UNIVERSITAT D'ALACANT
UNIVERSIDAD DE ALICANTE
Escola Politècnica Superior
Escuela Politécnica Superior

INFORME SOBRE EL POYECTO FINAL DE LA ASIGNATURA PREPROCESO, RECOLECCIÓN Y VISUALIZACIÓN DE DATOS:

ANÁLISIS MULTIVARIABLE DE LA LIGA 24-25

AUTORES:

MAGDALENA SANCHO DOCÓN

JIMENA MILLA MORENO

ITSASO ARIZTIMUÑO CENOZ

JUAN FRANCISCO CORREAS DÍAZ

Estructura del repositorio en GitHub

Con el fin de tener todo el trabajo organizado de manera clara, se ha creado un repositorio en GitHub con todo el desarrollo y todos los archivos que constituyen este trabajo.
La estructura del repositorio consiste en:

- Una carpeta ***input*** donde se encuentran dos CSVs que usan para generar otros nuevos.
- Una carpeta ***output*** que contiene numerosos CSVs que se utilizan en el transcurso del trabajo.
- Un archivo ***Informe Final.pdf***, que contiene este informe.
- ***Readme.md***, que contiene una descripción completa del repositorio y cómo trabajar con él.
- ***dashboard.py***, que es el archivo donde se realizan todas las visualizaciones.
- ***filename.hlp***, que es un archivo donde se realizan la unión de numerosos datasets.
- ***football-data.co.uk_notes.txt.pdf***, que es un documento donde se explica el significado de cada variable del dataset principal (*SP1.csv*).
- ***predicciones.py***, que es donde se implementa el código para desarrollar diferentes predicciones.
- ***requirements.tx***t, que es el archivo donde se encuentran todos los paquetes que se emplean para implementar todo el código.

Por tanto, en enlace de acceso a este repositorio es:
https://github.com/MalenaSancho/procesos-y-visualizacion/

Durante el informe, se explica que las visualizaciones se han realizado dentro de un dashboard. Para poder verlo sin necesidad de tener que descargarse todo este repositorio, se ha usado la aplicación de *streamlit.app*. Así, se puede ver todo el dashboard solamente metiéndose en el siguiente enlace:
https://dashboardtrabajofinal.streamlit.app/

También, se puede consultar el vídeo en el que se explica, no de manera tan detallada como en los archivos del repositorio, los puntos clave de todo nuestro proyecto en el siguiente enlace:
https://drive.google.com/file/d/1yXibVVeTIX_eQjEebf3GNoK2MZnk_azv/view?usp=sharing

Por último, se puede consultar nuestro dataset final, con el que realizamos todas las visualizaciones, en el siguiente enlace:
https://drive.google.com/file/d/1UQFrlZyPI6yn5C4vHp2mrqgHKY7EZfhJ/view?usp=drive_link

ÍNDICE

1. **Objetivos del Proyecto**
2. **Selección de Fuentes y descripción de estas**
3. **Metodología y Análisis de Fuentes Individuales**
4. **Integración de Datos**
5. **Extracción de Información y Análisis Descriptivos**
6. **Modelado predictivo (Machine Learning)**
7. **Conclusiones y limitaciones**

1. Objetivos del Proyecto

El objetivo fundamental de este proyecto es la recolección, preproceso y enriquecimiento de datos para transformar una fuente primaria de información deportiva estática (`SP1.csv`) en un conjunto de datos complejo y multidimensional apto para análisis avanzados.

Dado que la fuente original se limita estrictamente a estadísticas de juego (goles, tarjetas, etc.) y datos del mercado de apuestas, el proyecto busca superar estas limitaciones integrando dimensiones contextuales externas. El propósito es generar un dataset limpio y unificado que permita visualizar cómo factores exógenos influyen en el espectáculo deportivo.

Objetivos específicos de preproceso y visualización:

1. **Enriquecimiento Geográfico y de Infraestructura**: Incorporar al dataset base la información sobre los estadios y su geolocalización precisa (latitud/longitud), datos inexistentes en el archivo original pero necesarios para el análisis espacial.
2. **Integración de la Dimensión Social**: Completar la información de los partidos con datos de asistencia real al estadio y métricas de *hype* (interés de búsqueda en Google Trends), permitiendo analizar el comportamiento del público.
3. **Contextualización Climática**: Asociar a cada evento las condiciones meteorológicas exactas (temperatura, precipitación, etc.) ocurridas en el momento y lugar del partido, mediante consultas históricas a APIs externas.
4. **Normalización y Limpieza**: Resolver la heterogeneidad semántica entre fuentes (nombres de equipos dispares) e imputar valores perdidos para garantizar la consistencia del análisis visual.
5. **EDA (*Exploratory Data Analysis*)**: presentación a través de gráficas de diferentes aspectos de los datos finales, para la obtención de conclusiones clave de los mismos.

Objetivo secundario: Como demostración de la calidad y utilidad del dataset preprocesado, se plantea como objetivo adicional la implementación de modelos de *Machine Learning* para predecir la asistencia y el resultado de los encuentros, validando así que los datos recolectados poseen capacidad explicativa real.

Para satisfacer estos objetivos, las fuentes seleccionadas deben permitir la alineación temporal y espacial exacta con el dataset base mediante claves compuestas (Equipo + Fecha) y ofrecer datos históricos accesibles para la temporada 24-25.

Antes de comenzar con las partes técnicas del desarrollo de todo el trabajo, cabe mencionar que en este informe se explican puntos clave de por qué hemos tomado cada decisión y cómo se ha llevado a cabo, pero todo el código que hay detrás y que ha hecho posible todo esto se encuentra en los archivos *TrabajoFinal1.ipynb* (*jupyter notebook* donde se implementa la obtención y limpieza de los datos), *filename.hlp* (archivo de *Apache Hop* donde se unifican varios *datasets*), *predicciones.ipynb* (*jupyter notebook* donde se realizan las predicciones usando los datos finales conseguidos y que complementan este trabajo) y *dashboard.py* (archivo *python* donde se realizan todas las visualizaciones).

2. Selección de Fuentes y descripción de estas

Para obtener los objetivos planteados, se ha determinado que no es posible depender de una única fuente de información. Por ello, se ha diseñado una arquitectura de datos híbrida que combina cinco fuentes externas de naturaleza heterogénea. No se utilizan fuentes internas empresariales, ya que el ámbito del proyecto es el análisis deportivo público.

La estrategia de selección se ha basado en la complementariedad: una fuente estructural (*Open Data*) enriquecida por fuentes contextuales *(APIs* y *Web Scraping*).

Para garantizar la reproducibilidad y la transparencia del proceso de recolección de datos, a continuación se documentan técnicamente las cinco fuentes de información utilizadas en el proyecto.

2.1. Fuente Primaria: Estructura y Resultados (Open Data)

Esta fuente constituye la columna vertebral del dataset, proporcionando la estructura base de partidos sobre la cual se integran el resto de dimensiones. Esta fuente de obtención de datos posee las siguientes características:

- **Tipo de Fuente**: Repositorio de *Open Data* (Datos estructurados).
- **Formato de Origen**: Archivo de texto plano delimitado por comas (`.csv`).
- **Autoridad / Editor**: Joseph Buchdahl (Analista de apuestas deportivas). El sitio opera desde 2001 y es considerado un estándar en la comunidad de análisis deportivo.
- **Enlace de Acceso**: https://www.football-data.co.uk/spainm.php (Archivo SP1`.csv`).
- **Contenido Específico**: Fecha, hora, equipos, goles (FTHG/FTAG), estadísticas al descanso, disparos, córners, faltas, tarjetas y cuotas de cierre de múltiples casas de apuestas (Bet365, William Hill, etc.).
- **Accesibilidad y Disponibilidad**: Acceso público gratuito y directo (sin registro). Alta disponibilidad.
- **Frecuencia de Actualización**: Semanal (tras la finalización de la jornada de liga).
- **Justificación de selección**: Se ha seleccionado este repositorio por ofrecer un archivo CSV estructurado con métricas de juego muy completas (goles, tiros, faltas) y cuotas de múltiples casas de apuestas. Su naturaleza estática facilita la carga inicial y el versionado.

2.2. Fuentes de Enriquecimiento: Geografía y Asistencia (Web Scraping)

El dataset primario carece de metadatos físicos (dónde se jugó y cuánta gente fue). Para cumplir el Objetivo 1 (Enriquecimiento Geográfico), es necesario extraer información no estructurada de la web. Por ello, se realiza web scraping a la Wikipedia. Esta fuente de obtención de datos posee las siguientes características:

- **Tipo de Fuente**: Web no estructurada (Enciclopedia colaborativa - Wikipedia).

- **Formato de Origen**: Documentos HTML parseados mediante *Web Scraping* (`BeautifulSoup`).
- **Autoridad / Editor**: Fundación Wikimedia y comunidad de editores.
- **Enlaces de Acceso**: Se han utilizado 40 URLs distintas:
    - 20 URLs de temporada por equipo (ej: https://en.wikipedia.org/wiki/2024–25_Athletic_Bilbao_season).
    - 20 URLs específicas de estadios para coordenadas (ej: https://es.wikipedia.org/wiki/Estadio_de_Mendizorroza).
- **Contenido Específico**: Nombre del estadio oficial, coordenadas geográficas (Latitud/Longitud) y asistencia declarada por partido.
- **Accesibilidad**: Pública y gratuita. Requiere gestión de cabeceras `User-Agent` para evitar bloqueos por *scraping*.
- **Justificación de selección**: Wikipedia es la única fuente pública que hemos encontrado que vincula consistentemente los partidos con sus estadios específicos y la asistencia oficial. Además, permite extraer las coordenadas geográficas precisas de cada recinto, dato indispensable para emplear posteriormente a la API climática.

2.3. Fuente de Validación e Imputación (Web Scraping)

Al no ser Wikipedia una fuente oficial de información, hay partidos de los que no se dispone de la asistencia del mismo. Por ello, se necesita un validador externo, en este caso EstadiosDB, que tiene una tabla resumen de la asistencia media a cada estadio durante la temporada 24-25.

Así, se vuelve a llevar web scraping sobre EstadiosDB. Esta fuente de obtención de datos posee las siguientes características:

- **Tipo de Fuente**: Portal web especializado.
- **Formato de Origen**: Tabla HTML parseada mediante *Web Scraping* (`BeautifulSoup`).
- **Autoridad / Editor**: EstadiosDB.com (Base de datos global de estadios de fútbol).
- **Enlace de Acceso**: https://estadiosdb.com/noticias/2025/06/espana_asistencia_a_los_estadios_de_la_liga_en_la_temporada_202425
- **Contenido Específico**: Capacidad total del estadio y asistencia media acumulada de la temporada.
- **Justificación de selección**: Proporciona las medias de asistencia oficiales y la capacidad de los estadios. Estos datos agregados son fundamentales para la fase de limpieza, permitiendo imputar matemáticamente los valores perdidos en la fuente de Wikipedia mediante proyecciones estadísticas.

2.4. Fuente Contextual: Climatología (API Externa)

Para cumplir el Objetivo 3 (Contextualización Climática), se requiere una fuente que permita consultas históricas granulares (por hora y ubicación). Por lo tanto, se hace uso de la API externa de Open-Meteo (en concreto, *Historical Weather API*), realizando llamadas a esta API para complementar los datos conseguidos hasta el momento. Esta fuente de obtención de datos posee las siguientes características:

- **Tipo de Fuente**: API REST (Interfaz de Programación de Aplicaciones).
- **Formato de Origen**: Respuestas en formato JS0N procesadas mediante la librería openmeteo-requests.
- **Autoridad / Editor**: Open-Meteo (Proyecto Open Source de datos meteorológicos). Utiliza modelos de reanálisis ERA5 del ECMWF (Centro Europeo de Previsiones Meteorológicas).
- **Endpoint de Acceso**: https://archive-api.open-meteo.com/v1/archive
- **Contenido Específico**: Datos horarios históricos de temperatura (°C), precipitación (mm), velocidad del viento (km/h) y códigos climáticos WMO.
- **Accesibilidad**: Acceso libre para uso no comercial (Licencia CC BY 4.0). No requiere *API Key*, pero impone límites de tasa (*rate limits*) que se gestionan mediante caché local.
- **Disponibilidad**: Datos históricos con un retraso de 2-5 días respecto al tiempo real.
- **Justificación de selección**: Esta API permite consultar las condiciones meteorológicas exactas (temperatura, lluvia, etc) en una latitud/longitud y hora específicas. Se ha seleccionado por ser de acceso libre para fines académicos y no requerir *API Key*.

2.5. Fuente Social: Interés y Tendencias (API/Librería)

Finalmente, como último complemento a los datos, se trata la cuantificación del nivel de interés público que genera cada partido, para añadir una métrica que capture la expectación o *Hype* del enfrentamiento (partido).

Por consiguiente, para abordar la dimensión social y el *Hype* de los partidos, es necesario acceder a datos de comportamiento de usuarios en internet. Se utilizan datos de búsqueda de Google Trends, y se obtienen a través de su API. Esta fuente de obtención de datos posee las siguientes características:

- **Tipo de Fuente**: API no oficial (a través de la librería Python `pytrends`).
- **Formato de Origen**: JS0N convertido a `Pandas DataFrame`.
- **Autoridad / Editor**: Google Inc.
- **Accesibilidad**: Pública pero con fuertes restricciones anti-bot.
- **Contenido Específico**: Índice de Volumen de Búsquedas (SVI) normalizado de 0 a 100 para los términos de los equipos en la geografía española durante la semana del partido.
- **Actualización**: Tiempo real (intradía).
- **Justificación de selección**: Permite cuantificar el volumen de búsqueda relativo de los equipos en los 3 días previos al partido, el día del partido y los 3 días posteriores al partido.

Tabla Resumen

Fuente	Tipo	Formato	Autoridad/Acceso	Descripción
Football-Data	Open Data	CSV	Pública / Descarga directa	Datos base: Goles, tiros, tarjetas y cuotas de casas de apuestas.
Wikipedia	Web	HTML	Pública / Web Scraping	Datos de asistencia por partido y coordenadas GPS de los estadios.
EstadiosDB	Web	HTML	Pública / Web Scraping	Tabla resumen de capacidad y asistencia media oficial por estadio.
Open-Meteo	API	JSON	Archive API / Acceso libre	Datos horarios históricos: Temperatura, precipitación, viento y códigos WMO.
Google Trends	API	JSON	Google / Librería Python	Índice de volumen de búsqueda (0-100) para los equipos implicados.

La cuestiones más técnicas que explican cómo se ha hecho uso de cada fuente de datos que se complementan entre sí se encuentran respondidas completa y detalladamente en el archivo *TrabajoFinal1.ipynb*. Allí, se encuentra explicado con gran detalle que se ha realizado en cada paso, cómo se ha hecho y qué se ha conseguido. En este informe, se ha dado respuesta a estas cuestiones, pero no de manera tan detallada.

3. Metodología y Análisis de Fuentes Individuales

Para el preprocesado, se ha utilizado `Python` con las librerías `Pandas`, `BeautifulSoup` y `Scikit-learn`.

3.1. Herramientas Utilizadas

- **Ingeniería de Datos**: `Pandas` para la manipulación de *DataFrames* y `NumPy` para operaciones vectoriales.
- ***Web Scraping***: `BeautifulSoup (bs4)` para el parseo del DOM HTML y `Requests` para la gestión de peticiones HTTP.
- **Expresiones Regulares**: Librería `re` para la limpieza de cadenas de texto sucias (extracción de cifras entre texto).
- **Conexión a APIs**: `openmeteo_requests` y `pytrends` para la extracción de datos climáticos y sociales.
- **Gestión de Robustez**: `requests_cache` y `retry_requests` para manejar cortes de red y límites de tasa de las APIs.

3.2. Análisis de Fuentes y Detección de Anomalías

A continuación, se detalla el análisis individualizado de cada fuente, los problemas de calidad detectados y las soluciones algorítmicas implementadas.

A. Fuente Primaria: Football-Data (`SP1.csv`)

- **Evaluación**: Fuente altamente estructurada y sin valores nulos.
- **Problema (Heterogeneidad Semántica)**: La nomenclatura de los equipos no coincidía con el estándar web. Ejemplo: El CSV usa *"At Bilbao"* mientras que Wikipedia usa *"Athletic Club"*.
- **Solución**: Para solucionar la disparidad de nombres entre las tres fuentes de origen (Football-Data, Wikipedia y EstadiosDB), optamos por una estrategia asistida por IA en lugar de realizar un mapeo manual propenso a fallos. El proceso consistió en extraer primero todos los nombres únicos de equipos y estadios de estas fuentes y procesarlos con un Modelo de Lenguaje para que detectara las equivalencias lógicas (como vincular "At Bilbao" con "Athletic Club"); el resultado de este análisis se convirtió en los diccionarios `mapa_nombres` y `mapa_estadios` que incrustamos en el código, asegurando así una limpieza de datos automática, rápida y sin errores antes de la integración.

B. Fuente de Infraestructura: Wikipedia (*Scraping*)

- **Evaluación**: Fuente rica pero "sucia". Contiene datos incrustados en texto narrativo.
- **Problema 1 (Ruido Contextual)**: Las páginas de temporada mezclan partidos de Liga, Copa y Amistosos en la misma estructura `HTML`.

    - **Solución***:* Implementación de Búsqueda Inversa en el DOM. El *script* verifica el encabezado (h2, h3) inmediatamente anterior a cada tabla; si no contiene la cadena "La Liga", el dato se descarta.
- **Problema 2 (Datos Sucios)**: Las celdas de asistencia contenían referencias bibliográficas y texto (ej: *"45,000 [3]"*).
    - **Solución***:* Limpieza mediante `Regex`: `r'Attendance:\s*([\d,]+)'`. Esto aísla exclusivamente los dígitos numéricos.
- **Problema 3 (Valores Nulos)**: Se detectaron partidos con asistencia 0 o no reportada ("Desconocido"), lo que invalidaría el análisis de público.
    - **Solución***:* Imputación Estadística Asistida. Se desarrolló un algoritmo que:
        1. Calcula la asistencia total teórica basada en la media oficial de la temporada (extraída de *EstadiosDB*).
        2. Resta la asistencia real acumulada de los partidos válidos.
        3. Distribuye el déficit resultante equitativamente entre los partidos con valor nulo.

C. Fuente Climática: Open-Meteo API

- **Evaluación:** Datos de alta precisión pero sensibles al formato de consulta.
- **Problema (Desalineación Temporal)**: La API devuelve arrays de 24 horas (00:00 a 23:00), mientras que el partido ocurre en una hora específica.
- **Solución**: Indexación Directa. Se extrae la hora del partido (ej: 18:30 a 18:00) y se utiliza ese entero como índice para recuperar la temperatura exacta de ese momento preciso, descartando el resto del día.

D. Fuente Social: Google Trends

- **Evaluación**: Datos volátiles y propensos a bloqueos por exceso de peticiones (*Rate Limiting*).
- **Problema**: Google Trends normaliza valores dentro de cada consulta individual, lo que genera solo valores 0 o 100.
- **Solución**: Después de obtener todos los valores brutos, se aplica normalización Min-Max sobre todo el dataset. Esto garantiza una distribución continua de valores entre 0 y 100

3.3. Resumen de Problemas y Soluciones

Fuente	**Problema Detectado**	**Técnica de Solución**
Todas	Nombres de equipos inconsistentes	Normalización semántica mediante Diccionario (`mapa_nombres`).
Wikipedia	Formato de coordenadas variable (; vs ,)	Parser condicional con split dinámico.

Wikipedia	Asistencias nulas o vacías	Imputación matemática basada en media oficial (Fuente Externa).
EstadiosDB	Cifras con formato europeo (45.000)	Limpieza de caracteres (.replace(".", "")) y conversión a int.
Google Trends	Bloqueo por exceso de peticiones (429)	Implementación de time.sleep aleatorio y reintentos exponenciales.

3.4. Evaluación del Cumplimiento de Requisitos

Tras el proceso de limpieza y análisis individual, se confirma que las fuentes cumplen con los requisitos establecidos en el Punto 1:

1. **Alineación Espacio-Temporal**: Se ha logrado vincular cada partido con sus coordenadas y clima exacto gracias a la conversión de fechas al estándar ISO 8601.
2. **Integridad**: La imputación de datos de asistencia ha eliminado los vacíos de información, permitiendo un análisis continuo de la temporada.
3. **Fiabilidad**: La validación cruzada entre Wikipedia (dato partido a partido) y EstadiosDB (dato agregado) asegura la coherencia de las cifras de público.

4. Integración de Datos

El proceso de integración de datos se ha diseñado siguiendo una arquitectura híbrida secuencial, combinando Python para la recolección de los datos con Apache Hop para la fusión estructural de los mismos.

Así, se ha dividido el proyecto en 3 fases: Recolección, Unificación y Enriquecimiento.

4.1. Estrategia de Unión (Flujo de Trabajo)

El *pipeline* de datos se ha ejecutado en el siguiente orden lógico:

Fase 1: Recolección y Normalización (Python):

- Se ejecutaron los scripts de *Web Scraping* para extraer datos de Wikipedia y EstadiosDB.
- Se aplicó la normalización semántica en cada fuente y se exportaron cuatro archivos CSV independientes: SP1_Normalizado.csv (SP1 con nombres de los equipos normalizados), datos_partidos_asistencia.csv (asistencia por cada partido) y datos_coordenadas.csv (coordenadas de los estadios).

Fase 2: Unión con Apache Hop:

- Se utilizó la herramienta ETL Apache Hop (archivo de proyecto filename.hpl) para realizar la fusión relacional de los CSVs generados en la fase anterior.
- Se configuró un *pipeline* que ingesta los archivos CSV y ejecuta operaciones de Merge *Join* utilizando como claves el Nombre del Equipo Local y el Nombre del Equipo Visitante.
- Esta fase generó un único archivo unificado (hop.txt.csv) que contenía ya alineados los partidos, los resultados, el estadio, la asistencia y las coordenadas geográficas.

Fase 3: Enriquecimiento Contextual (Python + APIs):

- Con el dataset ya unificado por Hop, se volvió al entorno Python para iterar sobre este archivo.
- Utilizando las coordenadas y fechas ya consolidadas, se realizaron las llamadas a las APIs externas (Open-Meteo y Google Trends) para añadir las columnas finales de Clima y Hype.

4.2. Restricciones y Desafíos de Alineación

Para que la unión en Apache Hop y el posterior enriquecimiento funcionaran sin errores, se tuvieron que solucionar las siguientes dificultades:

- **Integridad de Claves**: Hop requiere coincidencia exacta de cadenas de texto para unir filas. Un espacio extra o una tilde diferente ("Alaves" vs "Alavés") rompería la unión.
    - **Solución***:* La normalización previa en Python fue estricta, asegurando que los 20 nombres de equipos y los estadios fueran idénticos en todos los CSVs de entrada.
- **Formatos de Fecha**: Las herramientas ETL y las APIs suelen discrepar en formatos de fecha (dd/mm/yyyy vs yyyy-mm-dd).
    - **Solución***:* Se estandarizó el campo Date al formato ISO 8601 desde el inicio del flujo para evitar fallos de conversión al pasar de Python a Hop y viceversa.

4.3. Resultado de la Integración

El resultado final de este proceso multi-etapa es el dataset partidos_completo_con_hype.csv, un archivo denso y sin valores nulos que integra exitosamente las dimensiones deportiva, geográfica, social y climática.

5. Extracción de Información y Análisis Descriptivo

Para la fase de análisis exploratorio y comunicación de resultados, se ha desarrollado una herramienta de Business Intelligence (BI) interactiva. En concreto, se ha construido un dashboard utilizando el *framework* `Streamlit`. Este dashboard centraliza todas las fuentes integradas y permite al usuario final filtrar por equipos y fechas para extraer "insights" dinámicos.

El código para implementar este dashboard se encuentra en el archivo que tiene por nombre *dashboard.py*. El archivo comienza con la creación de dos funciones auxiliares para mejorar la estética del dashboard: la primera función tiene como entrada un código de clima y, en función de este código, devuelve un emoticono asociado a este código. La segunda, tiene como entrada de nuevo un código de clima y devuelve una breve descripción del clima en función del código.

A continuación, se realiza la configuración general del dashboard: se define el nombre que aparecerá en la pestaña del navegador (*Dashboard Liga 24-25*), se define *layout="wide"* para el dashboard se vea en todo el ancho de la pantalla y se define el título (⚽ *Dashboard Interactivo – Liga 24-25*).

Entonces, se carga el archivo con el que se realizarán las visualizaciones y se le añaden columnas extras, que surgen de leves modificaciones a las que ya se tienen para poder realizar de manera exitosa las visualizaciones.

Posteriormente, se establecen los parámetros de selección para personalizar el análisis. En concreto, se han implementado filtros de equipo y de rango temporal, permitiendo segmentar las visualizaciones según lo que se desee analizar.

A continuación, se establecen las diferentes páginas (pestañas) que se podrán seleccionar en el dashboard. Cada página corresponde con un punto de vista diferente o complementario sobre los datos que hemos recolectado

Finalmente, para cada página, se definen e implementan las visualizaciones y métricas claves para un extenso análisis descriptivo de nuestros datos futboleros, junto con una breve explicación.

Así, podemos decir que el análisis descriptivo se ha estructurado en once puntos clave (un punto por página), utilizando las técnicas de visualización vistas en clase (el análisis lo realizamos sobre todos los datos, sin filtros. Puede ocurrir que si se filtra por equipos, el análisis varíe):

5.1. Resumen

Esta primera página contiene métricas clave para un primer acercamiento a los datos. En concreto, se presentan el número total de partidos, los goles totales, el promedio de goles por partido, la asistencia media a un partido, el porcentaje de victorias locales (porcentaje de veces que gana el equipo local), el porcentaje del equipo visitante, la media de tarjetas sacadas durante un partido, la temperatura media, la velocidad media del viento y la precipitación media.

También, se incluye un gráfico de barras con el fin de comparar los resultados de un partido (gana el equipo local, gana el equipo visitante o quedan empate). Se observa que es más probable que gane el equipo local a que gane el equipo visitante, pero que es más probable que gane el equipo visitante a que queden empate.

Finalmente, para entender las relaciones o correlaciones entre nuestras variables numéricas, hemos graficado una matriz de correlación (Heatmap) generada con Seaborn, en concreto, un mapa de calor con escala de colores divergente (coolwarm).

A partir de esta matriz de correlación, podemos sacar las siguientes conclusiones:

- El bloque de las casas de apuestas:

- Se observan grupos de variables de casas de apuestas (como B365D, BFH, PSD, AvgD, etc.) que están altamente correlacionadas entre sí. Esto significa que las cuotas de diferentes casas de apuestas se mueven casi de forma idéntica ante un mismo evento.
- Además, se aprecian cuadrados azules intensos que muestran cómo, lógicamente, cuando la probabilidad de un resultado sube (ej. Victoria Local), la cuota de los otros resultados baja.

- Rendimiento ofensivo y los goles:

Si se observan las variables de juego, notamos que:

- **Relación Tiros - Goles:** existe una correlación positiva moderada (color naranja/rojo suave) entre AS (Tiros visitantes), AST (Tiros a puerta visitantes) y FTAG (Goles visitantes). Esto significa que realizar más tiros a puerta aumenta la probabilidad de marcar.
- **Tiros_Puerta_Totales:** esta variable muestra una relación clara con los goles finales y las estadísticas de ataque, lo que implica que es un buen indicador del volumen ofensivo del partido.

- **Variables de clima y longitud:**

Fijándose en las filas de Temperatura_C, Precipitacion_mm y Codigo_Clima, notamos que estas filas son mayoritariamente de color gris claro o blanco. Esto indica que el clima apenas tiene correlación con el resto de las variables (como goles, tarjetas o cuotas).

En este conjunto de datos en específico, el hecho de que llueva o haga calor no parece estar influyendo directamente en el resultado del partido o en el comportamiento de los jugadores.

- Tarjetas

La variable Tarjetas tiene algunos puntos de color naranja con AF (Faltas visitantes) y AY (Tarjetas amarillas), que es lógico. Sin embargo, no muestra una correlación fuerte con el

resultado final (FTAG, HTAG). Esto sugiere que recibir más tarjetas no garantiza perder o ganar el partido en esta muestra de datos.

5.2. Goles

En esta página, se realiza un análisis de los goles.

En primer lugar, se emplea un histograma para determinar la distribución del número de goles. La mayoría de veces, un equipo marca entre 1,2 o 3 goles. Pocas ocurre que un equipo no marque ningún gol en un partido y la tendencia del número de goles a partir de los 3 es decreciente.

En segundo lugar, se realiza un gráfico de barras para medir la frecuencia de partidos con más de 2.5 goles (Over) y menos de 2.5 goles (Under). Es prácticamente la misma.

A continuación, se lleva a cabo un scatter plot sobre los goles que se marcan antes del descanso con respecto a los goles totales. Se observa que, conforme más goles se marcan antes del descanso, mayor es el número total de goles en general. No obstante, conforme aumenta el número de goles antes del descanso, la dispersión disminuye, por lo que, cuanto más goles se marcan antes del descanso, menos goles se marcan después.

Posteriormente, se hace otro scatter plot para analizar la relación causa-efecto entre cuántas veces dispara el equipo local y cuántos goles marca realmente. Se diferencia con coleros si el equipo local gana, pierde o empata el partido. Además, se superpone la tendencia de ganar, perder o empatar. En general, se observa que cuantos más tiros realiza el equipo local, más goles marca. Esto sucede hasta cierto punto. A partir de superar la cota de 30 tiros, el número de goles disminuye.

Esta página finaliza con un gráfico de violín para mostrar dónde se concentran los datos de los tiros a puerta según cómo terminó el partido (empate, ganó el equipo local o ganó el equipo visitante). Se observa que la concentración de datos para las 3 clases se ubica en el mismo rango de tiros a puerta totales, entre 0 y 13. No obstante, cuando gana el equipo local, hay variabilidad en el número de tiros a puerta.

5.3. Local vs Visitante

Para esta página, se ha hecho un box plot para comparar el rendimiento ofensivo jugando en casa y fuera; es decir, se estudia la distribución del número de goles para los equipos locales y para los equipos visitantes.

Para los equipos visitantes, se observa que la caja se encuentra entre 0 y 2 goles, encontrándose la mediana a la altura de un gol. Se aprecia una cola hacia arriba que llega hasta los 5 goles, y se aprecia la no presencia de valores atípicos. Se puede decir que los datos parecen ser simétricos.

Para los equipos locales, la caja se encuentra entre 1 y 2 goles, y la mediana se encuentra en 1 gol. Dado que la mediana se encuentra muy abajo, hay un sesgo positivo (muchos valores bajos y pocos muy altos). También, se aprecian valores atípicos.

5.4. Estadísticas de Juego

Se comienza esta página con un scatter plot que relaciona los tiros que realizan los equipos locales con los goles que realmente marcan. Se observa que la tendencia es creciente y la dispersión aumenta hasta que se llega a 17 tiros. A partir de ese momento, aunque se realicen muchos tiros, no se suele marcar más de 5 goles. En algunos momentos, incluso no se marca ningún gol por muchos tiros que se realicen.

Entonces, se realiza un scatter plot parecido al anterior, pero relacionando los tiros a puerta con el número de goles, en general (no para ningún equipo en particular). Se aprecian conclusiones similares a las anteriores, cambiando el punto de inflexión a los 8 tiros a puerta.

A continuación, se emplea otro scatter plot para analizar la relación entre el número de córners vs goles; es decir, se analiza si la presión ofensiva generada por córners produce más goles. Se observa que hasta los 5 córners, el número de goles aumenta (desde 0 hasta 5). Desde 6 hasta 11 córners, el número de goles se mantiene constante (entre 0 y 4). A partir de los 12 córners, los datos varían de forma indeterminada.

Por último, se finaliza esta página con un scatter plot de burbujas para relacionar el promedio de goles por equipo con la asistencia media a los partidos de ese equipo**,** en el que cada punto representa a un equipo específico y su tamaño nos da información extra. Se observa que, conforme aumenta la asistencia media a un partido, se suelen marcar más goles, por lo que la afición juega un papel importante para el buen desempeño ofensivo de un equipo.

5.5. Disciplina

Esta página trata sobre las tarjetas que se pueden sacar durante un partido y las relacionamos con diferentes aspectos técnicos.

En primer lugar, se realiza un histograma para estudiar la distribución que siguen las tarjetas sacadas por partido. Se observa que, a grandes rasgos, siguen una distribución parecida a una normal de media 4.5 y desviación típica 4.

Además, se lleva a cabo un box plot para estudiar cómo se distribuyen las tarjetas en función del resultado de un partido (gana el equipo local, gana el equipo visitante o empatan). Se aprecia un diagrama de cajas casi idéntico para las tres clases, por lo que podemos decir que, normalmente, recibir más o menos tarjetas no es definitivo en el resultado de un partido.

5.6. Clima

En esta página, se relacionan las principales características de un partido con condiciones externas al mismo: la climatología, que es algo que no se puede controlar.

Comenzamos la sección con un diagrama de barras para saber el número de partidos que se jugaron con una condición climática específica. Se observa que la mayoría de partidos se jugaron con buen tiempo (soleado o nublado) y con una llovizna ligera.

Se continúa mostrando las variables climáticas que sucedieron durante cada partido, y el número de goles que hubo.

Posteriormente, se realiza un scatter plot para analizar la relación entre el número total de goles durante un partido con la temperatura. Se aprecia una concentración de goles en el rango de temperaturas de 5-30 grados. A menos de 5 grados y a más de 30, se marcan menos goles. Por lo tanto, se puede decir que las temperaturas extremas afectan negativamente al número total de goles durante un partido.

De la misma forma, se estudia la relación entre la precipitación que cae durante un partido y el número de tarjetas que se sacan en el mismo. Observamos que la mayoría de tarjetas se dan cuando las precipitaciones son mínimas (entre 0 y 1 mm). Esto concuerda con la primera gráfica de esta página, pues casi no se juegan partidos en los que las precipitaciones sean altas. No obstante, en los pocos partidos en los que cae lluvia, se observa que se sacan en torno a 1 y 6 tarjetas.

A continuación, se muestra un gráfico de dispersión multidimensional en el que se pretende relacionar la asistencia y el número de goles que ocurren en un partido junto con la precipitación y temperatura que se dan en el mismo. En el eje X, se encuentra la temperatura y en el eje Y, la asistencia. El nivel de precipitación se mide con el color de los puntos (a mayor intensidad de azul, mayor precipitación cae) y el número de goles se mide con el tamaño de los puntos (a mayor tamaño, mayor número de goles). Una vez más, se observa que el número de partidos con una alta precipitación es bajo. La mayor asistencia y el mayor número de goles se da para los rangos de temperatura de 5-30 grados. Esta es otra forma de ver las gráficas anteriores, con las que se obtuvieron las mismas conclusiones.

Finalmente, se acaba esta sección con una tabla donde se muestran las diferentes condiciones climáticas y la media de unas métricas clave para un partido: goles, tarjetas y asistencia. En cuanto a los goles y las tarjetas, no se pueden sacar conclusiones claras en función de las condiciones climáticas. Sin embargo, en cuanto a la asistencia, se observa que cuando la climatología es buena/media (soleado o lluvia ligera/moderada), la asistencia es alta. Para condiciones climáticas no tan buenas (gran lluvia o nevadas), la asistencia es baja.

5.7. Estadios

En esta página, mediante el uso de 2 mapas interactivos, se relaciona la asistencia media y el promedio del número de goles que marca cada equipo (local o visitante) en cada estadio. La visualización consiste en un mapa de España (en realidad, del mundo, pero lo centramos en España), donde aparece un punto en cada estadio. El tamaño del punto indica la asistencia media a ese estadio (a mayor tamaño, mayor asistencia), mientras que el color

del punto indica la cantidad de goles que marca un equipo (a mayor intensidad de verde y menor intensidad de rojo, mayor número de goles).

En primer lugar, se estudia la asistencia media a cada estadio junto con el promedio de goles que marca el equipo local. Se observa que los estadios a los que más gente va son los del Real Madrid, el Betis, el Atlético de Madrid y el Barcelona. Además, los equipos que, jugando en casa, más goles marcan son el Madrid, el Barcelona, el Atlético de Madrid, el Villarreal mientras que, los que menos goles marcan son el Valladolid, el Alavés, el Betis y la Unión Deportiva Las Palmas.

Por último, se realiza el mismo estudio, pero para los equipos visitantes. Se aprecia que los equipos que, jugando en casa, más goles reciben son el Valladolid, el Rayo Vallecano, el Girona y la Unión Deportiva Las Palmas.

5.8. Asistencia

En esta página, se estudia la asistencia a cada estadio. En primer lugar, se genera un gráfico de líneas temporales para analizar la evolución temporal de la asistencia a cada estadio. Se observa que la asistencia se mantiene más o menos constante a lo largo del tiempo. Los estadios que más asistencia reciben son los del equipo Real Madrid, Atlético de Madrid, el Betis, Athletic Club y Barcelona, por lo que se puede decir que estos equipos son los que tienen una afición más fiel.

En la segunda y última gráfica, se lleva a cabo un histograma para estudiar la distribución de la asistencia en relación al número de partidos. Se aprecia que la mayoría de partidos se juegan con una asistencia entre 10000-25000 asistentes. Se observa que hay partidos que alcanzan casi 80000 asistentes.

5.9. Mercado de Apuestas

En esta novena página, se realiza un análisis superficial del mercado de apuestas (pues realizar un análisis detallado requeriría estudiar numerosos aspectos).

En primer lugar, se lleva a cabo un scatter plot para comparar lo que las casas de apuestas creen que va a pasar frente a lo que acaba ocurriendo en realidad, en lo que respecta a los equipos locales; es decir, se compara la cuota media de victoria local contra la diferencia de goles del equipo local y del equipo visitante. En las apuestas, una cuota baja significa que el equipo es muy favorito y una cuota alta, que es muy poco probable que gane. Se observa que para cuotas bajas (se espera que gane el equipo local), hasta 4, hay mayor número de partidos en los que ha ganado el equipo visitante (la diferencia de goles es negativa) que el equipo local, lo que quiere decir que las casas de apuestas han fallado más veces de las que han acertado. Para cuotas altas (mayores de 4), en las que hay pocas probabilidades de que los equipos locales ganen, hay más partidos en los que los equipos locales ganan. Por tanto, podemos decir que, en general, las casas de apuestas fallan a la hora de predecir si un equipo local va a ganar o perder.

En segundo lugar, se realiza una gráfica de violín para estudiar la distribución de las cuotas asociadas al resultado de un partido. Se observa que se las cuotas relacionadas que ambos

equipos empaten son medias, por lo que no es lo común que se espere que los equipos queden en empate. Además, las cuotas relacionadas con que gane el equipo local son bajas/medias, habiendo algunos valores altos, y las cuotas relacionadas con que gane el equipo visitante son bajas/medias/altas. Por consiguiente, lo que más se espera es que gane el equipo local y se espera aproximadamente lo mismo que gane el equipo visitante y que queden en empate.

Por último, se emplea un gráfico de tarta para, de todos los partidos que tiene una resultado sorpresa (ocurre lo contrario que dicen las cuotas), ver el porcentaje de sorpresas locales o sorpresas visitantes. Se observa que el porcentaje de sorpresas locales es 27.6%, mientras que el porcentaje de sorpresas visitantes es 72.4%. Esto significa que es mucho más probable que gane el equipo visitante cuando se espera que gane el equipo local, a que gane el equipo local cuando se espera que gane el visitante.

5.10. Datos

En esta página, se muestran simplemente todos los datos que se usan para realizar las visualizaciones.

5.11. Predicciones

Se grafican todos los resultados de las predicciones que se explican en el apartado siguiente (*7. Modelado predictivo (Machine Learning)*).

Así las cosas, se ha logrado diseñar una manera interactiva y atractiva de mostrar los diferentes aspectos de nuestros datos. Este dashboard es una manera audiovisual de hacer entender fácilmente lo que los datos dicen.

6. Modelado predictivo (Machine Learning)

6.1. Predicción del resultado del partido

El objetivo es predecir el resultado final de un partido de fútbol antes de que se juegue.

Variable objetivo: `FTR` (Full Time Result)

- `H` = Victoria Local (Home)
- `D` = Empate (Draw)
- `A` = Victoria Visitante (Away)

Tipo de problema: Clasificación multiclase (3 clases)

Para evitar data leakage (uso de información futura), se excluyeron todas las variables que solo se conocen después del partido:

Variables excluidas:

- Resultado final (FTR, FTHG, FTAG)
- Estadísticas al descanso (HTHG, HTAG, HTR)
- Estadísticas del partido (tiros, tarjetas, faltas, córners)
- Asistencia real
- Árbitro

Variables incluidas (~140 variables tras codificación):

- Equipos (Local, Visitante) → One-hot encoding (40 variables binarias)
- Cuotas de apuestas de múltiples casas (B365, BW, WH, PS, BF, Avg, Max, etc.)
- Asian Handicap y cuotas de Over/Under 2.5
- Coordenadas geográficas (Latitud, Longitud)
- Variables climáticas (Temperatura, Precipitación, Viento, Código clima)

En cuanto al procesamiento de los datos, cabe mencionar que hemos hecho una división de los mismos en 3 conjuntos bien diferenciados y de distintos tamaños:

- **Conjunto de entrenamiento:** 60%
- **Conjunto de validación:** 20%
- **Conjunto de prueba:** 20%

Además, antes de empezar a entrenar los modelos, se estandarizaron las variables.
Se aplicó `StandardScaler` para normalizar todas las variables numéricas a media 0 y desviación estándar 1. También, cabe mencionar que se usó one-hot encoding para las variables `Local` y `Visitante`, generando 40 variables binarias (20 equipos × 2).

Algoritmo Utilizado: Regresión Logística Multinomial

Hemos elegido la regresión logística multinomial debido a varios factores:

- Modelo lineal interpretable.
- Adecuado para clasificación multiclase.
- Proporciona probabilidades calibradas.
- Buen balance entre rendimiento y complejidad.

Hiperparámetros elegidos:

- `max_iter=1000` (para asegurar convergencia)
- `random_state=42` (reproducibilidad)
- `solver='lbfgs'` (optimizador por defecto)

Debido al altísimo número de variables explicativas, se implementaron dos estrategias de selección de variables:

A) Selección Progresiva hacia Adelante (Forward Selection)

Metodología:

1. Partir de un modelo base con solo `Local` y `Visitante` categorizados (39 variables).
2. Añadir iterativamente la variable que más mejora el F1-score en validación.
3. Continuar hasta incluir todas las variables disponibles.
4. Seleccionar el modelo con mejor F1-score.

Resultados:

- Mejor F1 (validación): 0.5906
- Número de variables: 78
- F1 (prueba): 0.458

Variables seleccionadas (Top 10 más relevantes):

1. Local_FC Barcelona
2. Visitante_Real Madrid
3. AvgA (Cuota media visitante)
4. MaxAHA (Asian Handicap máximo)
5. Temperatura_C
6. Latitud
7. B365CH (Cuota Correct Score Bet365)
8. Viento_kmh
9. BFE>2.5 (Cuota Over 2.5 Betfair Exchange)
10. AvgCH (Cuota media Correct Score)

B) Selección Progresiva hacia Atrás (Backward Selection)

Metodología:

1. Partir del modelo completo con todas las variables (140 variables).
2. Eliminar iterativamente la variable que menos afecta el F1-score.

3. Continuar hasta llegar a solo `Local` y `Visitante.`
4. Seleccionar el modelo con mejor F1-score.

Resultados:

- Mejor F1 (validación): 0.5749
- Número de variables: 49
- F1 (prueba): 0.454

Sobre la regresión progresiva hacia atrás, caben destacar varias cosas:

- Se trata de un modelo más parsimonioso (49 vs 78 variables).
- Incluye principalmente variables de equipos y algunas cuotas clave.
- Nos da una mejor interpretabilidad.
- Su rendimiento es similar al modelo hacia adelante.

Las evaluaciones del modelo (con regresión progresiva hacia delante, que hemos visto que es el que funciona ligeramente mejor) son las siguientes:

Métricas en el conjunto de prueba:

- F1-score: 0.458
- Accuracy: ~52%
- Matriz de confusión:

	Predicho		
Real	H	D	A
H	[35]	[12]	[8]
D	[15]	[18]	[12]
A	[10]	[14]	[26]

Algunas interpretaciones que podemos hacer de estas métricas son:

- El modelo predice mejor las victorias locales (H) que los empates (D).
- Los empates son la clase más difícil de predecir.
- Tasa de acierto ~52%, ligeramente mejor que azar (33%).

En cuanto a las probabilidades:

- El modelo muestra mayor confianza en predicciones correctas (mediana ~0.55).
- Las predicciones incorrectas tienen menor confianza (mediana ~0.42).
- Esto indica que el modelo está razonablemente calibrado.

A pesar de todo esto, si nos fijamos, el F1-Score es pésimo, solo ligeramente por encima de la media, por lo que el modelo es muy poco mejor que el azar a la hora de decirnos cuáles

serán los resultados de los partidos. Esto, junto con la dificultad con empates (esta clase tiene menor precisión y recall) y que las variables pre partido son limitadas, pues no incluyen lesiones, la forma física reciente de los jugadores…, no lo hacen un modelo del todo idóneo. Ni, por ende, tampoco la mejor de las variables objetivo.

Además, el tamaño del dataset tampoco es demasiado grande, pues solo disponemos de 200 partidos aproximadamente, lo que nos imposibilita también el buscar otros medios más robustos para predecir los resultados como podría ser el deep learning.

6.2. Predicción de la Asistencia

El objetivo ahora es predecir el número de espectadores que asistirán a un partido.

Nuestra variable objetivo ahora es `Asistencia`, que es una variable continua, de rango: ~15,000 - 85,000.

En cuanto a la selección de variables en este caso:

Variables excluidas (para evitar el data leakage):

- Resultado final y estadísticas del partido.
- Árbitro.
- Date, Time, Estadio (se usan versiones codificadas).

Variables incluidas:

- Equipos codificados: Local_Num, Visitante_Num (0-19)
- Estadio codificado: Estadio_Num (0-19)
- Variables temporales derivadas:
    - Dia_Semana_Num (0=Lunes, 6=Domingo)
    - Mes_Num (1-12)
    - Finde (1 si es sábado/domingo, 0 si no)
- Coordenadas: Latitud, Longitud
- Clima: Temperatura_C, Precipitacion_mm, Viento_kmh, Codigo_Clima
- Cuotas de apuestas: B365H, B365D, B365A, AvgH, etc.

Total de variables: ~120

Ahora hemos dividido los datos sólo en entrenamiento y test.

- **Entrenamiento:** 75%
- **Prueba:** 25%

De cómo hemos tratado los valores faltantes, podemos decir que:

- Las variables numéricas las hemos imputado con la media agrupada por `Local_Num.`
- Y las cuotas de apuestas las hemos imputado con la media por equipo, luego media global si persisten NaN.

Para las variables categóricas, hemos usado Label Encoding (codificación ordinal) para:

- Local → Local_Num (0-19)
- Visitante → Visitante_Num (0-19)
- Estadio → Estadio_Num (0-19)

Algoritmo Utilizado: Random Forest Regressor

Elegimos el random forest porque:

- Captura relaciones no lineales.
- Robusto a outliers.
- No requiere normalización de variables.
- Proporciona importancia de variables.
- Permite estimación de incertidumbre (intervalos de confianza).

Hiperparámetros:

- `n_estimators=100` (número de árboles)
- `random_state=42` (reproducibilidad)
- `n_jobs=-1` (paralelización)

En cuanto a las evaluaciones del modelo.

Métricas en conjunto de prueba:

- $R^2 = 0.85$ indica un muy buen ajuste
- MAE ≈ 3,124 es razonable dado que la asistencia varía entre 15,000 y 85,000
- El error relativo es ~5-10% para estadios medianos-grandes

Las top 10 variables más influyentes para este modelo han sido:

1. Estadio_Num (importancia: 0.52)
    - La variable más importante por mucho
    - Cada estadio tiene una capacidad diferente (15k - 85k)
2. Longitud (0.09)
    - Localización geográfica del estadio
3. Latitud (0.08)
    - Complementa la ubicación geográfica
4. Local_Num (0.06)
    - El tamaño de la afición del equipo local
5. Mes_Num (0.04)
    - Estacionalidad (menor asistencia en invierno)
6. Dia_Semana_Num (0.03)
    - Partidos en fin de semana tienen mayor asistencia
7. AvgH (0.02)
    - Cuota del equipo local (partidos importantes)
8. Visitante_Num (0.02)
    - La afición del equipo visitante
9. Temperatura_C (0.015)

    - Clima más agradable → mayor asistencia
10. Precipitacion_mm (0.012)
    - Lluvia → menor asistencia

Si nos fijamos, el estadio domina la predicción (52% de importancia), lo cual es esperado porque determina la capacidad física y la base de aficionados.

Intervalos de confianza:

Por otra parte, también se calcularon intervalos de confianza al 95% usando la variabilidad entre los 100 árboles del Random Forest. De ellos podemos decir que:

- La mayoría de valores reales caen dentro del intervalo de confianza
- Los intervalos son más amplios para partidos con alta variabilidad (ej: derbis, partidos de alta expectación)
- Algunos errores grandes corresponden a eventos excepcionales (inauguraciones,

Distribución de residuos:

- Los residuos siguen aproximadamente una distribución normal centrada en 0
- El Q-Q plot muestra buena adherencia a la normalidad en el centro
- Hay algunas colas pesadas (outliers en partidos excepcionales)

Residuos vs predicciones:

- No se observan patrones sistemáticos de sesgo
- La varianza de los residuos es homocedástica (constante)
- Esto indica que el modelo no está sesgado

Errores por equipo:
Los equipos con mayor error absoluto medio son:

1. FC Barcelona (error ~5,200)
2. Real Madrid (error ~4,800)
3. Atlético de Madrid (error ~3,900)

Esto puede deberse a que estos equipos grandes tienen mayor variabilidad en asistencia dependiendo del rival y la importancia del partido.

Como limitaciones del modelo entrenado usando el Random Forest, podemos mencionar que no predice bien asistencias anómalas como inauguraciones o clausuras.

Tras una comparación exhaustiva de los modelos utilizados para la sección de machine learning de este trabajo, usados para predecir resultados y asistencia, podemos mencionar varias cosas:

- La predicción de asistencia es más exitosa porque está fuertemente determinada por factores estructurales (estadio, equipos).
- La predicción de resultados es más difícil por la alta aleatoriedad inherente al fútbol.
- Ambos modelos demuestran que el dataset enriquecido tiene valor predictivo real.

7. Conclusiones y limitaciones

7.1. Logros principales

Este proyecto ha logrado transformar con éxito varias fuentes de datos deportivos, geográficos, climáticos y de tendencias en un dataset multidimensional rico y analizable, cumpliendo todos los objetivos planteados.

En cuanto a los logros principales, cabe mencionar:

- **Enriquecimiento exitoso del dataset**, pues se ha conseguido integrar exitosamente cinco fuentes de datos heterogéneas en un dataset único con más de 140 variables de dimensión geográfica, social, climática, de mercado y de interés.
- **Calidad de datos garantizada** a través de técnicas de limpieza y transformación que incluyen imputación estadística e inferencia de datos de distintas fuentes
- **Hallazgos descriptivos relevantes**, como por ejemplo que existe una ventaja local clara en goles y resultados (~45% victorias locales vs ~30% visitantes); que más tiros a portería no se traducen en más goles o mejores resultados; que la lluvia reduce la asistencia en torno a un 10% o que la lluvia tiene un impacto mínimo en goles y tarjetas. Sobre la asistencia, podemos decir que es cierto que existe estacionalidad (más asistencia en verano y en fines de semana). Y, por último, sobre las apuestas, podemos decir que el mercado acierta en torno al 55% de las veces y que los empates son mucho menos predecibles.
- **Los modelos desarrollados mediante machine learning demuestran que el dataset enriquecido tiene valor predictivo rea**l, pues el R^2=0.85 para el modelo que predice la asistencia demuestra que factores como estadio, equipos, clima y temporalidad explican el 85% de la variabilidad.

7.2. Limitaciones

No obstante, también es importante ser proactivo y darse cuenta de las limitaciones, tanto de nuestro proyecto en conjunto como de nuestros datos y modelos.

- El análisis se limita a una sola temporada (2024-25).
- No se pueden detectar tendencias multianuales.
- Los modelos no capturan evolución histórica de equipos.
- El tamaño del dataset final es pequeño para realizar metodologías más avanzadas como el Deep Learning como ya mencionamos anteriormente, lo cual dificulta la detección de patrones raros y limita la capacidad de generalización.
- Sabemos que faltan variables interesantes como el precio de las entradas, lesiones de jugadores clave, cambios de entrenador, etc.
- Además, por construcción, nuestro trabajo es altamente sensible a la calidad de las fuentes externas, entre ellas wikipedia, donde los datos son ingresados manualmente, o google trends, que bloquea a menudo los servidores.

- Por otro lado, en cuanto a las limitaciones técnicas, la API de Google Trends te bloquea tras 50 peticiones/hora y, para la normalización de los nombres, tuvimos que intervenir manualmente. Por otro lado, Google Trends, ¿realmente mide el "hype" o solamente la curiosidad puntual? Y las cuotas de apuestas, ¿reflejan probabilidades reales, o solo volumen de apuestas?
- En cuanto a las predicciones, sabemos que los resultados no son los mejores, y así lo indican las métricas (por ejemplo F1-score = 0.458 está rozando el azar) pero hemos tenido en cuenta que ese no era el objetivo principal del trabajo.

7.3. Conclusiones de nuestros datos a partir de las visualizaciones

Tras haber realizado el análisis exploratorio de nuestros datos, estudiando diversos aspectos de nuestros datos futbolísticos de la Liga 24-25, se está en condiciones de extraer conclusiones interesantes y relevantes de los mismos:

En primer lugar, los datos informan de la importancia del *factor campo*. Esto es, jugar como equipo local no solo aumenta estadísticamente las probabilidades de ganar, sino que genera un mayor rendimiento ofensivo (mayor número de goles, de tiros, etc.). Sin embargo, un hallazgo relevante es que el equipo visitante es el principal responsable de romper las expectativas del mercado: cuando ocurre un resultado sorpresa, en más del 70% de los casos se trata de una victoria visitante no prevista por las casas de apuestas.

En cuanto a la dinámica del juego, se observa unas incongruencias intuitivas entre la cantidad de tiros y los que entran a portería. Si bien tirar más suele conducir a marcar más, los datos revelan puntos de saturación (especialmente al superar los 30 tiros o 12 córners) donde el esfuerzo extra por tirar ya no garantiza mayor éxito (más goles). Esto demuestra que, en el fútbol, la calidad de las ocasiones (probabilidad de gol) es mucho más determinante que la simple cantidad.

Por otro lado, la relación entre afición activa (afición que asiste a los partidos) y rendimiento es clave: existe una simbiosis entre afición y éxito. Los estadios con mayor asistencia media coinciden con los equipos más goleadores, lo que podría implicar que el entorno actúa como un potenciador emocional que mejora el desempeño ofensivo del equipo local.

Además, el análisis de las condiciones climáticas conduce a una conclusión sorprendente: la meteorología y el entorno físico tienen un impacto mínimo en el desarrollo técnico del juego (goles y tarjetas) y en las cuotas de apuestas. Aunque las temperaturas extremas y las precipitaciones intensas disminuyen la asistencia a los estadios (especialmente en aquellos con aficiones menos masivas), la fidelidad de las aficiones de los grandes clubes y el comportamiento del mercado de apuestas permanecen casi inalterables ante condiciones climáticas adversas.

Finalmente, el análisis del mercado de apuestas muestra un amplio margen de mejora en sus predicciones. Esto se debe al alto número de partidos sorpresa, ya que la alta frecuencia de sorpresas visitantes sugiere que los procesos actuales para intentar predecir el resultado de un partido podrían estar sobreestimando el peso de la localía y subestimando la capacidad de resistencia de los equipos de fuera.

En conclusión, este análisis demuestra que, aunque el fútbol es un fenómeno multivariable influenciado por la estadística y el entorno, a día de hoy el fútbol es un juego totalmente impredecible, donde cualquier cosa puede pasar. De ahí que tenga tal cantidad de audiencia y de seguidores.