
UNIVERSIDAD FRANCISCO DE VITORIA

**ESCUELA POLITÉCNICA SUPERIOR**

**GRADO EN INGENIERÍA INFORMÁTICA**

PROYECTO FINAL DE GRADO

MODALIDAD INGENIERÍA

# Modelado Predictivo del Valor del Mercado de Jugadores Profesionales de Fútbol basado en Aprendizaje Automático y Análisis de Datos

Álvaro Salvador López

Convocatoria de Enero 2026

**CALIFICACIÓN DEL PROYECTO FINAL DE GRADO**

| | |
|---|---|
| CUALITATIVA: | |
| NUMÉRICA: | |

| Conforme Presidente: | Conforme Secretario: |
|---|---|
| | |
| Fdo.: | Fdo.: |

| Conforme Vocal: | Conforme Vocal: | Conforme Vocal: |
|---|---|---|
| | | |
| Fdo.: | Fdo.: | Fdo.: |

Lugar y fecha: Pozuelo de Alarcón, a ____ de _____ de 202__

# Agradecimientos

Quiero expresar mi agradecimiento a mi familia, por su apoyo incondicional y paciencia durante todos estos años de estudio. Sin su comprensión, ánimo y confianza, este logro no habría sido posible.

También quiero agradecer a mis amigos y compañeros, que me han acompañado y apoyado en los momentos más importantes de esta etapa. Su ayuda y motivación han sido clave para realizar este proyecto.

Por último, agradezco a todos aquellos que, de una u otra forma, han colaborado y contribuido en este proyecto. Sus aportes han sido esenciales para la culminación exitosa de este trabajo.

# Resumen

En este proyecto presento el desarrollo de modelos predictivos para estimar el valor de mercado de jugadores profesionales de fútbol utilizando técnicas de aprendizaje automático y análisis de datos. La motivación detrás de este proyecto está en la importancia de la valoración de jugadores en el mercado del fútbol y la necesidad de aplicar técnicas de ingeniería informática para mejorar la precisión en este proceso. Utilizando datos históricos de jugadores obtenidos de Transfermarkt y FBref (disponibles en Kaggle) para las temporadas 2017-2020, se ha construido un modelo predictivo sólido capaz de estimar con precisión el valor de mercado de los jugadores

# Palabras claves

Aprendizaje automático, análisis de datos, modelo predictivo, valor de mercado, fútbol.

# Abstract

In this project, I present the development of predictive models to estimate the market value of professional football players using machine learning techniques and data analysis. The motivation behind this work lies in the importance of player valuation in the football market and the need to apply computer engineering techniques to improve accuracy in this process. Using historical player data obtained from Transfermarkt and FBref (available on Kaggle) for the 2017–2020 seasons, a robust predictive model has been built to accurately estimate players' market values.

# Keywords

Machine learning, data analysis, predictive model, market value, football.

# Índice de Contenidos

# Índice de Tablas

# Índice de Figuras

# Lista de Acrónimos

| Acrónimo | Significado |
|---|---|
| **ACM** | Association for Computing Machinery |
| **AI** | Artificial Intelligence |
| **API** | Application Programming Interface |
| **CRISP-DM** | Cross Industry Standard Process for Data Mining |
| **CSV** | Comma-Separated Values |
| **CV** | Cross-Validation |
| **EDA** | Exploratory Data Analysis |
| **GPU** | Graphics Processing Unit |
| **IRLS** | Iteratively Reweighted Least Squares |
| **MAE** | Mean Absolute Error |
| **ML** | Machine Learning |
| **PFG** | Proyecto de Fin de Grado |
| **RMSE** | Root Mean Square Error |
| **SMART** | Specific, Measurable, Achievable, Relevant, Time-bound |
| **SVR** | Support Vector Regressor |
| **TPU** | Tensor Processing Unit |
| **UFV** | Universidad Francisco de Vitoria |
| **XAI** | Explainable Artificial Intelligence |

# 1. INTRODUCCIÓN

Este PFG se enmarca en el contexto del mercado del fútbol profesional, donde la valoración de jugadores es una tarea muy importante para clubes, agentes y analistas deportivos. Este mercado es altamente competitivo y multimillonario, y la capacidad de predecir el valor de los jugadores puede ofrecer una ventaja significativa en la toma de decisiones estratégicas.

La motivación detrás de este PFG surge de la necesidad de mejorar la precisión en la estimación del valor de mercado de los jugadores mediante el uso de técnicas avanzadas de aprendizaje automático y análisis de datos. Los métodos tradicionales de valoración pueden ser subjetivos y estar influenciados por múltiples factores externos. Sin embargo, con la disponibilidad de grandes cantidades de datos históricos y el avance en las técnicas de machine learning, es posible crear modelos que ofrezcan estimaciones más objetivas y precisas.

Para este proyecto, he recopilado datos de jugadores de fútbol de las temporadas 2017-2020 de fuentes reconocidas como Transfermarkt [1] y Fbref [2]. Estos datos incluyen estadísticas de rendimiento, edad, posición, historial de transferencias y otros atributos relevantes. A través de un riguroso análisis exploratorio de datos, he identificado las relaciones y patrones clave que influyen en el valor de mercado de los jugadores. Aunque se revisaron fuentes más recientes, no se han encontrado datasets públicos posteriores a 2020 que incluyan, con un nivel de detalle similar, datos de jugadores de las cinco grandes ligas junto con sus valores de mercado y estadísticas de rendimiento. Por este motivo, se decidió mantener este periodo (2017–2020) como la base más completa y representativa disponible para el desarrollo del proyecto.

Los modelos predictivos desarrollados emplean algoritmos de aprendizaje automático, como la regresión y técnicas de modelado predictivo, para crear una herramienta que permita predecir con precisión los valores de mercado de los jugadores.

El documento se estructura en varias secciones:

- **Investigación Previa**: revisa el estado del arte en la valoración de jugadores y justifico la necesidad del proyecto.
- **Objetivos**: detalla las metas específicas y los métodos de validación empleados.
- **El Plan de Desarrollo del Proyecto**: describe la metodología, tecnologías y planificación utilizada para llevar a cabo el proyecto.

- **Desarrollo de la Solución Técnica**: documenta el proceso de implementación del modelo predictivo, detallando cada fase del trabajo realizado y los resultados parciales obtenidos.
- **Resultados**: se presenta un análisis crítico de los hallazgos y se contrastan con los resultados esperados, justificando cualquier desviación producida.
- **Implicaciones Éticas e Impacto Social**: se reflexiona sobre las consecuencias éticas y el impacto en la sociedad.
- **Mi Recorrido en la UFV:** se comparte mi experiencia universitaria y cómo este proyecto culmina mi formación académica, vinculado a mi futuro profesional.
- **Conclusiones**: Por último, las conclusiones finales del proyecto y las posibilidades de evolución futura del trabajo presentado.

# 2. INVESTIGACIÓN PREVIA

## 2.1. CONTEXTO DEL PROYECTO

El mercado del fútbol profesional es uno de los más competitivos y lucrativos del mundo. La valoración precisa de los jugadores es crucial para la toma de decisiones estratégicas por parte de los clubes, agentes y analistas deportivos. Tradicionalmente, la valoración de jugadores se ha basado en la experiencia y el juicio subjetivo de expertos. Sin embargo, con el avance de las técnicas de análisis de datos y aprendizaje automático, es posible desarrollar modelos más objetivos y precisos para estimar el valor de mercado de los jugadores.

Este proyecto se enmarca en este contexto, buscando aprovechar el poder de los datos y la inteligencia artificial para mejorar la precisión en la **valoración de jugadores de fútbol**.

## 2.2. REVISIÓN DE PROYECTOS SIMILARES

En esta sección se presentan diversos estudios y proyectos previos relacionados con la estimación del valor de mercado de jugadores de fútbol mediante técnicas de aprendizaje automático y análisis de datos. Su revisión permite contextualizar el trabajo realizado, identificar los principales enfoques existentes y destacar las diferencias con la propuesta desarrollada en este proyecto.

1. Uso del Dataset Kaggle [3] y Trabajo Previo:

   El conjunto de datos principal utilizado en este proyecto proviene de la plataforma Kaggle, específicamente del archivo "Football Players Transfer Value Dataset", generado como parte del Trabajo de Fin de Grado titulado "Modelling Football Players Values and Their Determinants on a Transfer Market using Robust Regression Models" realizado por Rafał Stępień [4] . Además del dataset, el repositorio público de GitHub asociado al proyecto incluye tanto el código fuente como el documento completo de la tesis original.

   Los datos contenidos en el archivo provienen de dos fuentes principales: las páginas web de Transfermarkt y FBref. Transfermarkt proporciona información detallada sobre valores de mercado, contratos y transferencias de jugadores, mientras que FBref aporta estadísticas de rendimiento individual como goles, asistencias, minutos jugados, etc. La

combinación de ambas fuentes permite construir una base de datos rica y multidimensional.

El trabajo original de Stępień se enfoca en la estimación del valor de mercado mediante modelos de regresión robusta (IRLS) [5], con un enfoque estadístico y explicativo. En cambio, el presente PFG propone un enfoque predictivo con aplicación práctica, utilizando algoritmos modernos de aprendizaje automático como XGBoost [6] o LightGBM [7], los modelos Support vector Regressor (SVR) [8] y Random Forest [9], y técnicas de validación cruzada [10] para asegurar la generalización del modelo.

A continuación, se resumen las principales diferencias entre ambos trabajos:

| **Aspecto** | **Este PFG** | **Tesis original (R. Stępień)** |
|---|---|---|
| **Objetivo principal** | Predecir el valor de mercado de jugadores mediante modelos ML avanzados | Analizar variables que explican el valor mediante regresión |
| **Técnicas utilizadas** | Random Forest, SVR, XGBoost, LightGBM, validación cruzada | Regresión log-lineal robusta con IRLS |
| **Enfoque** | Predictivo | Explicativo |
| **Evaluación** | RMSE, GridSearchCV, pipeline de Scikit-learn | MAE, análisis por posición, validaciones estadísticas |
| **Modelos por posición** | No (modelo único para todos) | Sí (4 modelos: porteros, defensas, medios y delanteros) |
| **Uso de datos** | Reproducción de predicciones con visualizaciones | Análisis detallado de coeficientes e influencia de variables |

***Tabla 1: Comparación de proyectos. Fuente: Elaboración propia.***

El uso de los datos por parte de este PFG se realiza únicamente con fines académicos dentro del marco del Grado en Ingeniería Informática.

2. Predicting Market Value of Football Players Using Machine Learning [11]:

   Este estudio se centra en la **predicción del valor de mercado de jugadores de fútbol** utilizando diferentes algoritmos de aprendizaje automático, como regresión lineal [12], árboles de decisión [13] y modelos de *ensemble*. Los datos utilizados provienen de Transfermarkt y se analizan diversas características de los jugadores, como edad, posición

y estadísticas de rendimiento. El estudio destaca la importancia de seleccionar adecuadamente las características y de utilizar técnicas avanzadas para mejorar la precisión de las predicciones.

Se diferencia del presente proyecto no solo en el uso exclusivo de datos de Transfermarkt, sino también en el enfoque metodológico: mientras el estudio original emplea modelos clásicos sin optimización avanzada, este trabajo implementa técnicas modernas como Random Forest, XGBoost y validación cruzada estratificada, obteniendo un modelo más robusto y generalizable.

3. Predicting injury risk using machine learning in male youth soccer players [14]:

   Este estudio utiliza técnicas de aprendizaje automático para **predecir el riesgo de lesiones** en jugadores jóvenes de fútbol. Aunque el enfoque principal no es la valoración de jugadores, el uso de machine learning para analizar datos de rendimiento y condiciones físicas es relevante y aporta conocimientos sobre la aplicación de estas técnicas en el ámbito deportivo.

4. A machine learning framework for sport result prediction [15]:

   Este artículo presenta un marco de aprendizaje automático para **predecir resultados deportivos.** La metodología y los algoritmos utilizados son aplicables a la predicción de valores de mercado de jugadores, proporcionando una base teórica sólida para el desarrollo de modelos predictivos en deportes.

5. A study of forecasting tennis matches via the Glicko model [16]:

   Este estudio investiga la **predicción de resultados en partidos de tenis** utilizando el modelo Glicko. Aunque se centra en el tenis, las técnicas de predicción y análisis de datos presentadas pueden ser adaptadas y aplicadas a la valoración de jugadores de fútbol.

6. Predictive analysis and modelling football results using machine learning approach for English Premier League [17]:

   Este artículo se enfoca en el uso de machine learning para modelar y **predecir resultados de partidos en la Premier League** inglesa. La metodología utilizada para el análisis de datos de rendimiento y la predicción de resultados es relevante para el desarrollo de modelos de valoración de jugadores.

7. Proyecto Machine Learning: Predicción de Precios de Viviendas en Boston con Regresión [18]:

   Este proyecto de Medium utiliza técnicas de regresión para **predecir precios de viviendas**. La analogía entre la predicción de precios de viviendas y la valoración de jugadores de fútbol es evidente, y las técnicas de regresión utilizadas pueden ser aplicadas de manera similar en este proyecto.

8. Modelling Football Players Values and Their Determinants on a Transfer Market using Robust Regression Models [4]:

   Esta tesis de grado propone un enfoque basado en regresión robusta (IRLS) para **analizar las variables que determinan el valor de mercado de los jugadores de fútbol**. Utiliza

datos extraídos de Transfermarkt y FBref para construir modelos explicativos separados por posición (porteros, defensas, medios y delanteros), destacando la influencia de características como edad, nacionalidad, minutos jugados o goles. Aunque emplea aprendizaje automático desde una perspectiva estadística, el enfoque es analítico y no predictivo. Su metodología contrasta con el presente PFG, que aplica modelos avanzados de machine learning con validación cruzada y ajuste de hiperparámetros para realizar predicciones automáticas del valor de mercado.

Aunque, el desarrollo de este PFG culminó, según se refleja en la planificación, el 27 de mayo de 2024, al ser imperativo haber concluido todas las asignaturas del Grado en Ingeniería Informática, no pudo realizarse su defensa en el curso 2023-2024. No obstante se ha ampliado el estado del arte, con algunos trabajos posteriores a esa fecha:

9. Shen, Q. Predicting the value of football players: machine learning techniques and sensitivity analysis based on FIFA and real-world statistical datasets. Appl Intell 55, 265 (2025) [19]:

   Este artículo propone un **enfoque predictivo del valor de mercado de los jugadores** utilizando algoritmos avanzados de aprendizaje automático, incluyendo Random Forest, Support Vector Machines y XGBoost. A diferencia de este PFG, no se emplea LightGBM. El autor crea una base de datos personalizada combinando múltiples fuentes, como informes estadísticos y plataformas de *scouting*, lo que enriquece el conjunto de variables consideradas. La investigación destaca la utilidad de modelos ML para mejorar la precisión en la estimación del valor de mercado.

10. Hill, D. F., Skinner, J., & Grosman, A. (2025). A review of football player metrics and valuation methods: a typological framework of football player valuations. Managing Sport and Leisure, 1–24 [20]:

    Este trabajo presenta una revisión extensa de los **métodos de valoración de futbolistas**, clasificándolos en un marco tipológico. Se analizan enfoques estadísticos, económicos y basados en aprendizaje automático. La publicación aporta una visión general sólida sobre las principales métricas y metodologías utilizadas actualmente, sirviendo como una base teórica de referencia para futuras investigaciones en el ámbito de la valoración futbolística.

## 2.3. Valor Diferencial de mi Propuesta

Aunque existen varios estudios y proyectos que abordan la valoración de jugadores de fútbol utilizando técnicas de aprendizaje automático, mi propuesta se diferencia en varios aspectos clave:

1. **Integración de Datos de Múltiples Fuentes**: A diferencia de otros estudios que se centran en una única fuente de datos, este PFG utiliza datos de Transfermarkt y FBref para proporcionar una visión más completa y precisa de los jugadores y sus valores de mercado.

2. **Análisis Exploratorio Exhaustivo**: He realizado un análisis exploratorio de datos exhaustivo para identificar las características más influyentes en la valoración de los jugadores. Esto incluye la identificación de patrones y relaciones clave que pueden no ser evidentes a simple vista.
3. **Modelos Predictivos Avanzados:** Utilizo algoritmos avanzados de aprendizaje automático, incluyendo XGBoost y LightGBM para mejorar la precisión de las predicciones. Además, empleo técnicas de optimización de hiperparámetros para afinar los modelos.
4. **Aplicación Práctica:** El modelo predictivo desarrollado no solo se valida con datos históricos, sino que también se implementa en una aplicación práctica que permite a los usuarios ingresar datos de jugadores y obtener una estimación de su valor de mercado.
5. **Uso del Conjunto de Datos Más Completo Disponible:** Este proyecto emplea datos correspondientes a las temporadas 2017–2020 porque constituyen el conjunto más completo y detallado localizado hasta la fecha, incluyendo un elevado número de variables de rendimiento procedentes de Transfermarkt y FBref. Aunque existen datasets más recientes, su contenido es considerablemente más limitado y no reúne el mismo nivel de cobertura estadística para las cinco grandes ligas europeas. Una vez estén disponibles datos actuales con un grado de detalle equivalente, el modelo podrá actualizarse para generar predicciones más recientes y, además, mejorar su rendimiento gracias a un mayor volumen de información.

## 2.4. Tecnologías Utilizadas

En el desarrollo de este proyecto, se han utilizado una serie de tecnologías y herramientas avanzadas, seleccionadas por su capacidad para manejar grandes volúmenes de datos y ejecutar algoritmos de aprendizaje automático de manera eficiente:

- Scikit-learn [21]: Utilizada para la implementación de algoritmos de aprendizaje automático y técnicas de validación cruzada.
- XGBoost [6] y LightGBM [7]: Algoritmos avanzados de boosting utilizados para mejorar la precisión de las predicciones.
- Seaborn [22] y Matplotlib [23]: Herramientas de visualización de datos que facilitan el análisis exploratorio y la presentación de resultados.
- Plotly [24]: Utilizada para crear visualizaciones interactivas que mejoran la interpretación de los datos.

# 3. OBJETIVOS

## 3.1. OBJETIVO GENERAL

El objetivo general de este proyecto es desarrollar un **modelo predictivo que permita estimar con precisión el valor de mercado de jugadores de fútbol profesionales**. Este modelo pretende ser una herramienta útil para clubes, agentes y analistas deportivos, facilitando la toma de decisiones estratégicas en un mercado altamente competitivo y multimillonario. Al proporcionar estimaciones más objetivas y precisas, busca reducir la incertidumbre en las negociaciones de transferencias y contrataciones, mejorando así la eficiencia y la equidad en el mercado del fútbol profesional.

La capacidad de predecir el valor de mercado de los jugadores puede ofrecer una ventaja significativa, ya que permite a los clubes realizar inversiones más informadas y estratégicas. Además, un modelo predictivo preciso puede ayudar a los agentes a negociar mejores contratos para sus clientes y proporcionar a los analistas deportivos datos más fiables para sus estudios y reportes.

## 3.2. LISTA DE OBJETIVOS ESPECÍFICOS

Estos objetivos específicos cumplen las reglas SMART [25]: son objetivos específicos, medibles, alcanzables, realistas y limitados en el tiempo definido para el PFG.

1. **Realizar un análisis exploratorio de los datos**: Examinar los datos históricos de jugadores de las temporadas 2017–2018, 2018–2019 y 2019–2020 con el objetivo de identificar patrones y características relevantes que influyen en la valoración de mercado. Este objetivo se considera alcanzado mediante la generación de visualizaciones estadísticas y gráficas descriptivas que permitan analizar la distribución de las variables y sus relaciones con la variable objetivo.
2. **Implementar y comparar diferentes modelos de aprendizaje automático**: Desarrollar y entrenar varios modelos predictivos para estimar el valor de mercado de los jugadores. Los modelos incluyen XGBoost, LightGBM y RandomForestRegressor. La implementación de múltiples modelos permite comparar su rendimiento y seleccionar el más adecuado para el problema. Cada modelo será evaluado en términos de su precisión, robustez y capacidad para

generalizar a nuevos datos. El objetivo se valida mediante la correcta ejecución de los modelos y la obtención de predicciones evaluables sobre un conjunto de prueba

3. **Optimizar los hiperparámetros de los modelos**: Utilizar técnicas de optimización de hiperparámetros como RandomizedSearchCV para mejorar la precisión de los modelos predictivos seleccionados. La optimización de hiperparámetros es crucial para maximizar el rendimiento de los modelos, ajustando parámetros como la profundidad de los árboles, la tasa de aprendizaje y el número de estimadores. El cumplimiento de este objetivo se mide comparando el valor de la métrica RMSE antes y después del proceso de optimización.
4. **Validar el modelo predictivo final:** Evaluar el rendimiento del modelo optimizado utilizando un conjunto de datos de prueba independiente. Este paso es esencial para asegurar que el modelo tiene una buena capacidad de generalización y puede proporcionar predicciones precisas en datos no vistos durante el entrenamiento. La validación incluye la evaluación de métricas como la raíz del error cuadrático medio (RMSE). Este objetivo se considera cumplido cuando se obtiene y se documenta el valor final de dicha métrica como indicador del error medio de predicción.

## 3.3. Métodos de Validación

1. Para realizar un análisis exploratorio de datos, se utiliza un **conjunto de datos de prueba:** Validar el modelo final utilizando un conjunto de datos de prueba independiente, lo que me permitió evaluar su rendimiento en datos completamente nuevos. Esta validación es crucial para asegurar que el modelo no está sobre ajustado a los datos de entrenamiento y puede generalizar bien a nuevos datos.
2. Para implementar y comparar diferentes modelos de aprendizaje automático, se utiliza la **comparación de modelos**: Comparar los resultados obtenidos por diferentes modelos y configuraciones de hiperparámetros para seleccionar el modelo que mejor se ajuste a los datos y proporcione las predicciones más precisas. Esta comparación se basa en las métricas de evaluación y en la capacidad del modelo para generalizar a nuevos datos.
3. Para optimizar los hiperparámetros de los modelos, se realiza una **optimización de hiperparámetros**: Utilizar RandomizedSearchCV para encontrar la mejor configuración de hiperparámetros para cada modelo. RandomizedSearchCV realiza una búsqueda aleatoria en un espacio definido de hiperparámetros, evaluando diferentes combinaciones y seleccionando la que proporciona el mejor rendimiento. Esta técnica es eficaz para encontrar configuraciones óptimas en un tiempo razonable.
4. Para validar el modelo predictivo final, se utilizan **métricas de evaluación**: Utilizar la raíz del error cuadrático medio (RMSE) para cuantificar la precisión de los modelos predictivos. El RMSE es una métrica comúnmente utilizada en problemas de regresión que mide la diferencia entre los valores predichos por el modelo y los valores reales. Un RMSE más bajo indica un modelo más preciso.

# 4. PLAN DE DESARROLLO DEL PROYECTO

## 4.1. METODOLOGÍA

En este proyecto se ha utilizado la metodología **CRISP-DM** [26], un estándar ampliamente reconocido en el ámbito de la ciencia de datos. CRISP-DM divide un proyecto en seis fases iterativas —comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue— lo que permite estructurar de manera clara y coherente todo el proceso desarrollado en este PFG. Esta metodología ha sido elegida porque se adapta de forma precisa al flujo de trabajo que se necesitaba para el proyecto, desde la recopilación y exploración de los datos hasta la construcción y validación de modelos predictivos.

### 4.1.1. Justificación de la Metodología

La metodología CRISP-DM fue seleccionada desde por los siguientes motivos:

1. **Flexibilidad e Iteración:** CRISP-DM permite avanzar de forma iterativa entre fases, algo fundamental en proyectos de ciencia de datos donde es habitual tener que volver atrás para ajustar el preprocesamiento, revisar variables o repetir un modelo cuando aparecen nuevos patrones en los datos.
2. **Adecuación al Manejo de Datos Reales:** El proyecto trabaja con datos procedentes de distintas fuentes (Transfermarkt y FBref), con estructuras heterogéneas y valores incompletos. CRISP-DM pone un énfasis especial en la fase de preparación de datos, lo que encaja con las necesidades del proyecto: limpieza, integración de temporadas, codificación One-Hot y escalado de características.
3. **Apoyo al Análisis Exploratorio:** La metodología facilita dedicar una fase completa a la comprensión y exploración de los datos, lo cual encaja perfectamente con el uso de herramientas como matplotlib, seaborn y plotly para identificar patrones, distribuciones y relaciones antes de construir los modelos.
4. **Orientación al Modelado Predictivo:** CRISP-DM no limita los algoritmos utilizados y permite comparar distintos modelos dentro de un mismo marco. Esto se ajusta al enfoque del proyecto, en el que se emplean modelos como Random Forest, SVR,

XGBoost y LightGBM, junto con técnicas de ajuste de hiperparámetros como *RandomizedSearchCV*.

5. **Evaluación Estructurada:** La metodología incluye una fase específica dedicada a la evaluación, coherente con el uso de:
    a. StratifiedShuffleSplit para dividir los datos de forma consistente
    b. RMSE como métrica principal
    c. Búsquedas aleatorias de hiperparámetros para seleccionar el modelo óptimo

    Esto asegura una validación objetiva y comparable entre modelos.
6. **Reproducibilidad y Organización del Proyecto:** La estructura de CRISP-DM encaja con la forma en la que se ha organizado el código, especialmente mediante pipelines de scikit-learn, que permiten repetir y documentar todos los pasos del proceso sin ambigüedades.

### 4.1.2. Descripción de la Metodología

A continuación se describe cómo se han aplicado las fases de la metodología CRISP-DM en el desarrollo del proyecto:

1. **Comprensión del Negocio:** En esta fase se definió el objetivo principal del proyecto: desarrollar un modelo capaz de estimar el valor de mercado de futbolistas profesionales.
   También se identificó la importancia de este tipo de predicción en el ámbito deportivo y se establecieron los requisitos generales del sistema.
2. **Comprensión de los Datos:** En esta etapa se analizaron las fuentes de datos disponibles:
    - Se utilizaron datos procedentes de Transfermarkt y FBref, recopilados para tres temporadas: 2017–2018, 2018–2019 y 2019–2020.
    - Los archivos se leyeron en formato CSV desde la plataforma Kaggle.
    - Se revisó la estructura del dataset, las variables disponibles y la calidad de los datos.
3. **Preparación de los Datos:** Esta fase incluyó todas las tareas necesarias para dejar los datos listos para el modelado:
    - **Limpieza de Datos**
        - Eliminación de valores nulos.
        - Ajuste de índices.
        - Corrección de inconsistencias detectadas entre temporadas.
    - **Integración de Datos**

    - Creación de una columna de “temporada” para cada dataset antes de concatenarlos.
    - Unificación de las tres temporadas en un único conjunto de datos consolidado.
- **Transformación de Datos**
    - Creación de nuevas columnas relevantes (por ejemplo, nacionalidad o posición).
    - Codificación de variables categóricas mediante One-Hot Encoding.
    - Escalado de características numéricas para facilitar el entrenamiento de los modelos.
- **División del Dataset**
    - Separación en conjuntos de entrenamiento y prueba utilizando StratifiedShuffleSplit, garantizando una distribución equilibrada.

4. **Modelado:** Se han implementado y evaluado varios modelos de regresión: Random Forest, Support Vector Regressor (SVR), XGBoost, y LightGBM. Se aplicó RandomizedSearchCV para ajustar los hiperparámetros de cada modelo. Se probaron diversas combinaciones de parámetros para maximizar el rendimiento.
5. **Evaluación:** La evaluación se realizó utilizando:
    - RMSE (Root Mean Square Error) como métrica principal.
    - Validación estratificada mediante StratifiedShuffleSplit.
    - Comparación entre modelos para seleccionar el que mejores resultados ofreció en términos de error y generalización.
6. **Despliegue:** Aunque el modelo no se desplegó en un entorno productivo, se construyó:
    - Un **pipeline final** que integra el preprocesamiento y el modelo óptimo.
    - Un sistema de predicción capaz de estimar el valor de mercado de nuevos jugadores utilizando ese pipeline.
    - Una estructura reproducible que permite repetir el proceso o actualizar los datos en el futuro.

## 4.2. TECNOLOGÍAS

En el desarrollo del proyecto se han utilizado diversas tecnologías que han facilitado la manipulación, el análisis, la visualización y el modelado de los datos:

1. **Lenguajes de Programación**

- Python [27]: He elegido Python como el lenguaje principal debido a su versatilidad y la extensa cantidad de bibliotecas disponibles para la ciencia de datos, análisis y modelado predictivo.

2. **Bibliotecas y Frameworks**
   - **Pandas**: Utilizada para la manipulación y análisis de datos, pandas permite trabajar de manera eficiente con grandes conjuntos de datos y proporciona herramientas robustas para la limpieza y transformación de datos.
   - **NumPy**: Esencial para operaciones numéricas y manejo de arrays, NumPy facilita el cálculo de estadísticas y la realización de operaciones matemáticas complejas.
   - **Matplotlib y Seaborn**: Estas bibliotecas se emplean para la visualización de datos. Matplotlib es una herramienta fundamental para crear gráficos estáticos, mientras que Seaborn ofrece una interfaz de alto nivel para la visualización estadística.
   - **Plotly**: Utilizada para crear gráficos interactivos, Plotly permite una exploración más profunda de los datos a través de visualizaciones dinámicas.
   - **Scikit-Learn**: Esta biblioteca se ha utilizado para el preprocesamiento de datos, construcción y evaluación de modelos predictivos. Proporciona herramientas para la validación cruzada, la búsqueda de hiperparámetros y la construcción de pipelines.
   - **XGBoost y LightGBM**: Bibliotecas avanzadas de aprendizaje automático que se han empleado para el desarrollo de modelos de regresión. Estas herramientas son conocidas por su eficiencia y alto rendimiento en competiciones de ciencia de datos.
   - **Joblib** [28]: Utilizada para guardar y cargar modelos, Joblib facilita la persistencia de objetos Python.
3. **Entornos de Desarrollo**
   - **Google Colab** [29]: He utilizado Google Colab como el entorno de desarrollo para ejecutar el código Python. Google Colab proporciona recursos computacionales potentes y gratuitos, incluyendo GPUs, lo cual facilita el entrenamiento de modelos complejos.
   - **GitHub** [30]: He utilizado GitHub para el control de versiones y la colaboración, permitiendo el almacenamiento seguro y el acceso fácil a los archivos del proyecto.

## 4.3. PLAN DE DESARROLLO DEL PROYECTO

Aunque el proyecto completo ha seguido todas las fases definidas por la metodología CRISP-DM, en este apartado se describen únicamente los paquetes de trabajo asociados al desarrollo técnico principal del sistema. Antes de iniciar el desarrollo técnico del sistema, el proyecto incluyó una serie de fases previas orientadas a la definición del alcance, la

investigación inicial, el establecimiento de objetivos y la selección de la metodología y tecnologías a emplear. Estas fases, aunque no forman parte directa del desarrollo técnico del modelo, han sido fundamentales para estructurar correctamente el proyecto y se encuentran descritas en los apartados correspondientes de la memoria, así como reflejadas en el diagrama de Gantt del apartado 4.4.

Por este motivo, el presente apartado se centra en los paquetes de trabajo que abarcan las fases de comprensión de los datos, preparación, análisis exploratorio, modelado y evaluación, ya que constituyen el núcleo técnico del desarrollo del proyecto.

El proyecto se ha dividido en varios paquetes de trabajo (PT), cada uno con objetivos específicos y tareas detalladas. Aunque la metodología CRISP-DM se estructura en seis fases, para el desarrollo del proyecto se decidió agruparlas en tres paquetes de trabajo con el fin de organizar las tareas de manera más operativa y coherente con el flujo real del proyecto.

Cada PT integra varias fases de CRISP-DM: el PT1 recoge las actividades de comprensión y preparación de los datos, el PT2 corresponde a la fase de análisis exploratorio, y el PT3 integra las fases de modelado, evaluación y preparación del pipeline final.

Esta agrupación permite presentar el desarrollo del proyecto de forma clara sin alterar la estructura metodológica seguida. Los paquetes de trabajo son los siguientes:

- **PT 1: Recopilación y Preprocesamiento de Datos**
- **PT 2: Análisis Exploratorio de Datos (EDA)**
- **PT 3: Desarrollo y Evaluación de Modelos Predictivos**

### 4.3.1. PT 1: Recopilación y Preprocesamiento de Datos

| **Código** | **PT01** | **Nombre** | **Investigación Previa** |
|---|---|---|---|
| **Objetivo** | | En este paquete de trabajo, los esfuerzos se han centrado en la recopilación de datos y su preparación para el análisis y modelado posterior. | |
| **Entradas** | | N/A | |
| **Salidas** | | N/A | |
| **Tareas** | | ○ **PT01-T01:** Recopilación de Datos<br>○ **PT01-T02:** Limpieza de Datos<br>○ **PT01-T03:** Transformación de Datos<br>○ **PT01-T04:** Preparación de Conjuntos de Datos | |

***Tabla 2: PT 1: Recopilación y Preprocesamiento de Datos. Fuente: Elaboración propia***

En este paquete de trabajo, los esfuerzos se han centrado en la recopilación de datos y su preparación para el análisis y modelado posterior. Las principales tareas incluidas en este PT son:

- Recopilación de Datos: Obtención de datos de Transfermarkt y FBref para las temporadas 2017-2018, 2018-2019, 2019-2020. Los datos se han descargado desde Kaggle en formato CSV.
- Limpieza de Datos: Eliminación de valores nulos, ajuste de índices, y corrección de inconsistencias en los datos.
- Transformación de Datos: Creación de columnas adicionales, como year, nationality, y position, y manejo de variables categóricas mediante codificación One-Hot.
- Preparación de Conjuntos de Datos: División de los datos en conjuntos de entrenamiento y prueba utilizando StratifiedShuffleSplit para asegurar una representación balanceada de las clases.

### 4.3.2. PT 2: Análisis Exploratorio de Datos (EDA)

| **Código** | **PT02** | **Nombre** | **Investigación Previa** |
|---|---|---|---|
| **Objetivo** | | En este paquete de trabajo, se ha realizado la exploración visual y estadística de los datos para identificar patrones y relaciones significativas. | |
| **Entradas** | | N/A | |
| **Salidas** | | N/A | |
| **Tareas** | | ○ **PT02-T01:** Visualización de Datos<br>○ **PT02-T02:** Análisis Estadístico<br>○ **PT02-T03:** Identificación de Correlaciones | |

***Tabla 3: PT 2: Análisis Exploratorio de Datos. Fuente: Elaboración propia***

En este paquete de trabajo, se ha realizado la exploración visual y estadística de los datos para identificar patrones y relaciones significativas. Las principales tareas incluidas en este PT son:

- Visualización de Datos: Creación de gráficos de dispersión, gráficos de barras, histogramas, y gráficos de violín para explorar la distribución de valores de los jugadores, nacionalidades, posiciones, ligas y pie dominante.
- Análisis Estadístico: Cálculo de estadísticas descriptivas para entender las características y distribuciones de las variables.
- Identificación de Correlaciones: Uso de matrices de correlación para identificar relaciones significativas entre las variables y la variable objetivo.

### 4.3.3. PT 3: Desarrollo y Evaluación de Modelos Predictivos

| Código | PT03 | Nombre | Investigación Previa |
|---|---|---|---|
| **Objetivo** | | En este paquete de trabajo, se han implementado y evaluado varios modelos de regresión para predecir los valores de los jugadores. | |
| **Entradas** | | N/A | |
| **Salidas** | | N/A | |
| **Tareas** | | • **PT03-T01:** Implementación de Modelos<br>• **PT03-T02:** Búsqueda de Hiperparámetros<br>• **PT03-T03:** Evaluación de Modelos<br>• **PT03-T04:** Pipeline de Predicción | |

***Tabla 4: PT 3: Desarrollo y Evaluación de Modelos Predictivos. Fuente: Elaboración propia***

En este paquete de trabajo, se han implementado y evaluado varios modelos de regresión para predecir los valores de los jugadores. Las principales tareas incluidas en este PT son:

- Implementación de Modelos: Desarrollo de modelos de regresión utilizando Random Forest, SVR, XGBoost y LightGBM.
- Búsqueda de Hiperparámetros: Utilización de RandomizedSearchCV para encontrar los mejores hiperparámetros para cada modelo.
- Evaluación de Modelos: Evaluación de los modelos utilizando StratifiedShuffleSplit y RandomizedSearchCV, con la métrica de error cuadrático medio (RMSE) como principal criterio de rendimiento.
- Pipeline de Predicción: Definición de un pipeline final que incluye preprocesamiento y el mejor modelo encontrado para realizar predicciones sobre el conjunto de datos de prueba.

## 4.4. Plan de Trabajo

Para llevar a cabo este proyecto, se ha elaborado un plan de trabajo detallado que se visualiza en el diagrama de Gantt a continuación. Este plan calendarizado incluye plazos específicos e hitos importantes del proyecto, asegurando una gestión efectiva del tiempo y los recursos.

### 4.4.1. Diagrama de Gantt

El diagrama de Gantt presentado en esta sección muestra la planificación del proyecto a nivel de paquetes de trabajo (PT) con el objetivo de facilitar su lectura. El desglose completo de actividades asociado a cada PT se incluye en el anexo correspondiente, donde se presenta el diagrama de Gantt detallado.

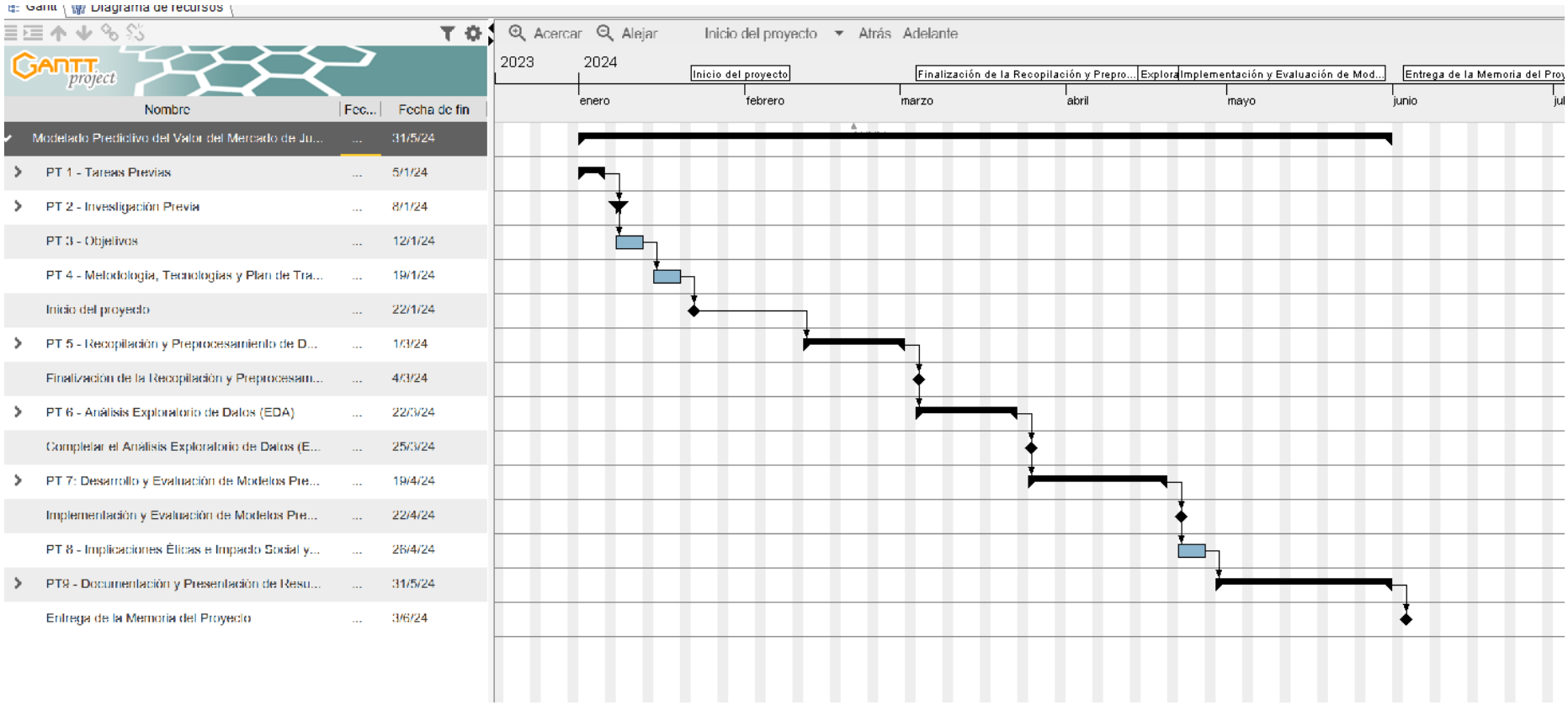



*Figura 1: GANTT. Fuente: Elaboración propia.*

### 4.4.2. Explicación del Plan de Trabajo

El diagrama de Gantt presentado refleja la planificación completa del proyecto, desde las primeras tareas de investigación hasta la entrega final de la memoria. Para organizar el trabajo se ha tenido en cuenta que un proyecto de 12 ECTS requiere aproximadamente 300 horas de dedicación, distribuidas a lo largo de varios meses. El cronograma combina tareas técnicas, actividades de documentación e hitos de control, lo que permite visualizar claramente el avance del proyecto.

A continuación se detallan las fases y paquetes de trabajo representados en el Gantt:

1. **PT 1 – Tareas Previas (1/1/24 – 5/1/24)**

   Esta fase recoge las actividades iniciales necesarias para poner en marcha el proyecto:

   - ACT 1.1 – Realización del Anteproyecto
   - ACT 1.2 – Presentación del Anteproyecto
   - ACT 1.3 – Revisión y Aceptación
   - ACT 1.4 – Asignación del Proyecto

   Incluye la preparación del documento inicial, su presentación y la aprobación por parte del tutor.

2. **PT 2 – Investigación Previa (8/1/24)**

   Incluye tareas iniciales de documentación:

   - ACT 2.1 – Revisión del Anteproyecto
   - ACT 2.2 – Identificación de Fuentes
   - ACT 2.3 – Otra actividad de investigación preliminar

3. **PT 3 – Objetivos (8/1/24 – 12/1/24)**

   Definición de los objetivos generales y específicos del proyecto con el fin de orientar toda la metodología posterior.

4. **PT 4 – Metodología, Tecnologías y Plan de Trabajo (15/1/24 – 19/1/24)**

   Redacción y selección de:

   - La metodología CRISP-DM como marco de referencia.
   - Las tecnologías, bibliotecas y herramientas necesarias.
   - El plan de desarrollo que estructura todo el proyecto.

5. **PT 5 – Recopilación y Preprocesamiento de Datos (12/2/24 – 1/3/24)**

En este paquete se agrupan las actividades correspondientes a la comprensión y preparación de los datos:

- ACT 5.1 – Recopilación de Datos (12–16 febrero): Obtención de datos de Transfermarkt y FBref para las temporadas 2017-2018, 2018-2019 y 2019-2020.
- ACT 5.2 – Limpieza de Datos (19–23 febrero): Eliminación de valores nulos, corrección de inconsistencias y ajuste de índices.
- ACT 5.3 – Transformación de Datos (26 febrero – 1 marzo): Creación de nuevas columnas, codificación One-Hot y escalado de características.
- ACT 5.4 – Preparación de Conjuntos de Datos (26 febrero – 1 marzo): División en conjuntos de entrenamiento y prueba mediante *StratifiedShuffleSplit*.

6. **Hito:** Finalización del preprocesamiento de datos (4/3/24)
7. **PT 6 – Análisis Exploratorio de Datos (EDA) (4/3/24 – 22/3/24)**

   Incluye todas las actividades de exploración visual y estadística:

   - ACT 6.1 – Visualización de Datos (4–8 marzo): Gráficos de dispersión, histogramas, barras, violín, etc.
   - ACT 6.2 – Análisis Estadístico (11–15 marzo): Cálculo de estadísticas descriptivas y revisión de distribuciones.
   - ACT 6.3 – Identificación de Correlaciones (18–22 marzo): Uso de matrices de correlación para detectar relaciones significativas.
8. **Hito**: Finalización del análisis exploratorio (25/3/24)
9. **PT 7 – Desarrollo y Evaluación de Modelos Predictivos (25/3/24 – 19/4/24)**

   Corresponde a las fases de modelado y evaluación de CRISP-DM:

   - ACT 7.1 – Implementación de Modelos (25 marzo – 5 abril): Entrenamiento de Random Forest, SVR, XGBoost y LightGBM.
   - ACT 7.2 – Búsqueda de Hiperparámetros (8–12 abril): Optimización mediante *RandomizedSearchCV*.
   - ACT 7.3 – Evaluación de Modelos (15–19 abril): Validación con *StratifiedShuffleSplit* y comparación mediante RMSE.
   - ACT 7.4 – Pipeline de Predicción (15–19 abril): Integración del mejor modelo y del preprocesamiento en un *pipeline* final.
10. **Hito:** Implementación y evaluación completadas (22/4/24)
11. **PT 8 – Implicaciones Éticas, Impacto Social y Conclusiones (22/4/24 – 26/4/24)**

    Elaboración de la reflexión ética, impacto social y redacción de conclusiones finales del proyecto.

12. **PT 9 – Documentación y Presentación de Resultados (29/4/24 – 31/5/24)**

Correspondiente a la parte final del proyecto:

- ACT 9.1 – Redacción de la Memoria (29 abril – 10 mayo)
- ACT 9.2 – Revisión y Corrección (13–24 mayo)
- ACT 9.3 – Preparación de la Presentación Final (27–31 mayo)

13. **Hito:** Entrega final de la memoria del proyecto: 3/6/24.

## 4.5. RECURSOS

En este apartado se describen los recursos utilizados, su función dentro del proyecto y cómo han contribuido al correcto desarrollo de cada una de las fases:

1. **Hardware:**
   - Ordenador Personal: Se utilizó el ordenador personal del alumno para el desarrollo del proyecto. Las especificaciones básicas de esta computadora son:
     - Procesador: Intel Core i5
     - Memoria RAM: 16 GB
   - Google Colab: Para el entrenamiento de modelos más intensivos en términos computacionales, he utilizado Google Colab. Google Colab proporciona acceso gratuito a GPUs y TPUs, lo que acelera significativamente el proceso de entrenamiento de modelos de machine learning.
2. **Software y Entornos de Desarrollo:**
   - Google Colab: Principal entorno de desarrollo para la ejecución de código Python. Google Colab permite compartir cuadernos, colaborar en tiempo real y aprovechar recursos computacionales avanzados como GPUs y TPUs.
   - GanttProject: Herramienta utilizada para la planificación y seguimiento del proyecto mediante la creación de diagramas de Gantt.
   - GitHub: Plataforma utilizada para el control de versiones y almacenamiento de archivos del proyecto. GitHub facilita la colaboración y la gestión de versiones del código.
   - Microsoft Word: Utilizado para la redacción y documentación de la memoria del proyecto.
3. **Librerías y Frameworks:**
   - Pandas: Para la manipulación y análisis de datos.
   - NumPy: Para operaciones numéricas y manejo de arrays.
   - Matplotlib y Seaborn: Para la visualización de datos.
   - Plotly: Para la creación de gráficos interactivos.

- Scikit-Learn: Para el preprocesamiento de datos, construcción y evaluación de modelos predictivos.
- XGBoost y LightGBM: Para el desarrollo de modelos de regresión avanzados.
- Joblib: Para guardar y cargar modelos.

4. **Fuentes de Datos:**

   1. Kaggle: Utilicé Kaggle para obtener los archivos CSV necesarios para el análisis de datos. Kaggle proporciona una amplia variedad de datasets que son accesibles y fáciles de integrar en proyectos de análisis de datos.

5. **Personas que han Colaborado en el Proyecto**

   Aunque he llevado a cabo la mayor parte del trabajo de forma individual, he recibido apoyo y orientación de mi tutor a lo largo del desarrollo del proyecto. Me ha proporcionado orientación académica, revisiones periódicas del progreso del proyecto y sugerencias para mejorar la metodología y los análisis.

## 4.6. COSTES

A continuación, se detallan los costes asociados al desarrollo del proyecto siguiendo las directrices habituales de presupuestación profesional. Se diferencian los costes de tipo **OPEX** (operativos y recurrentes) y **CAPEX** (inversiones en activos duraderos). Además, en cada partida se incluyen los márgenes aplicables y los impuestos correspondientes, tal como se recoge en la guía de costes del PFG

1. **Costes de Personal (OPEX)**

   El proyecto ha sido desarrollado por un único perfil técnico (Desarrollador de Datos/ML). Para calcular el coste profesional se sigue el procedimiento habitual:

   - Sueldo bruto anual estimado para un perfil junior: 24.000 €/año
   - Coste total para la empresa (incluye cotizaciones y cargas): 24.000 * 1,34 = 32.160€
   - Horas laborales anuales: 1.760 h
   - Cargabilidad estimada: 85 % → 1.496 horas imputables
   - Tarifa profesional resultante:
   - 32.160 € / 1.496 h = 21,49 €/h ≈ 22 €/h

2. **Horas imputadas al proyecto**

   El proyecto requiere aproximadamente 300 horas reales, equivalentes a las 25 horas/ECTS.

- **Cálculo del coste**
  - Coste base: 300 h * 22 €/h = 6.600 €

- Margen empresarial (40 %): 6.600 € * 0.40 = 2.640 €
- Subtotal (coste + margen): 6.600 + 2.640 = 9.240 €
- IVA (21 %): 9.240 * 0.21 = 1.940,40 €
- **Total Personal: 11.180,40 €**

2. **Costes Técnicos**

1. **Hardware (CAPEX)**

- El ordenador utilizado no se compra para el proyecto, sino que se imputa la parte proporcional correspondiente al período de uso. Precio estimado equipo: 1.200 €
  - Vida útil: 48 meses
  - Uso imputado: 4 meses del proyecto
  - **Amortización imputable:** 1.200 / 48 * 4 = **100 €**
- **Aplicación de márgenes e impuestos:**
  - Margen CAPEX (20%): 100 * 0.20 = **20 €**
  - Subtotal: 120 €
  - IVA 21%: 25,20 €
  - **Total Hardware:** 145,20 €

2. **Software por Suscripción (OPEX)**

| Herramienta | Tipo | Coste | Periodo | Coste imputado |
|---|---|---|---|---|
| Google Colab Pro (opcional) | OPEX | 10 €/mes | 4 meses | 40 € |
| Microsoft 365 (Word) | OPEX | 70 €/año | 4 meses | 23 € |

*Tabla 1 Costes Software por suscripción*

- **Cálculo total OPEX software (40 € + 23 € = 63 €)**
  - Margen 30 %: 18,9 €
  - Subtotal: 81,9 €
  - IVA 21 %: 17,20 €
  - **Total Software:** 99,10 €

3. **Costes de Infraestructura (OPEX)**

No se han utilizado plataformas cloud de pago, hosting o bases de datos externas, por lo que esta partida es **0 €**.

4. **Otros Recursos (OPEX)**

- Dataset de Kaggle: 0 €

- Reuniones, documentación, electricidad, etc.: se incluye un coste simbólico imputable:
  - Coste base: 30 €
  - Margen (30%): 9 €
  - Subtotal: 39 €
  - IVA 21%: 8,19 €
  - Total Otros: 47,19 €

5. **Resumen de Costes**

| Categoría | Tipo | Coste Base (€) | Margen (€) | IVA (€) | TOTAL (€) |
|---|---|---|---|---|---|
| **Personal** | OPEX | 6.600,00€ | 2.640,00€ | 1.940,40€ | 11.180,40€ |
| **Hardware** | CAPEX | 100,00€ | 20,00€ | 25,20€ | 145,20€ |
| **Software suscripción** | OPEX | 63,00€ | 18,90€ | 17,20€ | 99,10€ |
| **Otros recursos** | OPEX | 30,00€ | 9,00€ | 8,19€ | 47,19€ |
| **TOTAL GENERAL** | — | 6.793,00€ | 2.687,90€ | 1.990,99€ | 11.471,89 € |

***Tabla 6: Resumen de Costes. Fuente: Elaboración propia***

## 4.7. Condicionantes y Limitaciones

1. Tiempo Limitado Disponible

Uno de los condicionantes más significativos ha sido el tiempo limitado disponible para dedicar al proyecto, debido a la gran carga de trabajo de otras asignaturas del plan de estudios. Este factor ha tenido varios impactos:

- **Distribución del Tiempo**: La necesidad de equilibrar el tiempo entre el proyecto y las demás asignaturas ha reducido las horas semanales que pude dedicar exclusivamente al proyecto. Esto ha requerido una gestión del tiempo más estricta y una planificación detallada para asegurar el cumplimiento de los plazos.
- **Impacto en la Profundidad del Análisis**: Con más tiempo disponible, podría haber realizado análisis más detallados y exploraciones adicionales de los datos. Algunas áreas del proyecto podrían beneficiarse de un estudio más profundo y detallado que no fue posible debido a las restricciones de tiempo.

2. Recursos Computacionales

Aunque Google Colab ha proporcionado una excelente plataforma para el desarrollo y entrenamiento de modelos, también presentó algunas limitaciones:

- **Limitaciones de la Versión Gratuita**: La versión gratuita de Google Colab tiene restricciones en cuanto a tiempo de ejecución y uso de GPU, lo que a veces ha limitado la capacidad de realizar entrenamientos prolongados o pruebas exhaustivas de modelos.
- **Desconexiones y Tiempos de Espera**: Hubo ocasiones en las que las sesiones se desconectaron automáticamente o se requería esperar para acceder a recursos computacionales, lo que interrumpió el flujo de trabajo y extendió el tiempo necesario para completar ciertas tareas.

3. Complejidad de los Datos

Trabajar con datos de múltiples temporadas y fuentes ha presentado desafíos en términos de limpieza y unificación de los datos:

- **Integración de Datos**: La integración de datos de diferentes temporadas y fuentes requirió un esfuerzo considerable para asegurar la consistencia y la compatibilidad de las características. El proceso de limpieza y transformación de los datos fue más complejo y consumió más tiempo de lo inicialmente previsto.
- **Manejo de Valores Nulos**: La presencia de valores nulos en los datasets requirió una cuidadosa imputación y manejo para evitar sesgos en los modelos predictivos.

# 5. DESARROLLO DE LA SOLUCIÓN TÉCNICA

En este apartado se describen las distintas fases y paquetes de trabajo desarrollados a lo largo del proyecto, siguiendo la planificación definida en el diagrama de Gantt y alineadas con la metodología CRISP-DM.

## 5.1. PT 1 – TAREAS PREVIAS

Esta fase recoge las actividades iniciales necesarias para poner en marcha el proyecto y formalizar su desarrollo académico.

1. **Realización del Anteproyecto**
   - **Descripción:** Se elaboró el anteproyecto inicial, definiendo el tema, el contexto general y una primera aproximación a los objetivos del trabajo.
   - **Resultados:** Documento base del proyecto aprobado para su presentación.
   - **Dificultades:** No se presentaron dificultades relevantes en esta etapa.
2. **Presentación del Anteproyecto**
   - **Descripción:** Presentación formal del anteproyecto al tutor académico.
   - **Resultados:** Validación inicial del enfoque del proyecto.
   - **Dificultades:** No se detectaron incidencias.
3. **Revisión y Aceptación**
   - **Descripción:** Revisión del anteproyecto tras la presentación y aplicación de las correcciones sugeridas.
   - **Resultados:** Aprobación definitiva del anteproyecto.
   - **Dificultades:** Ajustes menores de redacción y enfoque.
4. **Asignación del Proyecto**
   - **Descripción:** Asignación oficial del proyecto y comienzo del trabajo planificado.
   - **Resultados:** Inicio formal del PFG.
   - **Dificultades:** No se encontraron dificultades.

## 5.2. PT 2 – INVESTIGACIÓN PREVIA

En esta fase se realizó una primera investigación documental con el objetivo de contextualizar el proyecto y conocer trabajos relacionados.

1. **Revisión del Anteproyecto**
    - **Descripción:** Análisis del anteproyecto aprobado para afinar el enfoque técnico.
    - **Resultados:** Definición más clara del alcance del proyecto.
2. **Identificación de Fuentes**
    - **Descripción:** Búsqueda de datasets, artículos científicos y proyectos similares relacionados con la predicción del valor de mercado de futbolistas.
    - **Resultados:** Selección de fuentes relevantes como Kaggle, Transfermarkt y FBref.
    - **Dificultades:** No se encontraron datasets más recientes con un nivel de detalle comparable.
3. **Investigación Preliminar**
    - **Descripción:** Revisión general de técnicas de machine learning aplicadas al deporte.
    - **Resultados:** Base teórica para la elección de modelos y metodología.

## 5.3. PT 3 – DEFINICIÓN DE OBJETIVOS

En esta fase se definieron los objetivos generales y específicos del proyecto, que guían todo el desarrollo posterior.

- **Descripción:** Definición de los objetivos técnicos y académicos del PFG.
- **Resultados:** Objetivos claros y alineados con la metodología CRISP-DM.
- **Dificultades:** No se presentaron dificultades.

## 5.4. PT 4 – METODOLOGÍA, TECNOLOGÍAS Y PLAN DE TRABAJO

Esta fase estuvo orientada a establecer el marco metodológico y organizativo del proyecto.

- **Descripción:**
    - Selección de la metodología CRISP-DM.
    - Elección de tecnologías, bibliotecas y herramientas.
    - Definición del plan de trabajo y del cronograma.

- **Resultados:** Marco metodológico sólido y planificación detallada del proyecto.
- **Dificultades:** No se encontraron dificultades relevantes.

## 5.5. PT5: Recopilación y Preprocesamiento de Datos

En esta fase, me centré en la recopilación, limpieza y transformación de los datos necesarios para el proyecto.

1. **Recopilación de Datos**
    - **Descripción**: Se obtuvieron los datos de las temporadas 2017-2018, 2018-2019 y 2019-2020 desde Kaggle. Estos datos se almacenaron en archivos CSV y contenían información detallada sobre los jugadores de fútbol.
    - **Resultados**: Se logró recopilar un conjunto de datos robusto con múltiples características relevantes para el análisis y el modelado predictivo.
    - **Dificultades**: No se encontraron mayores dificultades en esta etapa, ya que los datos estaban disponibles en un formato accesible y estandarizado.
2. **Limpieza de Datos**
    - **Descripción**: Se realizó la eliminación de valores nulos, ajuste de índices y corrección de inconsistencias en los datos. También añadí columnas adicionales como year para cada temporada.
    - **Resultados**: Se obtuvo un conjunto de datos limpio y consistente, listo para el análisis y modelado.
    - **Dificultades**: La presencia de valores nulos en algunas columnas requirió un manejo cuidadoso para evitar sesgos en los modelos predictivos.
3. **Transformación de Datos**
    - **Descripción**: Se crearon nuevas columnas para capturar información relevante, como la nacionalidad y las posiciones de los jugadores. Además, se manejaron las variables categóricas mediante codificación One-Hot.
    - **Resultados**: Se logro un conjunto de datos transformado y enriquecido, con todas las variables listas para ser utilizadas en el análisis y modelado.
    - **Dificultades**: La integración de datos de diferentes temporadas y fuentes presentó algunos desafíos, pero se resolvieron con técnicas de manipulación de datos adecuadas.
4. Preparación de Conjuntos de Datos
    - **Descripción**: Los datos fueron divididos en conjuntos de entrenamiento y prueba utilizando StratifiedShuffleSplit para asegurar una representación balanceada de las clases.

- **Resultados**: Se obtuvieron conjuntos de datos de entrenamiento y prueba bien definidos, listos para el análisis exploratorio y la construcción de modelos.
- **Dificultades**: No se encontraron mayores dificultades en esta etapa.

## 5.6. PT6: Análisis Exploratorio de Datos (EDA)

En esta fase, se realizó un análisis detallado de los datos para identificar patrones y relaciones significativas.

1. **Visualización de Datos**
   - **Descripción**: Se crearon diversas visualizaciones para explorar la distribución de los valores de los jugadores, nacionalidades, posiciones, ligas y pie dominante. Utilicé gráficos de barras, histogramas y gráficos de dispersión para entender mejor las características de los datos.
   - **Resultados**: Se obtuvieron gráficos que mostraban la distribución de los valores de los jugadores por temporada, la frecuencia de nacionalidades, posiciones y pie dominante.
   - **Dificultades**: Algunas visualizaciones requirieron ajustes para manejar adecuadamente los outliers y presentar los datos de manera clara.

**Valores jugadores**

En la Figura 2 se representa la distribución del valor de mercado de los jugadores en las temporadas 2017-2018, 2018-2019 y 2019-2020. En todas ellas se observa una distribución muy asimétrica, donde la mayoría de los jugadores se concentran en valores bajos, mientras que solo un pequeño número alcanza valores muy elevados. Este comportamiento refleja la realidad del mercado del fútbol profesional y justifica la necesidad de utilizar modelos capaces de manejar distribuciones no equilibradas.

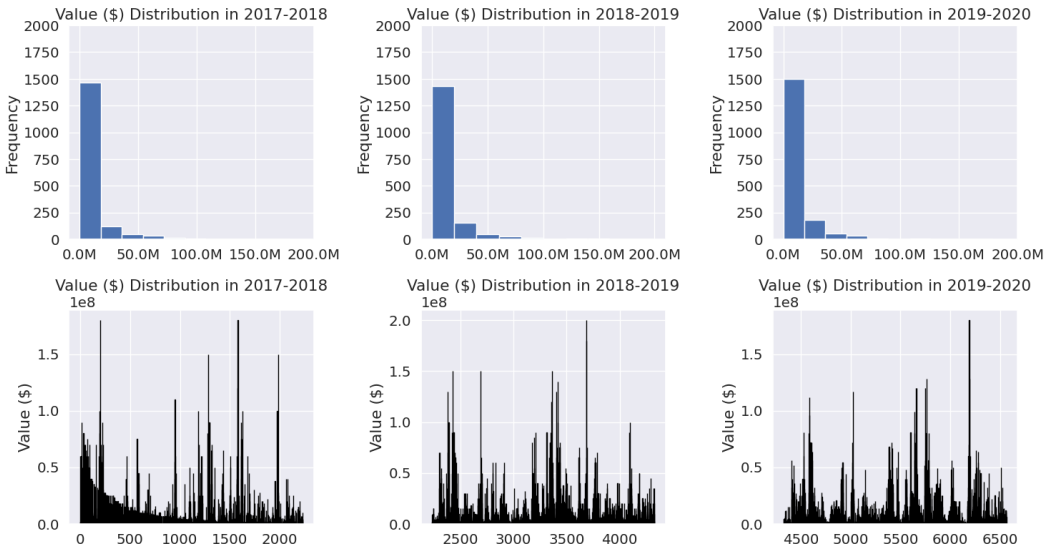



*Figura 2: Muestra la distribución del valor de los jugadores en las temporadas. Fuente: Elaboración propia*

### Atletas por nación

La Figura 3 muestra la distribución de jugadores según su nacionalidad. Se aprecia que determinadas nacionalidades cuentan con una mayor presencia en las ligas analizadas, lo que suele estar relacionado con una mayor tradición futbolística y una mayor proyección internacional. Esta información resulta relevante, ya que la nacionalidad puede influir indirectamente en el valor de mercado de un jugador.

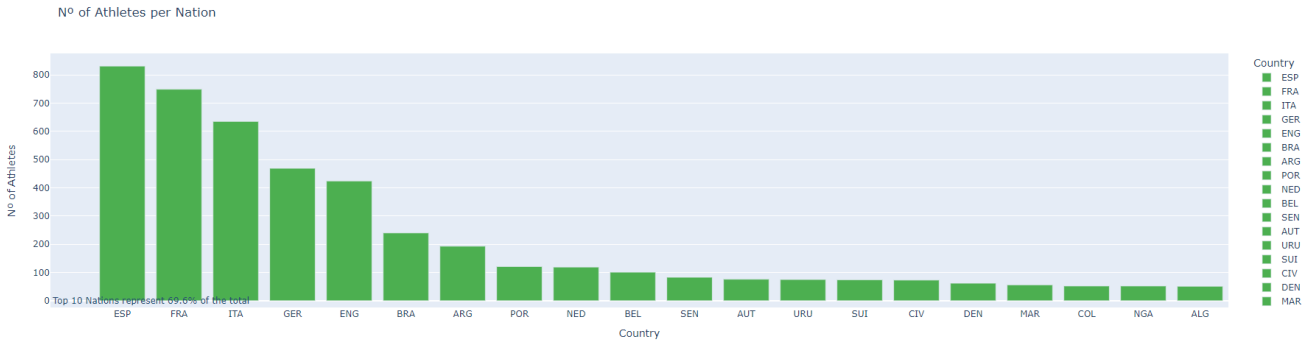



*Figura 3: Muestra la cantidad de atletas por nación. Fuente: Elaboración propia*

### Valor por posición

La Figura 4 compara el valor medio de mercado de los jugadores según su posición en el campo. Se observa que los jugadores de ataque presentan, de media, valores más elevados que los defensas y los porteros, lo que refleja la mayor relevancia ofensiva en el mercado del fútbol. Esta diferencia justifica la inclusión de la posición como una variable relevante en el modelo predictivo.

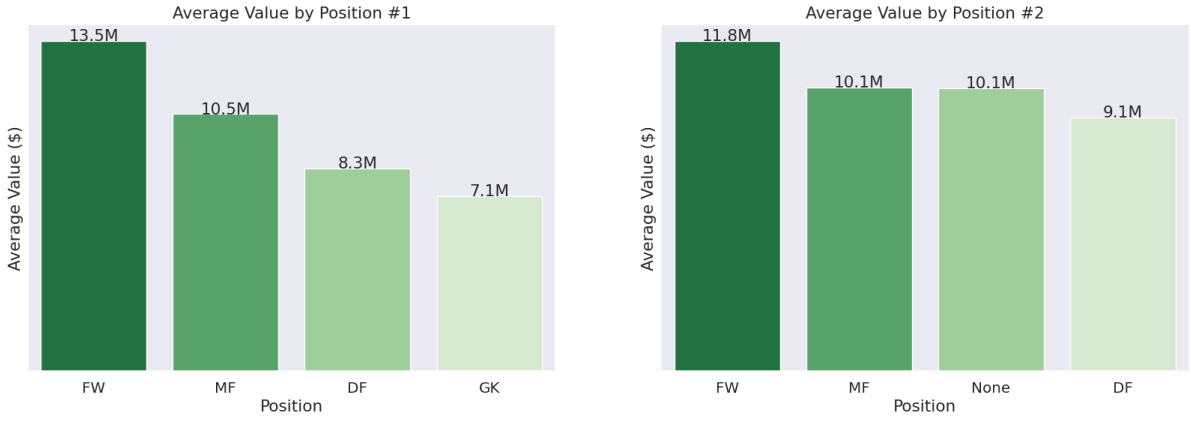



*Figura 4: Compara el valor promedio de los jugadores en diferentes posiciones. Fuente: Elaboración propia*

### Valor por liga

En la Figura 5 se muestra la distribución del valor de mercado de los jugadores por liga. Se aprecia que existen diferencias claras entre competiciones, tanto en la dispersión como en los valores máximos alcanzados. Este comportamiento indica que la liga en la que compite un jugador puede influir significativamente en su valor de mercado.

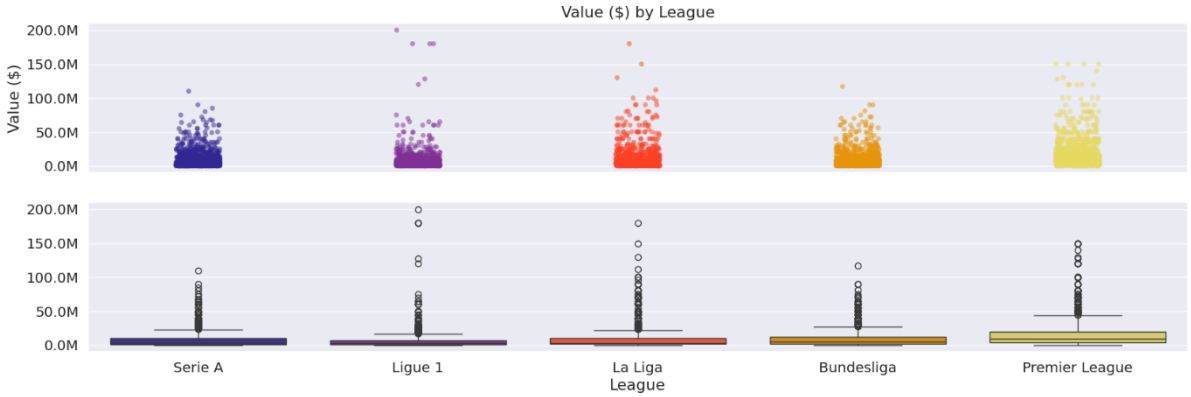



*Figura 5: Muestra la distribución del valor de los jugadores por liga. Fuente: Elaboración propia*

### Pie dominante de los jugadores

La Figura 6 presenta la distribución de jugadores según su pie dominante. Se observa que la mayoría de los futbolistas son diestros, mientras que los jugadores ambidiestros representan un porcentaje reducido. Esta variable se analiza posteriormente para estudiar si existe alguna relación con el valor de mercado.

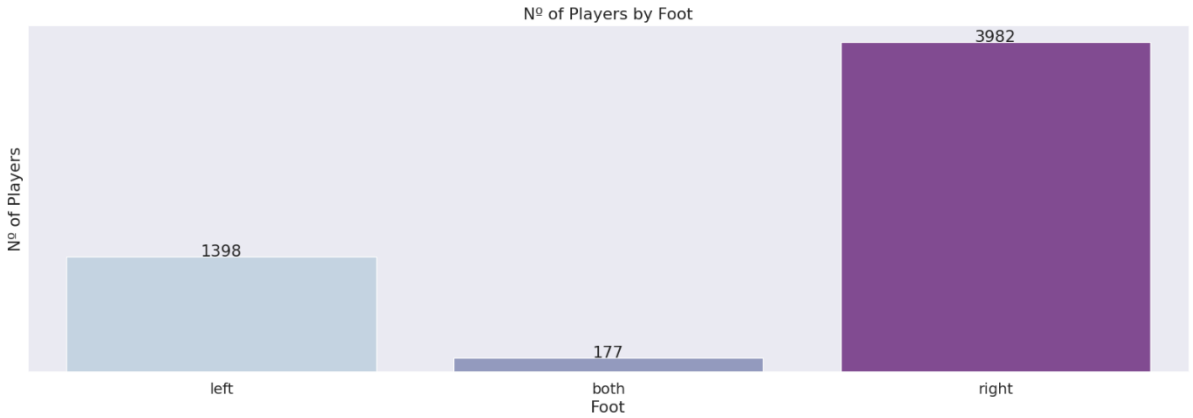



*Figura 6: Muestra el número de jugadores según su pie dominante. Fuente: Elaboración propia*

### Valor dependiendo del pie dominante

En la Figura 7 se analiza la distribución del valor de mercado en función del pie dominante del jugador. Aunque la mayoría de los jugadores son diestros, se aprecia que los futbolistas ambidiestros tienden a concentrar valores más elevados. Este patrón sugiere que la versatilidad puede tener un impacto positivo en la valoración de mercado.

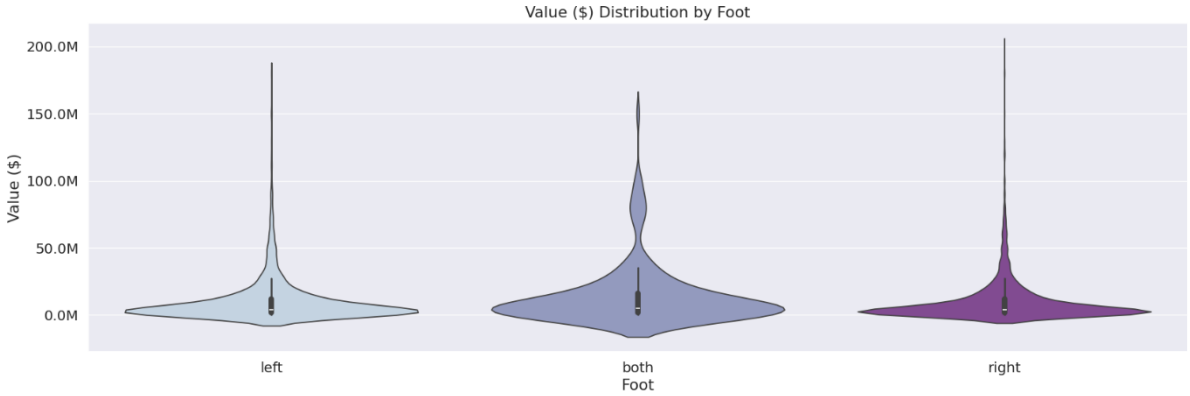



*Figura 7: Muestra la distribución del valor de los jugadores según su pie dominante. Fuente: Elaboración propia*

### Valor por asistencias

La Figura 8 muestra la relación entre el número de asistencias y el valor de mercado de los jugadores. Se observa una tendencia general creciente, donde mayores cifras de asistencias suelen asociarse con valores más altos. Esto indica que las métricas de rendimiento ofensivo son relevantes para explicar el valor de mercado.

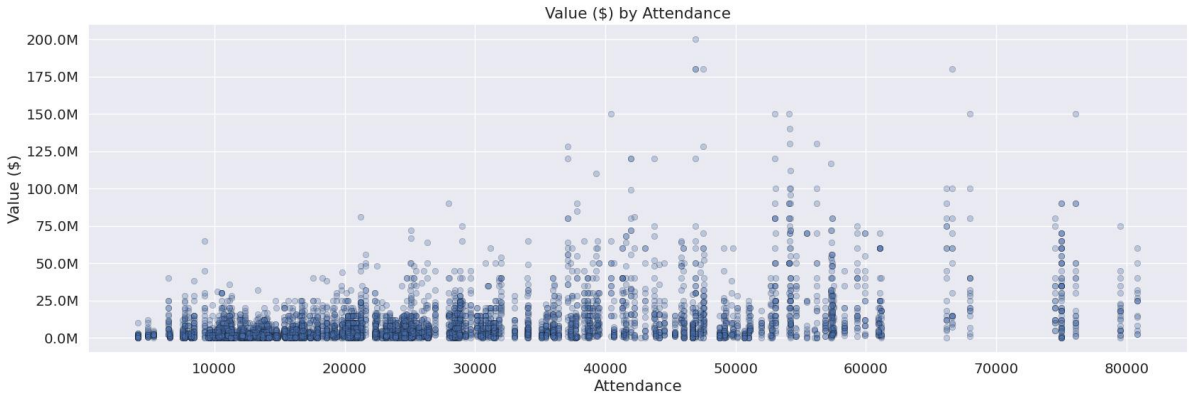



*Figura 8: Muestra la relación entre la asistencia y el valor de los jugadores. Fuente: Elaboración propia*

## 2. Análisis Estadístico

- **Descripción**: Se calcularon estadísticas descriptivas para entender las características y distribuciones de las variables.
- **Resultados**: Se obtuvieron un resumen estadístico de las principales variables, lo que permitió identificar tendencias y posibles anomalías en los datos.

- **Dificultades**: El manejo de grandes volúmenes de datos requirió optimizaciones para asegurar un análisis eficiente.

3. **Identificación de Correlaciones**

- **Descripción**: Se utilizaron matrices de correlación para identificar relaciones significativas entre las variables y la variable objetivo.
- **Resultados**: Se identificaron las variables más correlacionadas con el valor de los jugadores, lo que ayudó a seleccionar las características más relevantes para el modelado.
- **Dificultades**: Algunas correlaciones inesperadas requirieron un análisis más profundo para entender su origen y relevancia.

**Matriz de dispersión**

En la Figura 9 se representan las relaciones entre distintas variables del conjunto de datos. Se identifican correlaciones claras entre algunas características de rendimiento, lo que permite detectar variables potencialmente relevantes para el modelado. Este análisis sirve como base para la selección de variables en las fases posteriores.

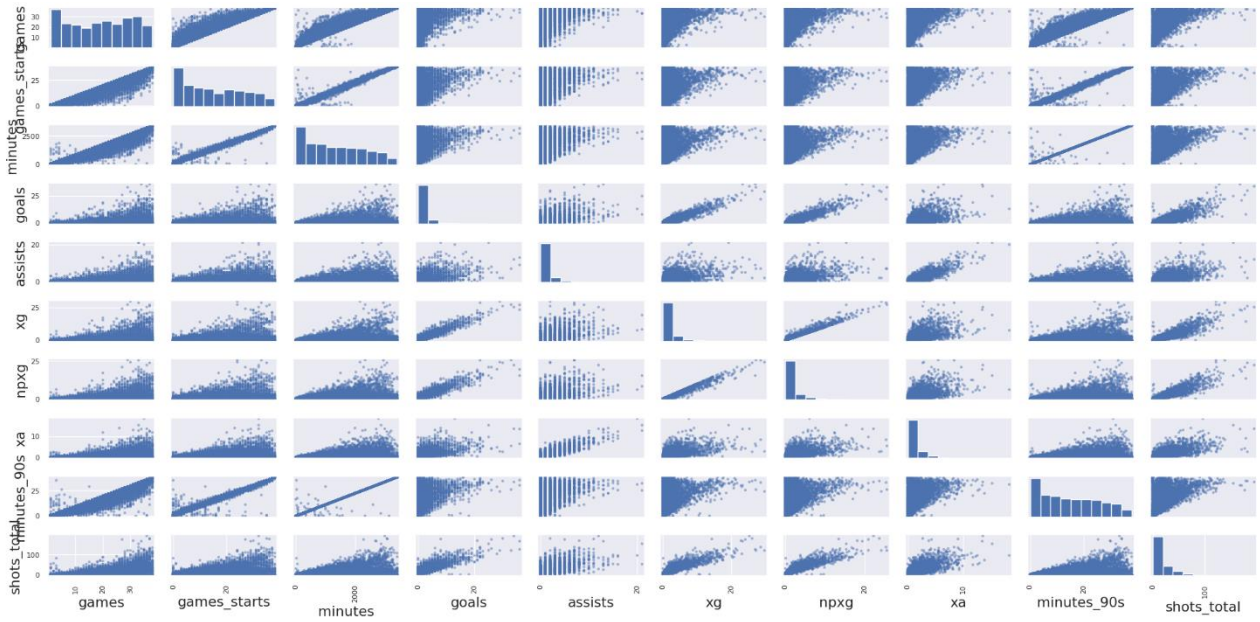



*Figura 9: Muestra las relaciones entre diferentes características. Fuente: Elaboración propia*

**Gráficos de dispersión**

La Figura 10 permite visualizar la dispersión y la relación entre múltiples variables y el valor de mercado. Se observa una alta variabilidad en muchas características, lo que refuerza la necesidad de emplear modelos no lineales capaces de capturar relaciones complejas entre variables.

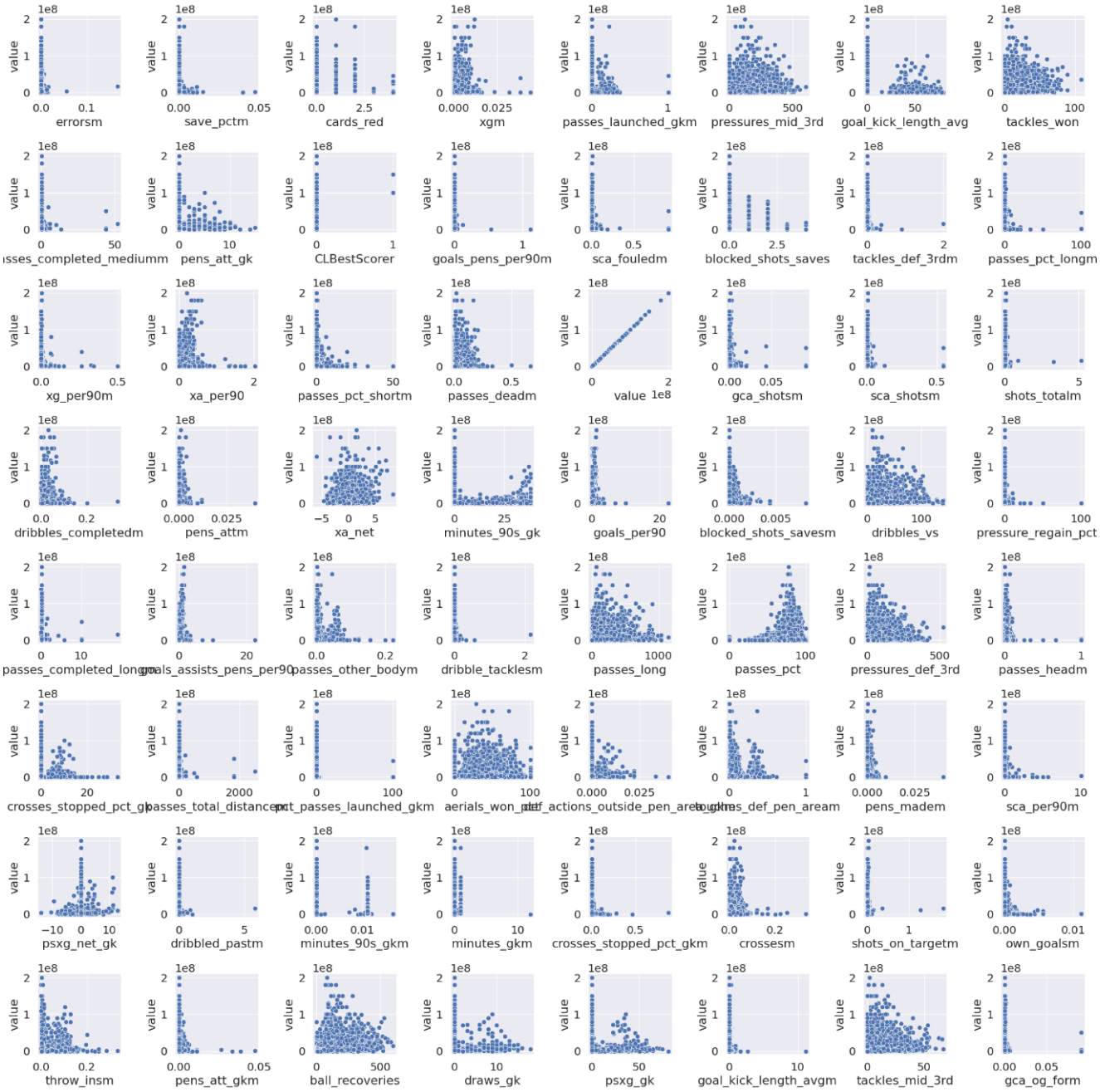



*Figura 10: Visualiza la relación y la dispersión entre diferentes variables. Fuente: Elaboración propia*

## 5.7. PT 7: Desarrollo y Evaluación de Modelos Predictivos

En esta fase, se implementaron y evaluaron varios modelos de regresión para predecir los valores de los jugadores.

1. **Implementación de Modelos**
    - **Descripción**: se desarrollaron modelos de regresión utilizando Random Forest, Support Vector Regressor (SVR), XGBoost y LightGBM.
    - **Resultados**: Se crearon modelos predictivos con diferentes enfoques y comparé sus rendimientos.
    - **Dificultades**: El ajuste de hiperparámetros y la selección del modelo óptimo requirió un enfoque iterativo y la realización de múltiples pruebas.
2. **Búsqueda de Hiperparámetros**
    - **Descripción**: Se utilizó RandomizedSearchCV para encontrar los mejores hiperparámetros para cada modelo.
    - **Resultados**: Se identificarón las configuraciones de hiperparámetros que optimizaban el rendimiento de cada modelo.
    - **Dificultades**: La búsqueda de hiperparámetros fue intensiva en términos computacionales, lo que limitó el número de combinaciones probadas.
3. **Evaluación de Modelos**
    - **Descripción**: Se evaluaron los modelos utilizando StratifiedShuffleSplit y RandomizedSearchCV, con la métrica de error cuadrático medio (RMSE) como principal criterio de rendimiento.
    - **Resultados**: El modelo de XGBoost mostró el mejor rendimiento, con el menor RMSE en el conjunto de datos de prueba.
    - **Dificultades**: La variabilidad en los resultados debido a diferentes divisiones de los datos requirió una evaluación cuidadosa para asegurar la robustez del modelo.
4. **Pipeline de Predicción**
    - **Descripción**: Se definió un pipeline final que incluye preprocesamiento y el mejor modelo encontrado para realizar predicciones sobre el conjunto de datos de prueba.
    - **Resultados**: Se desarrolló un pipeline reproducible y eficiente para la predicción de valores de jugadores, listo para su aplicación en nuevos datos.
    - **Dificultades**: Integrar todos los pasos en un solo pipeline requirió un diseño cuidadoso para asegurar la correcta ejecución de cada etapa.

# 6. RESULTADOS

En este capítulo, se describen e interpretan los resultados obtenidos durante el desarrollo del proyecto, realizando un análisis crítico de los mismos. Además, se contrastan estos resultados con los esperados y se justifica cualquier desviación producida. Se evalúa el grado de consecución de los objetivos planteados inicialmente, aportando evidencias que lo justifiquen.

## 6.1. RESULTADOS DEL ANÁLISIS EXPLORATORIO DE DATOS (EDA)

1. **Distribución del Valor de los Jugadores por Temporada**

En la fase de análisis exploratorio de datos, utilicé gráficos de histograma para visualizar la distribución del valor de los jugadores por temporada. Los resultados mostraron que la mayoría de los jugadores tienen un valor inferior a los 50 millones de dólares, con pocos jugadores de alto valor (superiores a 100 millones de dólares).

2. **Distribución del Valor de los Jugadores por Liga**

Se utilizaron gráficos de dispersión y gráficos de caja para analizar la distribución del valor de los jugadores en diferentes ligas. Se observó que la Premier League tiende a tener jugadores con valores más altos en comparación con otras ligas como la Serie A, Ligue 1, La Liga y Bundesliga.

3. **Distribución del Valor de los Jugadores por Nacionalidad**

En el análisis por nacionalidad, se identificó que España, Francia e Italia tienen el mayor número de jugadores en el dataset. Además, las diez principales naciones representan el 69.6% del total de jugadores.

4. **Distribución del Valor de los Jugadores por Posición**

Se analizó la distribución del valor de los jugadores según sus posiciones principales y secundarias. Los delanteros (FW) tienden a tener los valores promedio más altos, seguidos por los mediocampistas (MF) y los defensores (DF).

5. **Distribución del Valor de los Jugadores por Pie Dominante**

Finalmente, se examinó la distribución del valor de los jugadores según su pie dominante (izquierdo, derecho o ambos). Los jugadores que utilizan ambos pies tienden a tener un valor más alto, aunque representan una minoría en el dataset.

## 6.2. Resultados del Desarrollo y Evaluación de Modelos Predictivos

### 1. Evaluación de Modelos

Durante la fase de modelado, se implementaron y evaluaron varios modelos de regresión, incluyendo Random Forest, Support Vector Regressor (SVR), XGBoost y LightGBM. Utilicé la métrica de error cuadrático medio (RMSE) para comparar el rendimiento de los modelos. El modelo de XGBoost mostró el mejor rendimiento con un RMSE más bajo en el conjunto de datos de prueba.

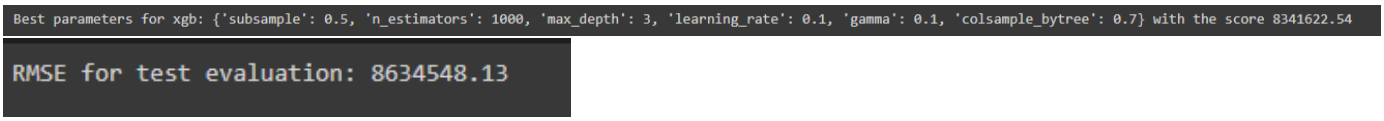



*Figura 11: Valor RMSE. Fuente: Elaboración propia*

### 2. Búsqueda de Hiperparámetros

Se realizó una búsqueda de hiperparámetros utilizando RandomizedSearchCV para optimizar los modelos. La búsqueda de hiperparámetros fue intensiva en términos computacionales, pero permitió identificar configuraciones que mejoraron significativamente el rendimiento de los modelos.

### 3. Predicción del Valor de Futbolistas

Para finalizar, realizaré una comprobación directa del modelo prediciendo el valor de algún futbolista específico. Esto permitirá validar aún más la efectividad del modelo en un contexto real.

## 6.3. Análisis Crítico y Contraste con los Resultados Esperados

### 1. Desviaciones y Justificaciones

Durante el desarrollo del proyecto se identificaron algunas desviaciones respecto al comportamiento inicialmente esperado, principalmente relacionadas con la variabilidad de los resultados obtenidos durante la fase de validación de los modelos. Esta variabilidad se debió a las diferentes particiones del conjunto de datos en los procesos de entrenamiento y prueba.

Para mitigar este efecto, se utilizó el método StratifiedShuffleSplit, que garantiza una distribución equilibrada de la variable objetivo entre los conjuntos de entrenamiento y prueba. Este enfoque permitió reducir la variabilidad entre ejecuciones y obtener resultados más estables y comparables, tal y como se describe en las secciones 3.3 (Métodos de Validación) y 5.7 (Desarrollo y Evaluación de Modelos Predictivos).

Asimismo, la presencia de valores nulos y la heterogeneidad de los datos procedentes de distintas temporadas y fuentes supuso un reto adicional durante la fase de preparación de los datos. Este aspecto fue abordado mediante procesos sistemáticos de limpieza, transformación y codificación, descritos en detalle en la sección 5.5 (Recopilación y Preprocesamiento de Datos), garantizando así la calidad del conjunto de datos final.

## 2. Alcance de los Objetivos

A continuación, se analiza el grado de consecución de cada uno de los objetivos específicos definidos en el apartado 3.2, indicando explícitamente el método de validación empleado y la sección de la memoria donde se documenta su cumplimiento:

- **Objetivo 1: Realizar un análisis exploratorio de los datos**

Este objetivo se considera alcanzado mediante la realización de un análisis exploratorio completo que incluye visualizaciones estadísticas, análisis descriptivo e identificación de patrones relevantes en los datos. La validación de este objetivo se apoya en las visualizaciones y resultados presentados en la sección **6.1 (Resultados del Análisis Exploratorio de Datos)**, así como en el desarrollo detallado del proceso en la sección **5.6 (Análisis Exploratorio de Datos)**. Las figuras incluidas permiten comprobar la correcta exploración de la distribución de variables, relaciones entre características y su impacto potencial sobre el valor de mercado.

- **Objetivo 2: Implementar y comparar diferentes modelos de aprendizaje automático**

Este objetivo se ha cumplido mediante la implementación de varios modelos de regresión (Random Forest, SVR, XGBoost y LightGBM) y su comparación utilizando una métrica común de evaluación. La validación se realiza a través de la comparación de los valores de RMSE obtenidos por cada modelo, documentados en la sección **6.2 (Resultados del Desarrollo y Evaluación de Modelos Predictivos)** y representados gráficamente en la **Figura 11**. El proceso completo de implementación y comparación se describe en la sección **5.7**, lo que permite verificar que el objetivo ha sido alcanzado de forma efectiva.

- **Objetivo 3: Optimizar los hiperparámetros de los modelos**

La optimización de los modelos se llevó a cabo mediante el uso de RandomizedSearchCV, permitiendo evaluar múltiples configuraciones de hiperparámetros de forma eficiente. Este objetivo se valida comparando el rendimiento de los modelos antes y después del proceso de optimización, utilizando la métrica RMSE como indicador cuantitativo. Los detalles de este proceso se encuentran documentados en la sección **5.7 (Búsqueda de Hiperparámetros)**, mientras que su impacto en el rendimiento final se refleja en los resultados presentados en la sección **6.2**.

- **Objetivo 4: Validar el modelo predictivo final**

La validación del modelo final se realizó mediante un conjunto de datos de prueba independiente, utilizando StratifiedShuffleSplit para asegurar una evaluación robusta y representativa. El cumplimiento de este objetivo se verifica mediante la obtención y documentación del valor final de RMSE, que actúa como métrica cuantitativa del error medio de predicción. Este proceso se describe en las secciones **3.3 (Métodos de Validación)**, **5.7 (Evaluación de Modelos)** y **6.2**, donde se justifica la selección del modelo final y su capacidad de generalización.

## 6.4. Evidencias de la Consecución de los Objetivos

Para demostrar la consecución de los objetivos, se incluyen las visualizaciones y los resultados de los modelos predictivos en el documento. Las gráficas y los resultados numéricos presentados evidencian el éxito del proyecto en términos de análisis de datos y predicción del valor de los jugadores.

En resumen, los resultados obtenidos en este proyecto son satisfactorios y cumplen con los objetivos planteados inicialmente. Las desviaciones encontradas fueron abordadas y justificadas adecuadamente, asegurando la calidad y validez del trabajo realizado.

# 7. Implicaciones Éticas e Impacto Social

## 7.1. Introducción

En este apartado, reflexionaré sobre las implicaciones éticas y el impacto social de mi proyecto, tomando como referencia el Código de Ética y Conducta Profesional de la ACM [31], así como otras normativas y leyes aplicables. Abordaré la viabilidad ética del proyecto, los posibles riesgos y su gestión, y el impacto social esperado.

## 7.2. Desarrollo

- **Valor del Proyecto**

Mi proyecto tiene un valor significativo al proporcionar un análisis justo y transparente del rendimiento de los futbolistas. Esto puede tener un impacto positivo en la toma de decisiones por parte de clubes y agentes deportivos, contribuyendo a una mayor equidad y justicia en el mercado de transferencias de jugadores. Al mejorar la precisión en la evaluación del valor de los jugadores, se puede optimizar la inversión en fichajes y la gestión de talento, promoviendo un uso más racional y efectivo de los recursos.

- **Alcance del Proyecto**

El ámbito de mi proyecto se enmarca principalmente en el sector deportivo, afectando directamente a clubes de fútbol, agentes, y jugadores. Indirectamente, puede tener repercusiones en los aficionados y en la industria del deporte en general, al mejorar la transparencia y la eficiencia en el mercado de transferencias. Al hacerlo, se promueve un entorno más justo y equitativo, donde las decisiones se basan en datos objetivos y precisos.

- **Responsabilidad e Impacto del Proyecto**
    - Riesgos Éticos: Uno de los principales riesgos de mi proyecto es la posible sobreestimación o subestimación del valor de los jugadores, lo que podría llevar a decisiones financieras inadecuadas por parte de los clubes y potencialmente perjudicar la carrera de los jugadores. Para mitigar este riesgo, he implementado técnicas avanzadas de validación y sugiero una reevaluación continua del modelo. Además, es crucial asegurar la transparencia en el proceso de evaluación para que todos los actores involucrados comprendan y confíen en los resultados.

- Justicia y Equidad: Es fundamental que el sistema no favorezca injustamente a ciertos jugadores o equipos. Por lo tanto, el modelo debe ser continuamente auditado para detectar y corregir cualquier sesgo. Esto garantiza que todos los jugadores sean evaluados con los mismos criterios, promoviendo la equidad.
- Privacidad y Protección de Datos: Dado que el proyecto maneja datos sensibles de los jugadores, es esencial cumplir con las normativas de protección de datos, como el RGPD [32] y la Ley Orgánica 3/2018 [33]. Esto asegura que los datos personales se manejen de manera ética y segura, respetando la privacidad de los jugadores.

- **Soluciones Propuestas para Minimizar y Controlar los Riesgos**
  - Transparencia y Responsabilidad: La implementación de un sistema transparente y responsable es esencial para minimizar los riesgos. Esto incluye documentar claramente todos los procesos y decisiones, así como permitir la revisión y auditoría externa del modelo y los datos utilizados.
  - Corrección de Sesgos: Para evitar injusticias, el modelo debe ser revisado y actualizado regularmente para detectar y corregir cualquier sesgo. Esto puede lograrse mediante la implementación de técnicas de aprendizaje automático justas y la colaboración con expertos en ética y derechos humanos.
  - Educación y Formación: Es fundamental educar a todos los usuarios del sistema sobre su correcto uso y las implicaciones éticas. Esto incluye la formación en la interpretación de los resultados y la conciencia sobre la importancia de la justicia y la equidad en las evaluaciones.
  - Cumplimiento Legal: Asegurar el cumplimiento de todas las leyes y normativas aplicables, como el RGPD y la Ley Orgánica 3/2018, es crucial para proteger la privacidad y los derechos de los jugadores. Además, seguir las directrices del Código de Ética de la ACM y las recomendaciones de las Directrices Éticas para una IA Confiable [34] garantiza un enfoque ético y responsable.

- **Impacto Más Allá del Proyecto**
  - A Nivel Social: Este proyecto tiene el potencial de democratizar el acceso a análisis avanzados de rendimiento, permitiendo a clubes más pequeños y jugadores sin grandes recursos acceder a herramientas de evaluación de alta calidad. Esto puede ayudar a nivelar el campo de juego y dar a todos los jugadores la oportunidad de ser evaluados de manera justa.
  - A Nivel Económico: Al mejorar la eficiencia en la gestión de fichajes, espero que los clubes puedan realizar inversiones más informadas, lo que podría llevar a un uso más racional y efectivo de los recursos financieros en el deporte. Esto puede contribuir a una economía deportiva más sostenible y justa.
  - A Nivel Medioambiental: Indirectamente, la optimización en la gestión de recursos puede reducir el desperdicio asociado con fichajes fallidos y la

consecuente rotación de jugadores. Esto puede tener un impacto positivo en la sostenibilidad ambiental del deporte.

## 7.3. CONCLUSIONES

En relación con todo lo expuesto, dado el valor que presenta este proyecto y que los riesgos detectados pueden evitarse o están controlados, se puede concluir que el proyecto no solamente es viable desde el punto de vista ético, sino que además es recomendable llevarlo a cabo. Este proyecto contribuye de manera positiva tanto a nivel social como económico, promoviendo un uso más eficiente y justo de los recursos en el fútbol profesional.

# 8. MI RECORRIDO EN LA UFV

## 8.1. EL PFG COMO CULMINACIÓN DE MI CAMINO UNIVERSITARIO

Al reflexionar sobre el recorrido del autor en la Universidad Francisco de Vitoria (UFV), es inevitable comparar cómo era cuando llegó a la universidad y cómo es ahora, después de varios años de aprendizaje y crecimiento personal y académico.

- **¿Cómo era cuando llegué a la Universidad?**

Cuando llegué a la UFV, venía de una experiencia no muy buena en la Universidad Politécnica, tanto en lo académico como en lo social. Por lo tanto, llegué con la ilusión de que esta vez la experiencia fuera mejor. Tenía la esperanza de encontrar una carrera que realmente me gustara y un ambiente en el que me sintiera a gusto. Mis objetivos eran claros: encontrar algo que me apasionara al 100% para poder dedicarme a ello con entusiasmo y disfrutarlo plenamente.

- **¿Cómo fue mi camino universitario?**

Al inicio, estaba bastante tranquilo y, debo admitir, con pocas ganas. Además, el primer año coincidió con la pandemia de COVID-19, lo que afectó significativamente la dinámica de la vida universitaria para todos. Sin embargo, a partir del tercer año, experimenté un cambio radical. Me volví mucho más dedicado y enfocado en mis estudios, con el claro objetivo de aprobar todas mis materias y avanzar hacia mi futuro profesional. Aunque conté con el apoyo constante de mi familia y mi novia, no hubo una persona en particular que provocara este cambio. Creo que fue el resultado de madurar y encontrar un propósito claro y definido para mi futuro.

Es importante destacar el papel de los compañeros del autor en este proceso. En especial, Álvaro y Diego han sido fundamentales en mi camino. Su apoyo y compañía han sido vitales para mantenerme motivado y enfocado, y juntos hemos superado muchos desafíos a lo largo de nuestra carrera universitaria.

- **¿Cómo me veo ahora?**

Actualmente, me considero una persona más madura y consciente de sus responsabilidades y objetivos. Estoy completamente centrado en mi meta personal y trabajo arduamente para alcanzarla. La transformación que he experimentado es evidente no solo en mi enfoque académico, sino también en mi perspectiva de la vida.

- **Motivación para la realización de este PFG**

Mi pasión y amor por el fútbol y el deporte en general fueron factores determinantes en la elección de este Proyecto de Fin de Grado (PFG). Siempre supe que quería hacer algo relacionado con el deporte. Después de realizar muchas búsquedas de investigación y

hablar con familia y amigos, surgió la posibilidad de este proyecto, y me pareció perfecto para combinar intereses personales con estudios.

- **Camino de conocimiento personal**

A lo largo de estos años, no me he formulado preguntas específicas, pero he experimentado un proceso natural de crecimiento personal. He aprendido a enfrentar desafíos, a encontrar un camino y un objetivo, y a comprometerme con ellos para alcanzarlos. Esta experiencia universitaria me ha preparado no solo académicamente, sino también en términos de desarrollo personal y profesional.

## 8.2. VINCULACIÓN CON MI FUTURO PROFESIONAL

Al reflexionar sobre el sentido que ha tenido para mí la realización de este Proyecto de Fin de Grado (PFG) en relación con su futuro profesional, puedo decir que ha sido una experiencia fundamental y reveladora en varios aspectos.

Este proyecto refleja claramente la pasión del autor por el fútbol y su interés en la tecnología y la ciencia de datos. Es coherente con lo que busco en la vida porque combina dos áreas que me apasionan profundamente: el deporte y la tecnología. Desde siempre, he sido un entusiasta del fútbol, y encontrar una forma de integrar este interés con mi carrera profesional ha sido una meta personal. La realización de este proyecto me ha permitido explorar y aplicar conocimientos técnicos en un contexto deportivo, lo cual es exactamente lo que quiero seguir haciendo en mi futuro profesional.

La realización de este PFG me ha llevado a preguntarme sobre cómo puede utilizar mis habilidades y conocimientos para generar un impacto positivo en el mundo del deporte. He descubierto que tengo la capacidad de aplicar la ciencia de datos y la tecnología para mejorar aspectos del rendimiento deportivo y la gestión de equipos. Este descubrimiento se produjo a través del proceso de investigación y análisis que implicó el proyecto, donde pude ver de primera mano cómo los datos pueden transformar la toma de decisiones en el fútbol.

Después de estos años de formación universitaria, se han abierto numerosas perspectivas para mi futuro profesional. Tengo una sólida base en ciencias de la computación y análisis de datos, y estoy preparado para continuar mi educación en esta área. Además, he obtenido una beca de fútbol para estudiar un máster en Computer Science en Estados Unidos, lo cual representa una oportunidad increíble para seguir desarrollándome tanto académica como deportivamente. Esta experiencia me permitirá no solo adquirir conocimientos avanzados, sino también ampliar mi red de contactos y explorar oportunidades laborales en el extranjero.

A partir de ahora, quiero aprovechar la oportunidad de estudiar y jugar al fútbol en Estados Unidos para completar mi formación con un máster en Computer Science. Después, mi objetivo es trabajar en Estados Unidos o Europa durante unos años, adquiriendo experiencia en el campo de la tecnología. Eventualmente, planeo regresar a España y aplicar todo lo aprendido para contribuir al desarrollo del deporte y la tecnología en mi país. Este PFG me ha dado una base sólida y una dirección clara para mi futuro profesional, y estoy convencido

de que me ayudará a alcanzar mis objetivos al haberme permitido combinar mis pasiones y habilidades en un proyecto tangible y relevante.

En conclusión, este Proyecto de Fin de Grado no solo ha sido una culminación de mi camino universitario, sino también un puente hacia mi futuro profesional. Me ha permitido descubrir mi verdadera vocación y preparar el terreno para los próximos pasos en mi carrera, combinando mi amor por el fútbol con la pasión por la tecnología.

# 9. CONCLUSIONES

## 9.1. PRINCIPALES CONCLUSIONES DEL PROYECTO

El objetivo principal de este Proyecto de Fin de Grado ha sido desarrollar un sistema capaz de estimar el valor de mercado de futbolistas profesionales mediante técnicas de aprendizaje automático, utilizando datos reales procedentes de distintas temporadas y fuentes especializadas. A lo largo del desarrollo del proyecto, se han alcanzado de forma satisfactoria tanto los objetivos generales como los objetivos específicos definidos inicialmente.

- En primer lugar, se llevó a cabo un **análisis exploratorio de los datos (EDA)** con el fin de comprender la estructura del conjunto de datos y analizar la relación entre las variables explicativas y la variable objetivo. Este objetivo se considera cumplido mediante la generación de visualizaciones estadísticas, análisis descriptivos y estudios de correlación, presentados en el apartado correspondiente de la memoria. Dicho análisis permitió identificar patrones relevantes, como la influencia de la posición, la liga, el rendimiento ofensivo o el pie dominante en el valor de mercado de los jugadores, y justificó la selección de las variables empleadas en el modelado posterior.
- En segundo lugar, **se implementaron y compararon diferentes modelos de aprendizaje automático para regresión**, concretamente Random Forest, Support Vector Regressor (SVR), XGBoost y LightGBM. La validación de este objetivo se realizó mediante la correcta implementación de los modelos y su evaluación objetiva utilizando la métrica RMSE, tal y como se describe en el apartado de desarrollo y evaluación de modelos. La comparación de resultados permitió analizar el comportamiento de cada algoritmo y seleccionar aquellos con mejor capacidad predictiva y de generalización.
- Asimismo, se llevó a cabo un **proceso de optimización de hiperparámetros** utilizando la técnica RandomizedSearchCV. Este objetivo se considera alcanzado al observar una mejora cuantificable del rendimiento de los modelos tras la optimización, reflejada en la reducción del valor del RMSE respecto a las configuraciones iniciales. Este proceso permitió ajustar parámetros clave como la profundidad de los árboles, el número de estimadores o la tasa de aprendizaje, maximizando así la eficiencia de los modelos.
- Por último, se realizó la **validación del modelo predictivo** final, empleando conjuntos de entrenamiento y prueba definidos mediante StratifiedShuffleSplit para garantizar una distribución equilibrada de los datos. La obtención y documentación del RMSE final permitió evaluar de forma objetiva la capacidad de generalización del modelo seleccionado. Como resultado, se construyó un pipeline reproducible que integra todas las fases de preprocesamiento y el

modelo óptimo, asegurando la coherencia del proceso y facilitando su posible reutilización o ampliación futura.

En conjunto, los resultados obtenidos demuestran que el enfoque seguido es válido y eficaz para abordar el problema planteado, confirmando que el uso de técnicas de machine learning, junto con una metodología estructurada y una validación adecuada, permite obtener estimaciones razonables del valor de mercado de futbolistas a partir de datos históricos reales.

Más allá de los resultados técnicos, la realización de este Proyecto de Fin de Grado ha supuesto una experiencia especialmente enriquecedora a nivel personal y profesional. A través de este trabajo he podido combinar dos ámbitos de gran interés para mí, como son el fútbol y la tecnología, aplicando conocimientos de ciencia de datos y aprendizaje automático a un contexto real del deporte profesional.

Durante el desarrollo del proyecto, he adquirido una mayor comprensión de la importancia de seguir procesos estructurados, de validar correctamente los modelos y de interpretar los resultados de forma crítica. Asimismo, el trabajo con datos reales ha puesto de manifiesto la relevancia del preprocesamiento y del análisis exploratorio, aspectos que resultan determinantes en la calidad final de cualquier sistema predictivo.

Este proyecto ha contribuido de manera significativa al desarrollo de mis competencias técnicas, así como a la mejora de habilidades relacionadas con la planificación, la resolución de problemas y la documentación de proyectos complejos. Todo ello ha supuesto un paso importante en mi madurez académica y profesional, reforzando mi interés por el ámbito del análisis de datos y el machine learning aplicado.

## 9.2. Posibilidades de Evolución Futura

El trabajo presentado en este proyecto tiene un gran potencial para evolucionar y expandirse en diversas direcciones. Algunas de las posibilidades futuras incluyen:

- **Mejora y Actualización del Modelo:**

Continuar refinando el modelo con datos más recientes y avanzados permitirá mejorar su precisión y relevancia. La incorporación de nuevas variables y técnicas de análisis puede llevar a una mejor valoración del rendimiento de los jugadores.

- **Aplicación en Otros Deportes:**

Aunque este proyecto se ha centrado en el fútbol, las técnicas y metodologías desarrolladas pueden aplicarse a otros deportes. La expansión del modelo a diferentes disciplinas deportivas podría proporcionar herramientas valiosas para una amplia gama de equipos y atletas.

- **Desarrollo de Herramientas Interactivas:**

Crear plataformas interactivas y aplicaciones basadas en los modelos desarrollados permitirá a los usuarios (clubes, agentes, jugadores) interactuar directamente con los datos y obtener

insights personalizados. Esto puede incluir dashboards interactivos, aplicaciones móviles y otras herramientas digitales.

- **Colaboraciones y Proyectos Multidisciplinarios:**

Colaborar con otros profesionales y académicos en áreas como la ciencia del deporte, la psicología y la economía deportiva puede enriquecer el proyecto y abrir nuevas oportunidades de investigación y aplicación práctica.

- **Exploración de Inteligencia Artificial y Machine Learning Avanzado:**

La integración de técnicas más avanzadas de inteligencia artificial, como el aprendizaje profundo (deep learning) y la inteligencia artificial explicable (XAI), puede llevar a una comprensión más profunda y precisa del rendimiento deportivo y sus determinantes.

# 10. OTROS MÉRITOS DEL PROYECTO

Durante el desarrollo de este Proyecto de Fin de Grado, se han conseguido una serie de logros adicionales que aportan un valor significativo al trabajo realizado. Estos méritos no solo mejoran la calidad del proyecto, sino que también amplían su impacto y relevancia en varios aspectos.

- **Resultados Obtenidos No Esperados**
  - Valoración y Clasificación de Jugadores:

    Además de desarrollar un modelo predictivo para la valoración económica de los futbolistas, se ha logrado clasificar a los jugadores según diversas métricas de rendimiento.

  - Insights Estratégicos:

    El análisis de datos ha revelado patrones y tendencias en el rendimiento de los jugadores que no eran inicialmente esperados. Por ejemplo, se han identificado factores clave que influyen en la valorización de los jugadores.

- **Uso de Software Libre**

El proyecto ha sido desarrollado utilizando software libre, como Python, pandas, scikit-learn, y Plotly. El uso de estas herramientas no solo reduce los costos asociados con el desarrollo del proyecto, sino que también promueve el acceso abierto y la capacidad de otros para utilizar y mejorar el código y los métodos desarrollados.

- **Integración de Disciplinas**

He integrado conocimientos y técnicas de diversas disciplinas, incluyendo ciencia de datos, estadística, informática y deporte. Esta interdisciplinariedad ha enriquecido el proyecto y ha permitido abordar el problema desde múltiples perspectivas, ofreciendo una solución más robusta y completa.

- **Elementos de Accesibilidad**

El proyecto incluye elementos que mejoran su accesibilidad y usabilidad. Por ejemplo, los gráficos y visualizaciones han sido diseñados para ser claros y comprensibles, utilizando colores y etiquetas que facilitan su interpretación por parte de diferentes usuarios, incluidos aquellos con discapacidades visuales.

- **Colaboración y Validación**

Aunque no he contado con colaboradores externos específicos para el desarrollo técnico del proyecto, el apoyo y las discusiones con mis compañeros Álvaro y Diego, así como con mi familia y amigos, han sido valiosos para la validación de ideas y la orientación general del proyecto. Este entorno de colaboración ha enriquecido mi enfoque y ha permitido un desarrollo más equilibrado y crítico.

# 11. BIBLIOGRAFÍA


[1] «Transfermarket,» [En línea]. Available: https://www.transfermarkt.es/.

[2] «FBREF,» [En línea]. Available: https://fbref.com/en/.

[3] RSKriegs, «Kaggle,» Student at SGH Warsaw School of Economics, [En línea]. Available: https://www.kaggle.com/datasets/kriegsmaschine/soccer-players-values-and-their-statistics?select=transfermarkt_fbref_201920.csv.

[4] R. Stępień, «Github,» 2021. [En línea]. Available: https://github.com/RSKriegs/Modelling-Football-Players-Values-on-a-Transfer-Market/blob/main/RS82640%20Modelling%20Footballers%20Values%20on%20a%20Transfer%20Market%20.pdf.

[5] G. G. P. J. P. K. Bhaskar Mukhoty, 2019. [En línea]. Available: https://arxiv.org/abs/2006.14211.

[6] 2025. [En línea]. Available: https://xgboost.readthedocs.io/en/stable/.

[7] 2026. [En línea]. Available: https://lightgbm.readthedocs.io/en/latest/#.

[8] M. Ghasri, 2026. [En línea]. Available: https://es.mathworks.com/matlabcentral/fileexchange/153391-support-vector-regression-svr.

[9] 2025. [En línea]. Available: https://www.inesdi.com/blog/random-forest-que-es/.

[10] «Scikit Learn,» 2025. [En línea]. Available: https://scikit-learn.org/stable/modules/cross_validation.html.

[11] G. (. Laros, «Predicting Market Value of Football Players Using Machine Learning,» Tilburg University, 2022. [En línea]. Available: https://arno.uvt.nl/show.cgi?fid=161188.

[12] «Scikit Learn,» 2025. [En línea]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html.

[13] J. R. Quinlan, «Induction of decision trees,» 1986. [En línea]. Available: https://link.springer.com/article/10.1007/BF00116251.

[14] J. M. P.-C. ,. J. A. G. M. D. S. C. A. C. F. S. P. S. d. B. F. A. Francisco Javier Robles-Palazón, «Predicting injury risk using machine learning in male youth soccer players,» Chaos, Solitons & Fractals, 2023. [En línea]. Available: https://www.sciencedirect.com/science/article/pii/S0960077922012589.

[15] F. T. Rory P. Bunker, «A machine learning framework for sport result prediction,» Applied Computing and Informatics, 2019. [En línea]. Available: https://www.sciencedirect.com/science/article/pii/S2210832717301485.

[16] E. P. C. M.-H. H. L.-C. H. Jack C Yue, «A study of forecasting tennis matches via the Glicko model,» PubMed, 2022. [En línea]. Available: https://pubmed.ncbi.nlm.nih.gov/35395047/.

[17] H. K. Rahul Baboota, «Predictive analysis and modelling football results using machine learning approach for English Premier League,» International Journal of Forecasting, 2019. [En línea]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0169207018300116.

[18] V. Roman, «Proyecto Machine Learning: Predicción de Precios de Viviendas en Boston con Regresión,» 2019. [En línea]. Available: https://medium.com/datos-y-ciencia/proyecto-machine-learning-predicci%C3%B3n-de-precios-de-viviendas-en-boston-con-regresi%C3%B3n-e8655e6c3655.

[19] Q. Shen, «Predicting the value of football players: machine learning techniques and sensitivity analysis based on FIFA and real-world statistical datasets,» 2025. [En línea]. Available: https://link.springer.com/article/10.1007/s10489-024-06189-0.

[20] J. S. A. G. Danny F. Hill, «A review of football player metrics and valuation methods: a typological framework of football player valuations,» 2025. [En línea]. Available: https://www.tandfonline.com/doi/full/10.1080/23750472.2025.2459727.

[21] [En línea]. Available: https://scikit-learn.org/stable/.

[22] «Seaborn,» [En línea]. Available: https://seaborn.pydata.org/index.html.

[23] «Matplotlib,» [En línea]. Available: https://matplotlib.org/stable/.

[24] [En línea]. Available: https://plotly.com/python/.

[25] «Ilerna,» 2024. [En línea]. Available: https://www.ilerna.es/blog/aprende-con-ilerna-online/comercio-marketing/definir-objetivos-con-la-regla-smart/.

[26] J. C. (. R. K. (. T. K. (. T. R. (. C. S. (. a. R. W. (. Pete Chapman (NCR), 2000. [En línea]. Available: https://mineracaodedados.files.wordpress.com/2012/12/crisp-dm-1-0.pdf.

[27] [En línea]. Available: https://www.python.org/.

[28] «Joblib: running Python functions as pipeline jobs,» [En línea]. Available: https://joblib.readthedocs.io/en/stable/.

[29] [En línea]. Available: https://colab.research.google.com/.

[30] [En línea]. Available: https://github.com/.

[31] 2018. [En línea]. Available: https://www.acm.org/code-of-ethics.

[32] E. P. E. Y. E. C. D. L. U. EUROPEA, EL PARLAMENTO EUROPEO Y EL CONSEJO DE LA UNIÓN EUROPEA, 2016. [En línea]. Available: https://www.boe.es/doue/2016/119/L00001-00088.pdf.

[33] J. d. Estado, 2018. [En línea]. Available: https://www.boe.es/buscar/act.php?id=BOE-A-2018-16673.

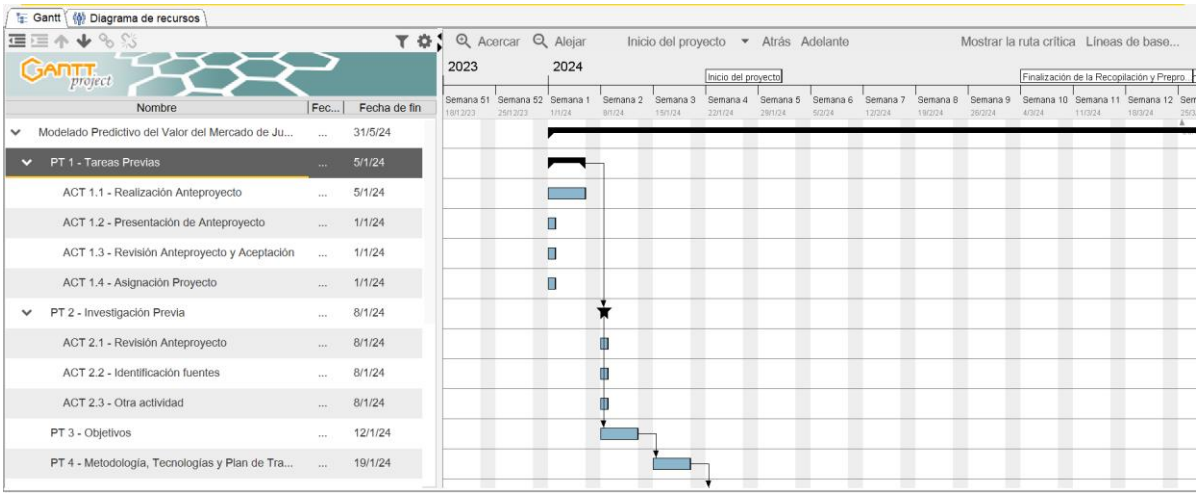
[34] 2024. [En línea]. Available: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

# ANEXO A: DIAGRAMA DE GANTT DETALLADO DEL PROYECTO

En este anexo se presenta el diagrama de Gantt completo y detallado, generado con la herramienta GanttProject, que recoge el desglose exhaustivo de tareas (ACT) asociadas a cada paquete de trabajo (PT) del proyecto.

El objetivo de este anexo es complementar la información mostrada en el apartado 4.4 de la memoria, donde se presenta el Gantt a nivel de paquetes de trabajo, proporcionando aquí una visión más granular del cronograma, con fechas concretas, dependencias e hitos intermedios.

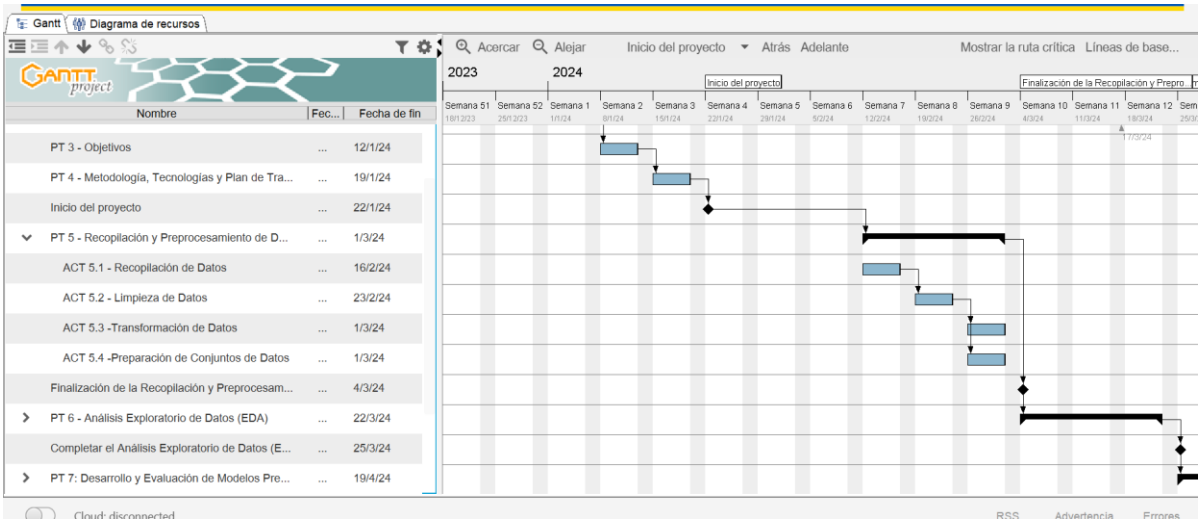
1. Detalle del diagrama de Gantt correspondiente al PT 1, PT 2, PT 3 y PT 4.




Esta primera captura muestra las **fases iniciales del proyecto**, correspondientes a los paquetes de trabajo **PT1 (Tareas Previas)**, **PT2 (Investigación Previa)**, **PT3 (Objetivos)** y **PT4 (Metodología, Tecnologías y Plan de Trabajo)**.

En esta etapa se incluyen actividades como la realización y presentación del anteproyecto, su revisión y aceptación, la identificación de fuentes de datos, así como la definición de los objetivos y del marco metodológico. Estas tareas permiten establecer las bases académicas y técnicas del proyecto antes de iniciar el desarrollo propiamente dicho.

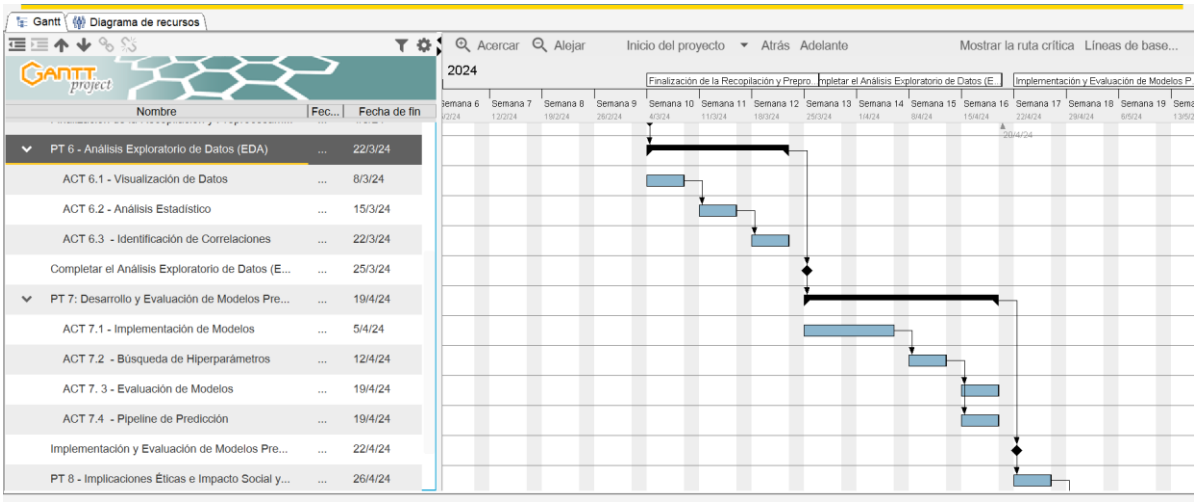
2. Detalle del diagrama de Gantt correspondiente al PT 5




La segunda captura corresponde al **PT5 – Recopilación y Preprocesamiento de Datos**, donde se detallan las actividades relacionadas con la obtención y preparación del dataset.

En esta fase se visualizan las tareas de recopilación de datos desde las fuentes seleccionadas, limpieza de valores nulos e inconsistencias, transformación de variables y preparación de los conjuntos de entrenamiento y prueba. Asimismo, se reflejan las dependencias entre tareas y el hito de finalización del preprocesamiento.

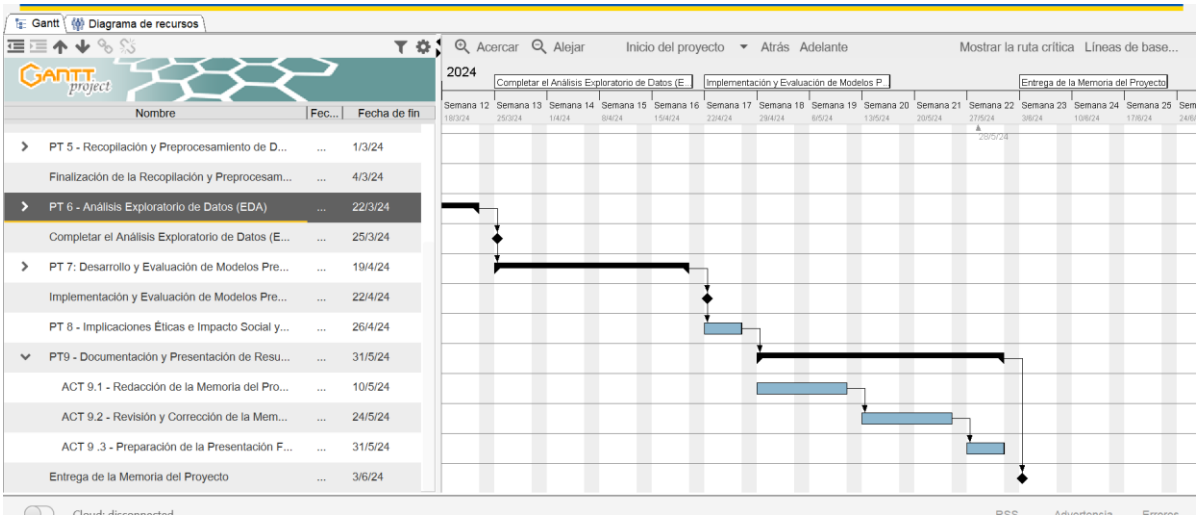
3. Detalle del diagrama de Gantt correspondiente al PT 6 y PT 7.




En esta figura se muestran los paquetes de trabajo **PT6 – Análisis Exploratorio de Datos (EDA)** y **PT7 – Desarrollo y Evaluación de Modelos Predictivos**.

Por un lado, se incluyen las actividades de visualización de datos, análisis estadístico e identificación de correlaciones. Por otro, se presentan las tareas asociadas a la implementación de modelos, la búsqueda de hiperparámetros, la evaluación mediante RMSE y la construcción del pipeline final. Esta fase constituye el núcleo técnico del proyecto.

4. Detalle del diagrama de Gantt correspondiente al PT 8 y PT 9.




La última captura recoge las fases finales del proyecto, correspondientes al **PT8 – Implicaciones Éticas e Impacto Social** y al **PT9 – Documentación y Presentación de Resultados**.

En esta etapa se incluyen las tareas de redacción de la memoria, revisión y corrección del documento final, preparación de la presentación y el hito de **entrega final del Proyecto de Fin de Grado**. Estas actividades aseguran el cierre correcto del proyecto tanto a nivel técnico como académico.