# Modelo de clasificación multiclases para la predicción de apuestas deportivas

Lina María Martínez Arias
Santiago Marulanda Vélez

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Tutor
Javier Fernando Botia Valderrama, Doctor (PhD)

Universidad de Antioquia
Facultad de Ingeniería
Especialización en Analítica y Ciencia de Datos
Medellín, Antioquia, Colombia
2023

| | |
|---|---|
| **Cita** | (Martínez Arias & Marulanda Vélez, 2023) |
| **Referencia**<br><br>**Estilo APA 7 (2020)** | Martínez Arias, L. M., & Marulanda Vélez, S. (2023). *Modelo de clasificación multiclases para la predicción de apuestas deportivas* [Trabajo de grado especialización]. Universidad de Antioquia, Medellín, Colombia. |



Especialización en Analítica y Ciencia de Datos, Cohorte IV.

Centro de Investigación Ambientales y de Ingeniería (CIA).

UNIVERSIDAD DE ANTIOQUIA
Vicerrectoría de Docencia

Sistema de Bibliotecas

Centro de Documentación de Ingeniería (CENDOI)

**Repositorio Institucional:** http://bibliotecadigital.udea.edu.co

Universidad de Antioquia - www.udea.edu.co

**Rector:** John Jairo Arboleda Céspedes

**Decano:** Julio César Saldarriaga Molina.

**Jefe departamento:** Diego José Luis Botia Valderrama

## Dedicatoria

A María Jenny, gracias mamá por enseñarme siempre a luchar, ser mejor y lo que es correcto.
A Noé y a Edi, por mostrarme que con esfuerzo y amor se puede obtener todo.

## Agradecimientos

A nuestra Alma Mater, a la cual siempre estaremos orgullosos de pertenecer
A nuestros profesores de la especialización, quienes nos compartieron su conocimiento,
A nuestras parejas, quienes fueron un apoyo muy valioso
y a nuestro asesor, por su ayuda fundamental para este proyecto desde el inicio.

.

**Tabla de contenido**

## Lista de tablas

## Lista de figuras

## Siglas, acrónimos y abreviaturas

| | |
|---|---|
| **KNN** | K-Nearest-Neighbor |
| **ROC** | Receiver Operating Characteristic |
| **TE** | Tree Explainer |
| **XGBoost** | Extreme Gradient Boosting |
| **HGBC** | Hist Gradient Boosting Classifier |
| **LGB** | Light Gradient Boosting |
| **ML** | Machine Learning |
| **PCA** | Principal Component Analysis |
| **TSNE** | T-distributed Stochastic Neighbor Embedding |
| **SVD** | Singular Value Descomposition |
| **FIFA** | Federación Internacional de Fútbol Asociación |

**Resumen**

El proyecto busca crear una herramienta de clasificación multimodal que permita identificar la probabilidad de un resultado en un evento deportivo, específicamente en la Serie A de Italia tomando información de las temporadas desde 2015 y hasta lo que va de la temporada 2023. Se busca predecir tres variables objetivo para los partidos utilizando la herramienta: los goles de local, los goles de visitante y el resultado del equipo local. El modelo utiliza técnicas de análisis de datos y aprendizaje automático para identificar patrones en los datos históricos de los equipos y predecir la probabilidad de cada resultado posible.

El contexto de las apuestas deportivas es un sector en constante crecimiento en el que los usuarios buscan obtener beneficios a través de sus conocimientos y habilidades en deportes. El problema de negocios radica en la falta de herramientas y modelos que permitan a los usuarios tomar decisiones informadas y seguras en sus apuestas deportivas. Para abordar este problema, se ha desarrollado un modelo de apuestas deportivas que utiliza algoritmos y análisis estadísticos para predecir los resultados de los partidos de fútbol.

Los datos fueron obtenidos de Understat (https://understat.com/) y se utilizan varias métricas de Machine Learning para evaluar el desempeño de los modelos de clasificación, como la exactitud (accuracy), la precisión, la tasa de verdaderos positivos (recall) y la curva característica operativa del receptor (ROC). Durante el desarrollo del proyecto, se enfrentaron algunos obstáculos relacionados con la calidad de los datos, la selección de variables y la elección de los algoritmos de aprendizaje automático más adecuados. Sin embargo, se lograron superar estos obstáculos y se obtuvo a través del modelo Hist Gradient Boosting Classifier (HGBC) una exactitud del 75%, cumpliendo con el rendimiento esperado. La realización de estos modelos se puede consultar en el repositorio de GitHub: https://github.com/lina-martinez/Modelo-clasificacion-multiclases-para-prediccion-de-apuestas-deportivas.git

*Palabras clave:* predicción, apuestas deportivas, clasificación, machine learning, goles.

## Abstract

The project proposes to create a multimodal classification tool that allows to identify the probability of a result in a sporting event, specifically in the Italian Serie A for the seasons from 2015 and until the 2023 season. Three objective variables are to be predicted for matches using the tool: home goals, away goals, and home team result. The model uses data analysis and machine learning techniques to identify patterns in historical team data and predict the probability of each possible outcome.

The context of sports betting is an ever-growing sector where users seek to profit from their knowledge and skills in sports. The business problem lies in the lack of tools and models that allow users to make informed and safe decisions in their sports betting. To address this problem, a sports betting model has been developed that uses algorithms and statistical analysis to predict the outcome of football matches.

The data was obtained from Understat (https://understat.com/) and several machine learning metrics are used to evaluate the performance of the classification models, such as accuracy, precision, true positive rate (recall) and the receiver operating characteristic (ROC) curve. During the development of the project, there were some obstacles related to the quality of the data, the selection of the variables and the choice of the most appropriate machine learning algorithms. However, these obstacles were overcome and an accuracy of 75% was obtained using the Hist Gradient Boosting Classifier (HGBC) model, in line with the expected performance. The implementation of these models can be consulted in the GitHub repository: https://github.com/lina-martinez/Modelo-clasificacion-multiclases-para-prediccion-de-apuestas-deportivas.git



*Keywords:* prediction, sports betting, ranking, machine learning, goals.

# 1. Introducción

La predicción de resultados de apuestas deportivas ha sido objeto de estudio en el ámbito académico y empresarial durante muchos años. A lo largo del tiempo, se han utilizado diferentes enfoques y técnicas para predecir los resultados deportivos y mejorar la rentabilidad de los apostadores. En los últimos años, los avances en el aprendizaje automático, la inteligencia artificial y el aumento de la capacidad de procesamiento de los computadores han permitido el desarrollo de modelos de predicción más sofisticados y precisos. Estos utilizan algoritmos de aprendizaje automático para analizar grandes cantidades de datos históricos de partidos y generar predicciones precisas de los resultados de los partidos futuros.

Hay varios casos relevantes sobre la utilidad de los modelos de aprendizaje automático para predecir las mejores apuestas posibles en diferentes ligas europeas. En un estudio de identificación basado en aprendizaje automático de las variables predictivas más fuertes de ganar y perder en el fútbol profesional belga, se desarrolló un modelo de predicción de resultados deportivos utilizando técnicas de aprendizaje automático. En el estudio, se construyó un modelo de aprendizaje automático predictivo basado en una amplia gama de variables (n = 100), utilizando un conjunto de datos que consta de 576 juego, se aplicaron diferentes técnicas de aprendizaje automático para predecir los resultados de los partidos futuros, se usó Extreme Gradient Boosting (XGBoost) para predecir ganar o perder, y se aplicó Tree Explainer (TE) para determinar la importancia de las características a nivel global y local (Youri et al., 2021). Los resultados del estudio sugieren que la inclusión de una amplia gama de variables puede mejorar la precisión de las predicciones y evaluaciones de los resultados del juego, además demostraron que el modelo propuesto mejoraba significativamente la precisión de las predicciones en comparación con los métodos tradicionales de predicción de resultados deportivos, el modelo tuvo una precisión de 89,6% ± 3,1% y clasificó correctamente 516 de 576 juegos.

En otro estudio publicado en el año 2020 sobre el uso de aprendizaje automático y patrones de velas japonesas para predecir los resultados de los partidos de fútbol americano, se utilizó varios métodos de aprendizaje automático, como el aprendizaje conjunto, las máquinas de vectores de soporte y las redes neuronales, para predecir los resultados de los partidos. Predecir el ganador y el perdedor implica un enfoque basado en resultados, que generalmente se llevan a cabo utilizando modelos basados en clasificación, donde el resultado previsto es una variable categórica, como

ganar, perder o empatar (Yu-Chia, 2020). La investigación concluye que los gráficos de velas basados en datos del mercado de apuestas pueden ser efectivos para predecir los resultados de los partidos utilizando el aprendizaje automático. El estudio muestra que este enfoque se puede aplicar a varios deportes sin necesidad de conocimientos específicos del deporte logrando la mejor tasa de precisión del 68,4 %.

Por otra parte, en un estudio sobre la incorporación del conocimiento del dominio en el aprendizaje automático para la predicción de resultados de fútbol, desarrollaron un modelo de predicción de resultados deportivos en función de los datos de partidos anteriores, a través de los métodos de selección de características y métodos de aprendizaje de clasificación, se usaron los siguientes dos algoritmos de aprendizaje para construir modelos predictivos a partir de nuestros conjuntos de datos: K-vecino más cercano (k-NN) y conjuntos de árboles potenciados por gradientes extremos para generar modelos predictivos a partir de los conjuntos de características seleccionados, el documento sugiere que la incorporación exitosa del conocimiento del dominio es un factor clave en la predicción de resultados de partidos de fútbol (Berrar et al. en el 2019).

La aplicación de técnicas de aprendizaje automático en las apuestas deportivas se ha vuelto cada vez más común en los últimos años. Los modelos de clasificación multiclases son un enfoque popular para la predicción de resultados deportivos debido a su capacidad para clasificar múltiples resultados en función de diversas variables (Zhang et al., 2020). Además, estos modelos pueden utilizarse para predecir resultados en diferentes deportes, lo que los hace altamente adaptables. Los estudios han demostrado que los modelos de clasificación multiclases pueden superar a los enfoques tradicionales de predicción en términos de precisión y confiabilidad en las apuestas deportivas (Zhao et al., 2021). Estos estudios demuestran que los modelos de predicción de apuestas deportivas basados en técnicas de aprendizaje automático pueden mejorar significativamente la precisión de las predicciones de resultados deportivos y, por lo tanto, mejorar la rentabilidad de las apuestas deportivas. Sin embargo, es importante tener en cuenta que el éxito de estos modelos depende en gran medida de la calidad y cantidad de los datos históricos utilizados para entrenarlos.

## 2. Planteamiento del problema

El problema de negocios radica en la falta de herramientas y modelos que permitan a los usuarios tomar decisiones informadas y seguras en sus apuestas deportivas. Los usuarios se enfrentan a una gran cantidad de información disponible, incluyendo estadísticas de equipos y jugadores, tendencias y noticias de última hora, lo que puede ser abrumador y difícil de procesar. Además, los resultados deportivos son impredecibles y pueden ser afectados por múltiples factores, lo que dificulta aún más la tarea de tomar decisiones informadas y seguras. Para abordar este problema, se plantea un modelo de apuestas que utiliza algoritmos y análisis estadísticos para predecir los resultados de eventos deportivos. En este caso, es necesario analizar la posibilidad de crear una herramienta de análisis de las apuestas de fútbol a través de los goles de local, los goles de visitante y el resultado del equipo local.

## 3. Justificación

El proyecto se desarrolla en el contexto de las apuestas deportivas, un sector en constante crecimiento en el que los usuarios buscan obtener beneficios a través de sus conocimientos y habilidades en deportes. Sin embargo, en este mercado altamente competitivo, la gran cantidad de información disponible y la variabilidad de los resultados de los eventos deportivos hacen que sea difícil para los usuarios tomar decisiones informadas y seguras a la hora de realizar sus apuestas. Por consiguiente, la necesidad de crear herramientas y modelos que permitan a los usuarios tomar decisiones informadas y seguras ayudaría a tomar mejores decisiones sobre el impacto de una apuesta y los escenarios de ganar o perder dinero. El modelo que se pretende generar debe utilizar un conjunto de datos históricos para usarlos en tiempo real, lo cual la capacidad de analizar tendencias y patrones es útil para proporcionar a los usuarios recomendaciones de apuestas seguras y rentables. El modelo beneficia a los apostadores brindándoles información precisa y confiable que les permite tomar decisiones informadas y seguras en sus apuestas deportivas.

Los modelos predictivos desarrollados servirán para predecir los resultados de eventos deportivos, en particular, para identificar la probabilidad de un resultado en un evento deportivo. Estos modelos ayudarían a los usuarios a tomar decisiones informadas al apostar en eventos deportivos considerando múltiples variables que influyen en los resultados de los eventos deportivos, como el rendimiento de los equipos en partidos anteriores, el rendimiento de los equipos siendo local y visitantes, entre otros factores relevantes.

La herramienta de clasificación será de gran utilidad para predecir los resultados de eventos deportivos de una manera más precisa y confiable, lo que permitirá a los fanáticos y apostadores hacer apuestas más conscientes y obtener mejores resultados en sus apuestas. Además, al utilizar datos históricos para predecir los resultados, se podrá tener en cuenta la forma actual del equipo, los resultados recientes y otros factores relevantes que pueden afectar el resultado final del evento deportivo.

Los modelos predictivos pueden resolver este problema al permitir la creación de herramientas que pronostiquen los resultados de los eventos suceso en función de los datos históricos disponibles. Estos modelos pueden utilizar diferentes técnicas de análisis de datos y

aprendizaje automático para identificar patrones en los datos históricos de los equipos y predecir la probabilidad de cada resultado posible.

Para lograr esto, se recopiló información histórica tomada de fuentes estadísticas deportivas con el objetivo de calcular la probabilidad de ocurrencia de cada uno de los escenarios posibles dados para un evento deportivo. En particular, se busca predecir tres variables objetivo para los partidos utilizando la herramienta: los goles de local, los goles de visitante y el resultado del equipo local.

Los modelos predictivos pueden resolver este problema al permitir la creación de herramientas que pronostiquen los resultados de los eventos deportivos en función de los datos históricos disponibles. Estos modelos pueden utilizar diferentes técnicas de análisis de datos y aprendizaje automático para identificar patrones en los datos históricos de los equipos y predecir la probabilidad de cada resultado posible.

# 4. Objetivos

## 4.1 Objetivo general

Desarrollar un modelo de clasificación de apuestas deportivas utilizando algoritmos de machine learning con el fin de lograr una precisión del 75% en los pronósticos, teniendo en cuenta diversas variables relevantes como estadísticas de equipos e información de los equipos locales y visitantes.

## 4.2 Objetivos específicos

- Analizar la base de datos de la apuesta deportiva seleccionada para explorar los posibles datos atípicos, tendencias o patrones en la información.
- Proponer un método para seleccionar el mejor modelo de clasificación teniendo en cuenta los hiperparámetros, costo-computacional y posibles limitaciones en su desempeño.
- Validar el desempeño del mejor modelo de clasificación seleccionado con la base de datos de prueba de la apuesta deportiva seleccionada, considerando la matriz de probabilidad de las etiquetas.

# 5. Marco teórico

## 5.1 Algoritmos de Clasificación

Para la solución del problema inicial se consideró las soluciones desde los modelos dados por la estadística clásica y así mismo se decidió convertir dicha situación en un problema de categorización, pues la decisión que dicho modelo soporta no está sesgada sobre el número exacto de goles, sino, el estado frente a un umbral, es decir, si marca o no un número de terminado de goles, se evaluaron los siguientes modelos con la información completa de las características, pero solo se explicarán los más importantes en términos de resultados:

- **Random Forest Classifier:** Es una técnica de aprendizaje automático utilizado netamente para temas de clasificación de etiquetas, permite que los modelos basados en árboles de decisión sigan la trazabilidad de los resultados y entender qué características son determinantes para la toma de decisiones en cada caso puntual, este método es una variante de árboles que genera varios de estos y para generar un resultado combina las predicciones, en su naturaleza limita el sobre ajuste generando árboles con datos aleatorios de la muestra, llevando el mismo desarrollo a tener una alta generalización, en términos prácticos, la decisión de la etiqueta precedida se crea al evaluar los resultados de todos los árboles y seleccionar la etiqueta con más árboles a favor (Breiman, L., 2001).
- **Gradient Boosting Classifier:** Como técnica de aprendizaje automático perteneciente a la familia de los árboles de decisión comparte las características fundamentales de este tipo de modelos, identificar trazabilidad en las respuestas y la clasificación eficaz de las etiquetas apoyado en condiciones, sin embargo enmarca diferencias claras, la construcción de reglas y elección de parámetros no se genera de forma aislada, la manera usada, es la generación de árboles en serie que a medida que avanza en el número de iteraciones aumenta la sensibilidad y ajusta los errores del árbol anterior, este ajuste se genera sobre la función de pérdida, buscando minimizar esta, se ajustan cada vez más, al tener una construcción en serie es más sensible al ruido y datos atípicos, y la manipulación de gran número de características generan diferentes conflictos o complicaciones a lo largo de las interacciones (Hastie et al., 2009)

- **Light Gradient Boosting (LGB):** Es una técnica soportado sobre el algoritmo definido anteriormente el Gradient Boosting Classifier, pero con mejoras significativas, su principal característica y ventaja respecto al anterior es el método leaf-wise, mismo que permite identificar y clasificar las características para una división más efectiva desde el comienzo, generando una convergencia más próxima y minimizando las iteraciones necesarias para el ajuste deseado, permite también la generación de clasificaciones en variables categóricas y tiene un rendimiento mejor en procesamiento si se contrasta con los demás métodos (Ke et al., 2017)
- **Histogram Gradient Boosting Classifier:** Es un algoritmo para problemas de clasificación por excelencia, se soporta en el Gradient Boosting como base para generar un histograma de características, y a través de este generar nuevas iteraciones y tomar mejores decisiones, este funcionamiento tiene una eficiencia superior al Gradient Boosting clásico porque agrupa los valores de los histogramas generados para procesar los datos, este algoritmo se apalanca en el muestreo basado en gradientes y es muy útil con grandes volúmenes de datos, tanto a nivel de registros como de variables y columnas (Müller et al., 2021)

### 5.2 Métricas de Validación para Tareas de Clasificación

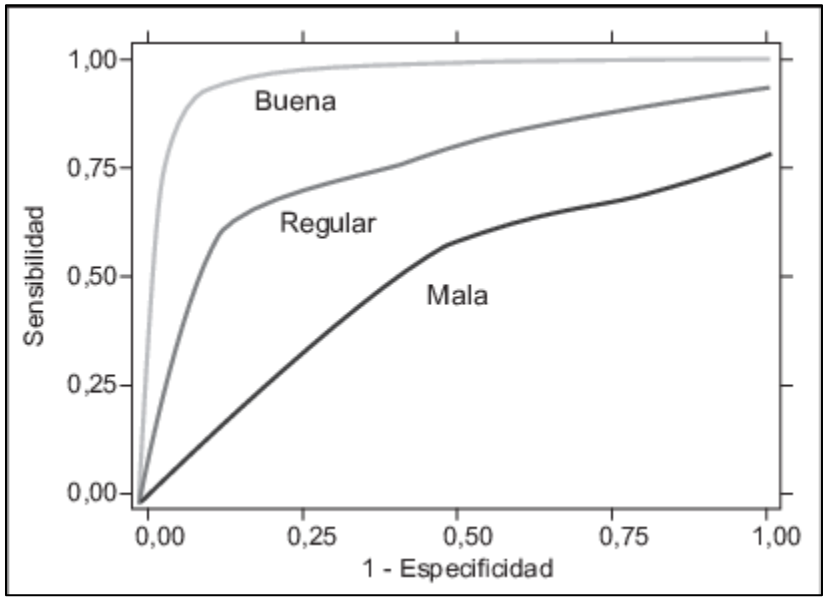
Para evaluar el desempeño de los modelos de clasificación, se utilizarán varias métricas de machine learning, entre ellas:

- **Exactitud (Accuracy):** indicará el porcentaje de casos en los que el modelo ha acertado. Se espera que el modelo tenga una alta tasa de accuracy o exactitud para garantizar la precisión de las predicciones.
- **Precisión y Recall:** son métricas que miden la calidad de las predicciones del modelo a través de la matriz de confusión, la precisión se refiere a la proporción de resultados positivos que son verdaderos positivos, mientras que el recall o la tasa de verdaderos positivos mide la proporción de resultados positivos que fueron identificados correctamente por el modelo.
- **Curva ROC:** es una representación gráfica de la capacidad de un modelo para distinguir entre clases positivas y negativas, como se puede observar en la Figura 1. La curva ROC muestra la tasa de verdaderos positivos (TPR) en el eje y la tasa de falsos positivos (FPR) en el eje x para

diferentes umbrales de clasificación. Cuanto más cerca esté la curva de la esquina superior izquierda del gráfico, mejor será la capacidad del modelo para distinguir entre clases positivas y negativas (García et al., 2008)

**Figura 1.**

*Esquema explicativo de distintas posibilidades de la curva ROC*




### 5.3 Métricas de Negocios

Las métricas de negocio utilizadas evaluarán los siguientes aspectos

- **ROI (retorno de inversión):** se espera que el modelo sea rentable en términos de apuestas deportivas y genere un retorno de inversión positivo. Se buscará determinar el umbral mínimo de acierto del modelo que permita obtener un ROI aceptable.
- **Apuestas ganadas:** se evaluará el porcentaje de apuestas ganadas del total de apuestas, gracias a la clasificación dada por el modelo. Se buscará un alto porcentaje de apuestas ganadas como indicador de un buen desempeño del modelo en términos de negocio.

En cuanto al valor mínimo de las métricas necesarias en el contexto de aplicación, reconociendo la naturaleza del proyecto, donde buscamos la probabilidad de que suceda un resultado específico de tres posibles para tomar una decisión de apuesta; es fundamental tener un

alto acierto en el resultado seleccionado, es decir, el ideal es pronosticar de forma acertada lo que sí sucederá, más allá de anticipar correctamente que los eventos no pasarán, así mismo, entre mayor sea la relación entre verdaderos positivos con los falsos positivos será más rentable para el proyecto, es aún más determinante esta relación que la existente entre los verdaderos y falsos negativos, pues, sobre estos últimos no se generarán apuestas a partir de los anticipados por el modelo.

# 6. Metodología

## 6.1 Descripción y Pretratamiento de la Base de Datos

El proceso de construcción de la base de datos de entrenamiento y validación involucra inicialmente la recopilación de los datos relevantes para el problema; como ya se explicó en el inciso anterior, se contaba con 84 características de las cuales 7 fueron eliminadas dado que la información no aportaba nada relevante para la variable objetivo, por contener valores únicos en todos los registros.

Con las 77 características del modelo, se procede a hacer un preprocesamiento de los datos donde se realiza la limpieza de la información, para esto se realizó una eliminación de los datos faltantes dado que no se tuvo en cuenta aquellos registros de partidos que no tienen historia entre los equipos y los registros de partidos donde no se tiene información de al menos los 3 últimos partidos jugados.

Adicionalmente, se realizó el análisis de los datos faltantes, realizando una imputación reemplazando esos valores faltantes por "-1" con el objetivo de mantener la integridad del conjunto de datos y asegurarse de que cada registro tenga un valor numérico válido y dado que este valor no es un valor realista para los datos faltantes permite distinguir claramente los valores que faltan de los valores numéricos reales. Además, algunos algoritmos de aprendizaje automático pueden manejar este tipo de valores especiales (-1) de manera adecuada buscando que no haya un impacto significativo en los resultados del análisis y en las conclusiones que se extraigan del conjunto de datos,

Para las variables categóricas se realizó una codificación de etiquetas donde se le asigna una etiqueta a cada clase de datos en el conjunto de datos, para esto las características se etiquetaron con la lógica de:

- Características "Local Historia" y "Visitante. Historia "si la columna contiene la 'a' de visitante se le asigna un 1 y 'h' de local se le asigna el cero, si tiene el -1 es porque no se tiene información
- Características "Visitante. Historia" si la columna contiene la 'a' de visitante se le asigna un 1 y 'h' de local se le asigna el cero, si tiene el -1 es porque no se tiene información

- La característica “Resultado Local” tiene tres posibles resultados, 'W' ganador, 'l' perdedor, 'd' empate, convirtiendo las etiquetas ['l' 'w' 'd'] a [1,2,0'].
- Para las variables objetivos “Goles Local” y “Goles Visita” se organizan los datos de tal forma que 0 o 1 goles obtenidos se etiquetan como 0, y más de 1 gol se etiquetan como 1.

Las características que brindan información temporal, como las fechas de juego históricas para los partidos como visitante y como local, se convierten en un valor numérico al realizar la operación entre la característica de Fecha de Juego y cada una de las fechas de los partidos históricos, consiguiendo a través de esta operación la cantidad de días que han pasado desde ese historial del partido hasta el partido actual evaluado, además de que ayuda a eliminar la característica de serie de tiempo que pueden generar problemas de escalabilidad o correlación disminuyendo la precisión del modelo.

Se realiza una operación matemática con característica de la “Fecha de juego” para que esta diga qué tan antiguo o nuevo es el partido del que se tiene información, a través de calcular la diferencia de días entre el día actual y el día del partido jugado.

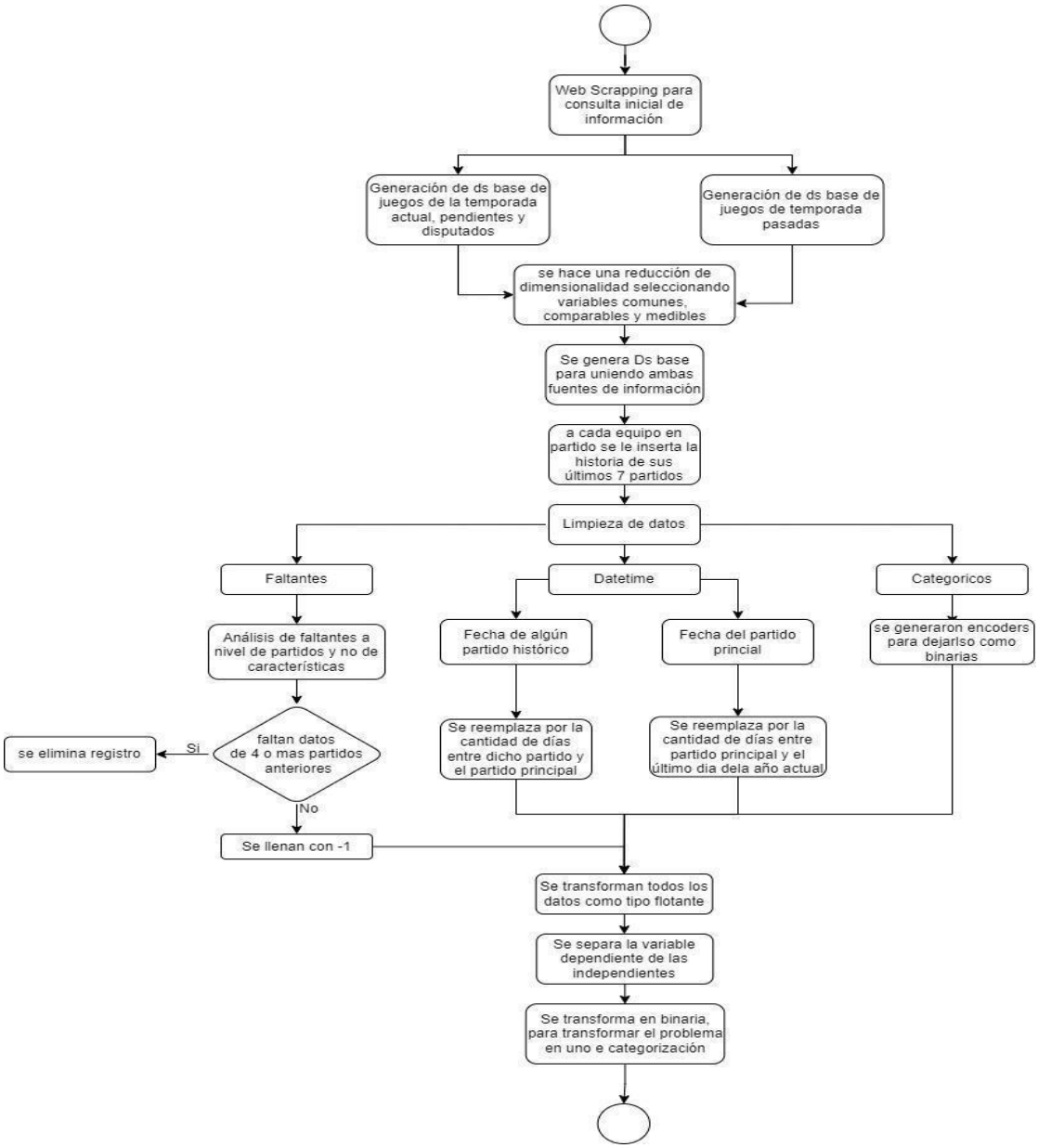
Finalmente, para la base de datos se realizó un análisis de los datos atípicos mediante el algoritmo Local Outlier Factor (LOF), este algoritmo se basa en la idea de que los puntos que están aislados o lejos de la mayoría de los puntos en un conjunto de datos son considerados valores atípicos, para la base de datos se encontró que solo había 210 datos atípicos, representando un 5% de los datos, debido a esto se decide no realizar la imputación de esta información.

Después de realizar el preprocesamiento de datos, se divide el conjunto de datos en dos conjuntos: el conjunto de entrenamiento y el conjunto de validación. La división típica es de 80/20, donde el 80% de los datos se utilizan para el entrenamiento y el 20% se utiliza para la validación.

### 6.2 Estrategia de Análisis de los Datos y Construcción de los Modelos

**Figura 2.**

*Pipeline principal del proceso de analítica para el modelo de predicción de apuestas deportivas.*




Como se puede observar en la **Figura 2.** para alcanzar el objetivo, el proyecto inicial con la **recopilación y preparación de datos,** donde se extrae datos históricos de los partidos de fútbol, incluyendo resultados de partidos para el equipo local y el equipo visitante, sobre estos datos se realiza la limpieza para la **construcción del data set**, organizando la información de manera adecuada para su posterior análisis.

Luego se realiza un análisis descriptivo de los datos, incluyendo un análisis detallado de los datos recopilados para identificar patrones, tendencias y relaciones relevantes. Esto permite comprender mejor la información disponible y seleccionar las variables más significativas para el modelo.

En el **preprocesamiento de la información**, se realizan una serie de tareas para preparar los datos antes de ser utilizados en un modelo de machine learning. Estas tareas son fundamentales para garantizar la calidad de los datos y mejorar el rendimiento y la precisión del modelo.

Para la **Selección datos de prueba y entrenamiento** y **la Iteración y evolución de los modelos**, se seleccionó los algoritmos de machine learning más apropiados para cumplir el objetivo. El modelo se entrenó utilizando datos históricos y se ajustó mediante técnicas de validación cruzada y optimización de hiper parámetros, se emplearon diversas técnicas, como regresión logística, KNN y árboles de decisión.

Para **la Evaluación del modelo**, se calculan las métricas de desempeño, como la precisión, la curva ROC y la matriz de confusión, para evaluar su capacidad predictiva y ajustar el modelo si es necesario, una vez que el modelo haya demostrado un rendimiento satisfactorio, permitirá realizar pronósticos en tiempo real y evaluar su rendimiento continuamente.

En la **implementación**, se tiene en cuenta el cálculo de la matriz de probabilidades para un seguimiento del rendimiento del modelo, además recopilando datos de pronósticos y resultados reales. Esto permitirá identificar posibles deficiencias o áreas de mejora, ajustar el modelo en consecuencia y realizar actualizaciones periódicas para mantener su precisión y eficacia.

### 6.3 Herramientas

En el proyecto de clasificación para predecir resultados de apuestas deportivas, se utilizaron varias bibliotecas y herramientas de ciencia de datos. Para el procesamiento de la información, se utilizaron bibliotecas como NumPy[1], Pandas[2] y Math[3]. Además, se utilizaron herramientas para la

[1] https://numpy.org/

[2] https://pandas.pydata.org/

[3] https://docs.python.org/es/3.10/library/math.html

visualización de datos como Seaborn[4] y Matplotlib[5], lo que permitió explorar los datos y comprender mejor los patrones y tendencias presentes en ellos. La librería Scikit-learn[6] se utilizó para el entrenamiento de modelos y su validación. También se utilizó la biblioteca BeautifulSoup[7] para realizar web scrapping y obtener la base de datos original. Los modelos se entrenaron en Google Colab, lo que permitió utilizar la potencia de cómputo de Google y acceder a varias bibliotecas y herramientas de ciencia de datos. Finalmente, se creó un repositorio en Github para almacenar el código fuente y los archivos de datos, lo que facilitó el control de versiones y la colaboración con otros desarrolladores.

[4] https://python-charts.com/es/seaborn/

[5] https://matplotlib.org/

[6] https://scikit-learn.org/stable/

[7] https://pypi.org/project/beautifulsoup4/

## 7. Resultados

### 7.1 Análisis descriptivo de los datos

Al realizar un análisis descriptivo de la base de datos se puede observar en la **Figura 3.** que las distribuciones de los datos históricos entre los partidos jugados por locales o visitantes los goles recibidos están en su mayoría sesgadas a la derecha dando a entender que las observaciones se concentran en valores bajos o moderados y sólo unas pocas observaciones tienen valores muy altos, este tipo de distribución puede encontrarse en diferentes contextos, en este caso el número de goles marcados por un equipo de fútbol en una temporada, donde la mayoría de los equipos marcan pocos goles y unos pocos equipos marcan muchos goles. Las características de los puntos obtenidos por el local y visitantes están sesgados a la izquierda como se observa en la **Figura 4.** Es decir, la mayoría de las observaciones se concentran en valores altos o muy altos, y solo unas pocas observaciones tienen valores bajos, permitiendo observar que en la mayoría de los partidos hay un resultado de ganar o perder, obteniendo así un alto puntaje.

**Figura 3.**

*Diagrama de cajas de las características "local_Historia_Goles_visitantes N-1' y 'local_Historia_Goles_visitantes N-2'*

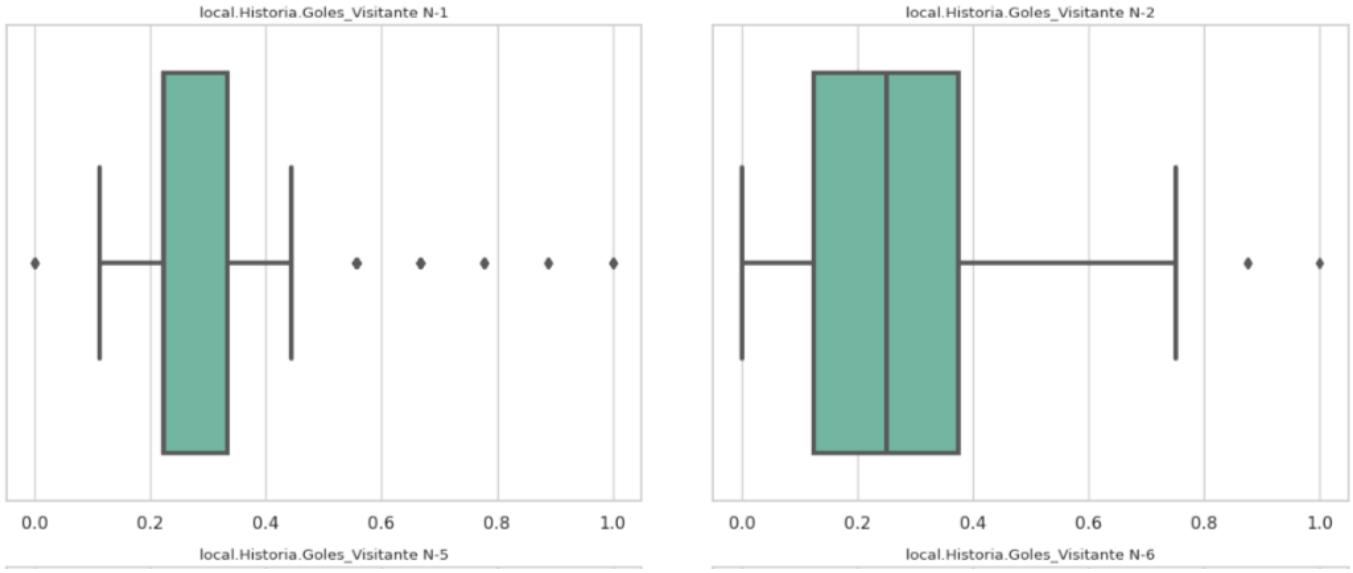
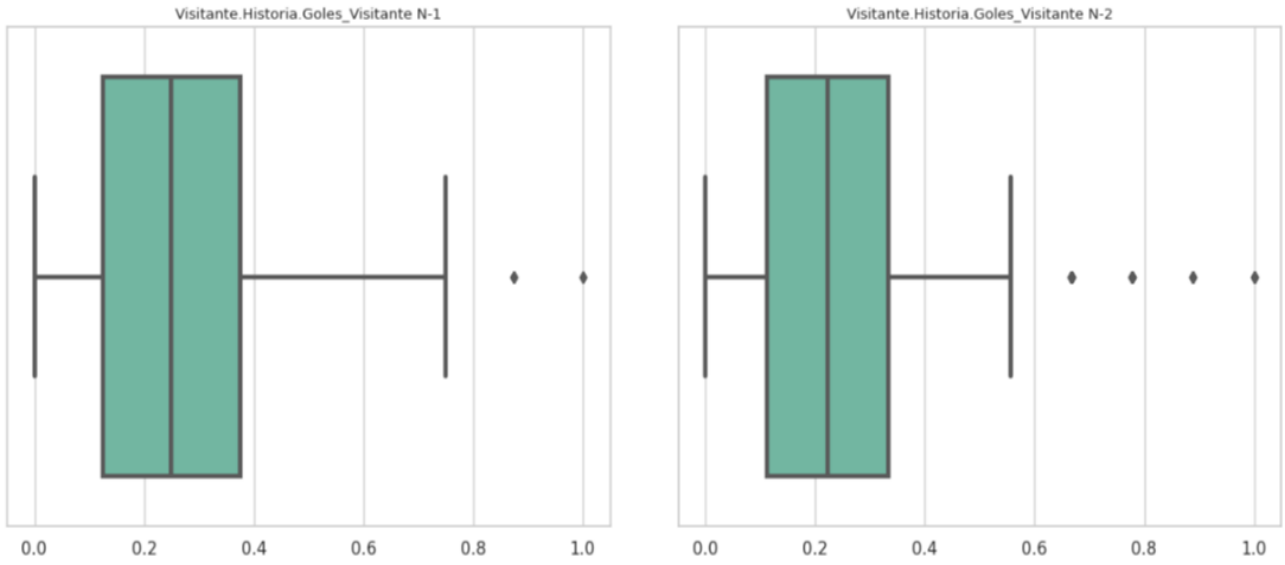


**Figura 4.**

*Diagrama de cajas de las características "visitante_Historia_Goles_visitantes N-1' y 'visitante_Historia_Goles_visitantes N-2'.*




En términos de distribución, las variables objetivo Goles local y Goles visitante son creadas como características binarias, que tienen una distribución sesgada a la derecha lo que significa que la mayoría de los valores se concentran en valores bajos o medios, mientras que los valores muy altos son menos comunes. Esto es común en el fútbol, ya que la mayoría de los partidos tienen un número relativamente bajo de goles (empate o 1 gol), pero en algunos partidos se pueden marcar muchos goles (más de 1 gol), esta distribución se representa en la **Figura 5 y Figura 6.**

**Figura 5.**

*Distribución de la variable objetivo 'Goles Local', la clase 0 representa el 60% y la clase 1 representa el 40%*

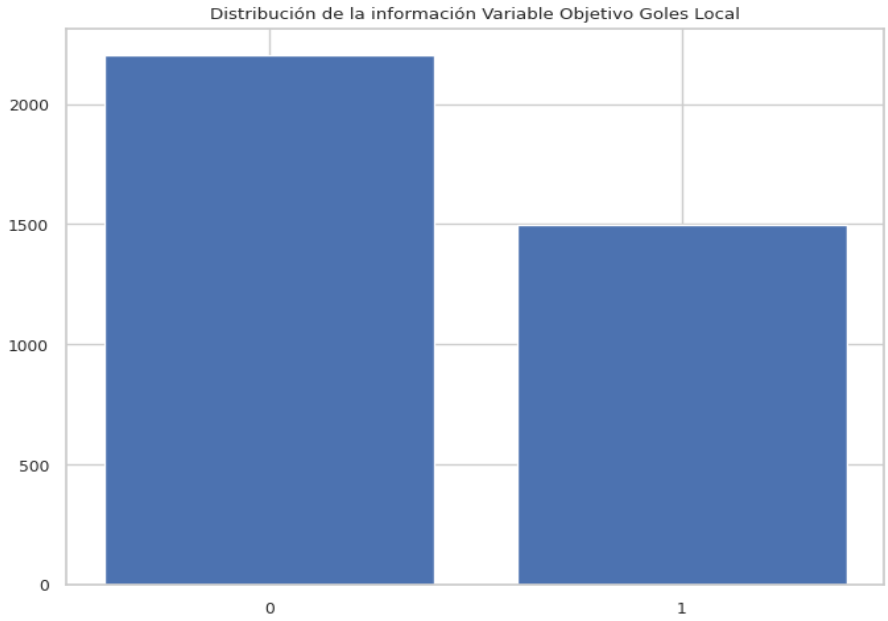
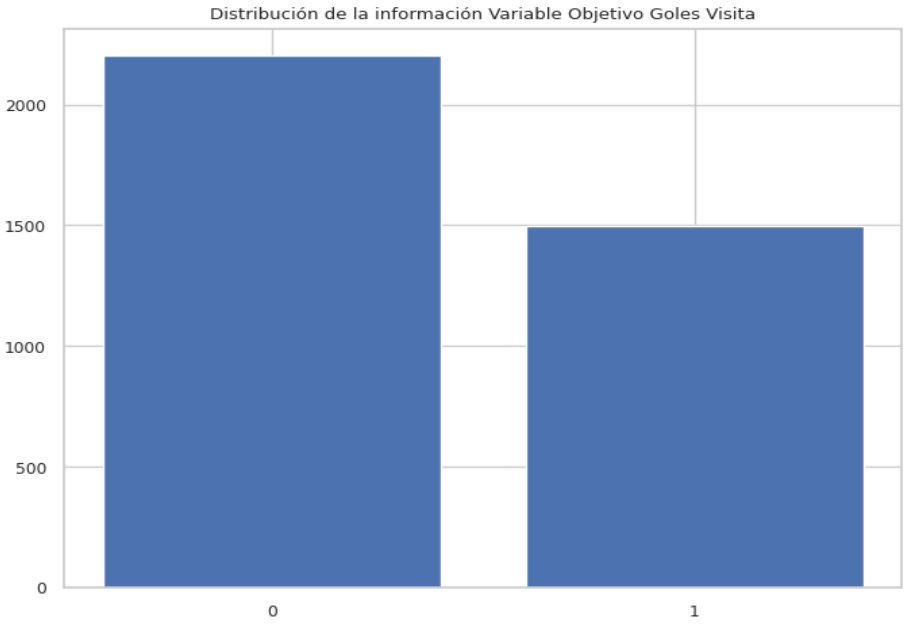



**Figura 6.**

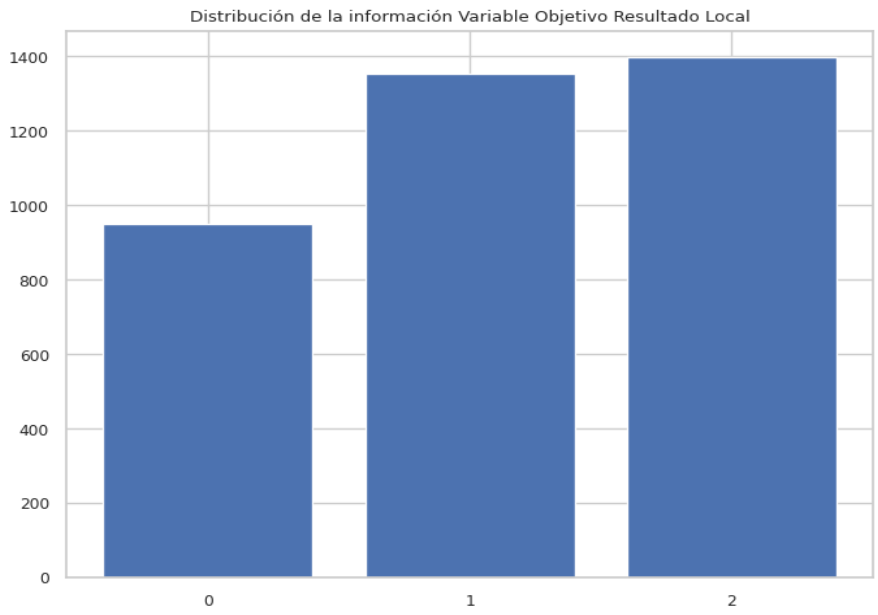
*Distribución de la variable objetivo 'Goles Visita', la clase 0 representa el 60% y la clase 1 representa el 40%*



En cuanto al resultado del partido, la característica “Resultado Local” es una variable discreta que puede tomar tres valores diferentes: victoria para el equipo local, victoria para el equipo visitante o empate. Esta variable a menudo tiene una distribución desequilibrada, ya que la mayoría de los partidos terminan con una victoria para uno de los equipos, mientras que los empates son menos comunes, por ello 2 que representa las victorias para el equipo local y 1 las victorias del equipo visitante están más equilibradas, mientras que cero que representa los empates es más baja como se observa en la **Figura 7.**

**Figura 7.**

*Distribución de la variable objetivo 'Resultado Local', la clase 0 representa el 25% y la clase 1 representa el 37% y la clase 2 representa el 38%*




## 7.2 Preprocesamiento de la información

Como alternativa de preprocesamiento de datos se enfrentaron diferentes tipos de retos, tanto a nivel estratégico como de forma operativa del manejo de la información, si bien se replanteó la cantidad de características que tomaron posición en el modelo como aportantes, la relación entre estas y la cantidad de registros tomados permitía la existencia de errores del sesgo pues notamos

que los registros eran  un poco cortos a la hora de relacionarlos con la cantidad de equipos y sus posibles combinaciones de resultados, es decir, para cada equipo se tenía un número pequeño de muestras en el total de registros, más aun considerando que los equipos entran y salen del torneo.

Aumentar la data uniéndose con información y tendencias de otros torneos fue una alternativa considerada pero desvalorizada, pues el inconveniente original se sostenía, se consideró también la generación sintética de datos y de características, luego de las primeras corridas y no obtener los resultados esperados se generó una base de datos con nuevas características, trayendo no 7 sino 10 juegos históricos, pero por la misma naturaleza del dataset se optó por un proceso de reducción de dimensionalidad, mismo que no mostró mejor rendimiento que el aumento de características. Los algoritmos de PCA y SVD no fueron eficaces por la no linealidad de los datos, al realizar dicho proceso por el algoritmo del T-SNE no hubo una mejor respuesta y dicho comportamiento se notó en los resultados de los modelos ejecutados con los datos originales y reducidos.

### 7.3 Selección datos de prueba y entrenamiento

La partición del modelo en las escalas entrenamiento y test tuvo distintas consideraciones, tales como tomar para el test únicamente los últimos datos, considerando que son los más similares a los próximos juegos, también se discutió la alternativa de usar esta información más reciente para entrenar y que el modelo genere pesos a partir de los datos más actuales, buscando estar más ajustado a los próximos resultados, pensando en los contrapesos de estas ideas y en la importancia que tienen los datos de la historia inmediata en los partidos, se decidió que los últimos datos debían ser protagonistas en la estabilización de parámetros y que a su vez algunos de estos datos deberían usarse para test, fue por ello que la partición se generó de manera aleatoria, más aún, se utilizó el recurso de validación cruzada para limitar el sobreajuste y que cada dato fuera tenido en cuenta para la optimización de los pesos.

La distribución entre datos de entrenamiento y de prueba fue de 80/20, teniendo en cuenta que la validación cruzada se iba a realizar, se generaron 10 pliegues, el rendimiento del modelo fue mucho mejor cuando se evalúa con los datos de prueba, la precisión tuvo una mejoría notable, también la matriz de precisión mostró un emparejamiento en el acierto de cada parámetro.

Luego de esta validación cruzada se limitan los riesgos de sobreajuste, además que la precisión del entrenamiento tiene una similitud clara con la de la prueba, además desde los parámetros dados en la rejilla de optimización se consideraron datos no muy extremos para limitar tal alternativa.

### 7.4 Iteraciones y evolución

Se llevó a cabo un proceso iterativo para crear un modelo de predicción de apuestas deportivas, que consistió en varias fases de prueba y ajuste. En la primera fase, se realizó el preprocesamiento de datos para limpiar y transformar los datos, y se decidió crear un dataframe alternativo con más información histórica, tomando los últimos 10 partidos en lugar de los 7 anteriores, lo que resultó en un conjunto de características adicionales y valores. También se separaron las variables de entrenamiento y prueba para utilizar los registros más nuevos en los modelos de predicción. En la segunda fase, se seleccionaron y entrenaron diferentes modelos de predicción para ambos datasets, y se evaluaron las métricas con la matriz de confusión y la curva ROC, además de realizar una validación cruzada para encontrar los mejores parámetros para cada modelo. En la tercera fase, se evaluó el rendimiento de los modelos y se tomaron decisiones para ajustar los hiper parámetros y reducir la dimensionalidad en busca de mejores métricas, utilizando técnicas como PCA, Kernel PCA y TSNE. Los datos se evaluaron con modelos como Histogram Gradient Boosting y Extreme Gradient Boosting. Cada iteración tuvo como objetivo mejorar el modelo en una fase específica del proceso, lo que eventualmente resultó en un modelo más preciso y confiable para la predicción de apuestas deportivas.

### 7.5 Evaluación

En la construcción del modelo y el ajuste de sus parámetros es importante reconocer la orientación de la información y el tipo de datos que enfrenta, interpretar que tipo de modelos puede ajustarse más a las características y patrones en la base de datos; se decidió optar por un enfoque clásico a través de las estructuras de los árboles, se buscó generar mejores resultados y ajustar cada uno mediante una rejilla de optimización buscando mediante la validación cruzada limitar el posible sobreajuste, tal como se describió en la metodología el Hist Gradient Boosting Classifier

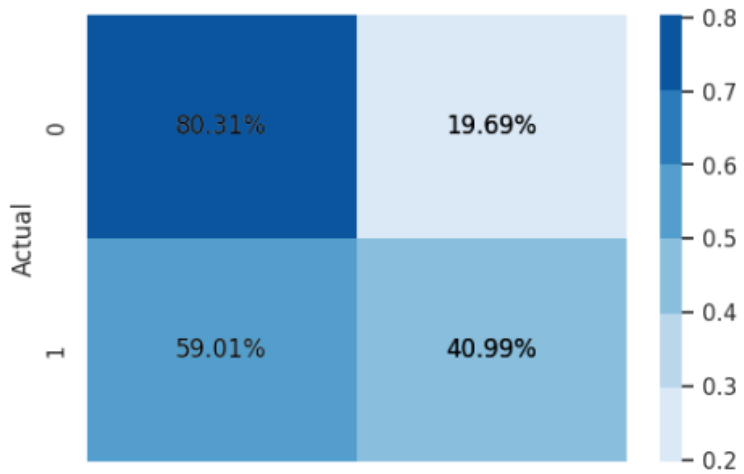
fue el modelo con mejor ajuste, por sus condiciones incrementales respecto a los demás esquemas de árboles.

La evolución desde la corrida inicial al modelo final tiene diferencias significativas vistas desde la métrica elegida por excelencia para la elección del modelo.

El primer modelo de tipo árbol probado fue DecisionTreeClassifier es un algoritmo de aprendizaje supervisado que crea un árbol de decisiones basado en las características de los datos de entrenamiento, probado con sus parámetros por defecto, además de max_depth= 5, criterion = 'gini', obteniendo una precisión del 65% y una matriz de confusión como se presenta en la **Figura 8.** Aunque este modelo proporcionó una precisión aceptable, se buscaron enfoques más sofisticados para mejorar aún más la precisión de la clasificación.

**Figura 8.**

*Matriz de confusión modelo Decision Tree Classifier.*




| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.68 | 0.80 | 0.74 | 452 |
| 1 | 0.57 | 0.41 | 0.48 | 283 |
| accuracy | | | 0.65 | 735 |
| macro avg | 0.63 | 0.61 | 0.61 | 735 |
| weighted avg | 0.64 | 0.65 | 0.64 | 735 |

Seguido de esto, se implementó un Random Forest, que es un conjunto de árboles de decisión entrenados con diferentes subconjuntos de características y datos de entrenamiento. Este

modelo mejoró la precisión a 0.73. Al utilizar múltiples árboles y combinar sus predicciones, el Random Forest reduce el sobreajuste y proporciona una mayor generalización para datos nuevos.

Posteriormente, se aplicó un Gradient Boosting, que alcanzó una precisión de 0.75. El Gradient Boosting es un método de ensamblaje que combina varios modelos más débiles para crear uno más fuerte. A diferencia del Random Forest, el Gradient Boosting construye los árboles de forma secuencial, mejorando iterativamente los errores cometidos por los modelos anteriores.

Luego, se evaluaron dos modelos adicionales. Se calculó un LightGBM Classifier, el cual obtuvo una precisión de 0.75. LightGBM es una implementación eficiente y de alto rendimiento del Gradient Boosting, que utiliza técnicas de optimización y estructuras de datos específicas para acelerar el entrenamiento y la predicción.

Finalmente, se aplicó un Histogram Gradient Boosting, el cual logró una precisión de 0.75. Este algoritmo utiliza histogramas discretos para construir árboles, lo que permite un procesamiento más rápido de los datos y una mayor escalabilidad en comparación con el Gradient Boosting tradicional, estos resultados se pueden observar en la **Tabla 1.**

**Tabla 1.**

*Resultados de los modelos y parámetros utilizados.*

| **Técnicas utilizadas** | **Exactitud** | **Parámetros** |
|---|---|---|
| Hist Gradient Boosting Classifier | 75% | *l2_regularization* = 0.001<br>*learning_rate* = 0.01<br>*max_depth* = 5<br>*max_iter* = 1000 |
| Gradient Boosting Classifier | 75% | *n_estimators* =100<br>*learning_rate* = 0.1<br>*max_depth* = 3 |
| LGB | 74% | *objective* = binary<br>*metric* = binary_logloss<br>*boosting_type* = gbdt<br>*num_leaves* = 31<br>*learning_rate* = 0.05<br>*n_estimators* = 100 |
| Random Forest Classifier | 73 % | *criterion* = gini<br>*max_depth* = 20<br>*max_features* = 9<br>*n_estimators* = 200 |

| | | |
|---|---|---|
| K-Nearest Neighbors (KNN) | 70% | *n_neighbors* =10 |
| Decision Tree Classifier | 65% | *max_depth* = 5<br>*criterion* = gini<br>*random_state* = 123<br>*min_samples_split* = 5 |
| Logistic Regression | 65% | *max_iter* = 1000<br>*solver* = sag |
| BernoulliNB | 64% | *Binarize* = True |
| SVC | 64% | *Kernel* = 'rbf'<br>*C* = 1 |
| GaussianNB | 63% | *Priors* = None<br>*var_smoothing* = 1e-09 |

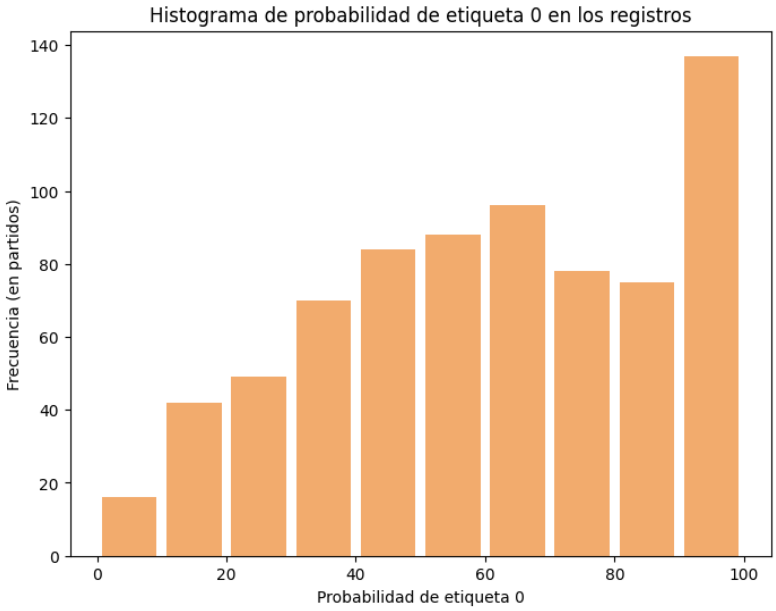
El modelo final seleccionado fue generado luego de evaluar una rejilla de optimización definidos con las siguientes posibilidades en los hiperparámetros definidos del modelo: 'learning_rate': [0.1, 0.01, 0.001], 'max_depth': [3, 5, 7], 'max_iter': [100, 500, 1000], 'l2_regularization': [0.1, 0.01, 0.001].

### 7.6 Implementación

Para evaluar las decisiones a tomar con el modelo se debe tener seguridad en cada decisión, basada en la parametrización del modelo y la cuota que aporta la casa de apuestas escogida, pensando en esta situación se genera el histograma de la probabilidad de cada decisión para la etiqueta 0, es decir con qué probabilidad determina el modelo que la etiqueta efectivamente es 0, un ejemplo claro sería, para el partido 1 se selecciona la etiqueta 0 con una probabilidad calculada de 85% es decir que el modelo determina con una seguridad del 85% que dicha etiqueta es 0, apoyado en este ejemplo se crea un histograma, agrupado la cantidad de partidos que están en seguridad entre 0% y 10% que la etiqueta sea 0 y así llegando a formar 10 grupos con posibilidades en intervalos de 10%.

**Figura 9.**

*Histograma de probabilidad de predicción.*




| rangos | frecuencias % |
|---|---|
| 0.0 | 2.18 |
| 10.0 | 5.71 |
| 20.0 | 6.67 |
| 30.0 | 9.52 |
| 40.0 | 11.43 |
| 50.0 | 11.97 |
| 60.0 | 13.06 |
| 70.0 | 10.61 |
| 80.0 | 10.20 |
| 90.0 | 18.64 |
| 100.0 | NaN |

En la **Figura 9.** se determina que en los rangos 0-10%, 10-20%, 30-40% y 70-80%, 80-90%, 90-100% son los que más aportan para tomar decisiones sobre las apuestas y se nota que son la gran mayoría de datos, más aún, generando la tabla de valores se nota que en las posibilidades más inseguras están acumulados son aproximadamente el 23% del total de los partidos, donde se sugeriría no apostar, por la ambigüedad de las respuestas.

Partiendo del anterior análisis para cada encuentro se centra la toma de decisiones, generando una herramienta para apostar de una manera más consciente, determinada por el tipo de riesgo del apostador, la ganancia esperada como resultado de la actividad y el tiempo de ejecución y estudio de los resultados.

## 8. Discusión

Desde el diseño de los modelos para iniciar el proceso de modelación y entrenamiento para el desarrollo del proyecto y sus rejillas de optimización se tuvo presente la idea de limitar el sobreajuste a lo largo del ejercicio, llevando el mismo a diferentes puntos convenientes y reflejados en los resultados, tanto de la matriz de confusión como en la precisión, además fue un factor clave en el modelo seleccionado para generar los pronósticos, pues el modelo de árbol elegido se seleccionó luego de optimizarse con unos parámetros posibles que limitaban la profundidad en sus ramas, buscando interpretar más la generalidad y menos la puntualidad de las características, así mismo el modelo fue llevado a su estado óptimo luego de un proceso de validación cruzada, pues esta práctica, al generar entrenamiento con cada porción de datos y permite una mejor adaptación al comportamiento completo de los datos.

Interpretando los resultados obtenidos por las métricas elegidas desde la medición de los modelos y llevándolos a la regla de negocio, se encuentra un producto que aporta desde la probabilidad y seguridad en decisiones a la toma de decisiones bajo diferentes circunstancias, genera valor para una interpretación del contexto de cada evento, unido a factores puntuales de cada condición conduce a un comportamiento sistemático orientado a los logros trazados, pues tomar decisiones basadas sólo en un modelo no es una idea de negocio efectiva, es fundamental interpretar la actividad de apuesta como una acción de inversión, como tal debe tener definida una clara gestión de riesgos y así mismo un perfilamiento y estrategias para lograr los resultados esperados apalancados en las herramientas disponibles, en este caso, el modelo en desarrollo

El desarrollo de la herramienta diseñada puede ser generada de manera completa, para este proceso es necesario destacar la importancia de la privacidad de los datos, tanto los generados para el entrenamiento y ajuste de parámetros, como los modelos e hiperparámetros encontrados, si bien es un desarrollo ambientado en una herramienta *open source*, es un modelo que por su misma naturaleza permite generar apuestas masivas basadas en el mismo evento probabilístico, llegando a afectar el rendimiento de las casas de apuestas en dichos escenarios deportivos.

El uso controlado de apuestas y del mecanismo diseñado para desarrollarlas debe ser supervisado constantemente para evitar caer en excesos y continuar tomando decisiones sobre eventos estadísticos y estocásticos.

Herramientas de monitoreo sobre el rendimiento generan valor a la hora del reentrenamiento, pues al ser un modelo sujeto a condiciones tan cambiantes mes a mes y sobre todo temporada a temporada, requiere un diseño de operaciones soportado para interpretar los cambios en tendencias y tener un acople ágil, evitando generar sesgos de eventos históricos que discrepan de los acontecimientos propios del presente.

En los servicios de la nube de grandes herramientas existen mecanismos para generar monitoreo sobre los resultados generados por los modelos desplegados en ellas, este tipo de herramienta generaría un escenario óptimo de reentrenamiento y estructuración sistemática de ajuste de parámetros para llegar de una manera más precisa al pronóstico de los resultados en juego.

## 9. Conclusiones

La herramienta de clasificación propuesta será valiosa para fanáticos y apostadores deportivos que desean tomar decisiones informadas y basadas en datos objetivos, la recopilación de históricos permitirá predecir con mayor precisión los resultados de eventos, lo que puede llevar a resultados más exitosos en las apuestas.

El soporte estadístico aportado desde los modelos clásicos aporta más información para la toma de decisiones conscientes, en el entorno de inversiones y apuestas deportivas las elecciones deben estar sustentadas por principios claros, establecidos previamente para mitigar la mayor cantidad de riesgos posibles y gestionar el capital de manera óptima.

El modelo elegido fue el un Histogram Gradient Boosting, logró una precisión de 0.75, su diseño permite un procesamiento de datos ágil, usando histogramas discretos para generar los árboles de decisión, es un algoritmo robusto, que a través de una rejilla de optimización generó valores esperados altos en los datos de test, es muy bien valorado en comparación con los otros métodos probados; es evidente la volatilidad y alta variabilidad en los resultados deportivos, por ello es clave tener un modelo que facilite el constante reentrenamiento bajo diversos parámetros, pero a su vez también restrinja naturalmente el sobreajuste, este tipo de herramienta se soporta en su interpretación sobre la importancia de algunas variables, generando lugar a la interpretación de los resultados del modelo y el análisis del contexto de cada situación.

La toma de decisiones en las inversiones de activos corrientes en sucesos deportivos o aleatorios representan gran atractivo para diferentes públicos actuales, las apuestas deportivas representan actualmente una gran competitividad en el mercado moderno; este tipo de herramientas diseñadas deben contar con un soporte adecuado al flujo de información relacionado con sus transacciones, luego el desempeño y el entorno de ejecución deben ser controlados según la demanda recurrente.

Las casas de apuestas tienen a su vez diferentes modelos que son generados para evitar pérdidas en sus operaciones, estos modelos generan cuotas según las probabilidades de cada evento deportivo, dichos modelos son constantemente retados y mejorados, desde allí toma la principal importancia la revisión continúa de las predicciones del modelo, su apoyo a la toma de decisiones y su vez la relación entre la seguridad en los pronósticos y las cuotas de las casas deportivas.

## 10. Recomendaciones

Soportados en la dinámica del mejoramiento constante de los modelos y búsqueda del método óptimo para el caso de uso actual, se tienen diferentes alternativas de robustez que podrían ser aplicadas al proyecto actualmente desarrollado, si bien fueron consideraciones que en su momento se tuvieron en cuenta, por las limitaciones en tiempo se decidieron dejar para próximas ocasiones.

Como en muchos proyectos de Machine Learning (ML) y analítica avanzada de datos, la información juega un papel fundamental para el desarrollo de modelos y consecución de los objetivos, este proyecto no es diferente, si bien hay gran parte de la información explicada en la base de datos obtenida, existen diversas variables que pueden llegar a ser aportantes y permitirían una mejora de los resultados y métricas del desarrollo generado, información como técnicos de cada partido y el resultado entre sus enfrentamientos, jugadores titulares y suplentes, fechas de partidos oficiales de la Federación Internacional de Fútbol Asociación (FIFA) cercanos y la información sobre la participación de los equipos en torneos diferentes al que está en análisis, muchas veces los árbitros juegan también papeles definitivos y este tipo de información puede agregar eficacia a las predicciones generadas.

Enfocados en mejorar algunos resultados se permite abrir las estructuras de ML para generar una herramienta multimodal y generar un desarrollo con estructuras más robustas por naturaleza, el aprendizaje profundo es una buena alternativa para llevar a cabo dicho propósito, pues las redes neuronales en muchas ocasiones implican aumentar los recursos para un funcionamiento más ajustado pero, es muy factible que este tipo de técnicas generen un mayor retorno visto desde un ámbito más general; bajo esta premisa de usar herramientas con más garantías y más exigencia en rendimiento, el AutoML toma lugar también como una técnica que propone un mejor desempeño desde su génesis conceptual, otro complemento funcional sería utilizar aprendizaje federado, dado que se convierte en una salida útil para fraccionar los escenarios del proyecto y desarrollar un trabajo colaborativo más fluido entre los diferentes actores que tienen incidencia en las estrategias de creación del modelo.

Las recomendaciones son las futuras y posibles líneas de investigación que llevarán a resolver problemas relacionados con la presente investigación.

## Referencias


Breiman, L. (2001). *Random forests.* Machine Learning, 45, 5-32.

Daniel, Berrar., Philippe, Lopes., Werner, Dubitzky. (2019). *Incorporating domain knowledge in machine learning for soccer outcome prediction.* Machine Learning, 108(1):97-126.

García, S., & Herrera, F. (2008). *An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons*. Journal of Machine Learning Research, 9, 2677-2694.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). *Lightgbm: A highly efficient gradient boosting decision tree.* In Proceedings of the 31st International Conference on Neural Information Processing Systems (pp. 3149-3157).

Müller, A. C., & Guido, S. (2016*). Introduction to machine learning with Python: a guide for data scientists*. " O'Reilly Media, Inc.".

Understat. (s. f.). *Understat: Expected goals (xG) statistics for EPL, La Liga, Bundesliga, Serie A, and more!*. Recuperado de https://understat.com/

Youri, Geurkink., Jan, Boone., Steven, Verstockt., Jan, Bourgois. (2021*). Machine learning-based identification of the strongest predictive variables of winning and losing in Belgian professional soccer.* Applied Sciences, 11(5):2378-.

Yu-Chia, Hsu. (2020*). Using Machine Learning and Candlestick Patterns to Predict the Outcomes of American Football Games*. Applied Sciences, 10(13):4484-.

Zhang, J., Lu, Y., Zhou, X., & Xie, H. (2020). *A deep learning model for multiclass classification problem in sports betting*. Journal of Intelligent & Fuzzy Systems, 38(6), 7153-7163.

Zhao, S., Yang, J., Liu, Y., & Wu, J. (2021*). A novel method of sports betting prediction based on deep learning model*. Journal of Ambient Intelligence and Humanized Computing, 12(4), 3327-3338.