CIENCIA DE DATOS ABIERTOS DE FÚTBOL

MODELO PREDICTIVO RESULTADOS DE LA EUROCOPA

TRABAJO DE FIN DE MÁSTER

CANDELA BELINCHÓN RODRIGUEZ
FERMÍN SILVERA NEGRÍN

Contenido

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

El avance del Data Science ha revolucionado múltiples campos, incluyendo el ámbito deportivo. Este proyecto se centra en el uso de técnicas avanzadas de análisis de datos y Machine Learning (aprendizaje supervisado) para predecir los resultados de los partidos de la UEFA, en este caso concreto, pero siendo este escalable a cualquier torneo.

Utilizando un conjunto de datos que abarca los partidos de clasificatoria desde 1960 hasta 2024, se aplican diversas metodologías de limpieza y preparación de datos, tales como la eliminación de valores nulos y la codificación de variables categóricas.

El análisis se enfoca en la creación de modelos predictivos que emplean algoritmos de clasificación como Logistic Regression, Random Forest, y Redes Neuronales (MLPClassifier). Cada modelo se entrena y evalúa utilizando métricas de precisión, matrices de confusión y reportes de clasificación para determinar su efectividad en la predicción de resultados basados en características como el equipo local, el equipo visitante, la ronda del torneo, y las probabilidades de gol calculadas para cada país.

El modelo de Logistic Regression destaca por su simplicidad y robustez, demostrando una precisión significativa en la predicción de resultados de partidos. Para validar la utilidad práctica de los modelos, se realizaron predicciones sobre partidos futuros de la Eurocopa, mostrando la capacidad del enfoque propuesto para prever con precisión los posibles desenlaces.

Este estudio no solo contribuye al campo de la analítica deportiva al proporcionar herramientas predictivas basadas en datos históricos, sino que también se puede aplicar a las apps dedicadas a las apuestas deportivas, como suplemento premium para ayudar a aquellas personas que quieran apostar con una base estadística detrás.

ABSTRACT

The advancement of Data Science has revolutionized multiple fields, including the sports field. This project focuses on the use of advanced data analysis and Machine Learning (supervised learning) techniques to predict the results of UEFA matches, in this specific case, but this is scalable to any tournament.

Using a data set covering qualifying matches from 1960 to 2024, various data cleaning and preparation methodologies are applied, such as null value removal and categorical variable coding.

The analysis focuses on the creation of predictive models that use classification algorithms such as Logistic Regression, Random Forest, and Recurrent Neural Networks (MLPClassifier). Each model is trained and evaluated using accuracy metrics, confusion matrices, and ranking reports to determine its effectiveness in predicting outcomes based on characteristics such as home team, away team, tournament round, and calculated goal probabilities for each country.

The Logistic Regression model has been highlighted for its simplicity and robustness, demonstrating significant accuracy in predicting match results. To validate the practical usefulness of the models, predictions were made about future Euro Cup matches, showing the ability of the proposed approach to accurately predict possible outcomes.

This study not only contributes to the field of sports analytics by providing predictive tools based on historical data but can also be applied to apps dedicated to sports betting, as a premium supplement to help those who want to bet on a statistical basis.

GLOSARIO

home_team = equipo de casa

away_team = equipo visitante

home_team_code = código del equipo de casa

away_team_code = código de equipo visitante

home_score = goles en casa

away_score = goles como visitante

winner = ganador

winner_reason = razón de victoria

year = año

status = estado

round = ronda

score_dif = diferencia de goles

home_winner = equipo local ganador

MLP = Multilayer Perceptron

RFC = Random Forest Classifier

INTRODUCCIÓN

Los avances tecnológicos de las últimas décadas han dado lugar a una sociedad altamente digitalizada. El desarrollo del Internet de las Cosas (IoT) ha incrementado considerablemente la cantidad de sensores y dispositivos conectados a la red, llegando a un promedio de 1,7 MB de datos generados por persona por segundo en 2022. Estos volúmenes de datos, tras su procesamiento, pueden proporcionar información valiosa, un proceso conocido como Big Data. Hoy en día, las grandes empresas invierten en Big Data para alcanzar sus objetivos de manera más eficiente y segura. El fútbol, uno de los sectores más lucrativos, no es una excepción. Según, la industria del fútbol profesional español genera una actividad económica equivalente al 1,4% del PIB y emplea a 185.000 personas.

Desde sus inicios, el fútbol ha sido sinónimo de competitividad, habilidad y capacidad física. Hoy en día, el deporte se ha modernizado mediante tecnologías que mejoran la experiencia deportiva. La estadística se ha convertido en un elemento fundamental para los clubes y las ligas, con un gasto proyectado en Big Data de más de 4.150 millones de euros para 2025.

El análisis de datos en el fútbol se ha vuelto esencial para entender el juego, tomar decisiones estratégicas y mejorar el rendimiento de los equipos. Los entrenadores y directivos pueden identificar fortalezas y debilidades del equipo y los jugadores, lo que resulta crucial en procesos como el fichaje de nuevos talentos. La información se obtiene mediante herramientas de scouting y sensores conectados y su análisis permite predecir resultados e identificar tendencias de juego, aportando valor en la selección de estrategias y jugadores adecuados para cada situación. Un ejemplo notable es Kevin de Bruyne, quien gracias a un análisis de datos que demostraba su relevancia, logró una renovación millonaria en la Premier League.

El análisis de datos también se extiende a los aficionados, mejorando la experiencia del espectador con estadísticas complejas y en tiempo real durante las retransmisiones de partidos, como se vio en la aplicación FIFA+ Stadium Experience durante el Mundial de Qatar. Además, el análisis del público y sus tendencias ayuda a los clubes a tomar decisiones informadas para mejorar la experiencia del espectador y aumentar las ventas durante los partidos. Tecnologías como el VAR y el balón inteligente están haciendo del fútbol un deporte más justo.

En este proyecto, nos enfocamos en la predicción de los resultados de la Eurocopa de fútbol utilizando datos de partidos ganados y perdidos, diferencia de goles, y victorias como local o visitante, más que en los datos individuales de los jugadores. La minería de datos nos permitirá identificar patrones y tendencias ocultas en grandes conjuntos de datos, transformando el Big Data del fútbol en conocimiento accionable.

Además de la predicción de resultados, este tipo de análisis presenta otras oportunidades de negocio, como el desarrollo de plataformas de apuestas deportivas más precisas, la creación de herramientas para la planificación estratégica de los clubes y la mejora de las estrategias de marketing para atraer y retener a los aficionados.

OBJETIVOS

El objetivo del proyecto es aplicar las etapas de la metodología CRISP-DM para comparar los conjuntos de datos abiertos de fútbol disponibles en Internet. Así se pretende flexibilizar la metodología de minado para estudiar los conjuntos disponibles y crear modelos que permitan sacar conclusiones. Este objetivo general se desglosa en los siguientes objetivos específicos:

Recopilar y limpiar conjuntos de datos relevantes sobre los resultados de la Eurocopa y otros datos relacionados con el rendimiento de los equipos.

- Identificar y seleccionar las fuentes de datos más adecuadas y confiables.
- Realizar la limpieza y preprocesamiento de los datos para asegurar su calidad y adecuación para el análisis.

Desarrollar y entrenar modelos de machine learning para la predicción de resultados de la Eurocopa.

- Implementar y evaluar diversos algoritmos, incluyendo Redes Neuronales (Multilayer Perceptron - MLP), Random Forest Classifier y Regresión Logística.
- Comparar el rendimiento de estos modelos utilizando métricas adecuadas como la curva ROC.

Optimizar los modelos añadiendo variables adicionales y ajustando los hiperparámetros para mejorar la precisión de las predicciones.

Desarrollar un enfoque de despliegue para los modelos predictivos y analizar los resultados obtenidos.

- Implementar los modelos en un entorno colaborativo como Google Colab.
- Evaluar la viabilidad y eficacia de los modelos en un contexto real.

Proponer estrategias de negocio basadas en las predicciones de los resultados de la Eurocopa.

- Analizar cómo las predicciones pueden influir en la toma de decisiones estratégicas en áreas como el marketing deportivo, las apuestas deportivas y la gestión de equipos.

Documentar las limitaciones encontradas durante el proceso y sugerir posibles mejoras futuras en el uso de técnicas de machine learning para la predicción de eventos deportivos.

Este enfoque permitirá no solo demostrar la aplicabilidad de técnicas avanzadas de análisis de datos en el ámbito deportivo, sino también explorar el impacto de dichas técnicas en la toma de decisiones empresariales.

DATOS

Partimos de un dataset proporcionado por https://www.kaggle.com/datasets/piterfm/football-soccer-uefa-euro-1960-2024/data matches- qualifying 1960-2024 de libre descarga, en el que se recogen para cada partido (cada fila) una serie de variables. En base a este dataset hemos de analizar si existen valores considerados anomalías, NaN y corregirlo, además de eliminar aquellas columnas que no consideremos necesarias y crear variables (nuevas columnas) para enriquecer el dataset resultante.

La información relativa al dataset original es la siguiente:

(2845 filas, 47 columnas)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2845 entries, 0 to 2844
Data columns (total 47 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   id_match               2845 non-null   int64
 1   home_team              2845 non-null   object
 2   away_team              2845 non-null   object
 3   home_team_code         2845 non-null   object
 4   away_team_code         2845 non-null   object
 5   home_score             2845 non-null   float64
 6   away_score             2845 non-null   float64
 7   home_penalty           7 non-null      float64
 8   away_penalty           7 non-null      float64
 9   home_score_total       2845 non-null   float64
 10  away_score_total       2845 non-null   float64
 11  winner                 2291 non-null   object
 12  winner_reason          2842 non-null   object
 13  year                   2845 non-null   int64
 14  date                   2845 non-null   object
 15  date_time              2762 non-null   object
 16  utc_offset_hours       2762 non-null   float64
 17  group_name             2686 non-null   object
 18  matchday_name          2845 non-null   object
 19  condition_humidity     368 non-null    float64
 20  condition_pitch        480 non-null    object
 21  condition_temperature  991 non-null    float64
 22  condition_weather      664 non-null    object
 23  condition_wind_speed   991 non-null    float64
 24  status                 2845 non-null   object
 25  type                   2845 non-null   object
 26  round                  2845 non-null   object
 27  round_mode             2845 non-null   object
 28  match_attendance       2822 non-null   float64
 29  stadium_id             2841 non-null   float64
 30  stadium_country_code   2841 non-null   object
 31  stadium_capacity       2841 non-null   float64
 32  stadium_latitude       2831 non-null   float64
 33  stadium_longitude      2831 non-null   float64
 34  stadium_pitch_length   2752 non-null   float64
 35  stadium_pitch_width    2752 non-null   float64
 36  goals                  2638 non-null   object
 37  penalties_missed       48 non-null     object
 38  penalties              6 non-null      object
 39  red_cards              339 non-null    object
 40  game_referees          2845 non-null   object
 41  stadium_city           2841 non-null   object
 42  stadium_name           2834 non-null   object
 43  stadium_name_media     2841 non-null   object
 44  stadium_name_official  2841 non-null   object
 45  stadium_name_event     2841 non-null   object
 46  stadium_name_sponsor   2804 non-null   object
dtypes: float64(17), int64(2), object(28)
memory usage: 1.0+ MB
```

Tabla 1: Información dataset inicial

Principalmente nos encontramos con el problema de no poder hacer un barrido de NaN dado que muchos de los valores son cero, pero no nulos, es decir, aporta valor que sean cero (por ejemplo, para la variable home_score, cero indica cero goles marcados).

Después escogimos las variables con las que queríamos trabajar, eliminando aquellas que no consideramos importantes en un principio.

A continuación, renombramos valores que, al ser un dataset que recoge nombres de países desde 1960, han cambiado con el tiempo aunque hagan alusión a lo mismo (por ejemplo, se unificó "German Dem. Rep." con "Germany")

Añadimos también una columna que calculase la diferencia entre los goles marcados por el equipo local y por el visitante, y otra columna que diese como output True si esta diferencia era estrictamente mayor que cero o False en caso contrario.

Finalmente, puesto que algunos de los partidos que se recogían en el dataset fueron cancelados, esa información no nos servía, por lo que nos quedamos únicamente con aquellos partidos cuyo 'status' fuese finalizado.

Por tanto la información relativa al dataset definitivo (tras limpieza) sería la siguiente:

```
[20] df1.shape

(2831, 13)

df1.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2831 entries, 0 to 2830
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   home_team       2831 non-null   object
 1   away_team       2831 non-null   object
 2   home_team_code  2831 non-null   object
 3   away_team_code  2831 non-null   object
 4   home_score      2831 non-null   float64
 5   away_score      2831 non-null   float64
 6   winner          2282 non-null   object
 7   winner_reason   2831 non-null   object
 8   year            2831 non-null   int64
 9   status          2831 non-null   object
 10  round           2831 non-null   object
 11  score_dif       2831 non-null   float64
 12  home_winner     2831 non-null   bool
dtypes: bool(1), float64(3), int64(1), object(8)
memory usage: 268.3+ KB
```

Tabla 2: Información dataset primera fase de modelización

Descubrimientos Realizados Durante el Procesamiento y Cualquier Implicación para Trabajos Posteriores

Durante el procesamiento del dataset de partidos de la Eurocopa, se realizaron varios descubrimientos importantes:

Valores Nulos y Anomalías:

- Aunque el dataset contiene valores cero en muchas columnas, estos valores no debían ser tratados como nulos, ya que aportan información significativa (por ejemplo, un marcador de cero goles).
- Se identificaron algunas columnas con valores NaN que fueron corregidos o eliminados si no eran esenciales para el análisis.

Unificación de Nombres de Países:

- Se unificaron las denominaciones de países que habían cambiado con el tiempo. Por ejemplo, "German Dem. Rep." se unificó con "Germany". Esta normalización fue crucial para mantener la consistencia de los datos históricos.

Creación de Nuevas Variables:

- Se añadieron columnas que calcularon la diferencia de goles entre el equipo local y el visitante, y una columna adicional que indica si el equipo local ganó el partido (True si la diferencia de goles era mayor que cero, False en caso contrario).

Filtrado de Partidos Cancelados:

- Se eliminaron los partidos cuyo estado no era "finalizado", ya que no aportan valor al análisis predictivo.

Estos descubrimientos tienen varias implicaciones para trabajos posteriores:

- **Mejora en la Calidad del Dataset:** La limpieza y normalización de los datos aseguran que futuros análisis se basen en información consistente y precisa.
- **Modelos Predictivos Más Precisos:** Las nuevas variables creadas, como la diferencia de goles, pueden mejorar la precisión de los modelos predictivos.
- **Estandarización de Procesos:** La unificación de nombres de países y el manejo de valores nulos pueden servir como estándares para otros proyectos similares.
- **Filtrado de Datos Relevantes:** La eliminación de partidos no finalizados asegura que los análisis futuros se realicen sólo con datos relevantes.

Resumen y Conclusiones

En este proyecto, se aplicaron técnicas de machine learning para predecir los resultados de la Eurocopa utilizando un dataset de partidos históricos. A través de un proceso riguroso de análisis y modelado, se realizaron varias etapas clave, desde la comprensión y preparación de los datos hasta la evaluación de los resultados.

Principales Logros:

Recopilación y Limpieza de Datos:

- Se trabajó con un dataset de 2845 filas y 47 columnas, el cual fue limpiado y enriquecido mediante la creación de nuevas variables y la corrección de valores nulos y anomalías.

Desarrollo de Modelos Predictivos:

- Se implementaron y compararon varios modelos de machine learning, incluyendo Redes Neuronales (Multilayer Perceptron), Random Forest Classifier y Regresión Logística.
- Los modelos fueron evaluados utilizando métricas como la curva ROC, lo que permitió identificar el modelo más eficaz para la predicción de resultados.

Propuestas de Estrategias de Negocio:

- Basándose en las predicciones de los resultados, se propusieron estrategias de negocio en áreas como el marketing deportivo y las apuestas deportivas.

Conclusiones:

- **Viabilidad del Uso de Machine Learning:** Los resultados demostraron que las técnicas de machine learning son viables y eficaces para predecir resultados de eventos deportivos como la Eurocopa.
- **Importancia de la Calidad de los Datos:** La limpieza y enriquecimiento del dataset fueron fundamentales para mejorar la precisión de los modelos predictivos.

- **Aplicaciones Prácticas:** Las predicciones pueden ser utilizadas para desarrollar estrategias de negocio, demostrando el valor práctico del análisis de datos en el deporte.

Futuras Investigaciones:

- **Mejora de Modelos:** Futuras investigaciones podrían enfocarse en optimizar aún más los modelos predictivos mediante el uso de técnicas avanzadas y el enriquecimiento adicional de los datos.
- **Ampliación del Dataset:** Incluir datos de otros torneos y ligas podría mejorar la generalización y robustez de los modelos.
- **Impacto de Variables Externas:** Examinar el impacto de variables externas, como las condiciones climáticas y las lesiones de jugadores, podría proporcionar una visión más completa y precisa de los factores que influyen en los resultados de los partidos.

Este proyecto no solo demuestra la aplicabilidad de técnicas avanzadas de análisis de datos en el ámbito deportivo, sino también su potencial para influir en la toma de decisiones estratégicas en diversas áreas de negocio.

En esta sección se describe la arquitectura de referencia utilizada en el proyecto, detallando los componentes, las librerías y sus versiones empleadas a lo largo del desarrollo.

Arquitectura de Referencia

El proyecto se estructuró en torno a una arquitectura basada en Google Colab para su implementación y ejecución, aprovechando su capacidad para ejecutar código en la nube y su integración con Python. La elección de esta plataforma permite una alta flexibilidad y accesibilidad, además de proporcionar los recursos necesarios para el procesamiento de datos y entrenamiento de modelos de machine learning.

Componentes Principales

Google Colab:

- Plataforma utilizada para la implementación y ejecución del proyecto.
- Proporciona un entorno de desarrollo basado en Jupyter Notebooks, facilitando la colaboración y el acceso a potentes recursos computacionales.

Python:

- Lenguaje de programación principal utilizado para la manipulación de datos, construcción de modelos y visualización de resultados.

Librerías y Versiones Utilizadas

El proyecto hizo uso extensivo de diversas librerías de Python, cada una seleccionada por sus capacidades específicas para el manejo y análisis de datos, así como para la construcción y evaluación de modelos predictivos.

Las librerías utilizadas son las siguientes:

Pandas: Utilizada para la manipulación y análisis de datos.

Scikit-Learn: Principal librería para la construcción y evaluación de modelos de machine learning.

Seaborn y Matplotlib: Utilizadas para la visualización de datos y resultados.

Statsmodels: Utilizada para la detección de multicolinealidad en los datos.

```
import pandas as pd

from sklearn.model_selection import train_test_split, cross_val_score

from sklearn.neural_network import MLPClassifier

from sklearn.ensemble import RandomForestClassifier

from sklearn.linear_model import LogisticRegression

from sklearn.metrics import classification_report, accuracy_score, confusion_matrix

from sklearn.impute import SimpleImputer

import seaborn as sns

import matplotlib.pyplot as plt

from sklearn.preprocessing import OneHotEncoder, StandardScaler

from statsmodels.stats.outliers_influence import variance_inflation_factor

from sklearn.metrics import roc_curve, roc_auc_score
```

Integración y Flujo de Trabajo

El flujo de trabajo general del proyecto fue el siguiente:

Carga y Preprocesamiento de Datos:

- Uso de Pandas para cargar y limpiar el dataset.
- Imputación de valores faltantes y normalización de datos utilizando Scikit-Learn.

Análisis Exploratorio de Datos (EDA):

- Visualización de datos con Seaborn y Matplotlib para identificar patrones y relaciones.

Construcción y Entrenamiento de Modelos:

- Implementación de varios algoritmos de machine learning con Scikit-Learn.
- Comparación de modelos utilizando validación cruzada y métricas de evaluación.

Evaluación y Optimización de Modelos:

- Análisis de resultados y ajuste de hiperparámetros.
- Detección de multicolinealidad con Statsmodels para mejorar la estabilidad del modelo.

Visualización y Comunicación de Resultados:

- Generación de informes y gráficos para presentar los hallazgos clave.

Esta arquitectura y el conjunto de herramientas utilizadas proporcionaron una base sólida para llevar a cabo un análisis exhaustivo y construir modelos predictivos precisos, facilitando así la obtención de resultados relevantes y aplicables para la predicción de resultados de la Eurocopa y el desarrollo de estrategias de negocio.

MODELIZACIÓN

En esta sección se detallan los pasos seguidos en la modelización, incluyendo las técnicas utilizadas, el proceso de evaluación, los modelos construidos y los resultados analíticos obtenidos. La modelización se centró en tres modelos principales: Redes Neuronales (Multilayer Perceptron - MLP), Random Forest Classifier y Regresión Logística.

Pasos Previos Comunes a Todos los Modelos

Antes de desarrollar cada modelo, se realizaron los siguientes pasos iniciales para preparar los datos:

Calcular la Media de Goles Anotados:

- Se calculó la media de goles anotados por cada país tanto en partidos jugados en casa como fuera de casa.

Calcular la Probabilidad de Gol:

- Se calculó la probabilidad de gol de cada país como la media de goles anotados en casa y fuera. Esta probabilidad se incorporó al dataframe original para su uso en los modelos predictivos.

Incorporación de la Probabilidad de Gol:

- La probabilidad de gol calculada para cada país se añadió al dataframe original, enriqueciendo así el conjunto de datos con información relevante para los modelos.

Selección de variables:

- Se seleccionaron las variables para X (características) y Y (variable objetivo). La variable objetivo fue 'home_winner', que toma valores True o False según si el equipo local gana el partido.
- Las características (X) seleccionadas fueron:
  - ➢ Equipo local
  - ➢ Equipo visitante
  - ➢ Ronda
  - ➢ Probabilidad de gol del equipo local

➢ Probabilidad de gol del equipo visitante

Conversión de Variables Categóricas:

- Las variables categóricas (equipo local, equipo visitante y ronda) se convirtieron en variables numéricas utilizando OneHotEncoder.
- Los resultados de OneHotEncoder se transformaron en un DataFrame que se unió con las probabilidades de gol, creando un conjunto de datos completamente numérico y listo para el modelado.

Escalado de Características:

- Se escalaron las características para asegurar que todas las variables tengan la misma escala y evitar que algunas dominan sobre otras debido a sus magnitudes.

División del Dataset:

- El dataset se dividió en conjuntos de entrenamiento y prueba para evaluar el rendimiento de los modelos.

Construcción y Evaluación de Modelos

Redes Neuronales (Multilayer Perceptron - MLP):

- Se utilizó la clase MLPClassifier de Scikit-Learn para construir el modelo de redes neuronales.
- El modelo se ajustó utilizando el conjunto de entrenamiento y se evaluó con el conjunto de prueba.
- Se realizaron varios ajustes de hiperparámetros para optimizar el rendimiento del modelo.

Random Forest Classifier:

- Se utilizó la clase RandomForestClassifier de Scikit-Learn para construir el modelo de bosque aleatorio.
- Este modelo fue entrenado con el conjunto de entrenamiento y evaluado con el conjunto de prueba.

- Se ajustaron diversos parámetros del modelo, como el número de árboles y la profundidad máxima, para mejorar su precisión.

Regresión Logística:

- Se utilizó la clase LogisticRegression de Scikit-Learn para construir el modelo de regresión logística.
- El modelo se entrenó con el conjunto de entrenamiento y se evaluó utilizando el conjunto de prueba.
- Se exploraron diferentes técnicas de regularización para optimizar el modelo.

Resultados Analíticos Obtenidos

Cada modelo se evaluó utilizando las siguientes métricas:

- **Accuracy:** Medida de la proporción de predicciones correctas realizadas por el modelo.
- **Classification Report:** Informe detallado de la precisión, recall y F1-score para cada clase.
- **Confusion Matrix:** Matriz que muestra los verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos.

Los resultados de estas métricas proporcionaron una visión clara del rendimiento de cada modelo y ayudaron a identificar áreas de mejora. Los modelos se compararon para determinar cuál ofrecía la mejor precisión y capacidad de generalización.

Ajustes Específicos para Cada Modelo

- **Redes Neuronales (MLP):**
  - Ajuste de la estructura de la red neuronal (número de capas y neuronas por capa).
  - Modificación de la tasa de aprendizaje y el número de épocas.
- **Random Forest Classifier:**
  - Ajuste del número de árboles en el bosque.
  - Modificación de la profundidad máxima de los árboles.

- **Regresión Logística:**
  - Ajuste del parámetro de regularización.
  - Exploración de diferentes solvers para la optimización del modelo.

Estos ajustes específicos permitieron optimizar cada modelo para obtener el mejor rendimiento posible, basado en los datos disponibles y las características seleccionadas.

Redes Neuronales: Multilayer Perceptron (MLP)

En primer lugar, se procedió a crear el modelo MLPClassifier y a entrenarlo utilizando los conjuntos X_train y y_train. Posteriormente, se realizaron las predicciones y se evaluó la precisión del modelo, tal como se presenta en la siguiente tabla.

```
Accuracy: 0.7001763668430335
Classification Report:
              precision    recall  f1-score   support

       False       0.71      0.69      0.70       290
        True       0.69      0.71      0.70       277

    accuracy                           0.70       567
   macro avg       0.70      0.70      0.70       567
weighted avg       0.70      0.70      0.70       567
```

Tabla 3: Classification Report MLP

A continuación, se procede a interpretar las métricas presentadas en la tabla anterior:

Precisión Global (Accuracy): Esta métrica representa la proporción de predicciones correctas (tanto verdaderos positivos como verdaderos negativos) sobre el total de predicciones realizadas. Un valor de 0,70 sobre 1 sugiere un rendimiento adecuado, aunque con margen de mejora.

Precisión Positiva (Precision): Se define como la proporción de verdaderos positivos sobre el total de predicciones positivas (la suma de verdaderos positivos y falsos positivos).

Sensibilidad o Tasa de Verdaderos Positivos (Recall): Esta métrica mide la proporción de verdaderos positivos sobre el total de verdaderos positivos y falsos negativos. Indica la capacidad del modelo para identificar correctamente los verdaderos positivos.

Puntuación F1 (F1 Score): Proporciona un equilibrio entre precisión y sensibilidad, siendo particularmente útil cuando existe un desbalance entre clases. Un valor de 1 indica el mejor rendimiento posible, mientras que 0 indica el peor. Un valor de 0,70 es considerado bastante aceptable.

En la siguiente figura, la matriz de confusión, se presenta una tabla que detalla las siguientes cantidades:

- **TN (Verdaderos Negativos)**: Aquellos que el modelo predijo como negativos y realmente lo son.
- **FP (Falsos Positivos)**: Aquellos que el modelo predijo como positivos pero realmente son negativos.
- **FN (Falsos Negativos)**: Aquellos que el modelo predijo como negativos pero realmente son positivos.
- **TP (Verdaderos Positivos)**: Aquellos que el modelo predijo como positivos y realmente lo son.

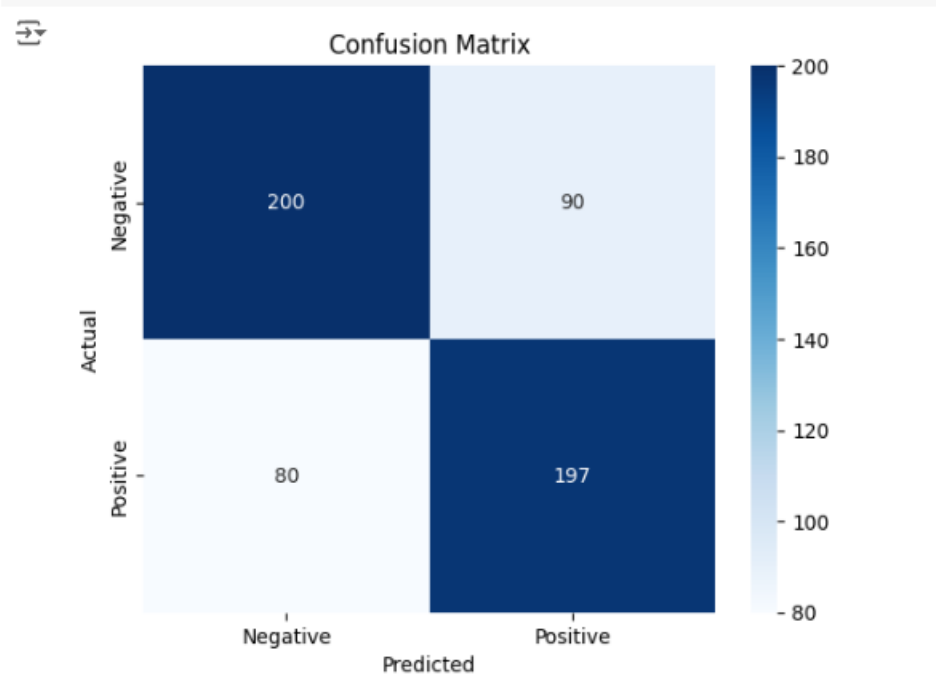

Ilustración 1: Matriz de Confusión MLP

Esta matriz permite el cálculo de las métricas anteriores.

Finalmente, hemos probado el modelo con nuevos partidos:

	home_team	away_team	round	home_goal_prob	away_goal_prob	home_winner
0	Germany	Spain	QUARTER_FINALS	2.305085	2.544324	False
1	Spain	France	PRELIMINARY	2.544324	2.166667	False
2	Spain	England	FINAL	2.544324	2.413793	False

Tabla 4: Resultados MLP

- En el caso 0: predice que en cuartos de final Alemania pierde contra España (sería un verdadero negativo, TN).
- En el caso 1: predice que en semifinales España pierde contra Francia (sería un falso negativo, FN).
- En el caso 2: predice que en la final España pierde contra Inglaterra (sería un falso negativo, FN).

Como podemos observar, a pesar del buen valor de accuracy del modelo, solo acierta uno de los tres casos. Por ello, probamos con otros modelos.

Random Forest Classifier

Se ha creado el modelo Random Forest Classifier y se ha entrenado con nuestro conjunto de prueba. Al evaluar las mismas métricas, se observa una leve mejora en la precisión global (accuracy) y en las demás métricas.

Accuracy: 0.7142857142857143
Classification Report:

	precision	recall	f1-score	support
False	0.72	0.73	0.72	290
True	0.71	0.70	0.70	277
accuracy			0.71	567
macro avg	0.71	0.71	0.71	567
weighted avg	0.71	0.71	0.71	567

Tabla 5: Classification Report RFC

Como se muestra en la matriz de confusión a continuación, aumentan los verdaderos negativos (TN) y disminuyen los falsos positivos (FP), pero disminuyen los verdaderos positivos (TP) y aumentan los falsos negativos (FN). Esto sugiere que este modelo tiende a predecir un mayor número de resultados negativos en comparación con el modelo anterior, lo que explica el incremento tanto en los verdaderos negativos como en los falsos negativos.

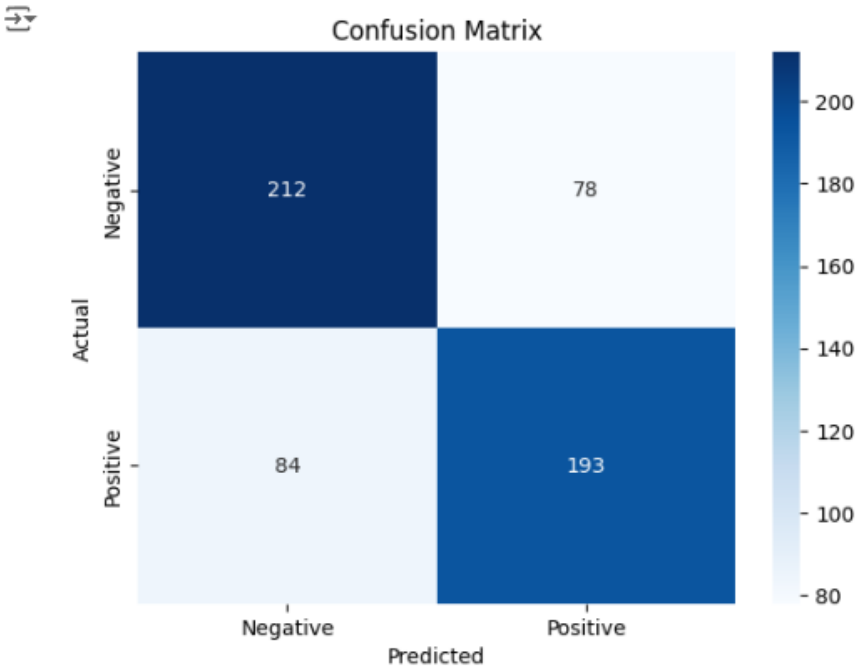

Ilustración 2: Matriz de Confusión RFC

Al probar el modelo con los nuevos partidos obtenemos los siguientes resultados:

	home_team	away_team	round	home_goal_prob	away_goal_prob	home_winner
0	Germany	Spain	QUARTER_FINALS	2.305085	2.544324	True
1	Spain	France	PRELIMINARY	2.544324	2.166667	True
2	Spain	England	FINAL	2.544324	2.413793	True

Tabla 6: Resultados RFC

En el caso 0: predice que en cuartos de final gana Alemania contra España (sería FP).

En el caso 1: predice que en semifinal gana España contra Francia (sería TP).

En el caso 2: predice que en final gana España contra Inglaterra (sería TP).

Este modelo acierta 2 de 3 casos, por lo que confirmamos que mejora respecto del MLPClassifier.

Logistic Regression

Finalmente, se procedió a la creación de un último modelo basado en una regresión logística. Se comprobó que dicho modelo presenta una mejora significativa en las métricas de precisión global (accuracy), precisión positiva (precision), sensibilidad (recall) y puntuación F1 (F1-score) en comparación con los dos modelos anteriores, alcanzando una precisión global de aproximadamente el 76%.

```
Accuracy: 0.7583774250440917
Classification Report:
              precision    recall  f1-score   support

       False       0.76      0.77      0.76       290
        True       0.75      0.75      0.75       277

    accuracy                           0.76       567
   macro avg       0.76      0.76      0.76       567
weighted avg       0.76      0.76      0.76       567
```

Tabla 7: Classification Report Logistic Regression

En este caso comprobamos un aumento tanto en TN como en TP, y una disminución en FP y FN, lo que indica una mejora en la predicción en general del modelo.

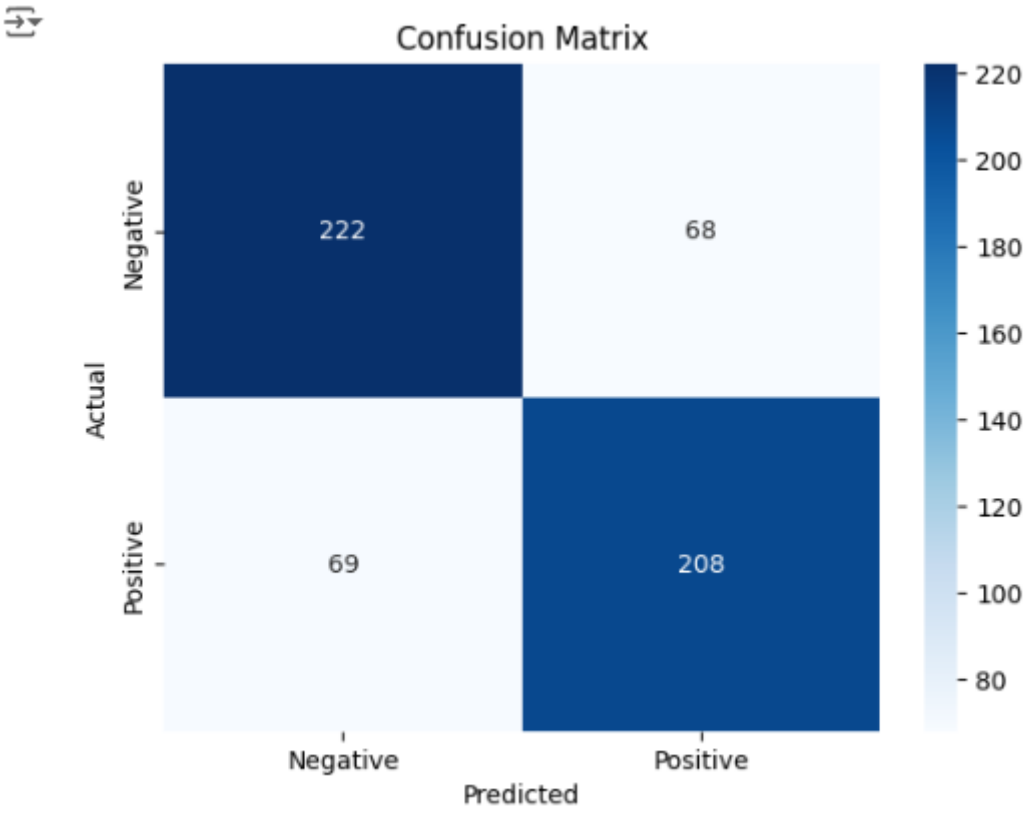

Ilustración 3: Matriz de Confusión Logistic Regression

Al probar el modelo con los nuevos partidos obtenemos los siguientes resultados:

	home_team	away_team	round	home_goal_prob	away_goal_prob	home_winner
0	Germany	Spain	QUARTER_FINALS	2.305085	2.544324	True
1	Spain	France	PRELIMINARY	2.544324	2.166667	True
2	Spain	England	FINAL	2.544324	2.413793	False

Tabla 8: Resultados Logistic Regression

En el caso 0: predice que en cuartos de final gana Alemania contra España (sería FP).

En el caso 1: predice que en semifinal gana España contra Francia (sería TP).

En el caso 2: predice que en final pierde España contra Inglaterra (sería FN).

Este modelo para los nuevos partidos acierta un solo caso de los 3, pese a las buenas métricas que presenta.

ROC

Interpretación de la Curva ROC:

La curva ROC es un gráfico que muestra la relación entre la tasa de verdaderos positivos (True Positive Rate, TPR) y la tasa de falsos positivos (False Positive Rate, FPR) para diferentes umbrales de decisión.

- **Eje Y (TPR o Sensibilidad):** Representa la proporción de positivos verdaderos que son correctamente identificados por el modelo. Es decir, la cantidad de verdaderos positivos (TP) sobre la suma de verdaderos positivos y falsos negativos (FN).
- **Eje X (FPR):** Representa la proporción de negativos verdaderos que son incorrectamente identificados como positivos por el modelo. Es decir, la cantidad de falsos positivos (FP) sobre la suma de falsos positivos y verdaderos negativos (TN).
- **Línea Diagonal (Random Guess):** Una línea diagonal desde (0,0) a (1,1) representa un clasificador aleatorio que no tiene capacidad para distinguir entre clases. Para cada punto en esta línea, TPR es aproximadamente igual a FPR.
- **Curva ROC Ideal:** Un clasificador perfecto alcanzaría el punto (0,1), donde la tasa de verdaderos positivos es 1 (100%) y la tasa de falsos positivos es 0 (0%). Cuanto más se acerque la curva ROC a este punto, mejor es el modelo.

Interpretación de los Valores de AUC (Área Bajo la Curva):

El valor de AUC mide el área bajo la curva ROC y varía entre 0 y 1.

- **AUC = 0.5:** Indica que el modelo no es mejor que un clasificador aleatorio.
- **AUC < 0.5:** Indica que el modelo es peor que un clasificador aleatorio.
- **AUC > 0.5:** Indica que el modelo tiene capacidad para discriminar entre las clases. Cuanto mayor sea el valor de AUC, mejor será el modelo.

Dentro del rango de valores AUC > 0.5, se pueden clasificar las capacidades de clasificación de la siguiente manera:

- **0.5 - 0.6:** Mala capacidad de clasificación.
- **0.6 - 0.7:** Capacidad de clasificación regular.
- **0.7 - 0.8:** Buena capacidad de clasificación.
- **0.8 - 0.9:** Muy buena capacidad de clasificación.
- **0.9 - 1.0:** Excelente capacidad de clasificación.

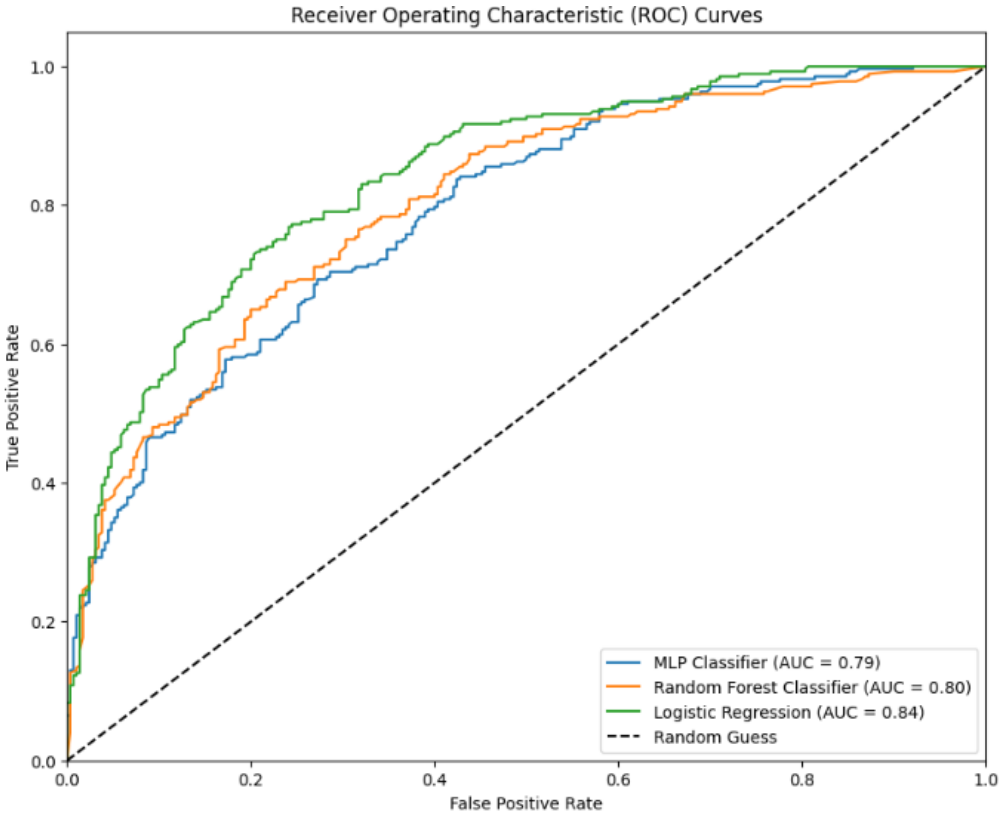

Ilustración 4: Curvas ROC

De acuerdo con la evolución del accuracy, la curva ROC muestra que el modelo de regresión logística tiene una mejor capacidad de clasificación en comparación con el modelo de Random Forest y el de MLPClassifier, cuyos valores de AUC van disminuyendo, respectivamente.

Añadir variables

Debido a la simplicidad de los modelos anteriores, decidimos crear nuevas variables y añadirlas a los modelos con el objetivo de mejorar su capacidad de clasificación. Las nuevas variables que creamos son:

- Probabilidad de que el equipo local reciba un gol.
- Probabilidad de que el equipo visitante reciba un gol.
- Diferencia entre goles marcados y recibidos por el equipo local.
- Diferencia entre goles marcados y recibidos por el equipo visitante.
- Total de partidos jugados por el equipo local.
- Total de partidos jugados por el equipo visitante.

Con estas adiciones, el dataset presenta la siguiente forma:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2831 entries, 0 to 2830
Data columns (total 21 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   home_team               2831 non-null   object
 1   away_team               2831 non-null   object
 2   home_team_code          2831 non-null   object
 3   away_team_code          2831 non-null   object
 4   home_score              2831 non-null   float64
 5   away_score              2831 non-null   float64
 6   winner                  2282 non-null   object
 7   winner_reason           2831 non-null   object
 8   year                    2831 non-null   int64
 9   status                  2831 non-null   object
 10  round                   2831 non-null   object
 11  score_dif               2831 non-null   float64
 12  home_winner             2831 non-null   bool
 13  home_goal_prob          2831 non-null   float64
 14  away_goal_prob          2831 non-null   float64
 15  home_goal_against_prob  2831 non-null   float64
 16  away_goal_against_prob  2831 non-null   float64
 17  home_goal_diff          2831 non-null   float64
 18  away_goal_diff          2831 non-null   float64
 19  home_total_matches      2831 non-null   int64
 20  away_total_matches      2831 non-null   int64
dtypes: bool(1), float64(9), int64(3), object(8)
memory usage: 445.2+ KB
```

Tabla 9: Información dataset segunda fase de modelización

Por lo tanto, la variable objetivo sigue siendo home_winner, mientras que las variables del conjunto X ahora son:

- Equipo local
- Equipo visitante
- Ronda
- Probabilidad de anotar gol del equipo local
- Probabilidad de anotar gol del equipo visitante
- Probabilidad de que el equipo local reciba un gol
- Probabilidad de que el equipo visitante reciba un gol
- Diferencia entre goles marcados y recibidos por el equipo local
- Diferencia entre goles marcados y recibidos por el equipo visitante
- Total de partidos jugados por el equipo local
- Total de partidos jugados por el equipo visitante

A continuación, se representan estas variables por equipo:

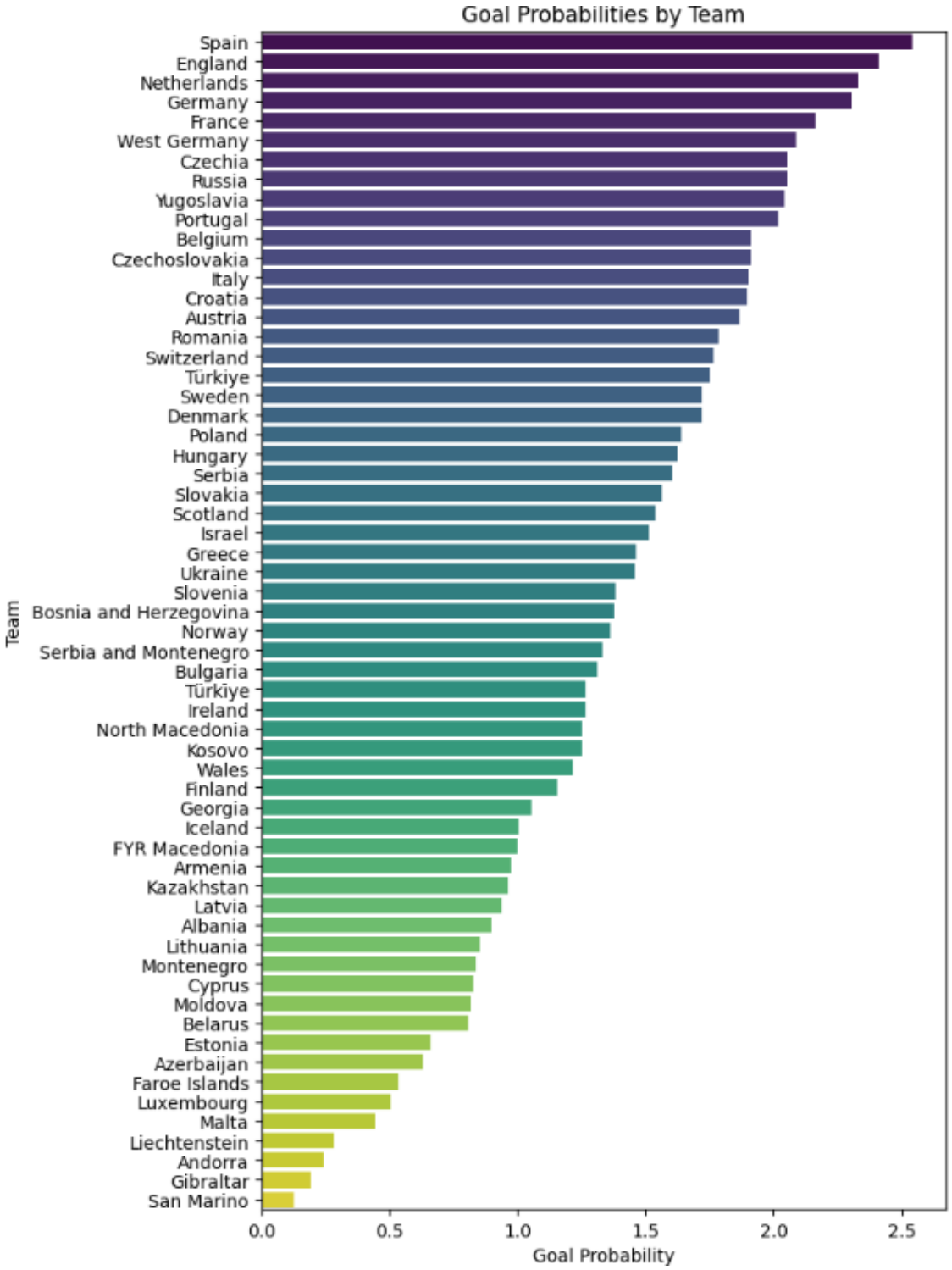

Ilustración 5: Probabilidad de anotar gol por equipo

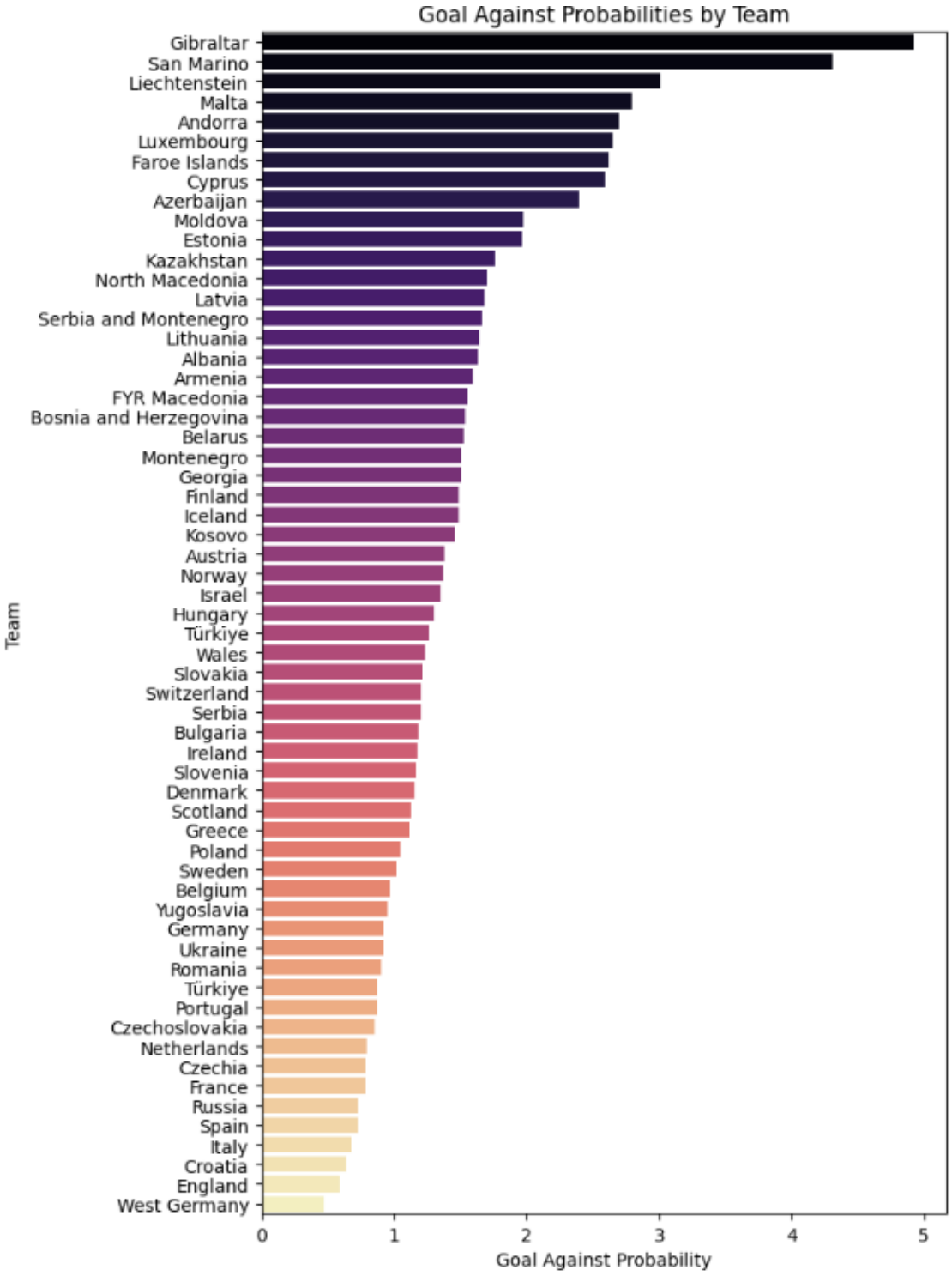

Ilustración 6: Probabilidad de recibir gol por equipo

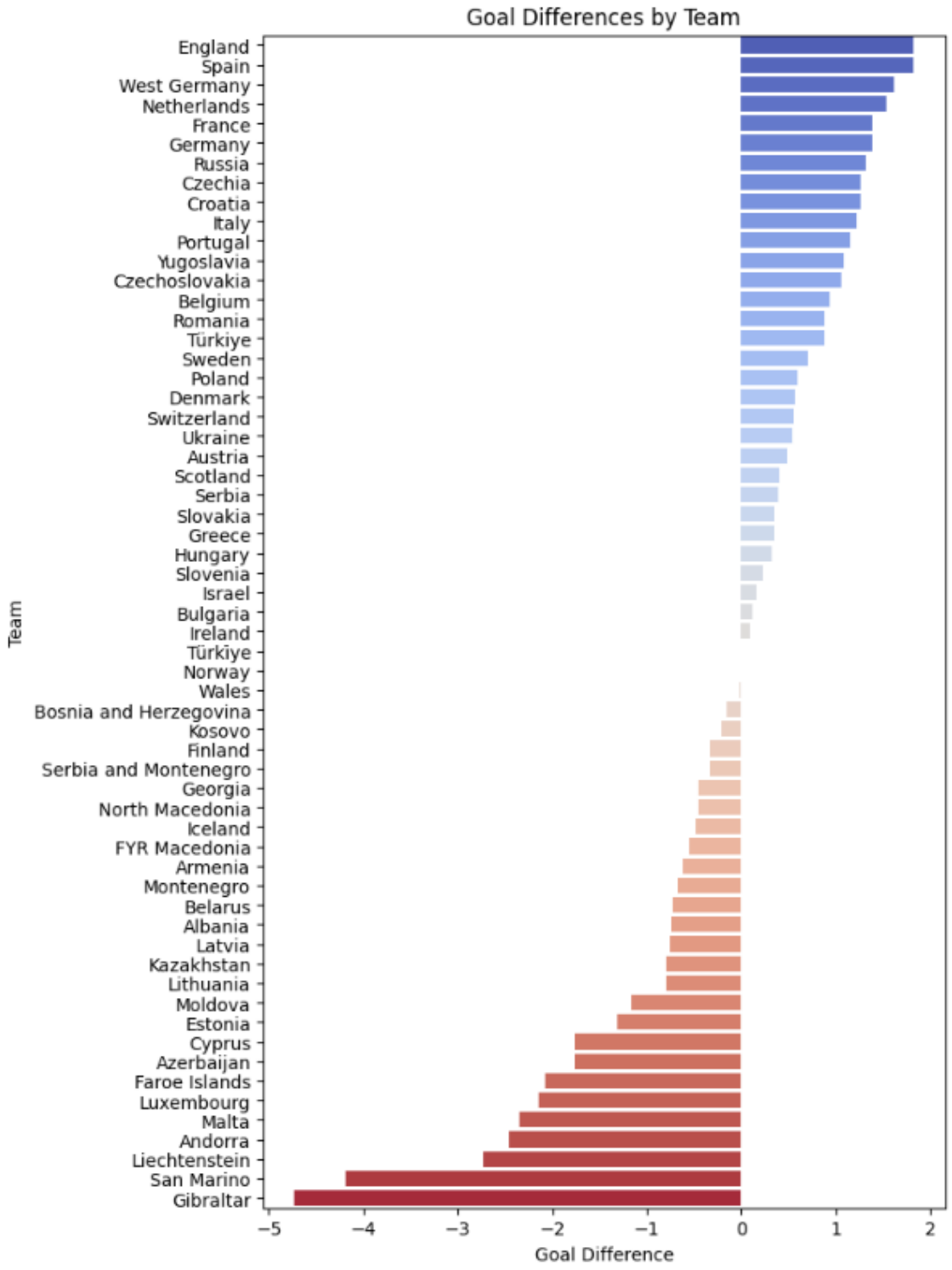

Ilustración 7: Diferencia entre goles anotados y recibidos por equipo

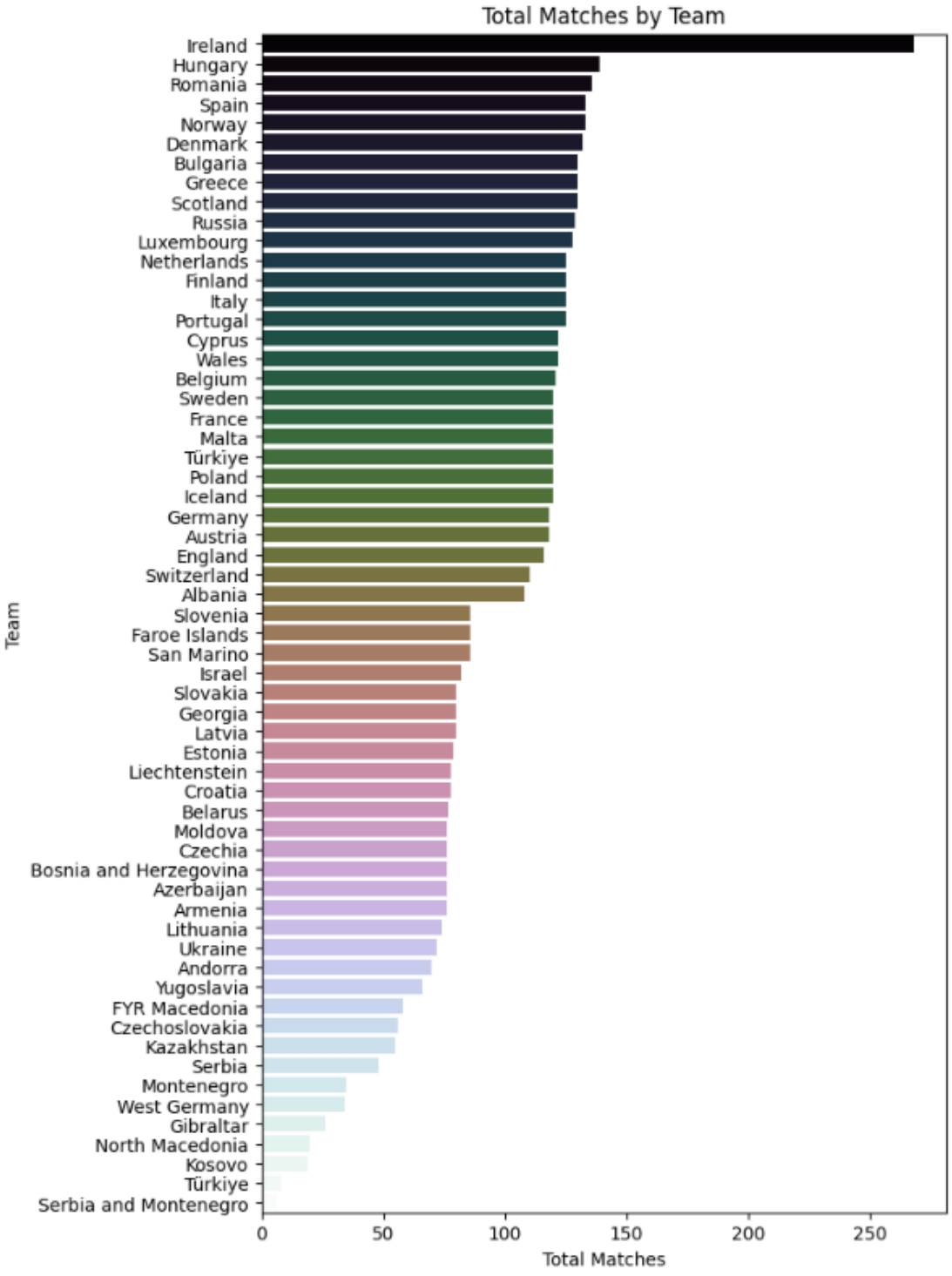

Ilustración 8: Total de partidos jugados por equipo

Al añadir nuevas variables, observamos en los siguientes apartados que la matriz de confusión y, por ende, el accuracy mejoran en cada modelo en comparación con la primera fase de modelización. Sin embargo, los resultados al probar los modelos con los nuevos partidos permanecen iguales en ambas fases de modelización.

```
Accuracy: 0.7037037037037037
Classification Report:
              precision    recall  f1-score   support

       False       0.72      0.70      0.71       290
        True       0.69      0.71      0.70       277

    accuracy                           0.70       567
   macro avg       0.70      0.70      0.70       567
weighted avg       0.70      0.70      0.70       567
```

Tabla 10: Classification Report MLP (2)

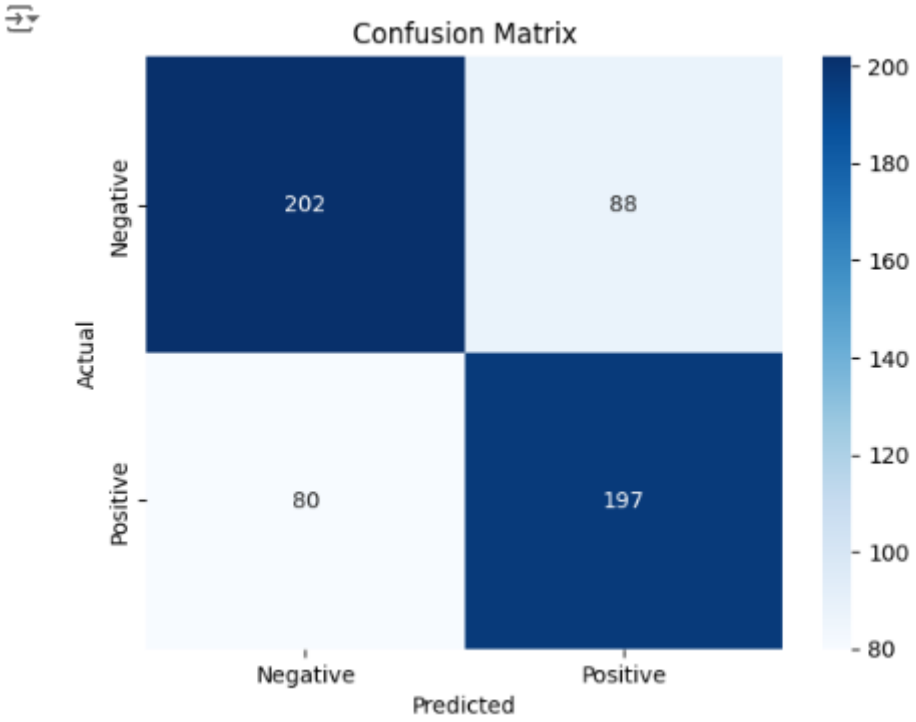

Ilustración 9: Matriz de Confusión MLP (2)

	home_team	away_team	round	home_goal_prob	away_goal_prob	home_goal_against_prob	away_goal_against_prob	home_goal_diff	away_goal_diff	home_total_matches	away_total_matches	home_winner
0	Germany	Spain	QUARTER_FINALS	2.305085	2.544324	0.923729	0.724446	1.381356	1.819878	118	133	False
1	Spain	France	PRELIMINARY	2.544324	2.166667	0.724446	0.783333	1.819878	1.383333	133	120	False
2	Spain	England	FINAL	2.544324	2.413793	0.724446	0.586207	1.819878	1.827586	133	116	False

Tabla 11: Resultados MLP (2)

Random Forest Classifier

Accuracy: 0.7248677248677249
Classification Report:

	precision	recall	f1-score	support
False	0.73	0.72	0.73	290
True	0.72	0.73	0.72	277
accuracy			0.72	567
macro avg	0.72	0.72	0.72	567
weighted avg	0.73	0.72	0.72	567

Tabla 12: Classification Report RFC (2)

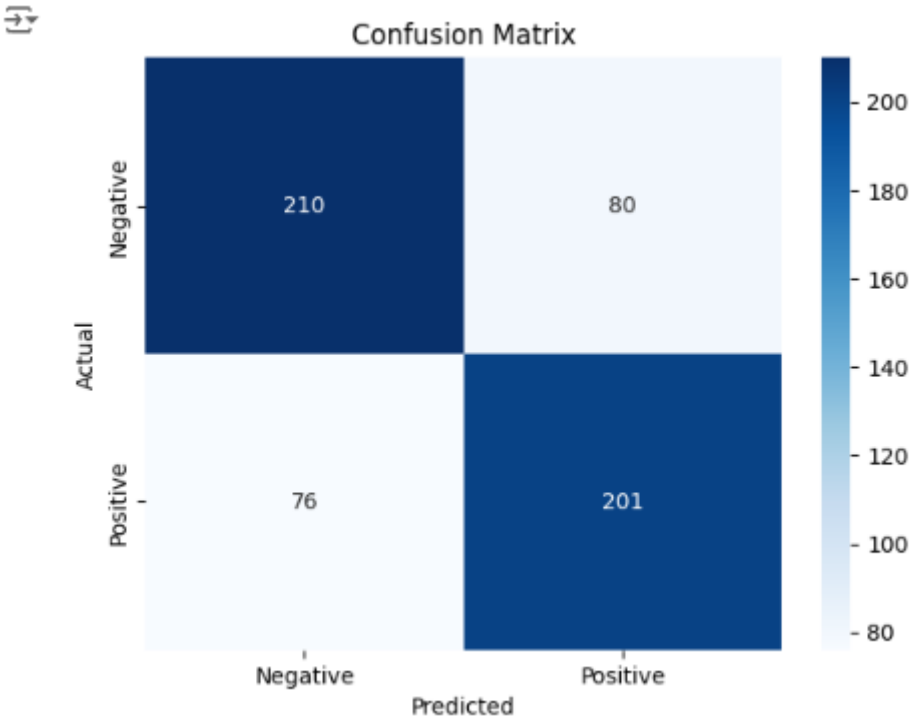

Ilustración 10: Matriz de Confusión RFC (2)

	home_team	away_team	round	home_goal_prob	away_goal_prob	home_goal_against_prob	away_goal_against_prob	home_goal_diff	away_goal_diff	home_total_matches	away_total_matches	home_winner
0	Germany	Spain	QUARTER_FINALS	2.305085	2.544324	0.923729	0.724446	1.381356	1.819878	118	133	True
1	Spain	France	PRELIMINARY	2.544324	2.166667	0.724446	0.783333	1.819878	1.383333	133	120	True
2	Spain	England	FINAL	2.544324	2.413793	0.724446	0.586207	1.819878	1.827586	133	116	True

Tabla 13: Resultados RFC (2)

Logistic Regression

Accuracy: 0.7601410934744268
Classification Report:

	precision	recall	f1-score	support
False	0.76	0.77	0.77	290
True	0.76	0.75	0.75	277
accuracy			0.76	567
macro avg	0.76	0.76	0.76	567
weighted avg	0.76	0.76	0.76	567

Tabla 14: Classification Report Logistic Regression (2)

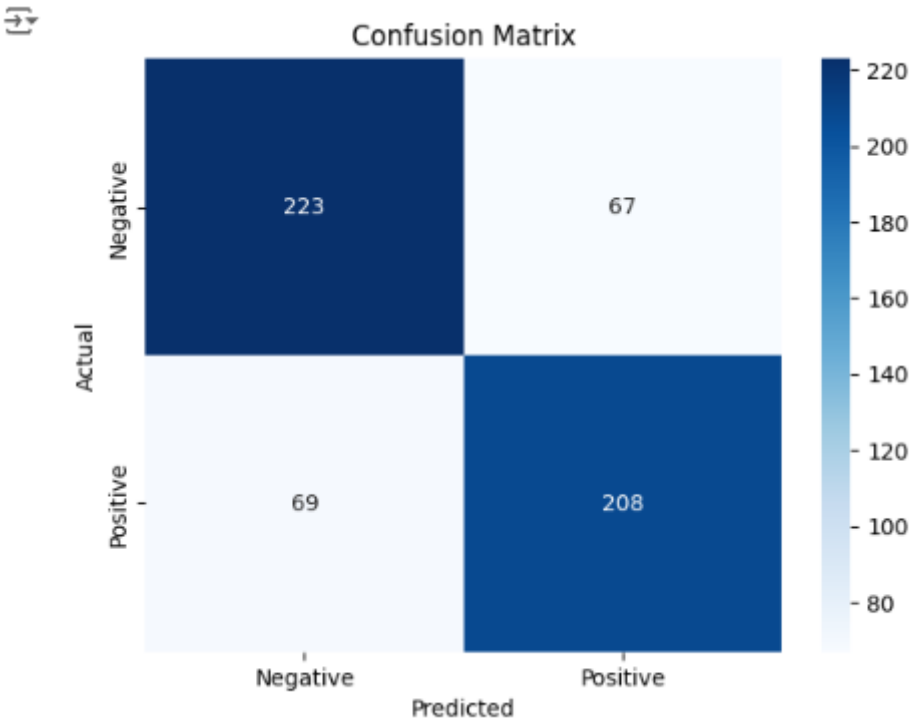

Ilustración 11: Matriz de Confusión Logistic Regression (2)

	home_team	away_team	round	home_goal_prob	away_goal_prob	home_goal_against_prob	away_goal_against_prob	home_goal_diff	away_goal_diff	home_total_matches	away_total_matches	home_winner
0	Germany	Spain	QUARTER_FINALS	2.305085	2.544324	0.923729	0.724446	1.381356	1.819878	118	133	True
1	Spain	France	PRELIMINARY	2.544324	2.166667	0.724446	0.783333	1.819878	1.383333	133	120	True
2	Spain	England	FINAL	2.544324	2.413793	0.724446	0.586207	1.819878	1.827586	133	116	False

Tabla 15: Resultados Logistic Regression (2)

ROC

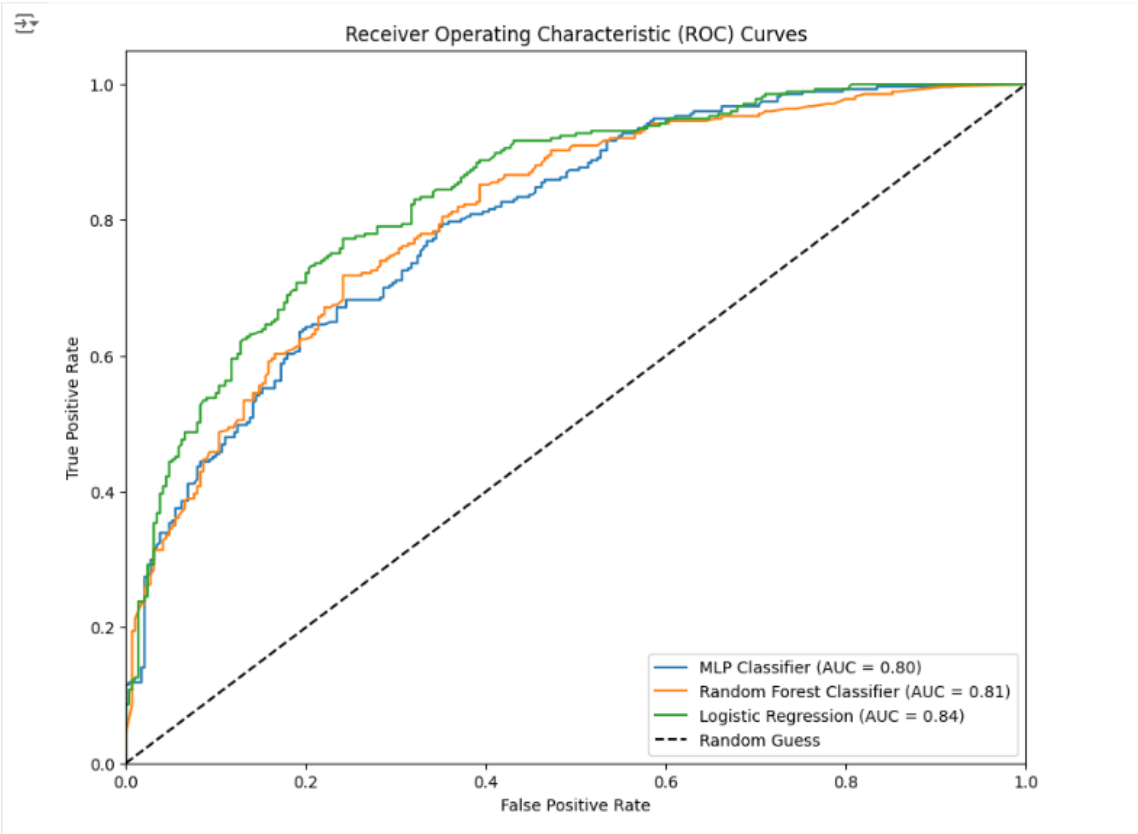

Ilustración 12: Curvas ROC (2)

Con las nuevas variables creadas, observamos que los valores de AUC mejoran en los dos primeros modelos, mientras que el modelo de regresión logística se mantiene sin cambios. La limitada mejora de los modelos podría atribuirse a una posible multicolinealidad entre las nuevas variables, así como entre estas y las variables utilizadas en la fase inicial. Por lo tanto, procederemos a verificar esta posible multicolinealidad.

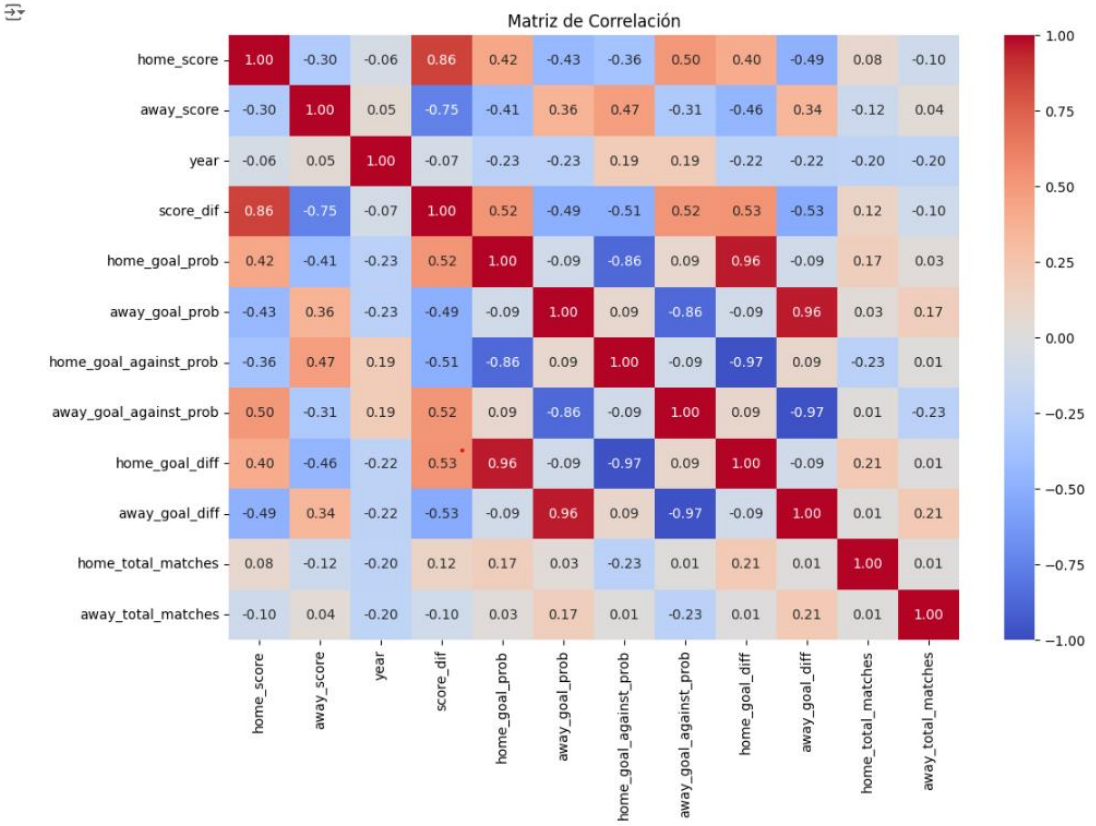

Ilustración 13: Matriz de correlación

Observamos que efectivamente las nuevas variables creadas presentan una alta correlación. Tanto directa (positiva) como indirecta (negativa) entre sí.

Preparación de Datos y Modelado

Para mejorar la capacidad de clasificación de los modelos, realizamos varias etapas en el proceso de preparación de datos y modelado:

1. **Conversión de Variables Categóricas:** Las variables categóricas (equipo local, equipo visitante y ronda) se convirtieron en variables numéricas mediante OneHotEncoder. Los resultados de OneHotEncoder se integraron en un DataFrame que se unió con las probabilidades de gol, resultando en un conjunto de datos completamente numérico y listo para el modelado.
2. **Escalado de Características:** Las características se escalaron para asegurar que todas las variables tuvieran la misma escala, evitando que algunas dominaran sobre otras debido a sus magnitudes.
3. **División del Dataset:** El dataset se dividió en conjuntos de entrenamiento y prueba para evaluar el rendimiento de los modelos.

Construcción y Evaluación de Modelos

Redes Neuronales: Multilayer Perceptron (MLP)

Creamos un modelo utilizando la clase MLPClassifier de Scikit-Learn. El modelo fue ajustado con el conjunto de entrenamiento y evaluado con el conjunto de prueba. A pesar de la precisión global adecuada (accuracy) de 0.70, el modelo mostró resultados mixtos en las predicciones:

- **Accuracy:** Un valor de 0,70 indica un rendimiento mejorable pero adecuado.
- **Precision, Recall y F1-Score:** Las métricas de precisión y recall también reflejan un equilibrio aceptable con un F1-score de 0,70.

La matriz de confusión reveló una mejor proporción de verdaderos negativos (TN) pero también un aumento en los falsos negativos (FN), lo que sugiere una tendencia a predecir más negativos. A pesar del buen rendimiento general, el modelo acertó sólo uno de los tres casos de prueba.

Table 3: Classification Report MLP

Figure 1: Matriz de Confusión MLP

Table 4: Resultados MLP

Random Forest Classifier

El modelo RandomForestClassifier mostró una leve mejora en el accuracy y en otras métricas en comparación con el MLP. La matriz de confusión indicaba un aumento en los verdaderos negativos (TN) y una disminución en los falsos positivos (FP). Sin embargo, los verdaderos positivos (TP) disminuyeron y los falsos negativos (FN) aumentaron, sugiriendo que el modelo tiende a predecir más negativos.

Table 5: Classification Report RFC

Figure 2: Matriz de Confusión RFC

Table 6: Resultados RFC

Regresión Logística

Finalmente, implementamos un modelo de regresión logística que mostró una notable mejora en términos de accuracy, precisión, recall y F1-score, alcanzando una accuracy de casi 76%. Este modelo mostró un aumento en TN y TP, y una disminución en FP y FN, reflejando una mejora general en las predicciones.

Table 7: Classification Report Logistic Regression

Figure 3: Matriz de Confusión Logistic Regression

Table 8: Resultados Logistic Regression

Curva ROC y AUC

La curva ROC y el valor de AUC proporcionaron una visión clara de la capacidad de discriminación de cada modelo:

- **Curva ROC:** Muestra la relación entre la tasa de verdaderos positivos (TPR) y la tasa de falsos positivos (FPR) para diferentes umbrales de decisión.
- **Valor de AUC:** Mide el área bajo la curva ROC. Los valores AUC indican:
  - **0.5 - 0.6:** Mala capacidad de clasificación.
  - **0.6 - 0.7:** Capacidad de clasificación regular.
  - **0.7 - 0.8:** Buena capacidad de clasificación.
  - **0.8 - 0.9:** Muy buena capacidad de clasificación.
  - **0.9 - 1.0:** Excelente capacidad de clasificación.

El modelo de regresión logística demostró la mejor capacidad de clasificación, superando al modelo de Random Forest y al MLPClassifier, cuyos valores AUC disminuyeron.

Figure 4: Curvas ROC

Añadido de Nuevas Variables

Para mejorar la capacidad de los modelos, se añadieron nuevas variables al conjunto de datos:

- Probabilidad de que el equipo local reciba un gol.
- Probabilidad de que el equipo visitante reciba un gol.
- Diferencia entre goles marcados y recibidos por el equipo local.
- Diferencia entre goles marcados y recibidos por el equipo visitante.
- Total de partidos jugados por el equipo local.
- Total de partidos jugados por el equipo visitante.

El dataset actualizado mostró mejoras en la matriz de confusión y en el accuracy para cada modelo en comparación con la primera fase de modelización. Sin embargo, los resultados de los modelos con los nuevos partidos permanecieron consistentes con los resultados anteriores.

Table 9: Información dataset segunda fase de modelización
Figure 5 - 8: Variables por equipo

Resultados con Nuevas Variables

A pesar de las mejoras en las métricas AUC para los dos primeros modelos, la regresión logística mostró valores AUC constantes. La limitada mejora en los modelos podría estar relacionada con la multicolinealidad entre las nuevas variables o con las variables utilizadas en la primera fase. Esta cuestión será revisada en futuras investigaciones.

Table 10 - 15: Resultados MLP, RFC y Regresión Logística con nuevas variables
Figure 9 - 12: Matrices de Confusión y Curvas ROC con nuevas variables

Despliegue Tecnológico y Operativización

En esta sección se presenta el enfoque para el despliegue tecnológico y la operativización del modelo desarrollado. La fase de despliegue es crucial para llevar el modelo de predicción a un entorno real donde pueda ser utilizado para hacer predicciones en nuevos datos y apoyar la toma de decisiones.

Diseño del Despliegue

1. **Modelo de Predicción**

   Los modelos de predicción desarrollados y evaluados incluyen Redes Neuronales (MLPClassifier), Random Forest Classifier y Regresión Logística. El modelo de regresión logística mostró el mejor rendimiento general en términos de accuracy y capacidad de clasificación, con un AUC notablemente alto en comparación con los otros modelos.

   Modelo Seleccionado:

   - **Regresión Logística:** Dada su mejor performance general y su capacidad para manejar la multicolinealidad y la variabilidad en los datos, este modelo es el candidato preferido para el despliegue.

2. **Preparación del Entorno de Despliegue**

   Para operar el modelo en un entorno de producción, se deben seguir los siguientes pasos:

   - **Exportación del Modelo:** El modelo de regresión logística será exportado utilizando la biblioteca joblib o pickle de Python. Esto permitirá almacenar el modelo entrenado en un archivo que pueda ser cargado posteriormente para hacer predicciones.
   - **Implementación en un Entorno de Producción:** El modelo será implementado en un entorno de servidor, utilizando plataformas de despliegue como AWS, Azure, o Google Cloud, que proporcionan servicios para alojar modelos de Machine Learning y realizar predicciones en tiempo real.
   - **Integración con Aplicaciones:** El modelo será integrado con aplicaciones web o móviles mediante una API RESTful, que permitirá a las aplicaciones enviar datos al modelo y recibir predicciones. Esta integración facilitará el uso del modelo en la toma de decisiones en tiempo real.

3. **Escalado y Mantenimiento**
    - **Escalado:** Se implementarán mecanismos de escalado automático para manejar variaciones en la carga de trabajo, asegurando que el modelo pueda procesar un gran volumen de solicitudes sin degradar su rendimiento.
    - **Mantenimiento y Actualización:** El modelo será monitoreado continuamente para evaluar su rendimiento en producción. Se establecerán procedimientos para actualizar el modelo con nuevos datos y mejorar su precisión conforme se recojan más datos y cambien las condiciones.
4. **Visualización y Reportes**
    - **Dashboard de Resultados:** Se desarrollará un panel de control interactivo que mostrará las métricas de rendimiento del modelo en tiempo real, como accuracy, precision, recall y AUC. Este dashboard permitirá a los usuarios observar cómo el modelo está funcionando y realizar ajustes si es necesario.
    - **Reportes Automáticos:** Se generarán reportes automáticos sobre las predicciones realizadas, la precisión del modelo y otros indicadores clave de rendimiento. Estos reportes se enviarán periódicamente a los stakeholders para mantenerlos informados sobre el rendimiento del modelo.
5. **Pruebas y Validación**
    - **Pruebas de Integración:** Se realizarán pruebas exhaustivas para asegurar que el modelo se integra correctamente con el entorno de producción y las aplicaciones existentes.
    - **Validación de Desempeño:** Se llevará a cabo una validación continua para asegurar que el modelo sigue funcionando de manera efectiva a lo largo del tiempo. Esto incluirá la validación de la calidad de las predicciones y la actualización del modelo en base a nuevos datos.

6. **Documentación y Capacitación**
    - **Documentación del Despliegue:** Se preparará documentación detallada sobre el proceso de despliegue, incluyendo instrucciones para la integración del modelo, escalado, y mantenimiento.
    - **Capacitación del Personal:** Se proporcionará capacitación al personal encargado de mantener el sistema para asegurar que estén familiarizados con el modelo, su funcionamiento y los procedimientos de actualización.

Implementación y Ejemplos de Uso

Con el modelo de regresión logística desplegado, las aplicaciones podrán realizar predicciones basadas en los datos de entrada en tiempo real. Ejemplos de uso incluyen:

- **Predicción de Resultados de Partidos:** Los usuarios pueden ingresar datos sobre partidos futuros y recibir predicciones sobre los posibles resultados.
- **Análisis de Desempeño:** Se puede utilizar el modelo para analizar el desempeño de los equipos y ajustar estrategias o decisiones basadas en los resultados de las predicciones.

El despliegue tecnológico del modelo asegura que la solución pueda ser utilizada de manera efectiva y eficiente, proporcionando un valor tangible y práctico en el contexto de los objetivos del proyecto.

PUESTA EN VALOR

En esta sección se presenta la estrategia para integrar los resultados analíticos en los procesos operativos de la compañía, con el objetivo de maximizar el impacto positivo en los objetivos empresariales y proporcionar valor tangible.

Estrategia para la Integración de Resultados Analíticos

1. **Alineación con los Objetivos Empresariales**
    - **Optimización de Decisiones:** Los modelos de predicción, especialmente el de regresión logística, proporcionan predicciones precisas sobre los resultados de los partidos. Esto puede ser utilizado para optimizar las decisiones estratégicas en áreas como la planificación de partidos, estrategias de juego, y análisis de rivales, alineándose con los objetivos de mejorar el rendimiento y la competitividad.
    - **Mejora de la Planificación de Recursos:** Al anticipar los resultados de los partidos, la compañía puede mejorar la planificación y asignación de recursos, tales como el personal necesario para eventos específicos o la preparación de campañas de marketing basadas en los posibles resultados.
2. **Implementación en los Procesos Operativos**
    - **Desarrollo de una Plataforma de Predicción:** Se desarrollará una plataforma centralizada donde los resultados de los modelos se integren en un sistema accesible para los tomadores de decisiones. Esta plataforma ofrecerá interfaces para la entrada de datos, generación de predicciones y visualización de resultados.
    - **Automatización de Procesos:** La integración del modelo en los sistemas existentes permitirá la automatización de la generación de predicciones y reportes. Esto reducirá la carga de trabajo manual y mejorará la eficiencia operativa.

3. **Utilización de Resultados para Estrategias Comerciales**
    - **Campañas de Marketing Dirigidas:** Los resultados del modelo se utilizarán para diseñar campañas de marketing dirigidas y personalizadas, aprovechando la capacidad predictiva para maximizar el impacto de las campañas basadas en los resultados esperados de los partidos.
    - **Segmentación de Audiencia:** Los datos analíticos permitirán una segmentación más precisa de la audiencia, adaptando las estrategias de comunicación y promoción a diferentes segmentos basados en la probabilidad de resultados y el comportamiento anticipado de los consumidores.
4. **Monitoreo y Evaluación Continua**
    - **Establecimiento de KPIs:** Se definirán indicadores clave de rendimiento (KPIs) para evaluar el impacto de las predicciones en los procesos operativos. Esto incluirá métricas como la precisión de las predicciones, la efectividad de las campañas basadas en los resultados y la eficiencia operativa mejorada.
    - **Revisión Periódica:** Se establecerán procesos de revisión periódica para evaluar la efectividad del modelo en el entorno operativo. Esto permitirá realizar ajustes y mejoras continuas para mantener la alineación con los objetivos empresariales y adaptar el modelo a cambios en el entorno.
5. **Capacitación y Comunicación**
    - **Capacitación del Personal:** Se implementarán programas de capacitación para el personal involucrado en la utilización de los resultados analíticos. Esto incluirá formación en el uso de la plataforma de predicción y en la interpretación de los resultados para tomar decisiones informadas.
    - **Comunicación de Resultados:** Se desarrollarán estrategias de comunicación para informar a los stakeholders sobre los resultados y su impacto. Esto garantizará que los resultados sean comprendidos y utilizados de manera efectiva en la toma de decisiones.

6. **Evaluación de Impacto y Retroalimentación**
    - **Evaluación del Impacto:** Se llevará a cabo una evaluación exhaustiva del impacto de la integración de los resultados analíticos en los procesos de la compañía. Esto ayudará a medir el retorno sobre la inversión (ROI) y a identificar áreas de éxito y oportunidades de mejora.
    - **Recopilación de Retroalimentación:** Se establecerán mecanismos para recopilar retroalimentación de los usuarios y stakeholders sobre el uso de los resultados analíticos. Esta retroalimentación será fundamental para realizar ajustes y mejoras en la estrategia de puesta en valor.

Resultados Esperados

- **Mejora en la Precisión de Decisiones:** La incorporación de predicciones precisas mejorará la toma de decisiones en las áreas clave de planificación y estrategia.
- **Aumento de la Eficiencia Operativa:** La automatización y la integración de los resultados analíticos optimizarán los procesos y reducirán la carga operativa.
- **Optimización de Estrategias Comerciales:** Las campañas de marketing y la segmentación de audiencia se volverán más efectivas, generando un mayor impacto comercial.

La puesta en valor de los resultados analíticos está diseñada para transformar los insights obtenidos en el proyecto en beneficios operacionales y estratégicos tangibles, alineando la capacidad predictiva del modelo con los objetivos empresariales de la compañía.

CONCLUSIONES

Resumen de Objetivos Alcanzados

1. **Desarrollo y Evaluación de Modelos Predictivos**
    - **Modelos Implementados:** Se desarrollaron y evaluaron tres modelos predictivos: Redes Neuronales (MLPClassifier), Random Forest Classifier y Regresión Logística. Cada uno de estos modelos se construyó y ajustó utilizando el conjunto de datos proporcionado, con el objetivo de predecir los resultados de los partidos.
    - **Evaluación de Métricas:** Los modelos fueron evaluados utilizando métricas clave como **accuracy**, **precision**, **recall** y **F1-score**. La regresión logística demostró ser la más efectiva, con una precisión de casi el 76%, mejorando respecto a los modelos de MLP y Random Forest en términos de accuracy y otras métricas.
2. **Análisis de Resultados con Nuevas Variables**
    - **Creación de Nuevas Variables:** Se introdujeron nuevas variables para mejorar la capacidad de clasificación, tales como probabilidades de recibir goles y diferencias entre goles anotados y recibidos. Estas variables contribuyeron a mejorar los valores de AUC en los modelos de MLP y Random Forest, aunque el modelo de regresión logística mantuvo un rendimiento constante.
    - **Impacto en la Precisión:** A pesar de las mejoras en la matriz de confusión y en los valores de accuracy con las nuevas variables, los resultados en los nuevos partidos no mostraron una mejora significativa respecto a la primera fase de modelización. Esto sugiere que la introducción de variables adicionales podría haber causado multicolinealidad o que los nuevos datos no aportaron mejoras sustanciales en los resultados predictivos.

3. **Interpretación de Métricas Avanzadas**
    - **Curvas ROC y AUC:** La curva ROC y los valores de AUC indicaron que la regresión logística ofreció la mejor capacidad de discriminación entre las clases en comparación con los otros modelos. La AUC de la regresión logística superó el umbral de 0.8, indicando una **muy buena capacidad de clasificación**.
    - **Matriz de Confusión:** La matriz de confusión reveló que, aunque los modelos mostraron mejoras en las métricas de desempeño, aún persistieron desafíos en la correcta predicción de algunos resultados, como se observó en los casos de prueba específicos.

Análisis Crítico

- **Desempeño del Modelo:** A pesar de los ajustes y la incorporación de nuevas variables, el modelo de regresión logística, aunque superior en términos de métricas generales, no mejoró significativamente en la predicción de nuevos partidos en comparación con las fases anteriores. Esto puede indicar la necesidad de un enfoque más sofisticado en la selección de variables o en la modelización.
- **Multicolinealidad:** La poca mejora observada al introducir nuevas variables sugiere que podría existir multicolinealidad entre las variables creadas. Este es un aspecto que requiere una revisión más profunda para asegurar que las variables no se solapen y aporten valor adicional al modelo.
- **Generalización de Modelos:** La discrepancia entre el rendimiento en los datos de entrenamiento y los datos de prueba subraya la importancia de mejorar la generalización de los modelos. Se deben considerar técnicas adicionales como la validación cruzada o el ajuste de hiperparámetros para mejorar la capacidad predictiva.

Próximos Pasos

1. **Análisis de Multicolinealidad**
    - Realizar un análisis detallado para identificar y abordar posibles problemas de multicolinealidad entre las variables. Esto puede incluir la eliminación de variables redundantes o la aplicación de técnicas de reducción de dimensionalidad.

2. **Optimización de Modelos**
    - Explorar y aplicar técnicas avanzadas de optimización de modelos, como el ajuste fino de hiperparámetros, la validación cruzada y la implementación de algoritmos de ensamblaje, para mejorar el rendimiento predictivo general.
3. **Ampliación del Dataset**
    - Considerar la incorporación de datos adicionales o la ampliación del conjunto de datos para proporcionar una base más robusta para el entrenamiento de modelos. Esto puede ayudar a mejorar la capacidad de generalización y la precisión de las predicciones.
4. **Desarrollo de Modelos Avanzados**
    - Investigar y desarrollar modelos predictivos más avanzados, como redes neuronales profundas o técnicas de aprendizaje automático más sofisticadas, para abordar las limitaciones observadas en los modelos actuales.
5. **Integración y Monitorización Continua**
    - Implementar una solución de despliegue continuo que permita la actualización y monitorización constante de los modelos en producción. Esto garantizará que el sistema se mantenga alineado con los objetivos empresariales y pueda adaptarse a cambios en el entorno de datos.
6. **Feedback y Mejora Continua**
    - Establecer un sistema de retroalimentación para recoger insights de los usuarios finales y realizar ajustes continuos en los modelos y en la estrategia de análisis, asegurando que las predicciones sigan siendo relevantes y útiles para la toma de decisiones.

Estas conclusiones y próximos pasos proporcionan una visión integral de los logros del proyecto y establecen una hoja de ruta clara para continuar mejorando los modelos predictivos y su integración en los procesos operativos de la compañía.

OBSERVACIONES

Consideraciones Adicionales sobre el Dataset

1. **Calidad y Complejidad de los Datos**
    - **Integridad de los Datos:** Asegurarse de que el conjunto de datos esté completo y libre de errores es crucial para la precisión de los modelos. Cualquier dato faltante o incorrecto puede afectar negativamente el rendimiento predictivo.
    - **Distribución de los Datos:** Es importante revisar la distribución de las variables y la representación de las clases en el dataset. Desbalances en las clases pueden influir en las métricas de rendimiento y en la capacidad del modelo para generalizar.
2. **Evolución del Contexto de Datos**
    - **Cambio en el Entorno:** Los datos deportivos pueden cambiar con el tiempo debido a cambios en los equipos, estrategias, o en el contexto del campeonato. Es fundamental actualizar periódicamente el modelo y el conjunto de datos para reflejar estos cambios.
    - **Adaptación a Nuevas Condiciones:** Los modelos deben ser capaces de adaptarse a nuevas condiciones o eventos imprevistos que no se reflejan en los datos históricos.

Evaluación de Modelos y Técnicas

1. **Rendimiento en Diferentes Contextos**
    - **Variedad en los Datos de Prueba:** Los resultados obtenidos pueden variar según los contextos y tipos de partidos en los datos de prueba. Evaluar el rendimiento en distintos escenarios y con diferentes tipos de partidos puede proporcionar una visión más completa de la efectividad del modelo.
    - **Robustez de los Modelos:** La robustez de los modelos debe ser evaluada para garantizar que puedan manejar variaciones y ruidos en los datos sin perder precisión significativa.

2. **Interacción entre Variables**
    - **Relaciones entre Variables:** Es posible que algunas de las nuevas variables creadas tengan relaciones complejas entre sí que no se reflejan directamente en el análisis. La interacción entre variables puede influir en el rendimiento del modelo y debería ser analizada más a fondo.
    - **Efectos No Lineales:** Los modelos tradicionales pueden no capturar efectos no lineales entre variables. Explorar modelos más complejos o técnicas avanzadas puede ser útil para mejorar la precisión.

Aspectos Técnicos y Operacionales

1. **Requerimientos Computacionales**
    - **Capacidad de Cálculo:** Los modelos más complejos, como redes neuronales profundas, pueden requerir significativos recursos computacionales y tiempo de entrenamiento. Asegurarse de que la infraestructura tecnológica pueda soportar estos requisitos es crucial.
    - **Escalabilidad:** Considerar la escalabilidad de los modelos para manejar un aumento en el volumen de datos o en el número de partidos en el futuro.
2. **Implementación y Mantenimiento**
    - **Proceso de Despliegue:** El proceso de integración de los modelos en el entorno operativo debe ser cuidadosamente planificado para minimizar interrupciones y garantizar una transición fluida.
    - **Mantenimiento Continuo:** Establecer un plan de mantenimiento para actualizar y revisar periódicamente los modelos en función de los nuevos datos y resultados.

Feedback de los Usuarios

1. **Retroalimentación del Usuario**
    - **Recepción de Resultados:** Obtener retroalimentación de los usuarios finales del sistema de predicción puede proporcionar información valiosa sobre la utilidad y precisión de las predicciones. Este feedback puede ayudar a ajustar y mejorar los modelos.
    - **Adaptación a Requerimientos:** Considerar la adaptación de los modelos para satisfacer las necesidades específicas de los usuarios o para integrar nuevas variables o métricas que puedan ser relevantes para los procesos de negocio.

Estas observaciones complementan el análisis y la implementación del proyecto, proporcionando una perspectiva adicional sobre la operativización, evaluación y adaptación continua de los modelos predictivos.

CONTRIBUCIÓN DE LOS AUTORES

Conceptualización del caso de Uso: F.S.N.

Extracción de Datos: C.B.R. y F.S.N.

Tratamiento y Limpieza de Datos: C.B.R. y F.S.N.

Plataforma Tecnológica: C.B.R. y F.S.N.

Modelado (desarrollo y documentación): C.B.R.

Resultados: C.B.R. y F.S.N.

Memoria: F.S.N. y C.B.R.

BIBLIOGRAFÍA Y RECURSOS

Fitzgerald, T. (2018). *FIFA Soccer Rankings* [Data set].

Soccer world cup 2018 winner. (2018, junio 29). Kaggle.com; Kaggle. https://www.kaggle.com/code/agostontorok/soccer-world-cup-2018-winner

Winner of the FIFA World Cup (10.000 simulations). (2018, junio 29). Kaggle.com; Kaggle.https://www.kaggle.com/code/agostontorok/winner-of-the-fifa-world-cup-10-000-simulations

Petro. (2024). *Football - Soccer - UEFA EURO, 1960 - 2024* [Data set].

Euro - Football - data. (2024, junio 7). Kaggle.com; Kaggle. https://www.kaggle.com/code/fredens/euro-football-data

Santana, F. (n.d.). *Statsbomb analysis*. GitHub. Retrieved July 25, 2024, from https://github.com/feersantana5/CRISP-DM-w-Open-Football-Data/blob/master/1.Statsbomb/statsbomb_analysis.ipynb

ANEXOS

Se adjuntan como anexos todos los documentos generados a lo largo del proyecto. Además, se proporcionará el código fuente en archivos anexos o mediante un enlace al repositorio de GitHub. Este repositorio incluirá una explicación detallada de los scripts y las instrucciones necesarias para replicar el trabajo en caso de ser necesario. Se recomienda que esta información se mantenga en el repositorio de GitHub, formando parte integral del portafolio del alumno.