el error de prueba en vez del de entrenamiento, y se puede "podar" el árbol eliminando "ramas" no esenciales (James et al., 2021). Aunque diseñados para modelar datos, muchos analistas los emplean en análisis exploratorio por su clara interpretación, aspecto que utilizaremos en este trabajo.

2.5. Random Forest

Los *Random Forest*, son un tipo de algoritmo de aprendizaje supervisado que combina varios árboles de decisión para poder generar, en este caso, una salida para clasificar la observación.

Este algoritmo lo propone Leo Breiman en 2001 y es una extensión de su anterior método llamado *“bagging”,* en el cual se usan múltiples árboles de decisión son entrenados en paralelo, donde cada árbol es construido a partir de una muestra aleatoria del conjunto de datos del modelo. En el caso de los *random forests*, se añade una nueva capa en la que al construir los árboles de clasificación en cada división solo se tiene en cuenta un subconjunto aleatorio de características en lugar de todas. Estos bosques son una gran técnica debido a que pueden manejar muchas entradas, prevenir el sobreajuste y ofrecer buenas métricas de rendimiento.

2.6. *Boosting - XGBoost*

Es un método de aprendizaje automático que se usa para mejorar la precisión de un modelo. La premisa inicial es tomar un algoritmo que produce una hipótesis aproximada al concepto objetivo y trata de mejorarla en diferentes distribuciones de datos.

La efectividad de este modelo se evidencia en la disminución significativa del error. Además, este proceso puede repetirse recursivamente para optimizar esta precisión del modelo, aunque hay que tener cuidado con el sobreajuste (Schapire, 1990).

XGBoost es un ecosistema de *boosting* de árboles escalable que es ampliamente utilizado en ciencia de datos y que ha cobrado bastante popularidad en los últimos años. Es bueno para manejar datos dispersos haciendo que maneje datos que no están uniformemente distribuidos (Chen & Guestring, 2016).

2.7. Evaluación de métodos de clasificación

Los diferentes métodos de evaluación se usan para poder elegir de manera informada qué modelo de clasificación es el que mejor clasifica cada conjunto particular de datos (Se puede ver de forma detallada en el Anexo 3.1), en este estudio se usan principalmente dos métodos:

- La matriz de confusión, también conocida como tabla de contingencia, representa cómo un algoritmo de clasificación consigue clasificar correctamente datos que no han sido usados para entrenarlo (Santra & Josephine, 2012). A partir de esta matriz, se pueden calcular varias métricas que proporcionan información sobre el desempeño del modelo de clasificación como la precisión, la exactitud, etc.
- La curva ROC (*Receiver Operating Characterics*) es un gráfico ampliamente utilizado en estadística para evaluar el rendimiento de un clasificador binario a medida que el umbral de clasificación varía. Representa la sensibilidad (tasa de verdaderos positivos) frente a la especificidad (1- tasa de falsos positivos) para todos los posibles umbrales de decisión. La eficacia del clasificador se compara mediante el área bajo la curva (AUC). Dando por clasificador ideal aquel cuya AUC sea igual a la esquina superior izquierda, indicando una total clasificación de verdaderos positivos sin existencia de falsos positivos.

3. Creación del modelo

En este capítulo, se expone la aplicación de la metodología CRISP-DM para abordar la estimación de la probabilidad de victoria en un partido a partir de los indicadores de rendimiento de los equipos. Inicialmente, se plantea un problema de clasificación que emplea diversos indicadores fundamentales del encuentro como variables de entrada para determinar si un partido culminará en victoria o no. Posteriormente, se aplican diversas técnicas de aprendizaje automático con el propósito de desarrollar modelos capaces de resolver esta clasificación con un nivel de desempeño adecuado, evaluado conforme a las métricas presentadas anteriormente.

3.1. Comprensión del problema

La probabilidad de victoria en deportes, como el balonmano, indica la posibilidad de un equipo ganar en un partido, considerando el rendimiento histórico en contextos similares. Dada la complejidad del juego y las múltiples variables en juego, el estudio busca crear un modelo estadístico que estime esta probabilidad con precisión. Se usarán variables globales del rendimiento del equipo, omitiendo factores del rival u otras circunstancias. Se seguirá el proceso CRISP-DM para preparar y transformar datos, calculando variables auxiliares y buscando el modelo más fiable.

3.2. Comprensión de los datos

Los datos utilizados pertenecen a la IHF (*International Handball Federation*) y a la EHF (*European Handball Federation*), concretamente pertenecen a las estadísticas de todos los partidos de los torneos masculino y femenino que serán tratados de forma separada con el fin de encontrar patrones distintos de juego. Las variables con las que se trabaja cada partido se pueden observar en la Tabla 3.1.

TABLA 3.1. LISTA DE LAS VARIABLES QUE SE TRATARÁN EN EL ESTUDIO

Campo	Descripción
Phase	Representa la fase del torneo en la que se juega el partido, como fase de grupos, cuartos de final o final del campeonato.
Match No.	Representación numérica de cada uno de los partidos
Match	Representa qué dos equipos se enfrentan.
Team	Representa de qué equipo se están viendo las estadísticas ofensivas, qué equipo está en posesión del balón.
Goals	Número de goles que ha marcado el equipo.
Attacks	Número de Ataques acometidos en el partido.
%Eff	Porcentaje de eficacia del equipo (Goals / Attacks).
%Lanz	Porcentaje de lanzamientos realizados en el partido (calculado como n° de lanzamientos / Attacks)
%GK	Porcentaje de paradas realizadas con respecto al número de lanzamientos del equipo contrario.
TO	Número de pérdidas de balón que se producen a lo largo de los ataques.
%TO	Porcentaje de pérdidas con respecto a los ataques totales (TO / Attacks).
¿GANA?	Representa una columna boolena que representa si el equipo acaba ganando con esa serie de estadísticas con "SI" y "NO".
Diff	Diferencia de goles en el partido (positivo si gana y negativo si pierde).

Los problemas o limitaciones presentan estos datos son:

1. La desigualdad de registros en el caso del balonmano masculino y femenino, ya que se han procesado 4 torneos en el caso de los equipos masculino y dos en el caso de los femeninos debido a problemas de formato y no pudieron ser agregados.
2. La mayoría de las variables disponibles vienen calculadas a raíz de otras más sencillas por lo que puede haber problemas de multicolinealidad.

3.3. Preparación de los datos

En primer lugar, los datos se cargarán a partir de ficheros Excel procedentes del procesamiento de los datos facilitados por las competiciones. Se realiza entonces el siguiente procesamiento de los datos:

- Se eliminan columnas innecesarias de varios *dataframes*.
- Se calcula y se añaden las columnas '*%TO'*, *'Goals' y 'Diff'* donde faltan.
- Se crea la columna *'Match No.'* a partir del procesamiento de la columna '*Match'*.
- Se renombran algunas columnas y se reorganizan para mantener la consistencia y disponer de un orden específico.
- Se combinan todos los *dataframes* masculinos y femeninos en uno sólo 'dfM' y 'dfW' respectivamente considerando sus columnas comunes. Posteriormente se dividen entre entrenamiento y test.

Posteriormente se mapea la columna '¿GANA?' a valores numéricos (1 para 'SI' y 0 para 'NO') y se eliminan las filas con valores nulos. También se analizan sin eliminarse los valores atípicos.

Finalmente, se incorporan variables de creación propia relevantes a partir de las directrices de varios entrenadores de nivel nacional, lo cual, a priori puede permitir describir patrones ocultos o más interesantes en la fase de modelado. Debido a que no hay una relación empírica de que la inclusión de estas variables se traduce en una mejor comprensión del devenir de un partido se probará a incluirlas y hacer análisis con y sin ellas. Estas variables son:

- 'RendimientoNeto': Esta variable puede interpretarse como una medida de la eficacia general de un equipo, teniendo en cuenta tanto su rendimiento ofensivo ('*%Eff'*) como defensivo ('*%GK'*), y penalizando las pérdidas de balón ('*%TO'*).
- 'ImpactoNeto': Se calcula de manera similar, pero utilizando el porcentaje de lanzamientos ('*%Lanz'*) en lugar del porcentaje de eficacia. Esta variable puede interpretarse como una medida del impacto general de un equipo en el partido.

3.4. Análisis exploratorio

A fin de comprender mejor los datos y su interacción se procede a realizar un análisis exploratorio, divididos por sexo a fin de analizar si existen diferencias notables.

3.4.1. Análisis univariante

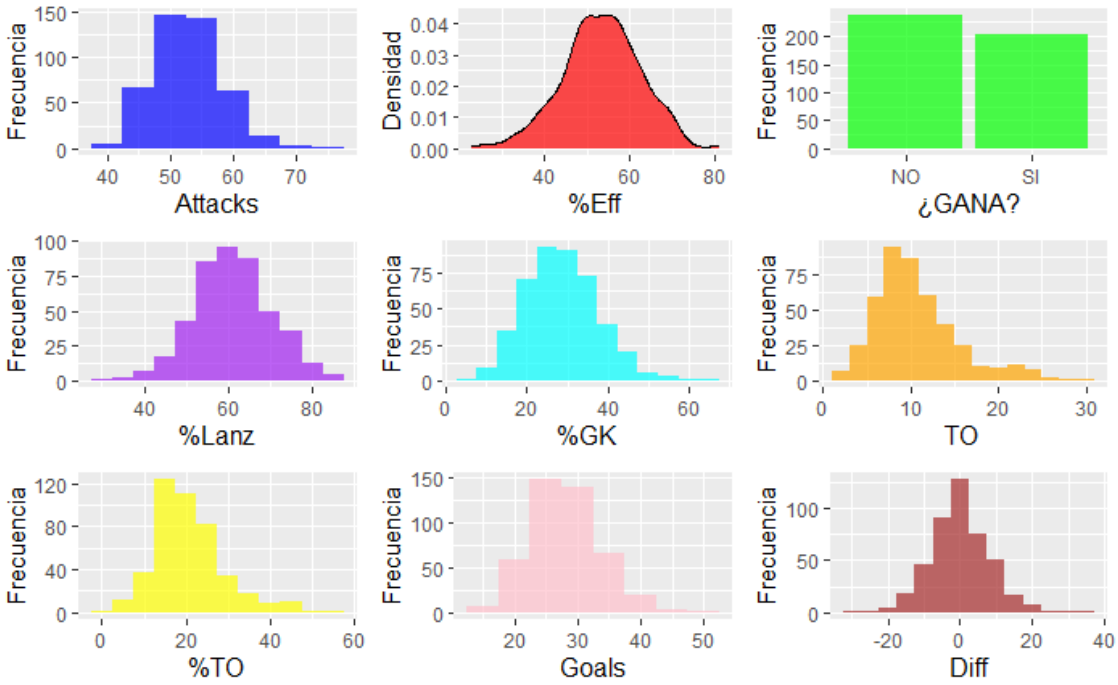
Este tipo de análisis es fundamental, ya que permite conocer la distribución, tendencia central y variabilidad de cada variable de manera individual. Así pues, se trata de la primera aproximación para comprender cómo se distribuyen los datos a fin de ver

anomalías o patrones interesantes. Para ello se generan histogramas para todas las variables numéricas para ver cómo se distribuyen y poder denotar algún tipo de anomalía.

En la Figura 3.1 se pueden observar los datos referentes a las competiciones masculinas de balonmano. la mayoría tiene la moda en torno a la mitad de la distribución, lo que hace que la media y la moda estén más cerca, las pérdidas suelen ser bajas por lo general y la diferencia de goles suele ser de 1 o 2 como mucho, denotando la gran cantidad de partidos igualados. Los goles por partido suelen estar entre 20 y 30, siendo los ataques totales de 50 a 55 generalmente, aunque hay casos atípicos.

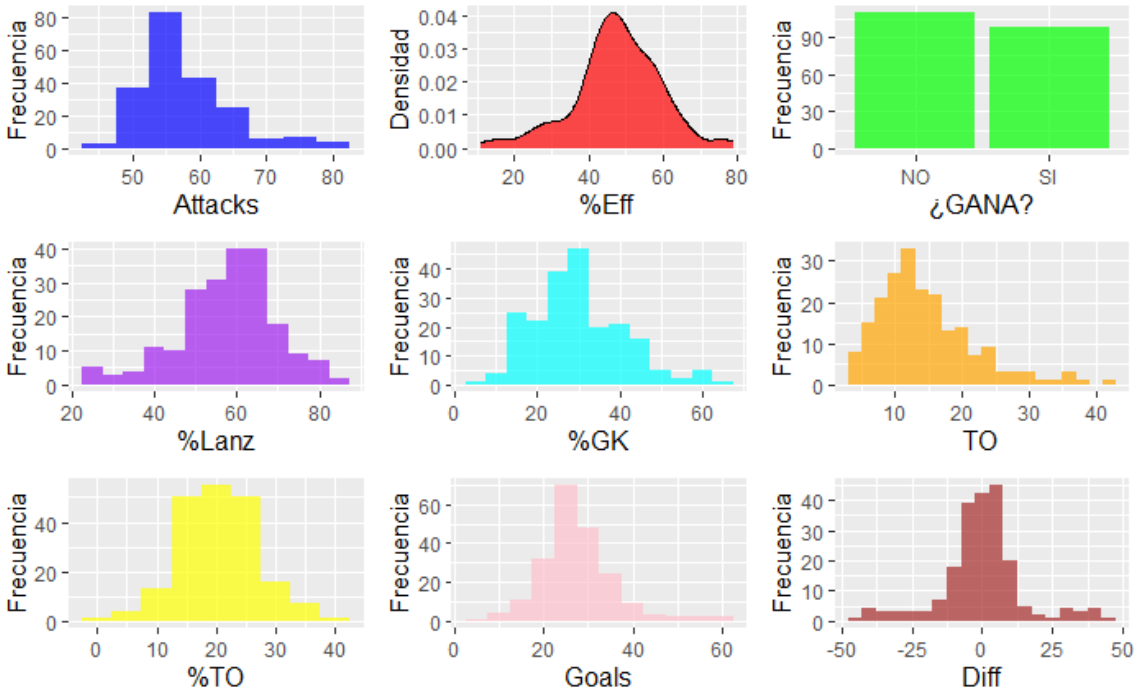
Datos de equipos de balonmano masculinos

FIGURA 3.1. ANÁLISIS UNIVARIANTE DE TODAS LAS VARIABLES, EQUIPO MASCULINO



En la Figura 3.2 se pueden observar los datos en cuento a las competiciones de balonmano femenino. En este caso, la mayoría de los partidos se resuelven entre 20 y 30 goles y unos 55 ataques, aunque ha habido equipos que han llegado a los 80. La eficacia ronda entre el 50% y 60% y el porcentaje de pérdidas en general suele ser mayor que en el caso de los hombres debido al propio estilo de juego tendente a ganar la línea de 6 metros.

FIGURA 3.2. ANÁLISIS UNIVARIANTE DE TODAS LAS VARIABLES, EQUIPO FEMENINO



3.4.2. Análisis multivariante

El análisis multivariante, examina las relaciones entre múltiples variables a la vez lo cual ayudando así a comprender cómo las variables interactúan entre sí, y si existen patrones que no serían evidentes al analizar cada variable de forma separada. Para ello se harán gráficos de violín primero con todas las variables respecto a '¿GANA?' y se podrán ver en Anexos (Anexo 4.1) las variables por pares en gráficos de dispersión, viniendo el color representado por la variable objetivo. Se ignorarán aquellas relaciones que son evidentes o carecen de importancia.

FIGURA 3.3. ANÁLISIS MULTIVARIANTE CON RESPECTO A LA OBJETIVO, EQUIPO MASCULINO

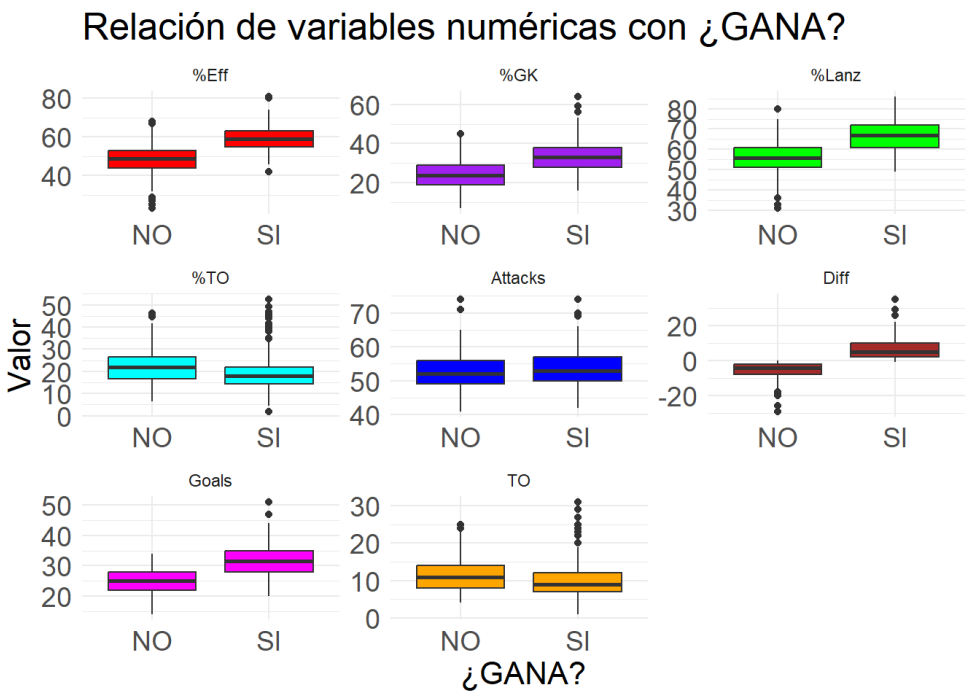
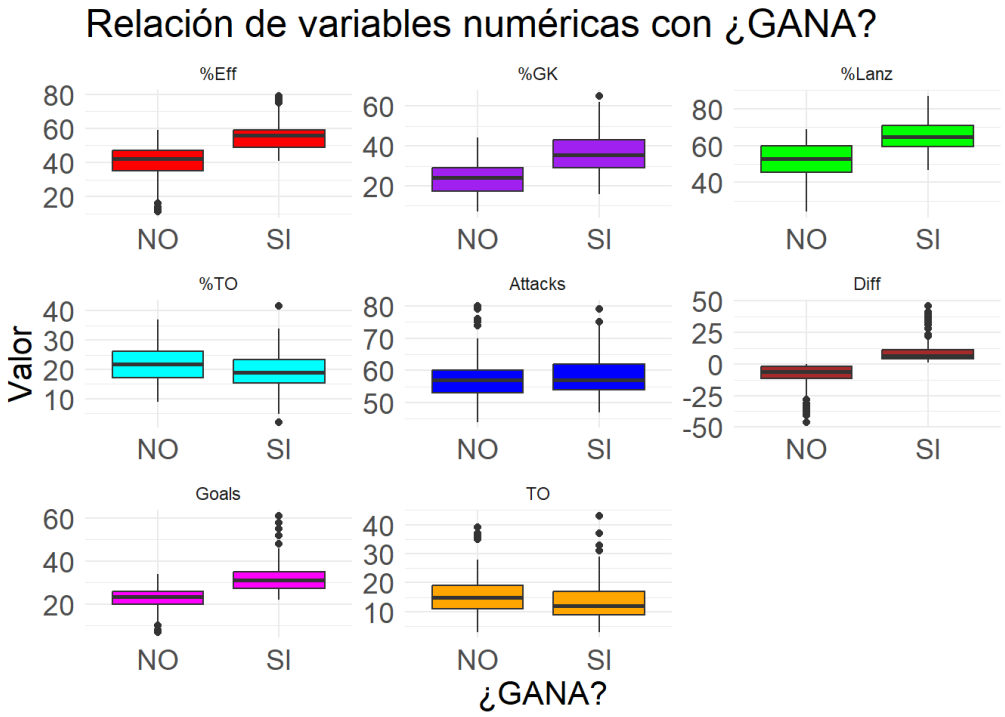


FIGURA 3.4. ANÁLISIS MULTIVARIANTE CON RESPECTO A LA OBJETIVO, EQUIPO FEMENINO



La Figura 3.3 muestra un análisis multivariante de competiciones masculinas de balonmano. Eficacias superiores al 75%, menos de 15 pérdidas de balón, paradas por

encima del 30% y al menos 30 goles, son indicativos de victoria, con excepciones. La Figura 3.4, sobre competiciones femeninas, destaca que equipos con eficacia superior al 50% tienen alta probabilidad de ganar, pero hay excepciones. Porcentajes de paradas entre 20%-40% y pérdidas similares dan resultados variados. Equipos femeninos con más del 60% en lanzamientos o más de 30 goles suelen ganar; menos de 20 pierden, siendo consistente en ambos géneros, aunque la muestra limita inferencias firmes.

3.4.3. Matriz de correlaciones

La matriz de correlaciones expresa la relación matemática entre las variables de un conjunto de datos. Una relación alta de forma directa o inversa no implica una causalidad, pero es importante analizarlo para ver qué variables a priori tienen más relación con la variable objetivo y entre cuáles de ellas podría haber problemas de multicolinealidad. Tanto en la Figura 3.5 correspondiente a los datos de balonmano masculino, como la 3.6 referida al femenino se pueden destacar correlaciones significativas entre ciertas variables derivadas de cálculos en un partido, como la eficacia y los goles, y la portería con *'Diff'*. Se puede inferir que un equipo con mayor eficacia y porcentaje de portería, y menos pérdidas, tendría más posibilidades de victoria. Se nota también una relación inversa entre el porcentaje de eficacia y pérdidas, y que incrementar los ataques no necesariamente altera otras variables. La matriz de correlaciones refleja tendencias similares entre equipos masculinos y femeninos, con la variable '*Diff'* mostrando la mayor relación con '%*Eff*', seguida por el número de goles y '%*Lanz'*.

FIGURA 3.5. MATRIZ DE CORRELACIONES, EQUIPO MASCULINO

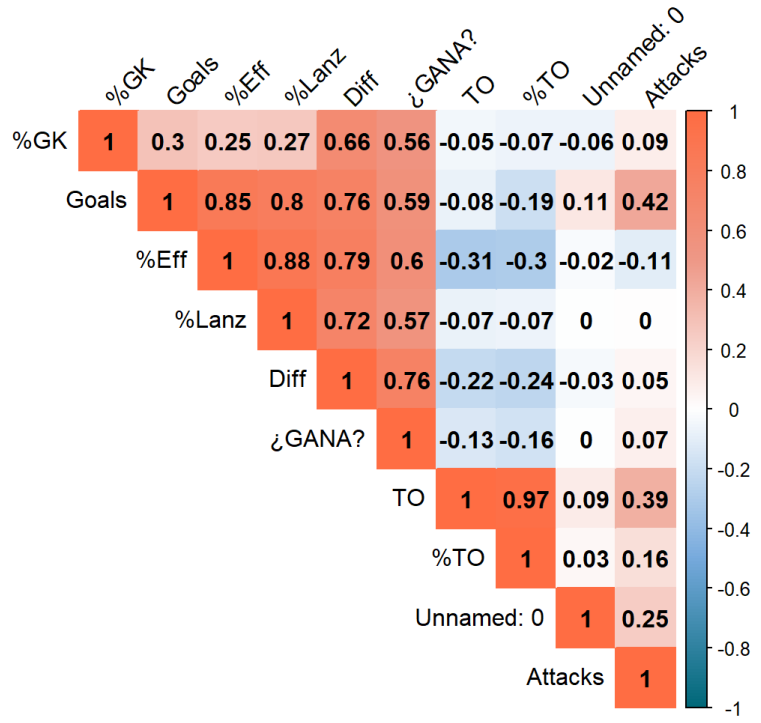
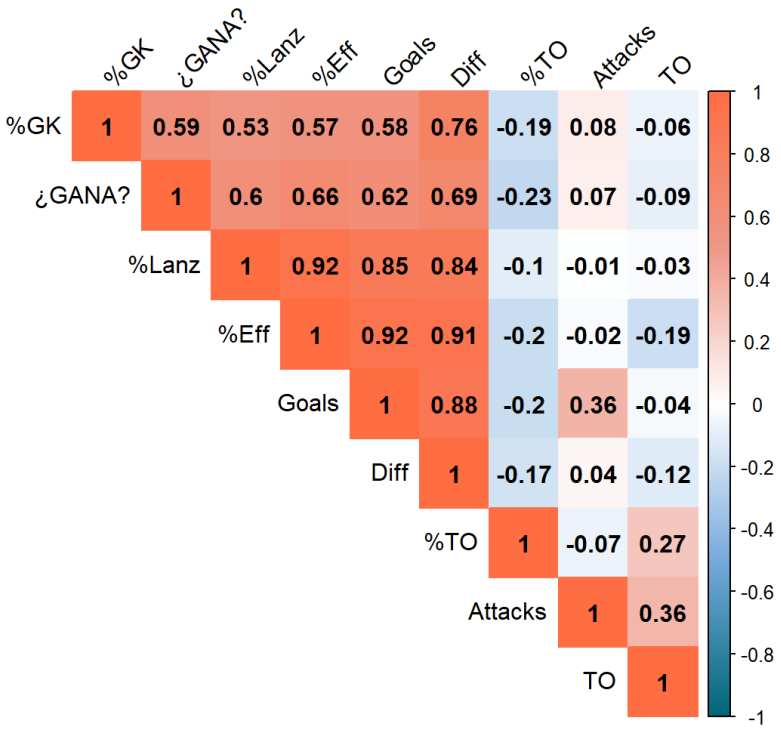


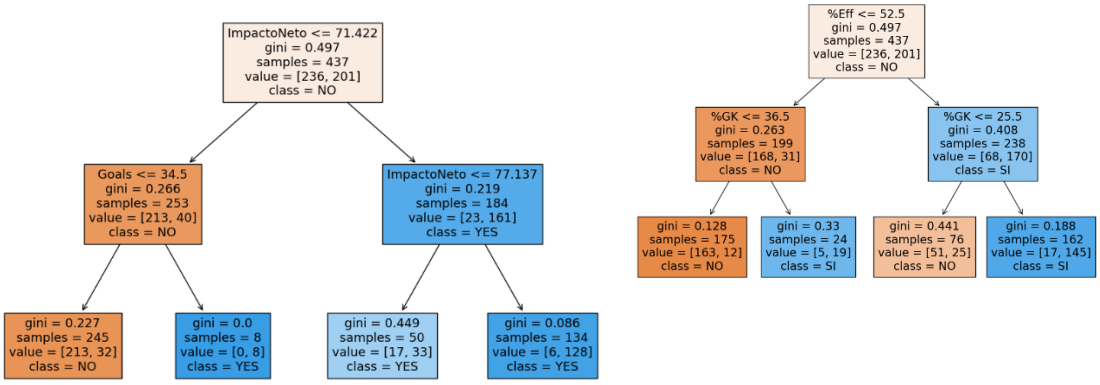
FIGURA 3.6. MATRIZ DE CORRELACIONES, EQUIPO FEMENINO



3.4.4. Análisis exploratorio usando árboles de clasificación

En este apartado se utilizan técnicas de árboles de decisión para análisis exploratorio con el fin de determinar qué características son más influyentes en la victoria de un equipo, para ello se usarán también las variables creadas y se enfrentarán a las originales para ver qué es más determinante para conseguir una victoria.

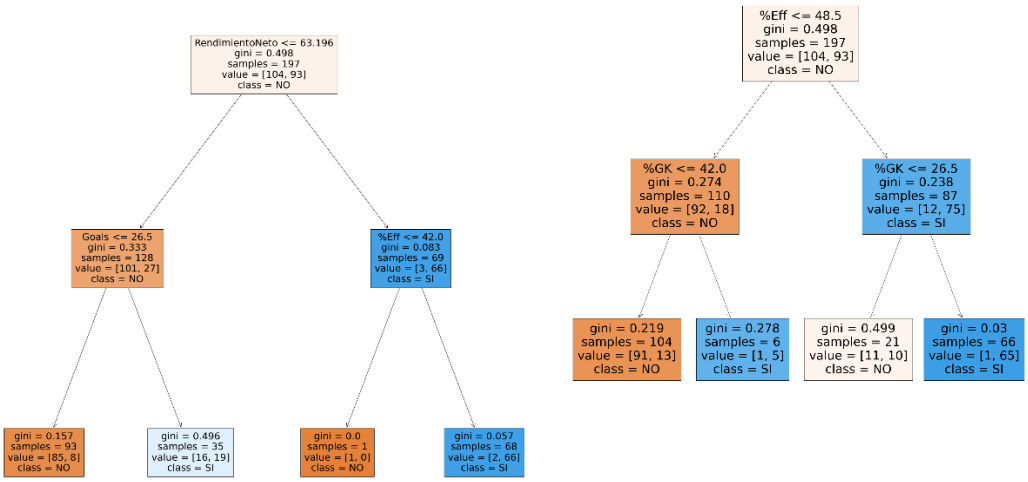
FIGURA 3.7. ANÁLISIS EXPLORATORIO CON TODAS LAS VARIABLES Y VARIABLES ORIGINALES, EQUIPO MASCULINO



En la Figura 3.7 se pueden observar los árboles resultantes del análisis de los datos de balonmano masculino. En la parte izquierda, en el que se analizan todas las variables, 'RendimientoNeto' e 'Impacto Neto' son los primeros indicadores de la probabilidad de ganar, tal y como se indicaba en la matriz de correlaciones. Ambas variables son clave para determinar la victoria de un equipo, siendo esencial mantener un alto porcentaje de eficacia y portería, y uno bajo de pérdidas. En niveles inferiores del árbol, en las ramas asociadas a mayores pérdidas, el número de ataques puede ser determinante para ganar, posiblemente debido a la mayor eficacia desde 9 metros que se tiene en el balonmano masculino. En el árbol de la derecha, con las variables originales, se revela que la eficacia y el porcentaje de portería son variables clave para determinar la victoria de un equipo, sugiriendo que más del 54.5% de eficacia y más del 30.5% de paradas inclinan hacia la victoria. Sin embargo, porcentajes inferiores en estas variables podrían resultar en derrota, a menos que se tengan pocas pérdidas. Al eliminar la eficacia y los goles del modelo, el número de goles y la portería se vuelven determinantes, con umbrales específicos que señalan la probabilidad de victoria o derrota. En este escenario simplificado, equipos con

mejores porcentajes de paradas y lanzamientos suelen ganar, aunque las pérdidas también juegan un rol crucial en niveles inferiores del árbol de decisión.

FIGURA 3.8. ANÁLISIS EXPLORATORIO CON TODAS LAS VARIABLES Y VARIABLES ORIGINALES, EQUIPO FEMENINO



En la Figura 3.8 se pueden observar los árboles resultantes del análisis de los datos de balonmano femenino. La eficacia y el impacto neto, influenciados por '*%Lanz'* y el porcentaje de portería, son determinantes en la victoria de un equipo tal y como se puede apreciar a la izquierda, en el árbol en el que se incluyen todas las variables. Esto sucede especialmente si superan el 48.5% y 26.5% respectivamente, aunque un bajo número de goles y lanzamientos puede resultar en derrota. Al eliminar las variables creadas, la eficacia sigue siendo clave, pero se complementa con el porcentaje de portería y el número de goles. Siguiendo la tendencia masculina, en el caso de eliminar la variable '*%Eff*' son los goles lo que determina en gran medida la victoria o derrota. Como se ha podido comprobar en los gráficos de violín, más de 28 goles hace que el equipo aumente sus probabilidades de victoria, si además la portería se sitúa por encima del 26.5% asegura victoria en la mayoría de las ocasiones. Si los goles son menos de 28 la portería debe aumentar 10 puntos porcentuales para poder tener oportunidades de victoria. Al eliminar de la ecuación los goles y la eficacia en el grupo masculino cobra mucha relevancia la portería y en el femenino son los lanzamientos. Esto se debe a que los equipos femeninos suelen asegurar más el tiro y la realización de un lanzamiento implica un gol de forma más segura.

3.5. Modelado

Tras entender en detalle los datos y las interacciones entre variables, se procederá a modelar y determinar qué algoritmo explica mejor los datos. Se utilizará el árbol de decisión, *Random Forest*, *Naive Bayes* y *XGBoost*. Estos modelos se entrenarán con y sin variables adicionales para evaluar su influencia en la explicación de los datos. Se divide la muestra en datos de entrenamiento y validación. Los primeros se usan para entrenar el modelo, captando relaciones y tendencias, mientras que los de validación testean su ajuste a nuevos datos. Luego, se analizarán los parámetros de los algoritmos empleados.

3.5.1. Árbol de decisión

Primero, se crea un modelo de árbol de decisión con parametrización por defecto, usando el criterio de Gini para las particiones y permitiendo expansión hasta hojas "puras". El árbol resultante refleja decisiones basadas en variables seleccionadas (Figuras 3.7, 3.8) y se aprecia la importancia de dichas variables en la Tabla 3.2. Un mayor valor indica mayor influencia en la decisión del árbol.

TABLA 3.2. IMPORTANCIA DE LAS VARIABLES EN LOS MODELOS DE ÁRBOL DE DECISIÓN (SIN PODA)

Variables	Masculino		Femenino	
	Creadas	Originales	Creadas	Originales
ImpactoNeto	0.2464	-	0.1687	-
RendimientoNeto	0.2095	-	0.2566	-
%Eff	0.1190	0.2102	0.2076	0.2398
%GK	0.1129	0.2878	0.0846	0.2165
%Lanz	0.0945	0.1398	0.0685	0.1391
Goals	0.0857	0.1746	0.1092	0.2463
%TO	0.0546	0.0732	0.0461	0.0609
TO	0.0387	0.0506	0.0314	0.0470
Attacks	0.0387	0.0639	0.0272	0.0503

3.5.2. Regresión logística

Se ha usado una Regresión Logística con *scikit-learn* en su configuración por defecto para clasificación binaria. La función logística convierte características de entrada en probabilidades entre [0, 1]. Utiliza la regularización L2 (Ridge) para prevenir sobreajuste, ajustada por el parámetro C; un valor menor refuerza la regularización. Se minimiza la

función de coste con métodos como Newton-Raphson para hallar coeficientes óptimos. El modelo predice probabilidades de pertenencia a clases y clasifica usando un umbral de 0.5.

3.5.3. Naive Bayes

El clasificador *Naive Bayes* opera bajo el principio de clasificación probabilística, utilizando el teorema de Bayes para estimar la probabilidad de pertenencia a una clase dado un conjunto de características. Su enfoque *'naive'* radica en la suposición de independencia condicional entre las características, lo que simplifica el cálculo de estas probabilidades. En este caso se ha utilizado el modelo Gaussiano.

3.5.4. Random Forest

Este utiliza múltiples árboles de decisión, cada uno entrenado en submuestras aleatorias de características y observaciones. La predicción se realiza por votación. La importancia de las características se calcula automáticamente, indicando su contribución al modelo. Resiste bien al sobreajuste. Los parámetros clave incluyen número de árboles (*n_estimators*, establecido en 100) y profundidad máxima (*max_depth*). La relevancia de las variables se muestra en la Tabla 3.3.

TABLA 3.3. IMPORTANCIA DE LAS VARIABLES EN LOS MODELOS DE *RANDOM FOREST*

Variables	Masculino		Femenino	
	Creadas	Originales	Creadas	Originales
ImpactoNeto	0.2464	-	0.1687	-
RendimientoNeto	0.2095	-	0.2566	-
%Eff	0.1190	0.2102	0.2076	0.2398
%GK	0.1129	0.2878	0.0846	0.2165
%Lanz	0.0945	0.1398	0.0685	0.1391
Goals	0.0857	0.1746	0.1092	0.2463
%TO	0.0546	0.0732	0.0461	0.0609

3.5.5. *XGBoost*

Se aplicó *XGBoost* con parámetros predeterminados. Utiliza árboles de decisión, optimizando la pérdida logarítmica mediante refuerzo de gradiente. Cada árbol subsiguiente corrige los errores del anterior. *XGBoost* incorpora regularizaciones L1 y L2 contra sobreajuste. La tasa de aprendizaje regula la contribución de cada árbol y se ajusta automáticamente. El número de árboles viene dado por el hiperparámetro *n_estimators*, ajustable según el rendimiento requerido. Es un punto de partida robusto para el problema.

3.6. Evaluación

En este paso se hará uso de los datos de validación para ver cuántos partidos consigue clasificar correctamente el modelo. Los modelos elegidos son regresión logística, debido a su mayor precisión general para clasificar, y *Random Forest* debido a que es muy buen clasificador con datos de esta naturaleza (correlaciones muy altas) y además arroja buenas precisiones. El árbol de decisión tiene un poder predictivo bastante bajo y su finalidad, pese a haberse evaluado, era de puro análisis exploratorio. *XGBoost* es un modelo que pese a ser una buena elección tiene menos poder de clasificación con pocos datos y tiende al sobreajuste. *Naive Bayes* asume la normalidad de los datos y la independencia de las variables, asunciones con no se cumplen en este caso. Sin embargo, se desplegará una tabla finalmente del resumen de las métricas de todos los modelos. Es importante destacar que para cada uno de los modelos se harán dos pruebas con dos conjuntos diferentes para cada género (conjunto con las variables originales y con las variables creadas). Los métodos escogidos para validar los modelos han sido la matriz de confusión, con sus métricas y la curva ROC, (Véase con más detalle en Anexo 4.2) comentadas en el capítulo de marco teórico.

3.6.1. Equipos de balonmano masculinos

En la Tabla 3.4 se reflejan los resultados obtenidos con todos los modelos obtenidos a partir de los datos de competiciones masculinas de balonmano pudiendo observar cómo en muchos casos los modelos basados en Regresión Logística y *Random Forest* son los que mejor resultados ofrecen. No se aprecian resultados demasiado distintos entre los modelos que incorporan las variables creadas con respecto a los modelos construidos a partir de los datos originales, ni en cuanto a indicadores de clasificación ni en cuanto a área bajo la curva.

TABLA 3.4. RESUMEN DEL RENDIMIENTO DE TODOS LOS MODELOS, EQUIPOS MASCULINOS.

Modelos	Variables	Precision (No)	Recall (No)	Precision (Si)	Recall (Si)	Accuracy	AUC
Árbol de decisión	Creadas	0.83	0.85	0.85	0.83	0.84	0.91
	Originales	0.87	0.88	0.88	0.87	0.88	0.92
Regresión Logística	Creadas	0.85	0.86	0.86	0.85	0.86	0.95
	Originales	0.85	0.88	0.88	0.85	0.87	0.95
XGBoost	Creadas	0.85	0.84	0.84	0.85	0.84	0.92
	Originales	0.85	0.89	0.88	0.74	0.82	0.92
Random Forest	Creadas	0.88	0.85	0.86	0.88	0.87	0.94
	Originales	0.87	0.86	0.86	0.87	0.87	0.94
Naive Bayes	Creadas	0.86	0.86	0.86	0.86	0.86	0.94
	Originales	0.87	0.86	0.86	0.87	0.87	0.94

En general, parece que todos los modelos mejoran un poco con las variables originales que, con la adición de las creadas, sobre todo en términos de *accuracy*. El modelo de Árbol de decisión parece beneficiarse más de esta mejora con una mejora del 4% de *accuracy*. Sin embargo, la diferencia en el rendimiento no es significativa, y cualquier elección entre usar las variables originales o las creadas probablemente debería basarse en consideraciones adicionales, como la interpretabilidad del modelo, el coste de cálculo o principios como el de parsimonia (Alfonso et al., 2013).

3.6.2. Equipos de balonmano femeninos

En la Tabla 3.5 se reflejan los resultados obtenidos con todos los modelos obtenidos a partir de los datos de competiciones femeninas de balonmano en los cuales, salvo algunos detalles se puede observar que los dos modelos principales, Regresión Logística y

Random Forest ofrecen buenos comportamiento y bastante similares en cuanto al uso de las variables creadas como el uso exclusivo de las originales.

TABLA 3.5. RESUMEN DEL RENDIMIENTO DE LOS MODELOS, EQUIPOS FEMENINOS.

Modelos	Variables	Precision (No)	Recall (No)	Precision (Si)	Recall (Si)	Acc.	AUC
Árbol de decisión	Creadas	0.84	0.85	0.83	0.83	0.84	0.88
	Originales	0.84	0.87	0.83	0.73	0.80	0.87
Regresión Logística	Creadas	0.87	0.87	0.85	0.85	0.86	0.96
	Originales	0.87	0.87	0.85	0.85	0.86	0.96
XGBoost	Creadas	0.85	0.89	0.87	0.83	0.86	0.93
	Originales	0.88	0.91	0.90	0.85	0.89	0.93
Random Forest	Creadas	0.87	0.89	0.88	0.85	0.87	0.93
	Originales	0.84	0.91	0.89	0.80	0.86	0.93
Naive Bayes	Creadas	0.89	0.91	0.90	0.88	0.90	0.95
	Originales	0.89	0.89	0.88	0.88	0.89	0.95

Se puede observar que los modelos evaluados exhiben ligeramente diferentes niveles de rendimiento: el modelo *Naive Bayes* con variables creadas lidera con un *accuracy* del 90%, seguido por *XGBoost* que alcanza el 89% con variables originales, pero cae al 86% con variables creadas. La Regresión Logística mantiene un rendimiento sólido y consistente en ambos casos con un 86%, mientras que *Random Forest* experimenta una leve disminución del 87% al 86% al utilizar variables originales. Por último, el Árbol de Decisión muestra el rendimiento más bajo, con un 84% de precisión para variables creadas y un 80% para las originales.

Hay que recordar que la precisión general no determina qué modelo es mejor ya que hay que tener en cuenta muchos otros factores como qué clase predice mejor y cuál conviene más al estudio, la minimización de falsos positivos, sobreajuste, etc.

3.7. Conclusiones

Tras analizar distintos modelos con variaciones, se extraen las siguientes conclusiones para guiar el estudio. Los modelos presentan buenos resultados, con precisiones y AUC superiores al 80%. No obstante, existe la preocupación del sobreajuste. Aunque los árboles de decisión y Naive Bayes muestran buenos desempeños, tienen limitaciones. El primero es simple y quizá no ofrezca el matiz requerido sobre las probabilidades, mientras que el segundo presupone la independencia de las variables, lo que no se cumple aquí.

La incorporación de nuevas variables no mejora significativamente los modelos. En la regresión logística, la prueba de razón de verosimilitud (Alba & Molina, 2017) sugiere que un modelo más complejo no aportaría mejoras significativas de rendimiento, por lo que no tendría sentido, en este caso, añadir las variables que se crearon.

Es relevante la distinción entre sexos en los análisis. Se evidencian diferencias significativas con la prueba *Mann-Withney*, particularmente en variables como 'Attacks', '%Eff', '%Lanz', 'TO', 'Goals' y 'RendimientoNeto' con un p-valor menor a 0.05. Esto contradice la hipótesis inicial de que las distribuciones entre partidos masculinos y femeninos son iguales. Sin embargo, para '%GK', '%TO', 'Diff' e 'ImpactoNeto', no hay diferencias estadísticas entre géneros.

4. Aplicación del modelo

Una vez analizado el rendimiento de los modelos se van a someter a datos nuevos para ver cómo se adaptan a datos de otra categoría como es el caso de un partido de la División Oro Española de Balonmano. Concretamente, se utilizarán datos "*play-by-play*" en el que se registran los eventos que ocurren a lo largo de un partido, el equipo que los ejecuta y los jugadores/as que están en pista en ese momento. El objetivo es lograr calcular una probabilidad de victoria del equipo a lo largo del tiempo y ver qué influye en ella.

4.1. Comprensión de los datos

Los datos para utilizar en este caso son es un registro de los eventos básicos que se producen durante un partido, obtenido a partir de la aplicación handball.ai[4], cuyo punto fuerte es la posibilidad de llevar un registro exhaustivo de los eventos que acontecen en un partido. Se utilizará un ejemplo perteneciente a un partido entre Balonmano Pozuelo y Balonmano Oviedo en División Oro femenina.

4.2. Preparación de los datos

En este caso, estos datos son de otra naturaleza, por lo que han de tener un proceso más largo hasta conseguir que los datos estén en el mismo formato que los anteriores y proceder a la aplicación. Este proceso se explica con detalle en el Anexo X.

4.3. Cálculo de las probabilidades de ganar *in-game*

El siguiente paso será usar los modelos entrenados anteriormente para calcular la probabilidad de victoria en los partidos cuyos datos han sido procesados en el paso anterior. Después se compararán una con la otra para ver una imagen del partido. Además, si un profesional que haya visto ese partido analiza las gráficas puede determinar si se ajustan a la realidad y muestran una imagen fiel del enfrentamiento, lo que serviría para dar coherencia a los modelos a parte de la validación vista anteriormente.

En este caso se cuenta con el análisis de Eusebio Angulo, entrenador del BM Pozuelo de Calatrava, División Oro, además de profesor de estadística en la Universidad de Castilla-La Mancha (UCLM) que tendrá el papel de comprobar el ajuste de los resultados de los

[4] https://handball.ai/

modelos. Se han analizado sólo los modelos elegidos tras las conclusiones del punto anterior, es decir, Regresión Logística y *Random Forest* dado que el resto ofrecen resultados escasamente interpretables de la evolución de las probabilidades de ganar.

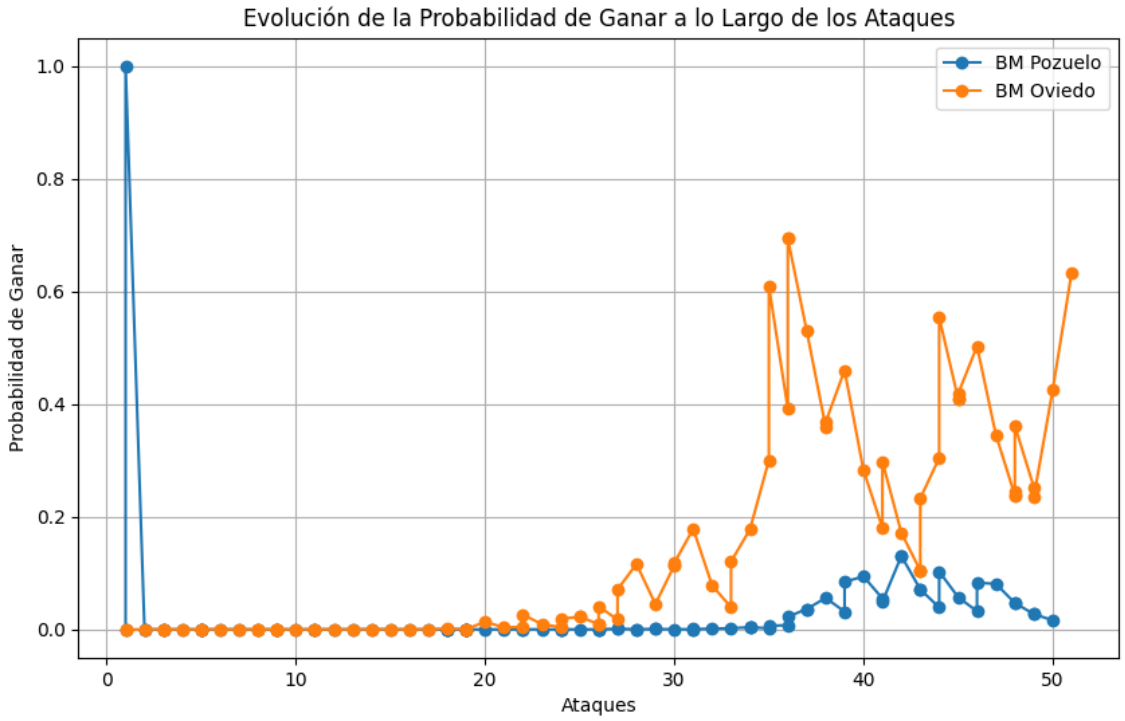
Primero, se compararán varias estadísticas clave de los dos equipos para ver si el resultado tiene coherencia con ellas (véase en Anexo 5). Después, se usarán los modelos para determinar si el equipo gana o pierde dependiendo de las estadísticas que tenga y se creará una nueva columna con ellas en cada momento del tiempo ('Probabilidad_Ganar'). Una vez completado este paso se podrán comparar las gráficas de probabilidad de victoria a lo largo del tiempo. También es interesante utilizar Lime para analizar momentos clave del partido. Concretamente, se elegirán los ataques en los que se suele introducir un tiempo muerto del modelo que menor coeficiente de variación (CV) consiga. Estos son:

- Ataque 25: Allá por el final de la primera parte.
- Ataque 35: Suele coincidir con el minuto 15-20 (dependiendo del ritmo).
- Ataque 45: Coincide con los últimos minutos del encuentro.

4.3.1. Regresión Logística

La regresión logística es uno de los modelos que mejor consigue clasificar los datos, pero al introducir los datos del *play-by-play* se pueden apreciar aspectos no del todo concluyentes. El partido en cuestión fue bastante igualado, con un desempeño generalmente bueno por los dos equipos, pero el modelo otorga probabilidades generalmente bajas en los dos equipos.

FIGURA 4.1. PROBABILIDADES MODELADAS POR REGRESIÓN LOGÍSTICA CON VARIABLES ORIGINALES.

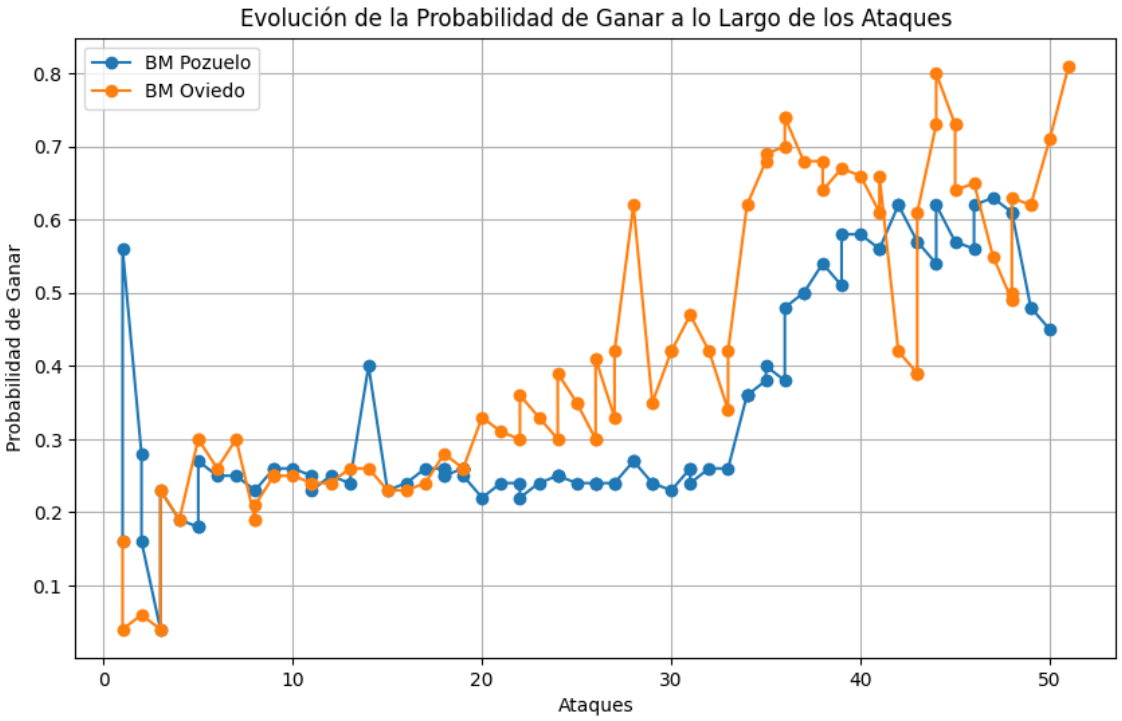


Como se puede observar al principio las probabilidades de victoria son muy altas al principio del partido (con marcar el primer gol y parar el primer balón habría un 100% de eficacia y de paradas por lo que probabilidad de victoria será 1), esto se dará en todos los modelos, pero luego se va a ajustando y se aprecia que las probabilidades de ambos equipos de conseguir la victoria son muy bajas durante el todo el partido, llegando en el mejor momento de Oviedo a un 60% de posibilidades, lo cual no se puede considerar representativo del partido.

4.3.2. Random Forest

Puede que sea debido a que es más robusto ante la multicolinealidad o que se ajusta mejor a estos datos, pero el modelo de *Random Forest* representa de una manera más fiel que cualquier otro la realidad que se vivió durante el enfrentamiento.

FIGURA 4.1. PROBABILIDADES MODELADAS POR *RANDOM FOREST* CON VARIABLES ORIGINALES.



Se aprecia mucha menos volatilidad, aunque se ajusta bien a las variaciones en las estadísticas y muestra un acercamiento entre los dos equipos más fiel a lo que realmente se vivió en la pista. Ambos equipos tienen bastantes probabilidades de ganar porque ambos hicieron un buen partido, pero el algoritmo se decanta por BM Oviedo, que fue el que finalmente se acabó por llevar la victoria.

4.4. Análisis Comparativo

A continuación, se muestra una tabla con los CV de todo el partido, primera mitad y segunda mitad. Esto se hace debido a que se necesita que el modelo consiga otorgar probabilidades de ganar en todos los momentos del partido, aunque esta sea baja para poder hacer un método de ponderación para evaluar jugadores a raíz del modelo y porque es no puede haber demasiada volatilidad o variación a lo largo del partido, se necesita una ponderación más o menos estable. Este criterio se establece por lógica, ya que, cuando se ha transcurrido medio partido, marcar o fallar un gol no debería incurrir en una variación escandalosa en su probabilidad de ganar.

TABLA 4.1. COMPARACIÓN DE LOS CV DE TODOS LOS MODELOS Y VARIACIONES DE ESTOS

Modelo	Variables	Equipo B			Equipo A		
		Toda la serie	**1ª Parte**	**2ª Parte**	**Toda la serie**	**1ª Parte**	**2ª Parte**
Árbol de Decisión	Originales	1,719	6,481	1,047	1,366	No Aplica	0,658
	Creadas	4,050	6,481	3,162	1,179	No Aplica	0,441
Regresión Logística	Originales	3,861	6,473	1,113	1,288	1,930	0,609
	Creadas	3,860	6,474	1,122	1,299	1,958	0,619
Random Forest	Originales	**0,434**	**0,276**	**0,335**	**0,466**	**0,304**	0,231
	Creadas	**0,507**	**0,485**	**0,391**	**0,581**	**0,741**	**0,137**
XGBoost	Originales	1,706	6,067	1,041	1,325	2,342	0,644
	Creadas	1,833	5,812	1,032	1,241	1,403	0,571
Naive Bayes	Originales	0,553	0,681	0,426	0,556	0,860	0,159
	Creadas	0,787	0,981	0,595	0,662	1,237	**0,059**

Se puede apreciar que, *Random Forest* es el modelo que mejor responde en términos de variación en ambos equipos. Por lo que esto unido a la coherencia que da la evolución que representa la gráfica de probabilidades es lo que lo convierte en el modelo elegido como el que mejor es capaz de clasificar datos de estas características, además de otorgar probabilidades de victoria coherentes.

4.5. Explicabilidad

Por último, se va a hacer uso de LIME para explicar los tres momentos comentados anteriormente en los dos equipos del modelo elegido con las variables originales, debido a que son momentos que podrían ser de gran importancia a un entrenador para la toma de decisiones y poder ajustar u optimizar el rendimiento de su equipo.

Correspondiente al BM Pozuelo de Cva. (Equipo B) los momentos clave seleccionados son explicados de esta manera:

- Ataque n.º 25: La probabilidad de ganar en este ataque es muy baja (10%) debido en mayor medida por la baja eficacia del equipo hasta el momento y en menor medida por el escaso número de goles (lo cual es normal por el momento del partido)
- Ataque n.º 35: En esta ocasión hay algo más de probabilidades de victoria, aunque no son muchas (30%). La librería analiza que se debe al escaso porcentaje de lanzamientos unido a su aún escaso número de goles.
- Ataque n.º 45: En el ataque 45 la situación se torna diferente debido a que las probabilidades están bastante igualadas (47% de probabilidades de ganar). Esto se debe a su eficacia por encima del 50%, su mayor número de goles y de porcentaje de lanzamiento. El problema es la portería, en la que guardan un porcentaje de paradas demasiado baja, lo que le hace aumentar mucho sus probabilidades de perder.

Las explicaciones correspondientes al BM Oviedo (Equipo A) son las siguientes:

- Ataque n.º 25: Su 22% de probabilidades de ganar se debe a que, a pesar de que su porcentaje de portería es bueno, su eficacia y su escaso número de goles hace que el modelo se decante por la derrota de momento.
- Ataque n.º 35: En este caso ya han cambiado mucho sus posiciones, teniendo un 75% de probabilidades de ganar en este punto debido en mayor medida a su casi 50% de eficacia, 35% de balones parados y su bajo porcentaje de pérdidas.
- Ataque n.º 45: Ahora se le sigue otorgando más probabilidades de ganar que de perder, pero menos que 10 ataques atrás (65% en este caso). Esto se debe a su peor porcentaje de portería, pese a que su eficacia ha subido.

5. Método de valoración de jugadores

El modelo seleccionado anteriormente se puede utilizar para valorar la contribución de los jugadores al equipo durante el partido, dado que en todo momento se conoce qué jugadores están en cancha y la variación de probabilidad de victoria que da cada una de las jugadas. Esto es de gran utilidad debido a que se puede saber qué jugadores son clave en un equipo en cualquier instante del partido.

5.1. ¿Qué es Plus/Minus?

El método seguido es una variación del tradicional Plus/Minus (Magnus, 2019), el cual tiene en cuenta las acciones de los equipos y si son positivas o negativas mientras el jugador está en pista sin necesariamente fijarse solo en el jugador que anota. Esto abre nuevas aplicaciones al trabajo realizado como proporcionar una evaluación objetiva, informada y basada en datos del rendimiento de los jugadores de un equipo. El Plus/Minus debe ser adaptado deporte en el que se aplique (Kharrat et al., 2017).

5.2. Desarrollo de Plus/Minus WinProbability (PMWP)

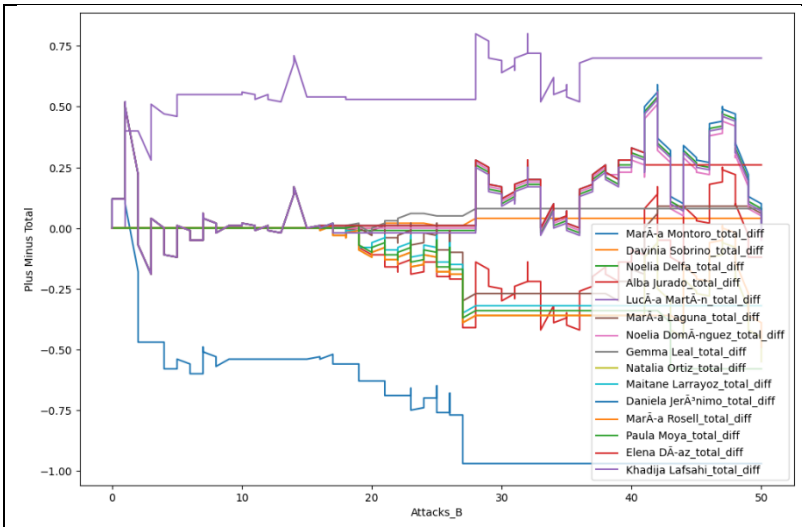
El balonmano tiene una característica particular, hay jugadores que solo se encargan de atacar, o solo de defender y otros que se encargan de ambas tareas. Para extraer una valoración justa que tenga en cuenta a todos los jugadores y evalúe además la defensa (principal sesgo de los métodos de evaluación en balonmano) se usa la probabilidad de victoria de los dos equipos. Se escoge el equipo objetivo (que se pretende analizar) y se extraen en qué momento está cada jugador en pista y se les evaluará según aumente o disminuya la probabilidad de victoria al atacar. Mientras que se analizará la inversa de la probabilidad de victoria de su rival para poder evaluar la defensa. A la suma de ambas variaciones se la llamará Plus/Minus *WinProbability*.

El proceso de cálculo se basa en el uso de varias fuentes de datos diferentes para realizar los cálculos del PMWP, estas son: estadísticas de ambos equipos y el tiempo transcurrido, las fuentes de datos resultantes de la aplicación de los modelos, es decir, con las probabilidades de ganar del equipo local y visitante, y finalmente otro conjunto de datos con cálculos de plus/minus para evaluación.

A partir de estos datos, se ejecutan una serie de transformaciones de datos, como las siguientes:

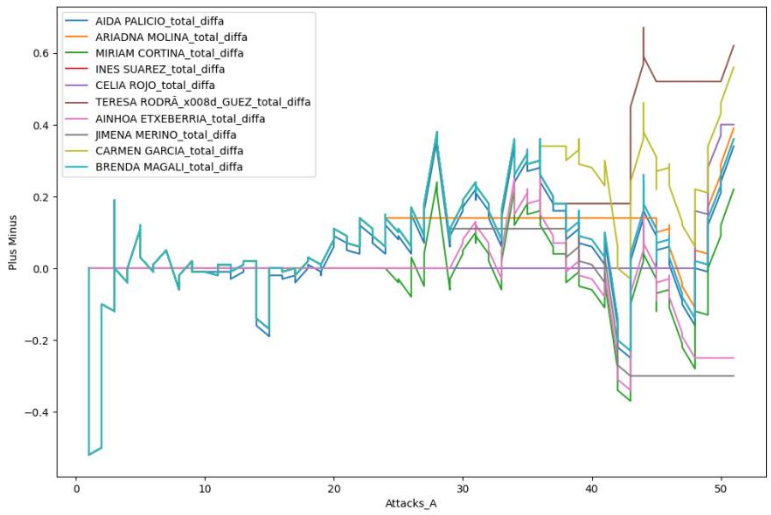
1. Procesamiento de Datos: Se generando nuevas columnas booleanas que marcan la presencia en pista de cada jugadora a partir de lo registrado en la aplicación.
2. Combinación y Sincronización: Los diferentes conjuntos de datos se combinan utilizando variables temporales (‘*Start’* y ‘*End’*) para sincronizar eventos con las probabilidades de cada equipo. Se manejan los valores faltantes y se crea una columna “diferencia” esencial para la asignación de puntajes.
3. Cálculo Acumulativo de Ataques: Se calculan ataques acumulativos para cada jugadora, ajustando las diferencias según la variación en las probabilidades de victoria del equipo.
4. Evaluación de Defensa y Valoración Final: Se realiza un procedimiento similar para evaluar la defensa de las jugadoras, considerando el negativo de la diferencia de probabilidad del equipo rival. La valoración final se determina sumando los puntos obtenidos en ataque y defensa.
5. Visualización: Se facilita una visualización gráfica de la evolución de las puntuaciones de cada jugadora y evaluación de su desempeño según lo requiera el analista.

FIGURA 5.1. EVOLUCIÓN DE PUNTUACIÓN DE TODAS LAS JUGADORAS DEL EQUIPO VISITANTE (POZUELO)



En la Figura 5.1 se puede apreciar como claramente hay una pequeña distinción entre jugadoras que valoran por encima de 0 y jugadoras que puntúan por debajo. Varias de ellas se han visto bastante penalizadas porque en este partido la defensa no fue muy buena.

FIGURA 5.2. EVOLUCIÓN DE PUNTUACIÓN DE TODAS LAS JUGADORAS DEL EQUIPO LOCAL (OVIEDO)



Aquí se puede apreciar como la mayoría de las jugadoras están por encima de la puntuación de 0, aunque algunas hacia el final del partido bajan mucho su puntuación debido a que coincide cuando Pozuelo consigue recortar la diferencia y mejorar las estadísticas.

5.3. Evaluación e Interpretación

Se va a utilizar la correlación de *Spearman* para evaluar la semejanza de los rankings hechos por otros plus/minus calculados de forma diferente para el mismo partido, de esta forma se pueden validar unos a otros, esto se conoce como dar consistencia al método. Este plus/minus se denominan como PMP y PMB.

El PMB o Plus/Minus Básico se utiliza para medir el comportamiento de un equipo con un jugador en pista. Se registra el desempeño de los jugadores en cada posesión de defensa y ataque en la que estén presentes. Cada vez que en defensa no encajen gol, los jugadores suman 1 punto, y cada vez que encajen un gol, se les resta 1 punto. En ataque, cada vez que consigan meter un gol, suman 1 punto, y cada vez que el ataque no termine en gol, se les resta 1 punto. Por otro lado, el PMP o Plus/Minus Promedio se utiliza para para medir el rendimiento de un jugador en comparación con el promedio del equipo. Mientras que PMI se calculará para cada equipo de forma individual dada la especificidad de su fórmula y con PMB se obtendrá una valoración general.

TABLA 5.1. CORRELACIONES ENTRE LOS PLUS/MINUS.

Correlaciones	PMI		PMB
	Pozuelo	Oviedo	Ambos Equipos
PMWP	90,71%	78,33%	81,61%

Ante estos resultados se puede concluir de forma preliminar que el modelo de valoración es capaz de evaluar la actuación de las jugadoras de forma bastante precisa, consiguiendo resultados parecidos a métodos más específicamente pensados para ello,

6. Conclusiones y trabajo futuro

Para finalizar este estudio, se va a dar paso a explorar una serie de puntos clave de este, así como limitaciones que han surgido en su realización.

Como conclusiones, en este trabajo se han recopilado y procesado diversas fuentes de datos de partidos y estadísticas de balonmano para satisfacer los requisitos del análisis. Durante el proceso, se identificaron variables clave para la creación de modelos; sin embargo, a pesar de su importancia, fueron eliminadas por su baja interpretabilidad y porque no mejoraban significativamente el rendimiento de los modelos, optándose por modelos más sencillos conforme al principio de parsimonia. Se llevó a cabo un exhaustivo análisis exploratorio, revelando diferencias significativas y patrones relevantes entre sexos. Diferentes modelos de clasificación fueron entrenados, evaluados y comparados, seleccionando el más adecuado para modelizar los datos mediante criterios estadísticos uniformes. Posteriormente, se calculó la probabilidad de victoria "*in-game*" y se implementó un sistema de valoración, cuyos resultados fueron validados estadísticamente y corroborados por la opinión de un experto, lo que fortalece su validez.

Este estudio, aunque prometedor, posee limitaciones y abre nuevas líneas de investigación. Una limitación reside en la utilización de la función "*predict_proba*" de *scikit-learn* para estimar la probabilidad de pertenencia a una clase, que podría no ser precisa y requerir técnicas de calibraBBción, como la de Platt o isotónica, para reflejar de manera más exacta la verdadera probabilidad (Johansson, 2021). Además, existe una oportunidad significativa de probar y validar la robustez del modelo con más datos, especialmente masculinos, e incorporar nuevas variables con menor correlación, lo que contribuiría a integrar mayor robustez y reducir el error no modelizable.

Bibliografía

Alfonso, C., García, L., & García, N. B. (n.d.). *El principio de parsimonia en la ciencia cognitiva actual: Riesgos y soluciones*. www.cienciacognitiva.org

Anguera, M. T., & Hernández Mendo, A. (2015). Técnicas de análisis en estudios observacionales en ciencias del deporte. Cuadernos de Psicología del Deporte, 15(1), 13–30. Recuperado a partir de https://revistas.um.es/cpd/article/view/223011

Angulo, E., Romero, F.P., & López-Gómez, J.A. (2022). A comparison of different soft-computing techniques for the evaluation of handball goalkeepers. *Soft Computing*, *26*(6), 3045-3058. https://link.springer.com/content/pdf/10.1007/s00500-021-06440-7.pdf

Aoki, R. Y. S., Assunção, R. M., & Vaz De Melo, P. O. S. (2017). *Luck is hard to beat: The difficulty of sports prediction. Paper presented at the Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, , Part F129685 1367-1375. doi:10.1145/3097983.3098045 Retrieved from www.scopus.com

Autor desconocido. (2020*). ¿Para qué sirve Python? ¿Qué es y cuáles son sus usos? Universia*. https://www.universia.net/es/actualidad/orientacion-academica/para-que-sirve-phyton-que-es-y-usos-1154393.html

Breiman, L. (2001). *Random Forests* (Vol. 45).

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-August-2016*, 785–794. https://doi.org/10.1145/2939672.2939785

Granitto, P. M., Furlanello, C., Biasioli, F., & Gasperi, F. (2006). Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. *Chemometrics and Intelligent Laboratory Systems*, *83*(2), 83–90. https://doi.org/10.1016/j.chemolab.2006.01.007

Hill, S. E. (2022). In-game win probability models for canadian football. Journal of Business Analytics, 5(2), 164-178. doi:10.1080/2573234X.2021.2015252

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R Second Edition*.

Johansson, U., Löfström, T., & Boström, H. (2021, September). Calibrating multi-class models. In *Conformal and Probabilistic Prediction and Applications* (pp. 111-130). PMLR.

Kovalchik, S. A. (2023). *Annual Review of Statistics and Its Application Player Tracking Data in Sports*. https://doi.org/10.1146/annurev-statistics-033021

Mirabal Sosa, MayelÍn, Robaina García, Maytee, & Uranga Piña, Rolando. (2010). R: una herramienta poco difundida y muy útil para la investigación clínica. *Revista Cubana de Investigaciones Biomédicas*, *29*(2), 302-308. Recuperado en 09 de septiembre de 2023, de http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0864-03002010000200012&lng=es&tlng=es.

Müller, O., Caron, M., Döring, M., Heuwinkel, T., & Baumeister, J. (n.d.). *PIVOT: A Parsimonious End-to-End Learning Framework for Valuing Player Actions in Handball using Tracking Data*. https://www.sg-flensburg-handewitt.de

López-Gómez, J.A., Romero, F.P., & Angulo, E. (2022). A feature-weighting approach using metaheuristic algorithms to evaluate the performance of handball goalkeepers. *IEEE Access*, *10*, 30556-30572. https://ieeexplore.ieee.org/abstract/document/9726167

Oytun, M., Tinazci, C., Sekeroglu, B., Acikada, C., & Yavuz, H. U. (2020). Performance prediction and evaluation in female handball players using machine learning models. IEEE Access, 8, 116321-116335. doi:10.1109/ACCESS.2020.3004182

Romero, F.P., Angulo, E., Serrano-Guerrero, J., & Olivas, J.A. (2020). A fuzzy framework to evaluate players' performance in handball. *International Journal of Computational Intelligence Systems*, *13*(1), 549-558. https://www.atlantis-press.com/article/125941017.pdf

Silva Fuente-Alba, C, & Molina Villagra, M. (2017). *Likelihood ratio (razón de verosimilitud): definición y aplicación en Radiología. Revista argentina de radiología*, *81*(3), 204-208. https://dx.doi.org/10.1016/j-rard.2016.11.002

Santra, A. K., & Christy, C. J. (2012). *Genetic Algorithm and Confusion Matrix for Document Clustering*. www.IJCSI.org

Schapire, M. C. R., & da Silva, A. N. R. (2022). Barrier Effect in a Medium-Sized Brazilian City: An Exploratory Analysis Using Decision Trees and Random Forests. *Sustainability (Switzerland)*, *14*(10). https://doi.org/10.3390/su14106309

Schapire, R. E. (1990). *The Strength of Weak Learnability* (Vol. 5).

Schwenkreis, F. (2020). *Why the concept of shopping baskets helps to analyze team-handball. Paper presented at the 2020 International Conference on Intelligent Data Science Technologies and Applications*, IDSTA 2020, 4-10. doi:10.1109/IDSTA50958.2020.9264068 Retrieved from www.scopus.com

Thiago-gsantos. (2020). *Google Colab: ¿Qué es y cómo usarlo? Alura Cursos.* https://www.aluracursos.com/blog/google-colab-que-es-y-como-usarlo

Van, J., Ku Leuven, H., Robberechts, P., van Haaren, J., & Davis, J. (2019). *Who Will Win It? An In-game Win Probability Model for Football*. http://espn.com/nba/game?gameId=401071795

Wagner, H., Hinz, M., Melcher, K., Radic, V., & Uhrmeister, J. (2023). The PlayerScore: A Systematic Game Observation Tool to Determine Individual Player Performance in Team Handball Competition. *Applied Sciences (Switzerland)*, *13*(4). https://doi.org/10.3390/app13042327

Wirth, R., & Hipp, J. (2000, April). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical aplicaciones of knowledge discovery and data mining* (Vol. 1, pp. 29-39).

Anexos

En este último capítulo se explorarán más en profundidad algunos de los apartados del cuerpo del estudio para facilitar y solidificar su comprensión.

Anexo 1. Metodología

La metodología utilizada para la realización del estudio está basada en un proceso estándar de actuación para procesos basados en minería de datos. El concepto CRISP-DM (Wirth, 2000), son las siglas que dan lugar a *Cross-Industry Standard Process for Data Mining*, es decir, hace referencia a un método en el que basarse para un correcto trabajo de minería de datos. En lo que a metodología se refiere, provee de descripciones de las fases necesarias de un proyecto de este estilo, así como las tareas necesarias en cada fase. Como modelo de proceso, otorga un resumen del ciclo vital de minería de datos.

Este ciclo representado en la imagen contiene las fases que indican los ámbitos más importantes y su relación. La secuencia entre fases no es estricta, aunque es recomendable llevar un orden, comenzando con el entendimiento del negocio hasta llegar al despliegue del modelo en cuestión.

Las fases de esta metodología son:

1. Comprensión del negocio (*Business Understanding*): Es la fase inicial normalmente y se define el objetivo del proyecto, en este caso se ha hecho acopio de información para poder entender el caso y definir los objetivos detallados anteriormente.
2. Comprensión de los datos (*Data Understanding*): En esta etapa se formulan hipótesis sobre qué analizar para descubrir información valiosa que puedan ayudar a la consecución de objetivos, así como las

limitaciones que surjan para saber qué datos recopilar.

3. Preparación de los datos (*Data Preparation*): Se recopilan los datos relevantes y se preparan para la manipulación eficaz de minería de datos. Esto incluye reducción de dimensionalidad, tratamiento de valores nulos, filtrado, generación de nuevas variables, transformaciones, etc.

4. Modelado (*Modeling*): En esta fase se construye un flujo de trabajo para encontrar los algoritmos deseados y ejecutar la tarea de minería de datos de entrenamiento.

5. Evaluación (*Evaluation*): En esta etapa, el modelo entrenado se prueba contra conjuntos de datos reales dentro de un escenario real o simulado, con datos nuevos o que no hayan sido usados para el entrenamiento o modelado previo.

Despliegue (*Deployment*): Después de una evaluación exitosa del modelo entrenado, se implementa en la producción y se realiza un seguimiento de cómo actúa el modelo en situaciones reales.

Anexo 2. Marco tecnológico

Python[5]: Las librerías más utilizadas para la manipulación, modelado y visualización de datos han sido:

- Pandas[6]: Pandas es una biblioteca de manipulación y análisis de datos que proporciona estructuras de datos flexibles y eficientes, como *DataFrames*.
- Numpy[7]: Numpy es utilizada para el cálculo numérico y científico.
- Seaborn[8]: Seaborn es una biblioteca de visualización de datos construida sobre matplotlib.
- Matplotlib.pyplot[9]: Es una biblioteca ampliamente utilizada para la generación de gráficos. Pyplot es un módulo de matplotlib que proporciona una interfaz similar a MATLAB para crear gráficos y visualizaciones.
- Lime[10]: Permite ver en detalle una de las observaciones del *data.frame* y explica en qué se ha basado el modelo para categorizarlo en una u otra clase.

R: Las librerías utilizadas para este trabajo han ido enfocadas al análisis exploratorio de datos, otro de los fuertes del programa:

- readxl[11]: Esta biblioteca facilita la lectura de archivos Excel. (tanto .xls como .xlsx).
- ggplot2[12]: Es uno de los paquetes más populares y utilizados de R. Permite crear visualizaciones complejas y personalizadas.
- dplyr[13]: Proporciona funciones que facilitan operaciones como seleccionar, filtrar, agrupar y modificar *dataframes*.
- tidyr[14]: Se utiliza para transformar datos a formatos *"tidy"* o limpios.

[5] https://docs.python.org/es/3/tutorial/
[6] https://pandas.pydata.org/docs/user_guide/index.html
[7] https://numpy.org/doc/stable/user/whatisnumpy.html
[8] https://seaborn.pydata.org/
[9] https://matplotlib.org/stable/tutorials/introductory/pyplot.html
[10] https://lime.readthedocs.io/en/latest/
[11] https://readxl.tidyverse.org/reference/read_excel.html
[12] https://www.rdocumentation.org/packages/ggplot2/versions/3.4.3
[13] https://dplyr.tidyverse.org/
[14] https://tidyr.tidyverse.org/

- corrplot[15]: Es una biblioteca para realizar matrices de correlación.
- gridExtra[16]: Permite combinar gráficos individuales en un solo gráfico compuesto.
- plotly[17]: Facilita la creación de gráficos interactivos basados en la biblioteca de JavaScript Plotly.
- GGally[18]: Proporciona funciones para crear gráficos de pares y otras visualizaciones relacionadas que no están disponibles en ggplot2.
- skimr[19]: Ayuda a proporcionar estadísticas resumidas de tus datos de una manera más amigable y comprensible.

[15] https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html
[16] https://www.rdocumentation.org/packages/gridExtra/versions/2.3
[17] https://plotly.com/r/
[18] https://www.rdocumentation.org/packages/GGally/versions/1.5.0
[19] https://www.rdocumentation.org/packages/skimr/versions/2.1.5

Anexo 3. Marco Teórico

En este anexo se ahondarán más en ciertos conceptos que se han visto de forma más superficial en el cuerpo del estudio, por si es de interés para el lector profundizar en estos.

Anexo 3.1. Evaluación de métodos de clasificación

Se usan para poder elegir de manera informada qué modelo de clasificación es el que mejor clasifica cada conjunto particular de datos, en este estudio se usan principalmente dos métodos:

Anexo 3.1.1. Matriz de confusión

Una matriz de confusión, que también es conocida como tabla de contingencia, es una herramienta que permite analizar el rendimiento de algoritmos de clasificación. Esta representa cómo un algoritmo de clasificación consigue clasificar correctamente datos que no han sido usados para entrenarlo (Santra & Josephine, 2012).

En una matriz de confusión de dos clases, se definen los siguientes elementos:

- "a": es el número de predicciones correctas de que una instancia es negativa.
- "b": es el número de predicciones incorrectas de que una instancia es positiva.
- "c": es el número de predicciones incorrectas de que una instancia es negativa.
- "d": es el número de predicciones correctas de que una instancia es positiva.

A partir de esta matriz, se pueden calcular varias métricas que proporcionan información sobre el desempeño del modelo de clasificación:

1. Exactitud (*Accuracy*): Proporción del total de predicciones que fueron correctas. *Accuracy* = (a + d) / (a + b + c + d)
2. *Recall o True Positive Rate* (TP): Proporción de casos positivos que se identificaron correctamente. TP = d / (c + d)
3. *False Positive Rate* (FP): Proporción de casos negativos que se clasificaron incorrectamente como positivos. FP = b / (a + b)
4. *True Negative Rate* (TN): Proporción de casos negativos que se clasificaron correctamente. TN = a / (a + b)

5. *False Negative Rate* (FN): Proporción de casos positivos que se clasificaron incorrectamente como negativos. FN = c / (c + d)

6. *Precision* (P): Proporción de los casos positivos predichos que se clasificaron correctamente. P = d / (b + d)

Anexo 3.1.2. Curva ROC

La curva ROC *(Receiver Operating Characterics)* es un gráfico ampliamente utilizado en estadística para evaluar el rendimiento de un clasificador binario a medida que el umbral de clasificación varía. Representa la sensibilidad (tasa de verdaderos positivos) frente a la especificidad (1- tasa de falsos positivos) para todos los posibles umbrales de decisión.

La eficacia del clasificador se compara mediante el área bajo la curva (*AUC)*. Dando por clasificador ideal aquel cuya *AUC* sea igual a la esquina superior izquierda, indicando una total clasificación de verdaderos positivos sin existencia de falsos positivos.

Anexo 4. Creación del modelo

En este capítulo, se van a explorar más a profundidad aspectos que no se pueden incluir por problemas de extensión en el cuerpo del estudio.

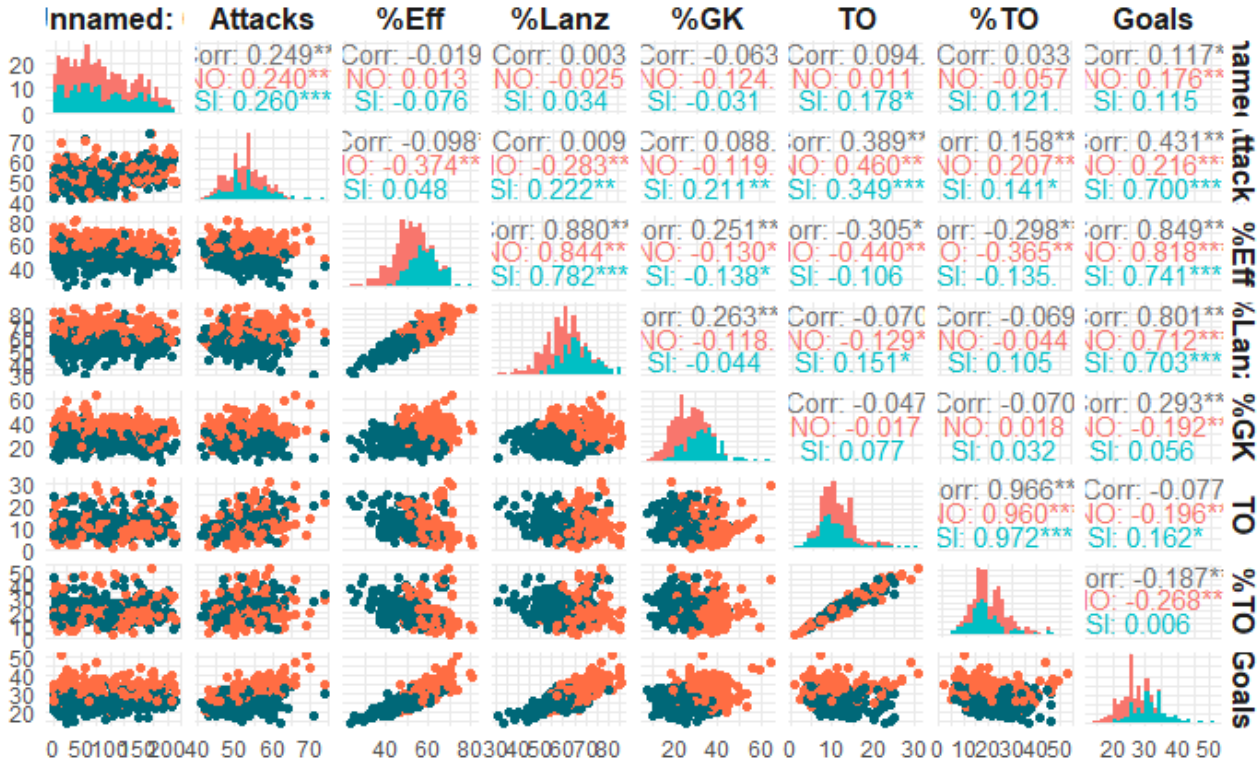
Anexo 4.1. Análisis exploratorio

En este anexo se hará una exploración algo más profunda de algunos de los apartados del análisis exploratorio para aportar a su riqueza.

Anexo 4.1. Análisis multivariante

Se plasma una visión general tanto en masculino como femenino de la relación de todas las variables cuyo color viene determinado por la variable objeto del estudio.

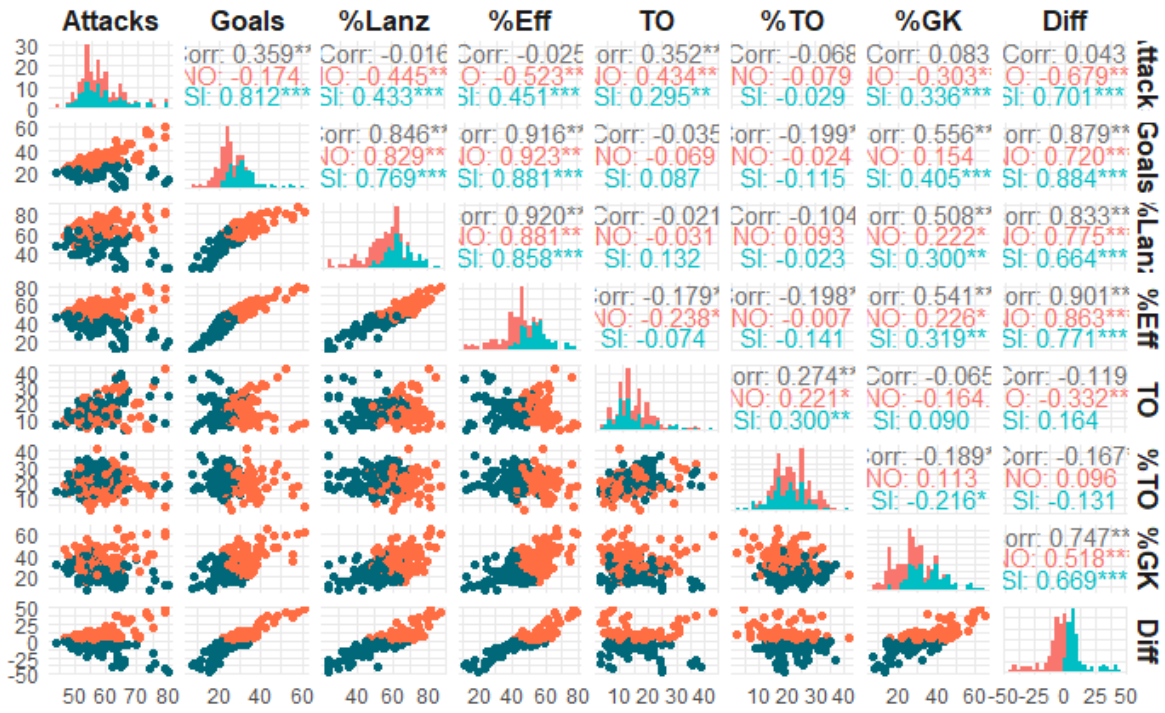
Figura anexa 1. Análisis multivariante por pares con respecto a variable objetivo. equipo masculino



Se puede observar de nuevo que, en general, a partir de una combinación dada todos los equipos ganan o pierden en la mayoría de los pares reflejados. La clave está en intentar

determinar qué influye en que pierda o gane en las franjas que coexisten ambas situaciones y que no se fruto de la pura aleatoriedad.

FIGURA ANEXA 2. ANÁLISIS MULTIVARIANTE POR PARES CON RESPECTO A VARIABLE OBJETIVO. EQUIPO MASCULINO



Siguiendo el ejemplo del caso masculino, se puede ver que en general se pueden distinguir franjas claras de victoria y derrota, pero hay intervalos en los que se entrelazan las dos situaciones, son en esos intervalos en los que interesa determinar en buena medida qué hace que se consiga una victoria o una derrota

Anexo 4.2. Evaluación

En este apartado se verán las figuras correspondientes a la evaluación de cada modelo junto a sus comentarios por si es de interés para el receptor.

En este paso se hará uso de los datos de validación para ver cuántos partidos consigue clasificar correctamente el modelo. Se mostrarán los dos mejores modelos para estos datos, el resto se podrán revisar en anexos o en el cuaderno de Colab.

Los modelos elegidos son regresión logística (debido a su mayor precisión general para clasificar) y *Random Forest* debido a que es muy buen clasificador con datos de esta

naturaleza (correlaciones muy altas) y además arroja buenas precisiones. El árbol de decisión tiene un poder predictivo bastante bajo y su finalidad, pese a haberse evaluado, era de puro análisis exploratorio. *XGBoost* es un modelo que pese a ser una buena elección tiene menos poder de clasificación con pocos datos y tiende a sobreajuste. *Naive Bayes* asume la normalidad de los datos y la independencia de las variables, asunciones con se cumplen en este caso. Sin embargo, se desplegará una tabla finalmente del resumen de las métricas de todos los modelos. Es importante destacar que para cada uno de los modelos se harán dos pruebas con dos conjuntos diferentes para cada género (conjunto con las variables originales y con las variables creadas).

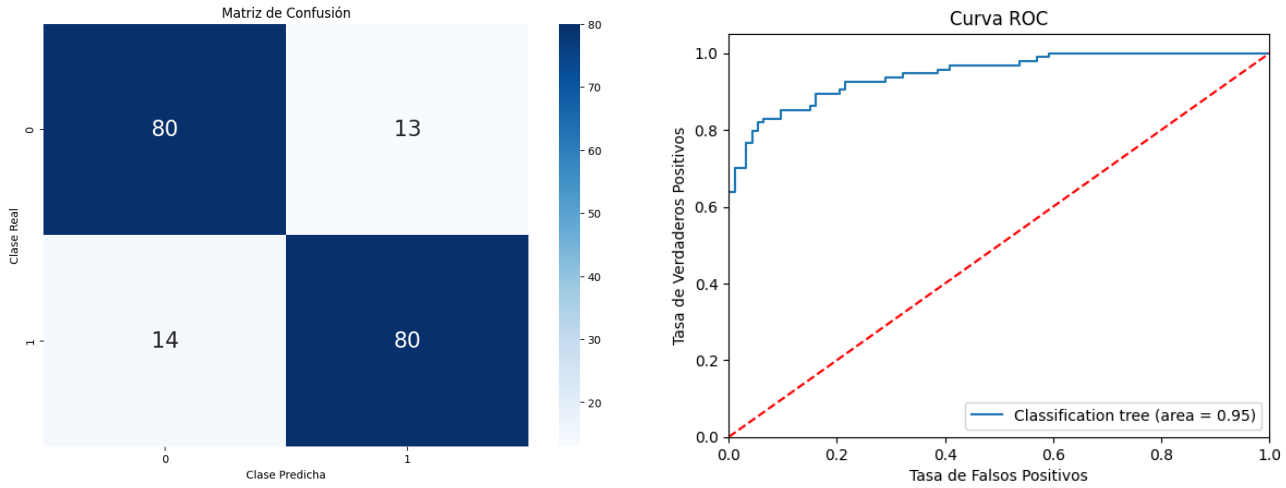
Los métodos escogidos para validar los modelos han sido la matriz de confusión, con sus métricas y la curva ROC, comentadas en el capítulo de marco teórico.

Anexo 4.2.1. Equipos de balonmano masculinos

Se van a ver en detalle los rendimientos de los modelos de regresión logística, debido a que tiene un rendimiento bastante bueno y es un modelo clásico y ampliamente utilizado. Por otro lado, *Random Forest*, debido a que es un algoritmo que puede modelizar muy bien unos datos de estas características y tiene uno de los mejores rendimientos.

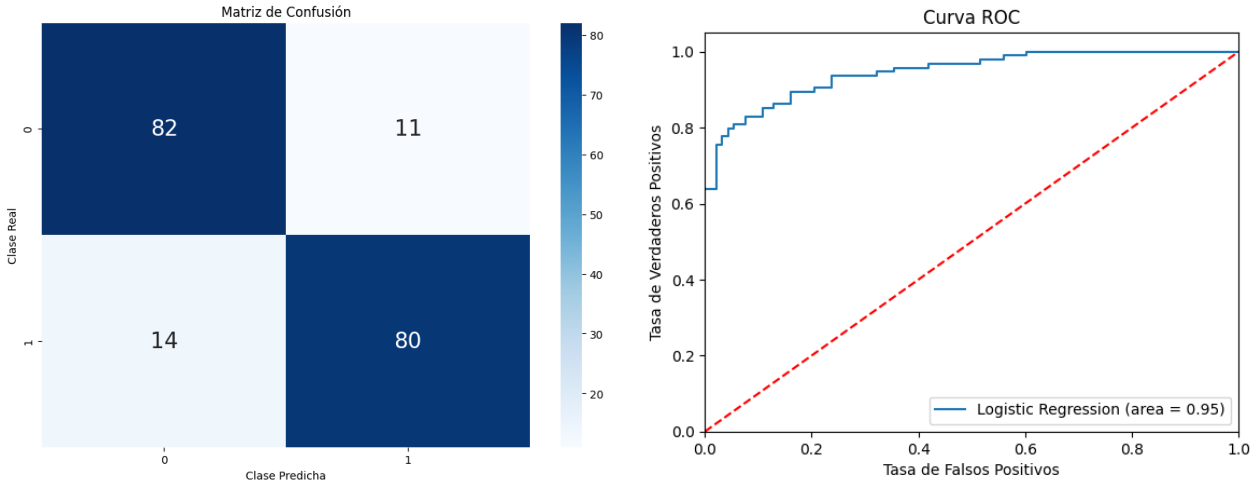
Regresión logística.

FIGURA ANEXA 3. REPORTE DE CLASIFICACIÓN CON LAS VARIABLES CREADAS



Reporte de clasificación				
	precision	recall	f1-score	support
No	0.85	0.86	0.86	93
Si	0.86	0.85	0.86	94
Accuracy			0.86	187
Macro avg	0.86	0.86	0.86	187
Weighted avg	0.86	0.86	0.86	187

FIGURA ANEXA 4. REPORTE DE CLASIFICACIÓN CON LAS VARIABLES ORIGINALES



Reporte de clasificación				
	precision	recall	f1-score	support
No	0.85	0.88	0.87	93
Si	0.88	0.85	0.86	94
Accuracy			0.87	187
Macro avg	0.87	0.87	0.87	187
Weighted avg	0.87	0.87	0.87	187

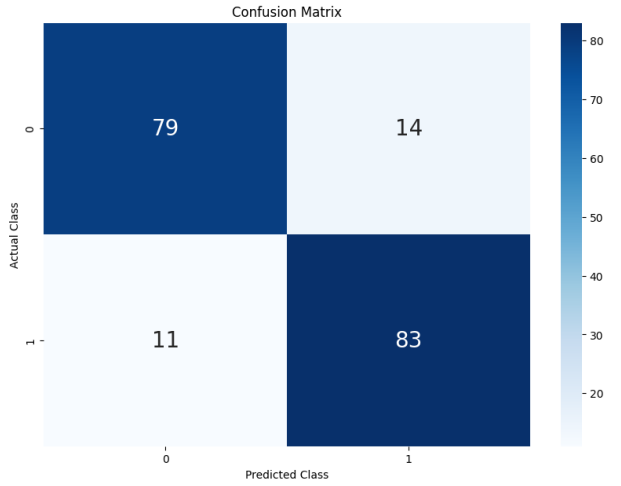
Ambos modelos de regresión muestran un rendimiento bastante similar en casi todas las métricas para ambas clases, así como su precisión global.

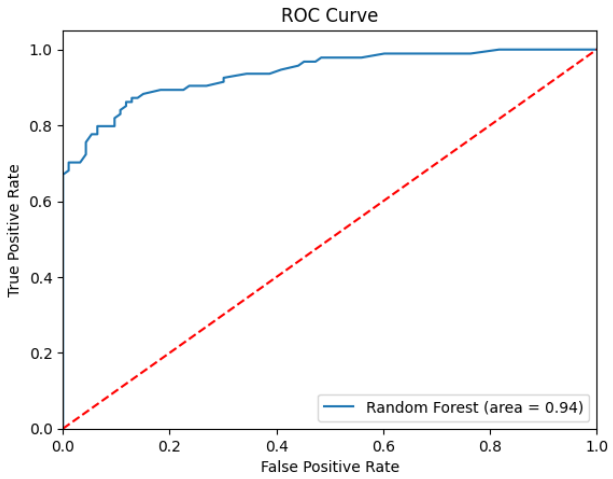
En ambos casos, el modelo tiene una mayor capacidad para predecir correctamente la clase "NO" en cuanto a *recall* (0.86 para el primer modelo y 0.88 para el segundo), lo que se traduce como que se identifica con mayor eficacia los casos de verdaderos "NO". En cuanto a la clase "SI", los modelos tienen una precisión superior (0.86 para el primer modelo y 0.88 para el segundo), es decir, que cuando el modelo predice que se va a ganar un partido, es más probable que esté en lo correcto.

Las curvas ROC, así como sus áreas bajo la curva (AUC) también tienen un rendimiento muy similar para los dos modelos, siendo la AUC ligeramente superior en el modelo de las variables originales (difiriendo en el cuarto decimal)

Random Forest

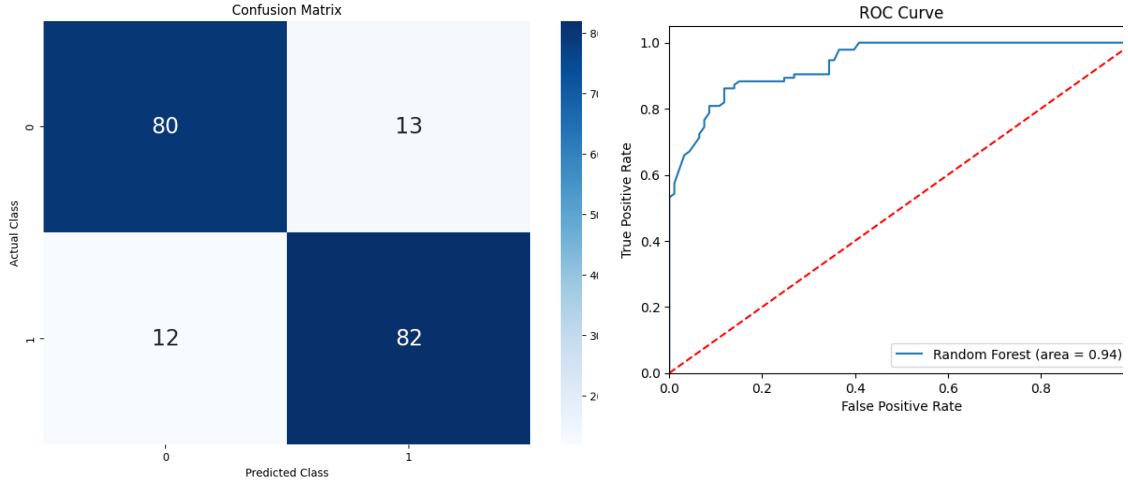
FIGURA ANEXA 5. REPORTE DE CLASIFICACIÓN CON LAS VARIABLES CREADAS





Reporte de clasificación				
	precision	recall	f1-score	support
No	0.88	0.85	0.86	93
Si	0.86	0.88	0.87	94
Accuracy			0.87	187
Macro avg	0.87	0.87	0.87	187
Weighted avg	0.87	0.87	0.87	187

FIGURA ANEXA 6. REPORTE DE CLASIFICACIÓN CON LAS VARIABLES ORIGINALES



Reporte de clasificación				
	precision	recall	f1-score	support
No	0.87	0.86	0.86	93
Si	0.86	0.87	0.87	94
Accuracy			0.87	187
Macro avg	0.87	0.87	0.87	187
Weighted avg	0.87	0.87	0.87	187

Los modelos de *Random Forest* también presentan rendimientos muy similares. Las matrices de confusión presentan un comportamiento similar al clasificar positivos y negativos

El informe de clasificación respalda esta observación, ya que las métricas de precisión, *recall* y F1-score son similares para ambas clases ('SI' y 'NO') en ambos modelos.

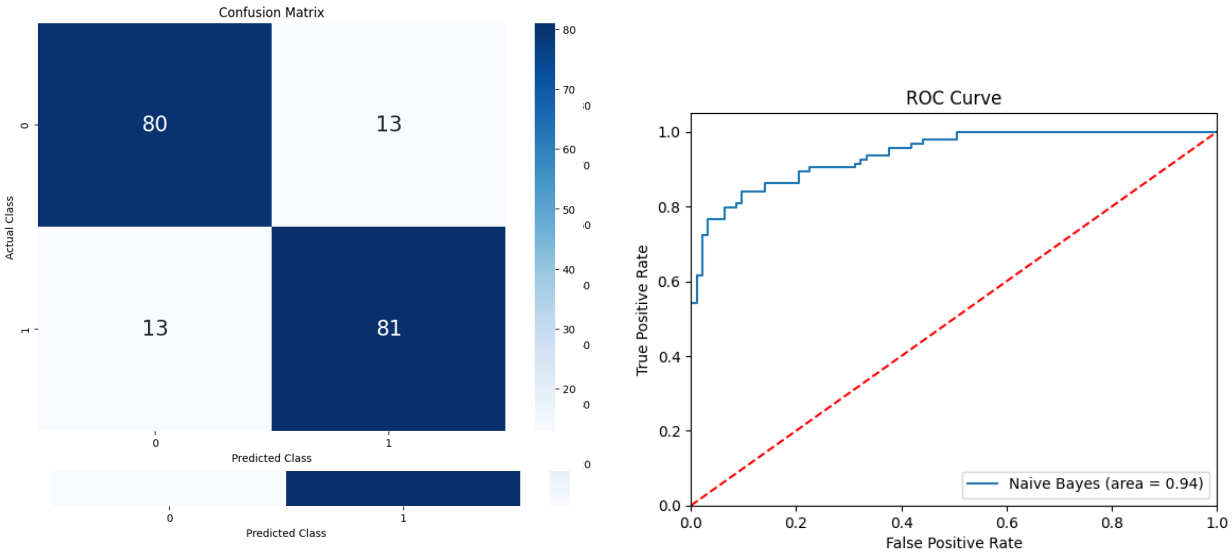
En cuanto a las métricas de la curva ROC, ambos modelos tienen tasas de verdaderos positivos (TPR) y falsos positivos (FPR) similares. El área bajo la curva ROC (AUC), una medida de la capacidad del modelo para distinguir entre las clases también es muy similar en ambos modelos (0.9384 para el modelo con variables creadas y 0.9383 para el modelo con variables originales).

En cuanto a la curva ROC, ambos modelos de nuevo tienen una tasa de verdaderos y falsos positivos bastantes similares y una AUC que difiere en el cuarto decimal a favor del modelo con las variables creadas.

Naive Bayes

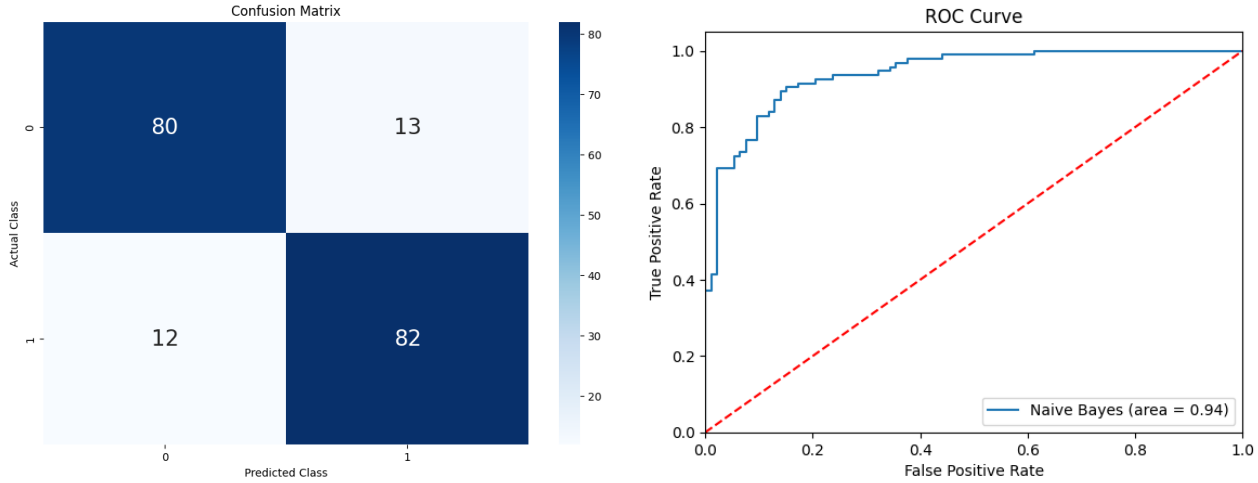
Los modelos de Naive Bayes con todas las variables y sin las variables creadas muestran una mejora significativa en su rendimiento cuando se eliminan las variables creadas.

FIGURA ANEXA 7. REPORTE DE CLASIFICACIÓN CON LAS VARIABLES CREADAS



Reporte de clasificación				
	precision	recall	f1-score	support
No	0.86	0.86	0.86	93
Si	0.86	0.86	0.86	94
Accuracy			0.86	187
Macro avg	0.86	0.86	0.86	187
Weighted avg	0.86	0.86	0.86	187

FIGURA ANEXA 8. REPORTE DE CLASIFICACIÓN CON LAS VARIABLES ORIGINALES



Reporte de clasificación				
	precision	recall	f1-score	support
No	0.87	0.86	0.86	93
Si	0.86	0.87	0.87	94

Accuracy			0.87	187
Macro avg	0.87	0.87	0.87	187
Weighted avg	0.87	0.87	0.87	187

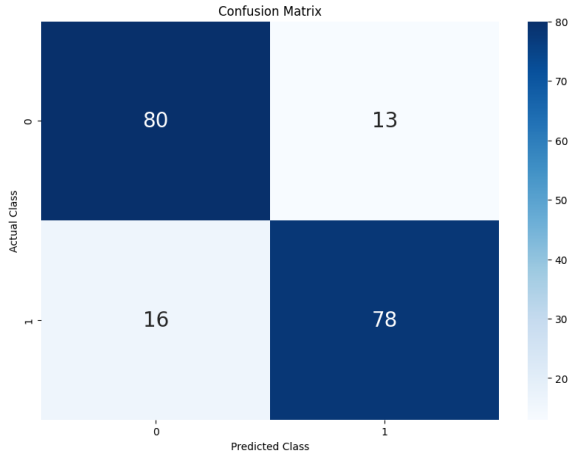
Con todas las variables: La precisión del modelo es del 86%. La matriz de confusión muestra que el modelo ha predicho correctamente 80 instancias de la clase "NO" y 81 instancias de la clase "SI", mientras que ha hecho errores en 13 instancias de ambas clases. El informe de clasificación indica que el modelo tiene una precisión y *recall* decentes para ambas clases. La Tasa de Verdaderos Positivos (TPR) y la Tasa de Falsos Positivos (FPR) muestran un comportamiento bueno en general, y el Área Bajo la Curva ROC (AUC) es bastante alta (0.94), lo que indica un buen rendimiento del modelo en la clasificación.

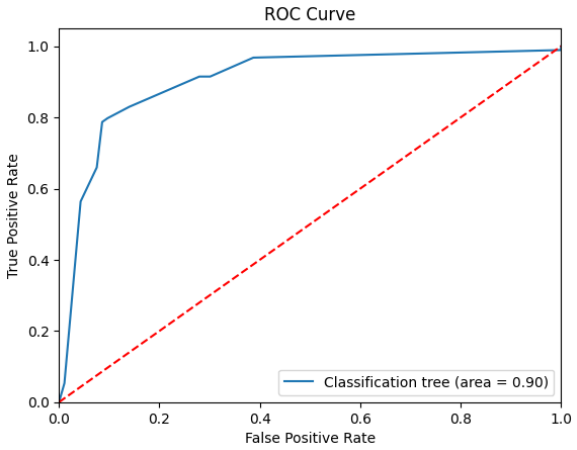
Sin las variables creadas: La precisión del modelo aumenta a un punto porcentual en precisión para la clasificación de 'SI' y en el *recall* de 'NO' lo que indica una mejora en el rendimiento. Por lo demás, tiene un comportamiento muy similar en todo.

Concluyendo, la eliminación de las variables creadas ha mejorado el rendimiento del modelo de Naive Bayes, aumentando su precisión y mejorando su capacidad para predecir ambas clases. Esto podría indicar que las variables creadas podrían haber introducido algún ruido o complejidad innecesaria en el modelo, y que su eliminación ha permitido al modelo centrarse en las variables más informativas para hacer sus predicciones.

Árboles de decisión

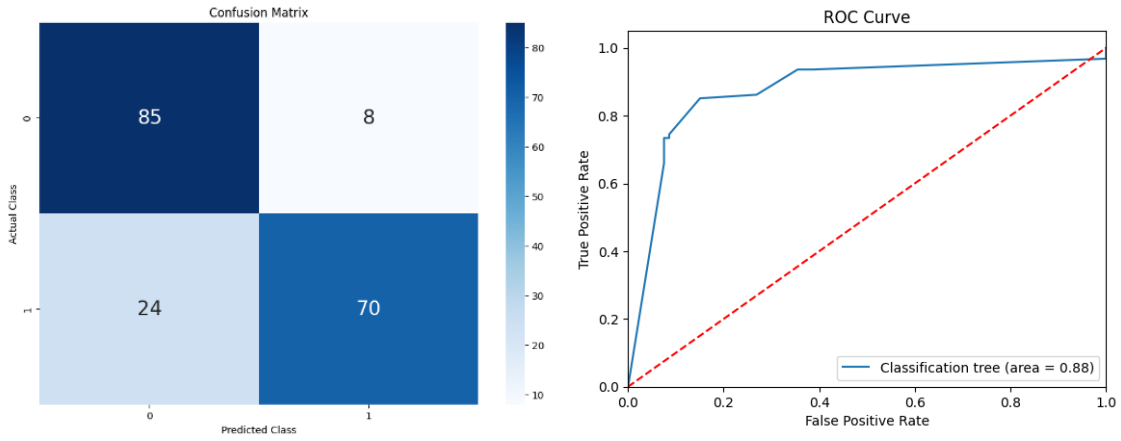
FIGURA ANEXA 9. REPORTE DE CLASIFICACIÓN CON LAS VARIABLES CREADAS





Reporte de clasificación				
	precision	recall	f1-score	support
No	0.83	0.86	0.65	93
Si	0.86	0.83	0.84	94
Accuracy			0.84	187
Macro avg	0.85	0.85	0.84	187
Weighted avg	0.85	0.84	0.84	187

FIGURA ANEXA 10. REPORTE DE CLASIFICACIÓN CON LAS VARIABLES ORIGINALES



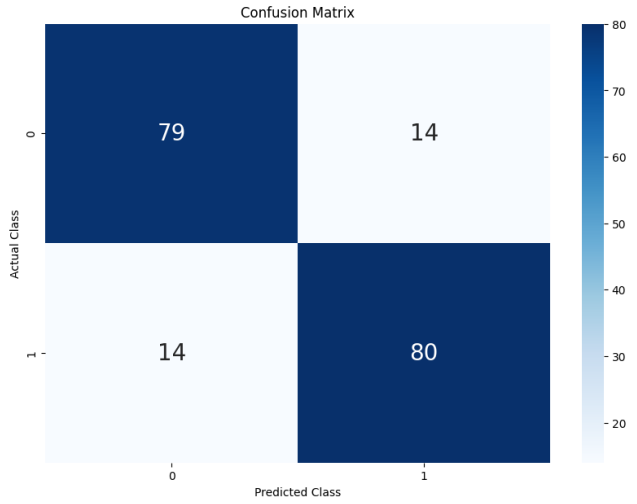
Reporte de clasificación				
	precision	recall	f1-score	support
No	0.78	0.91	0.84	93
Si	0.90	0.74	0.81	94
Accuracy			0.83	187
Macro avg	0.84	0.83	0.83	187
Weighted avg	0.84	0.83	0.83	187

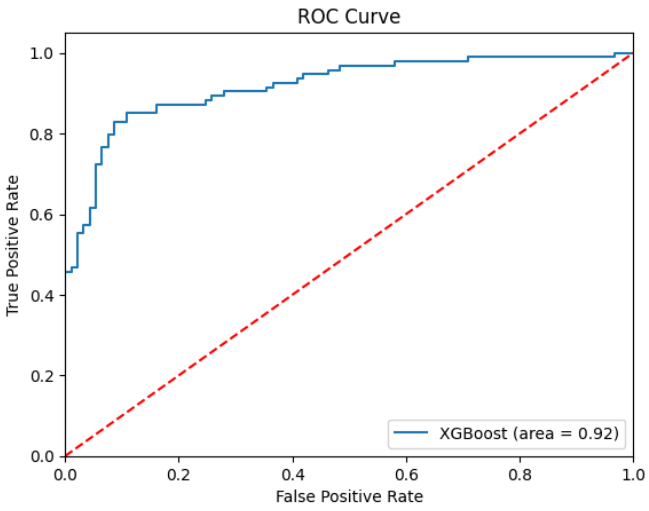
Al comparar las métricas, hay diferencias sutiles en términos de *precision* y *recall*. En el primer modelo, se obtiene una mayor precisión para la clase 'SI' y un mayor *recall* para la clase ‘NO y viceversa. Esto puede indicar una diferencia en cómo cada modelo clasifica las observaciones entre las dos clases.

Sin embargo, se observa que ambos modelos tienen un rendimiento similar en términos del área bajo la curva ROC (AUC), lo cual es una capacidad de clasificación moderadamente elevada. Esto indica que, en términos generales, ambos modelos son capaces de distinguir entre las clases de manera bastante similar

XGBoost.

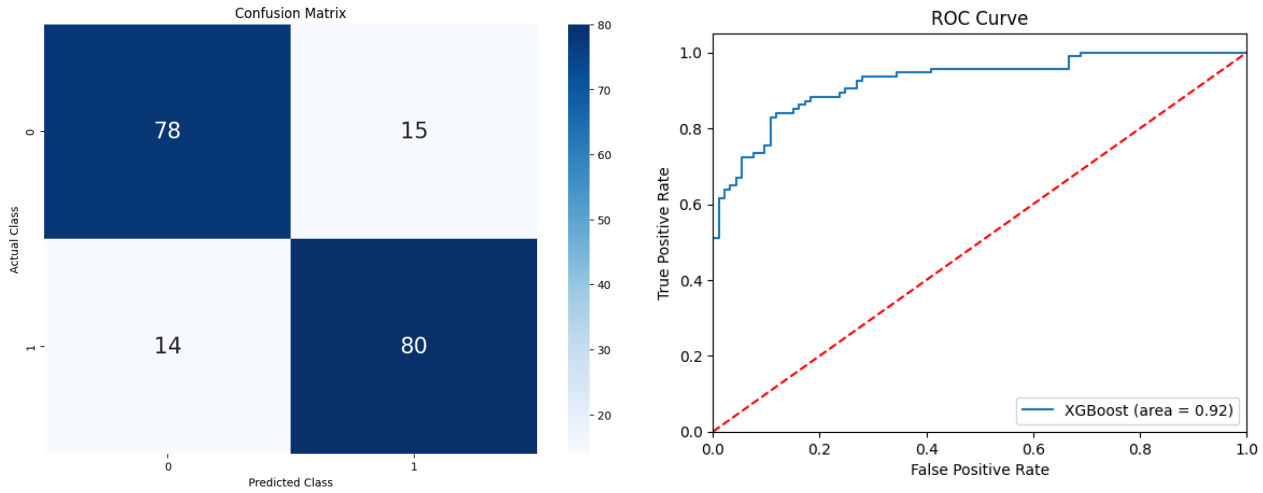
FIGURA ANEXA 11. REPORTE DE CLASIFICACIÓN CON LAS VARIABLES CREADAS





Reporte de clasificación				
	precision	recall	f1-score	support
No	0.85	0.85	0.85	93
Si	0.85	0.85	0.85	94
Accuracy			0.85	187
Macro avg	0.85	0.85	0.85	187
Weighted avg	0.85	0.85	0.85	187

FIGURA ANEXA 12. REPORTE DE CLASIFICACIÓN CON LAS VARIABLES ORIGINALES



Reporte de clasificación				
	precision	recall	f1-score	support
No	0.77	0.89	0.83	19
Si	0.88	0.74	0.80	19
Accuracy			0.82	38
Macro avg	0.82	0.82	0.81	38
Weighted avg	0.82	0.82	0.81	38

Los modelos XGBoost presentados aquí, tanto el que utiliza todas las variables como el que excluye las variables creadas, muestran un rendimiento bastante similar en términos de precisión, *recall* y F1-score para ambas clases, así como en la precisión general del modelo.

La matriz de confusión también muestra resultados similares para ambos modelos, siendo ligeramente mejor el modelo de las variables añadidas, aunque son prácticamente idénticos

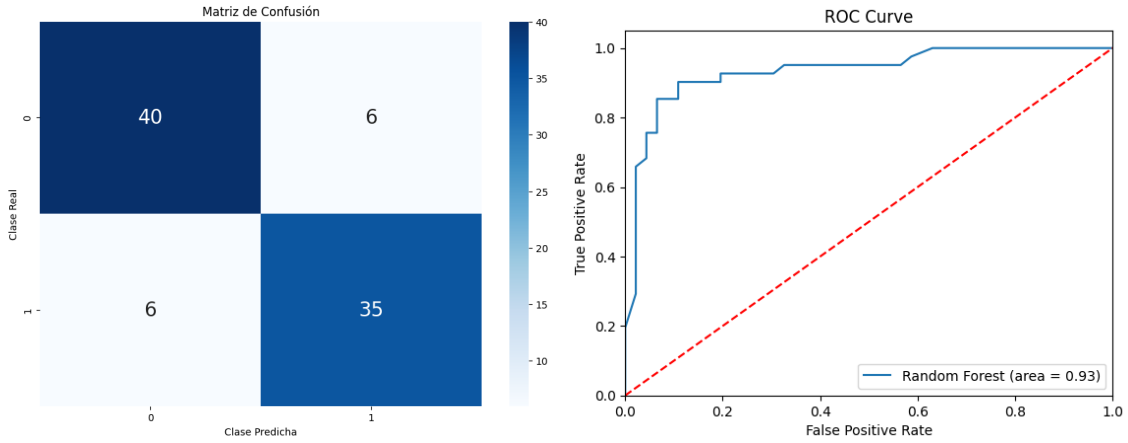
En cuanto a las curvas ROC y las áreas bajo la curva (AUC), ambos modelos tienen exactamente los mismos valores. Esto sugiere que ambos modelos tienen una capacidad de discriminación muy similar. Sin embargo, estos valores son un poco más bajos que los de los modelos de regresión logística anteriores.

Anexo 4.2.2. Equipos de balonmano femeninos

En este caso, se seguirá el mismo procedimiento que en el conjunto de datos relacionado con balonmano masculino para después comparar resultados.

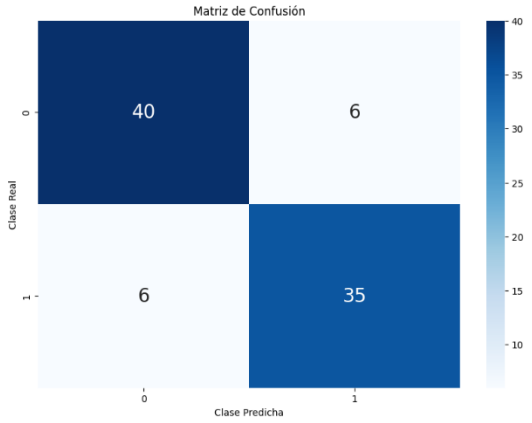
Regresión logística

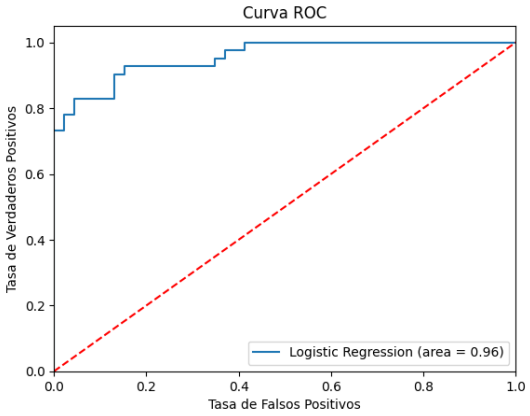
TABLA 3.9. REPORTE DE CLASIFICACIÓN CON LAS VARIABLES CREADAS



Reporte de clasificación				
	precision	recall	f1-score	support
No	0.87	0.87	0.87	46
Si	0.85	0.85	0.85	41
Accuracy			0.86	87
Macro avg	0.86	0.86	0.86	87
Weighted avg	0.86	0.86	0.86	87

TABLA 3.10. REPORTE DE CLASIFICACIÓN CON LAS VARIABLES ORIGINALES





Reporte de clasificación				
	precision	recall	f1-score	support
No	0.87	0.87	0.87	46
Si	0.85	0.85	0.85	41
Accuracy			0.86	87
Macro avg	0.86	0.86	0.86	87
Weighted avg	0.86	0.86	0.86	87

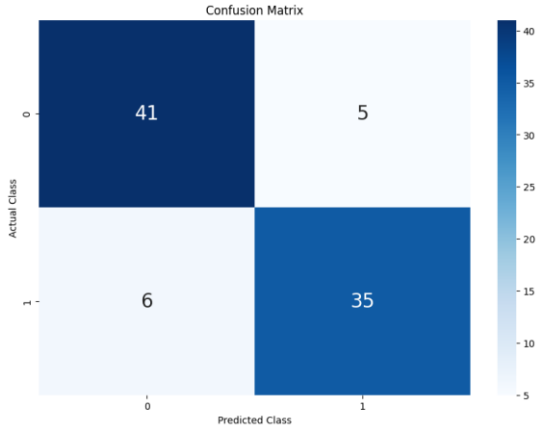
Los dos modelos de regresión logística presentan un rendimiento similar en términos de precisión, recuperación y F1, con valores en torno al 0.86. Sin embargo, se pueden destacar algunas diferencias al observar las tasas de verdaderos positivos (TPR) y falsos positivos (FPR), y el Área Bajo la Curva ROC (AUC).

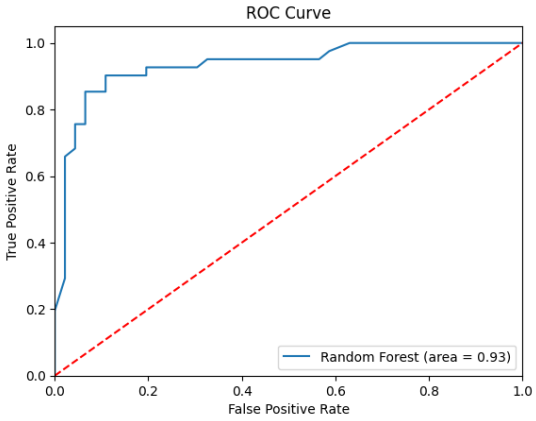
Ambos modelos presentan un rendimiento similar en cuanto a precisión, recuperación y F1, rondando el 0.86. Sin embargo, sí que hay algunas diferencias destacables en términos de TPR y FPR. En el modelo con todas las variables, la TPR y FPR varían más frecuentemente y su AUC es de 0.934, lo que indica un buen rendimiento general del modelo. Por otro lado, el modelo con las variables originales tiene una TPR que aumenta de manera más constante y también llega a 1. Su FPR tiene menos variación que el primer

modelo y un valor máximo más bajo. Además, su AUC es de 0.956, lo que sugiere un rendimiento algo superior al primer modelo.

Random Forest

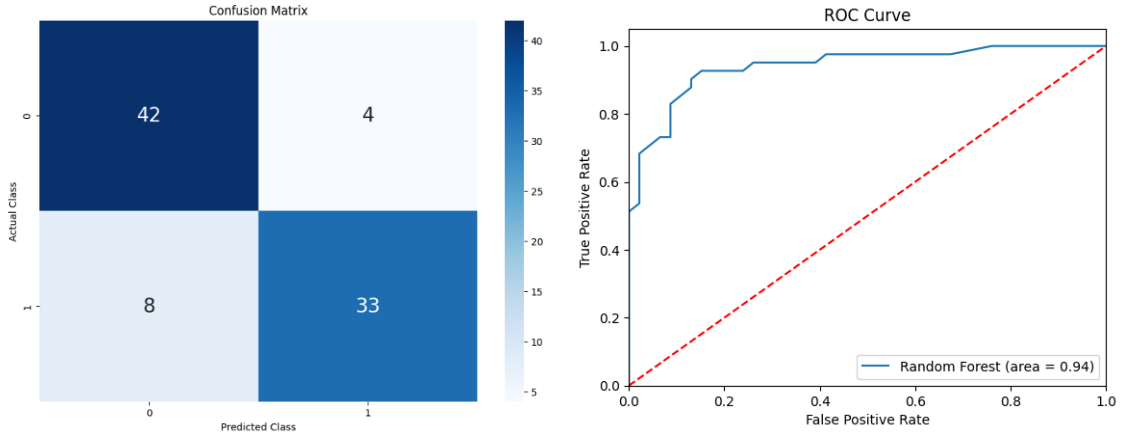
TABLA 3.11. REPORTE DE CLASIFICACIÓN CON LAS VARIABLES CREADAS





Reporte de clasificación				
	precision	recall	f1-score	support
No	0.87	0.89	0.88	46
Si	0.88	0.85	0.86	41
Accuracy			0.87	87
Macro avg	0.87	0.87	0.87	87
Weighted avg	0.87	0.87	0.87	87

TABLA 3.12. REPORTE DE CLASIFICACIÓN CON LAS VARIABLES ORIGINALES



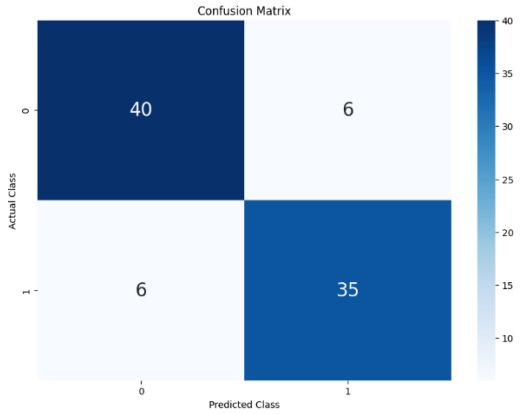
Reporte de clasificación				
	precision	recall	f1-score	support
No	0.84	0.91	0.87	46
Si	0.89	0.80	0.85	41
Accuracy			0.86	87
Macro avg	0.86	0.86	0.86	87
Weighted avg	0.86	0.86	0.86	87

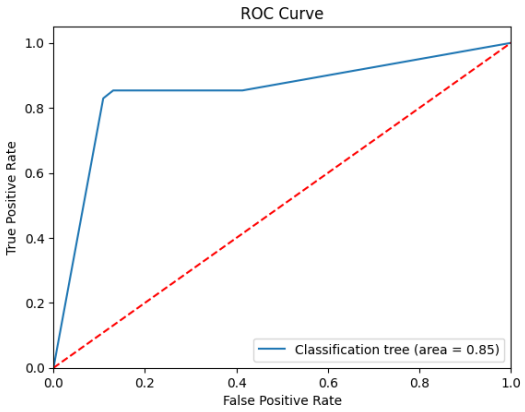
Los dos modelos de bosques aleatorios (*Random Forest*) muestran una tendencia muy similar, teniendo unos rendimientos parecidos, teniendo algo más de precisión general el modelo con las variables añadidas (0.87 vs 0.86).

En cuanto a la curva ROC, el modelo con las variables originales tiene un FPR más bajo y un área bajo la curva algo mayor.

Árbol de clasificación

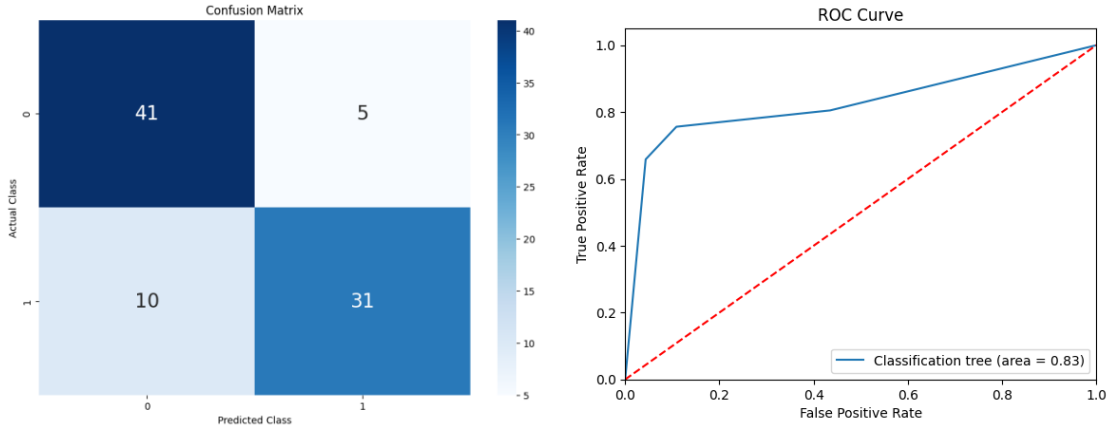
TABLA 3.13. REPORTE DE CLASIFICACIÓN CON LAS VARIABLES CREADAS





Reporte de clasificación				
	precision	recall	f1-score	support
No	0.87	0.87	0.87	46
Si	0.85	0.85	0.85	41
Accuracy			0.86	87
Macro avg	0.86	0.86	0.86	87
Weighted avg	0.86	0.86	0.86	87

TABLA 3.14. REPORTE DE CLASIFICACIÓN CON LAS VARIABLES ORIGINALES



Reporte de clasificación				
	precision	recall	f1-score	support
No	0.80	0.89	0.85	46
Si	0.86	0.76	0.81	41
Accuracy			0.83	87
Macro avg	0.83	0.82	0.83	87
Weighted avg	0.83	0.83	0.83	87

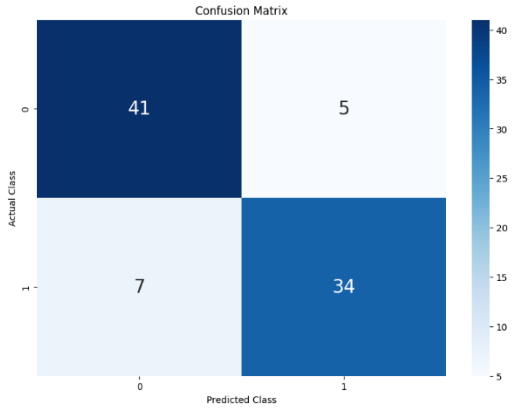
El modelo que emplea variables creadas muestra una precisión para las clases 'NO' y 'YES' de 0.87 y 0.85, respectivamente, y un *recall* de 0.87 y 0.85 para las mismas clases. Su F1-Score es de 0.87 y 0.85 y logra una precisión global del 86%. En términos del Área bajo la curva ROC (AUC), este modelo alcanza un valor de 0.8486.

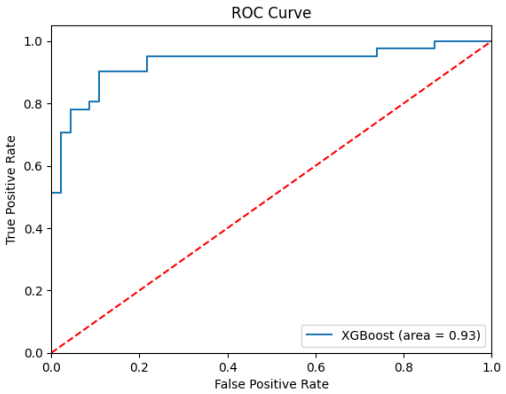
Por otro lado, el modelo con variables originales presenta una precisión de 0.80 para la clase 'NO' y 0.86 para 'SI'. El *recall* para estas clases es de 0.89 y 0.76 respectivamente, y el F1-Score se sitúa en 0.85 y 0.81. La precisión global del modelo es del 83% y su AUC es de 0.8250.

Se observa que el basado en variables creadas muestra un rendimiento ligeramente superior en términos de precisión, *recall*, F1-Score y AUC en comparación con el modelo que utiliza variables originales.

XGBoost

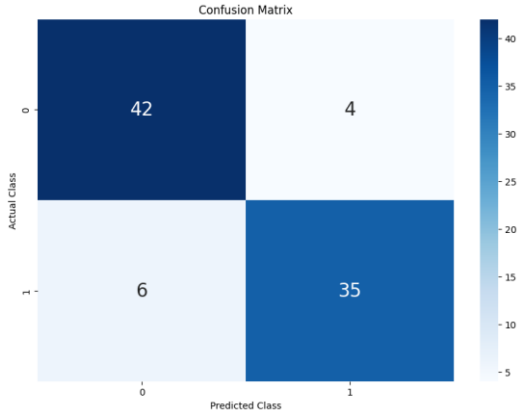
TABLA 3.15. REPORTE DE CLASIFICACIÓN CON LAS VARIABLES CREADAS

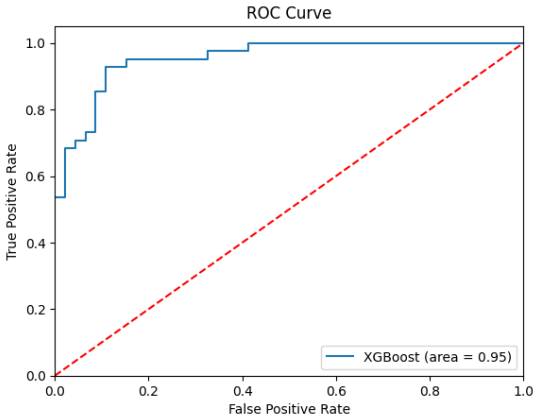




Reporte de clasificación				
	precision	recall	f1-score	support
No	0.85	0.85	0.87	46
Si	0.87	0.87	0.85	41
Accuracy			0.86	87
Macro avg	0.86	0.86	0.86	87
Weighted avg	0.86	0.86	0.86	87

TABLA 3.16. REPORTE DE CLASIFICACIÓN CON LAS VARIABLES ORIGINALES





Reporte de clasificación				
	precision	recall	f1-score	support
No	0.88	0.91	0.89	46
Si	0.90	0.85	0.88	41
Accuracy			0.89	87
Macro avg	0.88	0.88	0.88	87
Weighted avg	0.89	0.89	0.88	87

Con respecto a la implementación basada en variables creadas, se observa una precisión de 0.85 para la clase 'NO' y 0.87 para 'SI'. Su *recall* es de 0.89 y 0.83 para 'NO' y 'SI', respectivamente. El F1-Score, que combina *precision* y *recall*, es de 0.87 para 'NO' y 0.85 para 'SI'. La precisión global del modelo es del 86%. Una métrica destacable es el Área bajo la curva ROC (AUC), que alcanza un valor notablemente alto de 0.9300.

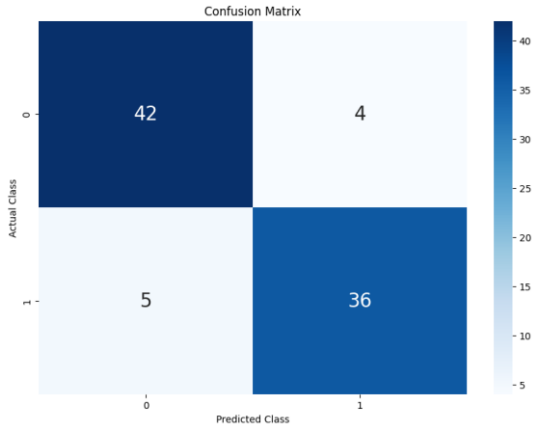
Por otro lado, el modelo basado en variables originales muestra una precisión de 0.88 y 0.90 para las clases 'NO' y 'SI', respectivamente. Su *recall* es de 0.91 para 'NO' y 0.85 para 'SI'. En cuanto al F1-Score, se obtiene un 0.89 para 'NO' y un 0.88 para 'SI'. La precisión global para este modelo es del 89%. Sorprendentemente, el AUC para este modelo es aún más elevado, llegando a 0.9539.

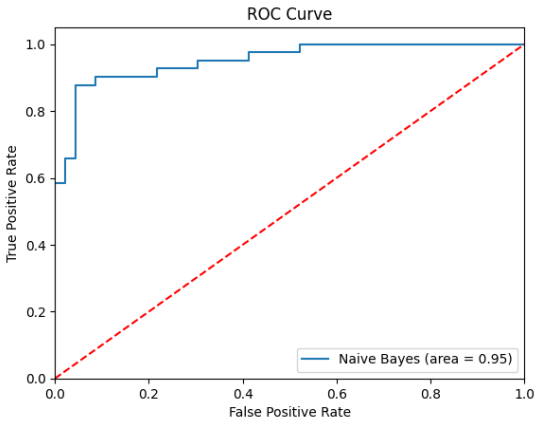
Comparando ambas implementaciones, el modelo que emplea variables originales presenta un rendimiento ligeramente superior en términos de precisión, *recall*, F1-Score y AUC en comparación con el modelo con variables creadas. Aunque ambas implementaciones alcanzan métricas altas, la diferencia en el AUC, una métrica crucial para evaluar la capacidad discriminatoria de un modelo destaca notablemente, favoreciendo al modelo con variables originales.

Si bien ambas implementaciones del XGBoost muestran un rendimiento robusto, el modelo basado en variables originales supera ligeramente al basado en variables creadas en las métricas clave.

Naive Bayes

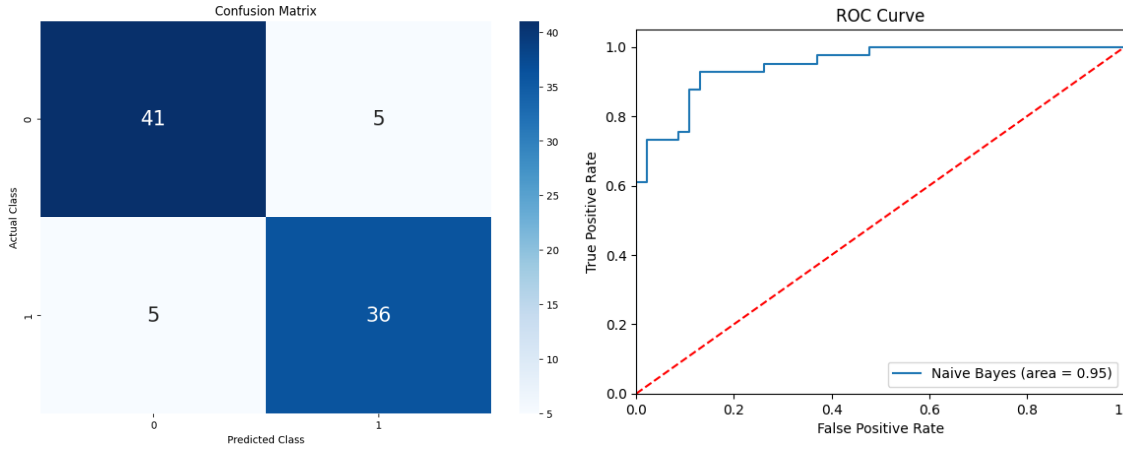
TABLA 3.17. REPORTE DE CLASIFICACIÓN CON LAS VARIABLES CREADAS





Reporte de clasificación				
	precision	recall	f1-score	support
No	0.89	0.89	0.89	46
Si	0.88	0.88	0.88	41
Accuracy			0.89	87
Macro avg	0.88	0.88	0.88	87
Weighted avg	0.89	0.89	0.89	87

TABLA 3.16. REPORTE DE CLASIFICACIÓN CON LAS VARIABLES ORIGINALES



Reporte de clasificación				
	precision	recall	f1-score	support
No	0.89	0.89	0.89	46
Si	0.88	0.88	0.88	41
Accuracy			0.89	87
Macro avg	0.88	0.88	0.88	87
Weighted avg	0.89	0.89	0.88	87

El modelo que utiliza variables creadas exhibe una precisión de 0.89 para la clase '0' y 0.90 para la clase '1'. El *recall* para estas mismas clases se sitúa en 0.91 y 0.88, respectivamente. El F1-Score, que es una métrica que combina tanto la precisión como el *recall*, es de 0.90 para ambas clases, y la precisión global alcanza un 0.90. Notablemente, el Área bajo la curva ROC (AUC) para esta configuración es de 0.9512, indicando una excelente capacidad discriminatoria del modelo.

Por su parte, la configuración con variables originales muestra una precisión de 0.89 para la clase '0' y 0.88 para la clase '1'. En cuanto al *recall*, ambos valores son idénticos a la precisión, situándose en 0.89 y 0.88 respectivamente. El F1-Score para estas clases es de 0.89 y 0.88. La precisión global es ligeramente menor que la anterior, con un valor de 0.89. El AUC, aunque sigue siendo alto, es de 0.9486.

Al evaluar ambas configuraciones, se puede observar que el modelo basado en variables creadas posee un ligero margen de superioridad en términos de precisión, *recall*, F1-Score y AUC en comparación con el modelo que utiliza variables originales. Sin embargo, es esencial señalar que ambas implementaciones alcanzan métricas destacadas y cercanas entre sí.

En conclusión, aunque ambos modelos de Naive Bayes ofrecen un rendimiento robusto y comparable, la implementación que utiliza variables creadas presenta ventajas ligeras en las métricas evaluadas. Es imperativo contextualizar estos hallazgos dentro del marco del estudio y considerar pruebas adicionales para una interpretación más profunda y aplicada de los resultados.

Anexo 5. Aplicación del modelo

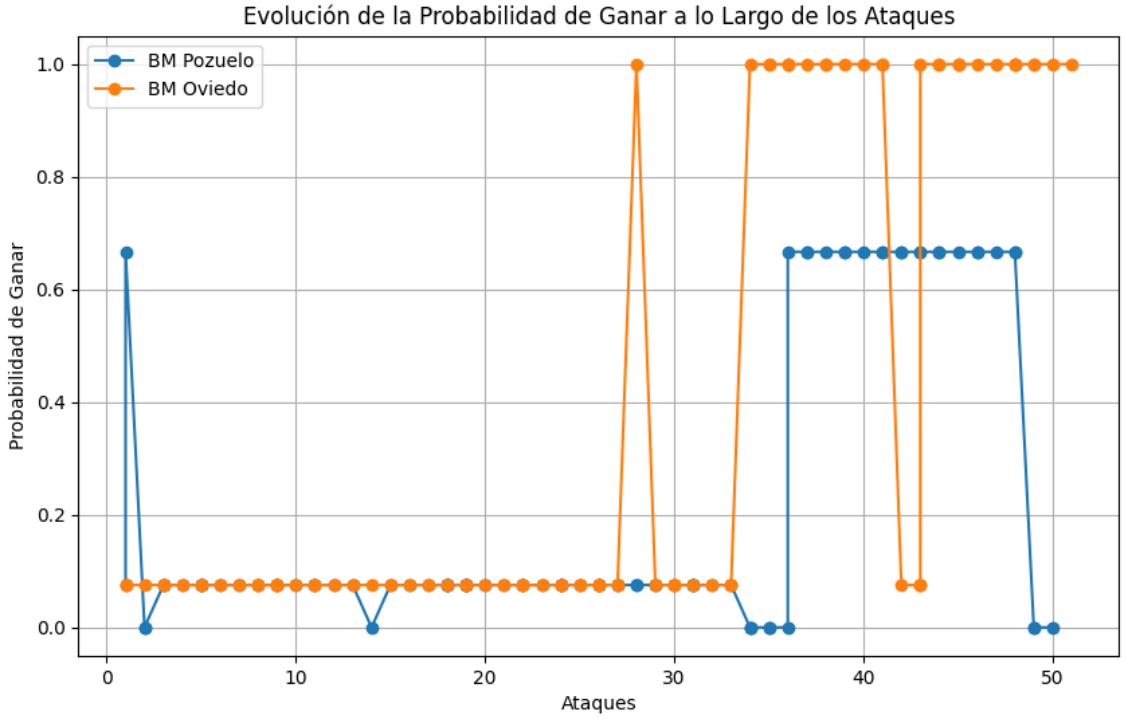
En este Anexo se dispondrá de contenido relevante a la hora de profundizar en la evaluación realizada para elegir el modelo que mejor puede ajustarse a los datos.

Anexo 5.1. Cálculo de las probabilidades de ganar *in-game*

En este apartado se mostrarán todas las aplicaciones de los modelos evaluados anteriormente, comentando cómo se ajustan a datos reales

Árbol de clasificación

FIGURA ANEXA 17. PROBABILIDADES MODELADAS POR *ÁRBOL DE CLASIFICACIÓN* CON VARIABLES ORIGINALES.

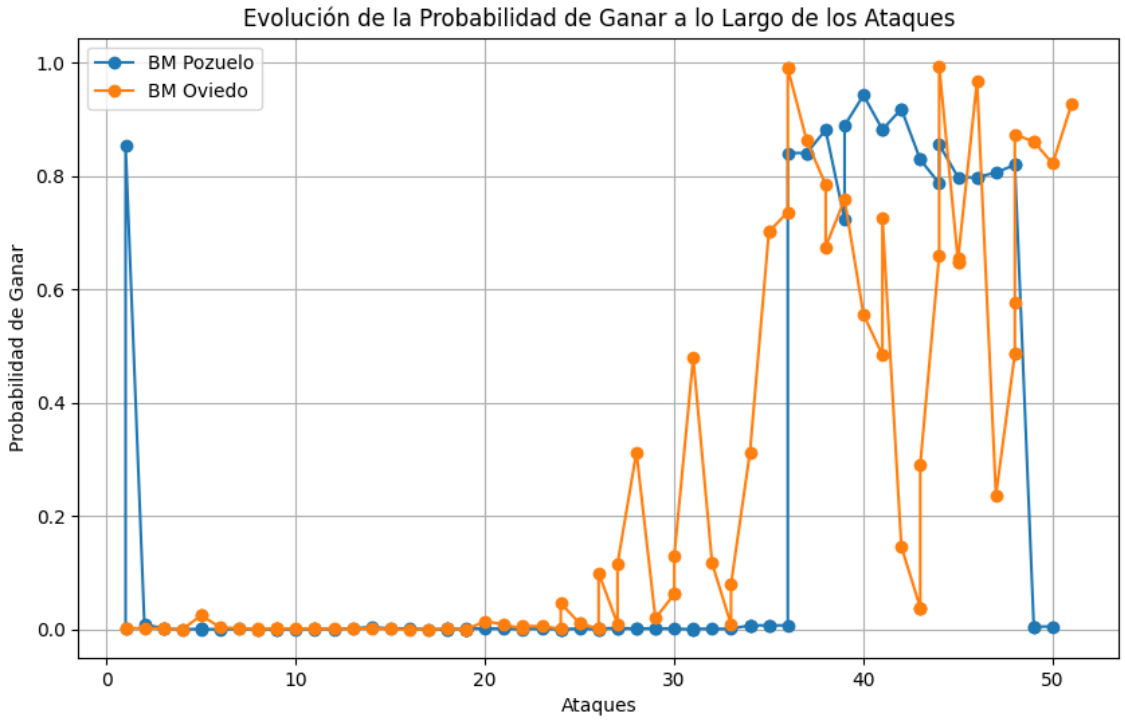


Como se puede apreciar, el modelo no es para nada realista. El modelo otorga o probabilidades muy cercanas a 0% o 100% de probabilidades de ganar, indistintamente de las variaciones que se producen a lo largo del partido de las estadísticas de un equipo. Esto se puede deber a que a partir de cierto corte dado por una combinación de estadísticas en la división del árbol se considera que el equipo ganará, y por debajo perderá, indistintamente. Por lo que la variación, además, aunque se produce en contadas ocasiones, es muy alta.

XGBoost

Se trata de un modelo más complejo que normalmente requiere de una gran cantidad de datos para el ajuste, lo que no es el caso, aun así, tras su aplicación se obtienen resultados interesantes como los que se ven en la figura.

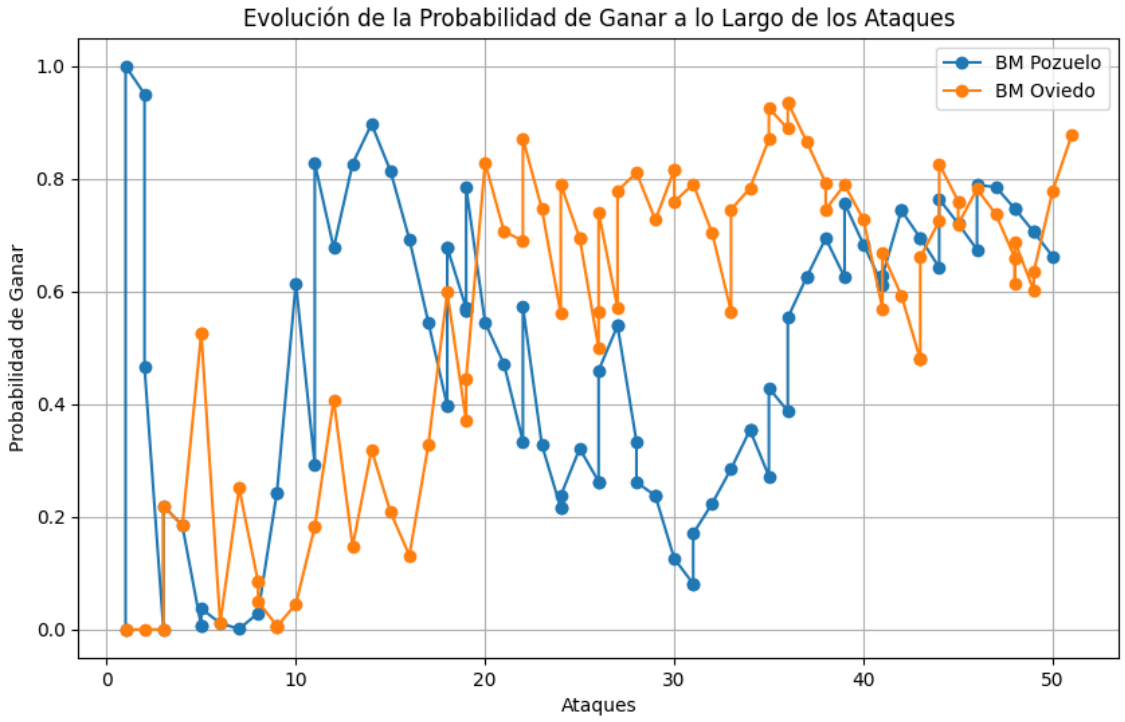
FIGURA ANEXA 18. PROBABILIDADES MODELADAS POR *XGBOOST* CON VARIABLES ORIGINALES.



Parece ser que se ajusta de mejor manera a los eventos que sucedieron en el partido. Otorga más probabilidad al equipo que acabó ganando (con una diferencia de un gol) pero hay una brecha bastante grande porque a Pozuelo sigue otorgándole probabilidades bastante bajas y además es bastante volátil. La volatilidad es muy alta, en un ataque se tiene un 80% de probabilidades de ganar y en el siguiente 0%, lo cual no se ajusta a la realidad.

Naive Bayes

FIGURA ANEXA 18. PROBABILIDADES MODELADAS POR *NAIVE BAYES* CON VARIABLES ORIGINALES.



Es el modelo más curioso porque pese a que teóricamente es un modelo que no debería ser bueno clasificando este tipo de datos, tanto en la evaluación como gráficamente parece que se ajusta bastante bien a los datos. Sin embargo, se puede ver un fallo claro y es su variación. Parece ser bastante volátil sobre todo en los tres primeros cuartos del partido, porque en el final del partido otorga realidades bastante realistas.