# Proyecto BigData: Predicción del Valor de Mercado de Futbolistas

**Autores:**
Esteban Bernal - Cód. 271930
Juan Montes - Cód. 272113
Juan Gomez - Cód. 286774

**Asignatura:**
Herramientas de Big Data (Coterminal)

**Institución:**
Universidad de La Sabana
Facultad de Ingeniería

**Programa:**
Maestría en Analítica Aplicada

**Profesor:**
Hugo Franco, Ph.D.

**Fecha de Entrega:**
Septiembre 5, 2025

# Índice

# 1. Resumen ejecutivo (Brief)

En el contexto de la creciente profesionalización y mercantilización del fútbol, la valoración precisa de los jugadores se ha convertido en un pilar estratégico para clubes y agencias. Sin embargo, las valoraciones de mercado actuales, dominadas por plataformas como Transfermarkt, a menudo carecen de transparencia metodológica. Este proyecto aborda dicha problemática mediante el diseño y la implementación de un pipeline de datos ETL (Extract, Transform, Load) completo y reproducible para predecir el valor de mercado de futbolistas. Se integran datos de rendimiento avanzados extraídos mediante web scraping de *FBref.com* con datos de mercado históricos de un compendio de Transfermarkt disponible en Kaggle. El sistema, desarrollado íntegramente en Python, automatiza la recolección, limpieza, fusión y preprocesamiento de datos de más de cinco temporadas de las principales ligas y competiciones europeas. Finalmente, se entrena y evalúa un modelo de Machine Learning (XGBoost) que, con un **$R^2$ de 0.854**, demuestra una excelente capacidad predictiva. El análisis valida la hipótesis de que el rendimiento en el campo, especialmente las contribuciones a gol y los goles esperados, son los factores más determinantes del valor económico de un jugador.

# 2. Abstract

In the context of football's increasing professionalization and commodification, accurate player valuation has become a strategic cornerstone for clubs and agencies. However, current market valuations, dominated by platforms like Transfermarkt, often lack methodological transparency. This project addresses this issue by designing and implementing a complete and reproducible ETL (Extract, Transform, Load) data pipeline to predict the market value of football players. Advanced performance data extracted via web scraping from *FBref.com* is integrated with historical market data from a Transfermarkt compendium available on Kaggle. The system, developed entirely in Python, automates the collection, cleaning, merging, and preprocessing of data spanning over five seasons of major European leagues and competitions. Finally, a Machine Learning model (XGBoost) is trained and evaluated, achieving an **$R^2$ score of 0.854** which demonstrates strong predictive capabilities. The analysis validates the hypothesis that on-field performance, particularly goal contributions and expected goals, are the key determinants of a player's economic value.

# 3. Introducción

## 3.1. Formulación del problema y las necesidades de información asociadas

El mercado de fichajes del fútbol profesional moviliza miles de millones de euros cada año, convirtiendo la valoración de jugadores en una tarea de alto impacto financiero. A pesar de su importancia, el proceso de valoración es notablemente opaco. Plataformas como *Transfermarkt.com* se han erigido como el estándar de facto, pero sus estimaciones se basan en una combinación de análisis de expertos y consensos de la comunidad, sin una metodología cuantitativa, pública y reproducible. Esta falta de transparencia genera incertidumbre para los stakeholders (clubes, agentes, inversores) y crea una oportunidad para el desarrollo de modelos de valoración alternativos basados en datos empíricos.

La necesidad de información es clara: se requiere un modelo que pueda estimar el valor de mercado de un futbolista basándose en su rendimiento objetivo en el campo. Para construir dicho modelo, es indispensable superar el desafío técnico de integrar múltiples fuentes de datos heterogéneas. Por un lado, se necesitan métricas de rendimiento detalladas y, por otro, datos históricos de valoraciones de mercado. El problema central que este proyecto resuelve es el diseño y la construcción de un pipeline de datos robusto que automatice este complejo proceso de integración, creando un dataset de alta calidad listo para ser explotado por algoritmos de Machine Learning.

## 3.2. Marco conceptual (definiciones y conceptos fundamentales)

Este proyecto se fundamenta en un sólido marco conceptual que combina técnicas de ingeniería y ciencia de datos, en línea con los temas abordados en el curso:

- **Data Pipeline ETL (Extract, Transform, Load):** Siguiendo el modelo de la Sesión 4, el proyecto implementa un pipeline ETL clásico. La **Extracción** se realiza mediante web scraping (*FBref*) y carga de archivos (*Kaggle*). La **Transformación** es la fase más crítica, donde se limpian, unifican y enriquecen los datos. La **Carga** consiste en la persistencia del dataset final procesado en un archivo `.csv`, que actúa como nuestro "data mart"para el análisis.

- **Web Scraping:** Técnica utilizada para la extracción de datos de *FBref.com*. Se emplean librerías de Python como `requests` y `BeautifulSoup` para navegar la estructura HTML del sitio y extraer las tablas de estadísticas de manera sistemática.

- **Métricas Avanzadas en Fútbol:** Más allá de goles y asistencias, se utilizan métricas que ofrecen una visión más profunda del rendimiento, como los Goles Esperados (xG) y las Asistencias Esperadas (xA), que miden la calidad de las oportunidades en lugar de solo el resultado final.

- **Análisis Exploratorio de Datos (EDA):** Se aplican técnicas de estadística descriptiva y visualización (histogramas, gráficos de dispersión, mapas de calor de correlación) para entender la distribución de las variables, identificar relaciones y detectar outliers, tal como se exploró en la Sesión 4.

- **Machine Learning (Regresión Supervisada):** Se utiliza un algoritmo de Gradient Boosting (XGBoost) para predecir una variable continua (el valor de mercado) a partir de un conjunto de variables predictoras. El modelo se entrena y evalúa utilizando métricas estándar como MAE, RMSE y $R^2$.

## 3.3. Antecedentes (trabajos previos y relacionados)

La aplicación de la analítica de datos al fútbol, popularizada por el libro y la película "Moneyball."[en] el béisbol, ha ganado un impulso significativo. Investigaciones académicas y proyectos independientes han explorado la relación entre el rendimiento y el salario o valor de mercado. Sin embargo, muchos de estos trabajos se centran en una única liga o utilizan datos limitados.

Este proyecto se distingue por su escala y su enfoque en la reproducibilidad de extremo a extremo. Al construir un pipeline que extrae datos de múltiples competiciones (las 5 grandes ligas, Champions League, etc.) y múltiples temporadas, se crea un dataset de una riqueza y granularidad excepcionales. El énfasis en un pipeline modular y completamente programado en scripts de Python no solo produce un resultado, sino un **proceso** robusto que puede ser re-ejecutado y extendido en el futuro, abordando una de las principales críticas a los análisis de datos ad-hoc.

## 3.4. Objetivo del proyecto

Desarrollar un pipeline de datos ETL automatizado y reproducible en Python para construir un dataset unificado que integre métricas avanzadas de rendimiento y datos de mercado, con el fin de entrenar y evaluar un modelo de Machine Learning capaz de predecir el valor de mercado de futbolistas profesionales, identificando los factores de rendimiento más influyentes.

# 4. Datos empleados

## 4.1. Descripción detallada de las fuentes de datos

El proyecto se basa en la sinergia de dos fuentes de datos complementarias:

1. **FBref.com (vía Web Scraping):**
   - **Descripción:** Es el proveedor de datos de rendimiento. Utiliza estadísticas de StatsBomb, una de las fuentes más reputadas en el análisis de fútbol. Se extrajeron múltiples tablas por jugador y temporada.
   - **Cobertura:** Ligas "Big 5"de Europa (Premier League, LaLiga, Serie A, Bundesliga, Ligue 1), UEFA Champions League, Europa League y Conference League, para las temporadas desde 2017-18 hasta 2022-23.
   - **Categorías de datos:** Estadísticas estándar, de tiro, de pases, de creación de juego, defensivas y específicas de porteros.
2. **Transfermarkt (vía Kaggle Dataset):**
   - **Descripción:** Es el proveedor de los datos económicos. Se utilizó un dataset precompilado de Kaggle que contiene información histórica de Transfermarkt.
   - **Variables clave:** Valor de mercado histórico (`market_value_in_eur`), datos demográficos (edad, nacionalidad) e información contractual (club, posición).

## 4.2. Contenido y variables del dataset integrado

El pipeline de datos consolida más de 150 variables. A continuación, se detallan las más relevantes para el modelado:

| Variable | Tipo | Descripción |
|---|---|---|
| `market_value_in_eur` | Cuantitativa Continua | **Variable Objetivo**. Valor de mercado. |
| `age` | Cuantitativa Continua | Edad del jugador en la temporada. |
| `position` | Cualitativa Nominal | Posición principal del jugador. |
| `minutes_90s` | Cuantitativa Continua | Número de partidos de 90 minutos jugados. |
| `goals` | Cuantitativa Discreta | Goles anotados. |
| `assists` | Cuantitativa Discreta | Asistencias realizadas. |
| `xg` | Cuantitativa Continua | Goles Esperados. |
| `npxg` | Cuantitativa Continua | Goles Esperados sin contar penales. |
| `xa` | Cuantitativa Continua | Asistencias Esperadas. |
| `sca` | Cuantitativa Continua | Acciones creadoras de tiros. |
| `gca` | Cuantitativa Continua | Acciones creadoras de goles. |
| `tackles` | Cuantitativa Discreta | Entradas defensivas realizadas. |

Cuadro 1: Diccionario de datos de las variables más relevantes.

# 5. Materiales y Métodos

## 5.1. Arquitectura del Pipeline de Datos (ETL)

El proyecto se articula en torno a un pipeline de datos modular, donde cada script de Python cumple una función específica dentro del flujo ETL. Esta arquitectura garantiza la mantenibilidad y la claridad del código. A continuación se describe el flujo de datos:

```
Fuentes de Datos Brutas
  |
  |-- Kaggle (Transfermarkt Data)
  |-- FBref.com (StatsBomb Data)
  |
  V
[ETAPA 1: EXTRACCION]
  |
  |-- script: fbref_scraper_all.py -> [Datos de rendimiento en /data/raw]
  |-- carga manual -> [Datos de mercado en /data/raw]
  |
  V
[ETAPA 2: TRANSFORMACION]
  |
  |-- script: clean_transfermark.py -> [Datos de mercado limpios]
  |-- script: merge_quick.py -> [Dataset de rendimiento unificado]
  |      |
  |      V
  |-- script: merge_transfermarkt.py -> [Dataset integrado]
  |      |
  |      V
  |-- script: prepare_dataset.py -> [Dataset final preprocesado]
  |
  V
[ETAPA 3: CARGA]
  |
  |-- Archivo: /data/processed/dataset_ready.csv
  |
  V
[FASE DE ANALISIS]
  |
  |-- notebook: 01_EDA.ipynb (EDA y Modelado)
  |
  V
Resultados y Conclusiones
```

## 5.2. Análisis Exploratorio de Datos (EDA)

Realizado en el notebook `01_EDA.ipynb`, el EDA se centró en:

- **Análisis de Distribución:** Se generó un histograma del valor de mercado, confirmando su fuerte sesgo a la derecha, lo que justifica la transformación logarítmica.

- **Análisis de Correlación:** Se calculó y visualizó una matriz de correlación de Pearson para identificar las variables de rendimiento con mayor asociación lineal con el (logaritmo del) valor de mercado.
- **Análisis de Outliers:** Se identificaron jugadores como Mbappé y Haaland como outliers de alto valor, validando que los datos reflejan la realidad del mercado.

## 5.3. Modelado de Machine Learning

- **Selección de Características:** Se seleccionó un subconjunto de las variables más relevantes y con menor colinealidad, priorizando métricas avanzadas (xG, xA, SCA) y datos demográficos (edad, minutos jugados).
- **Algoritmo:** Se eligió **XGBoost**, un algoritmo de Gradient Boosting conocido por su alto rendimiento en datos tabulares y su robustez frente al sobreajuste.
- **Entrenamiento y Validación:** El dataset se dividió en conjuntos de entrenamiento (80 %) y prueba (20 %). El modelo se entrenó en el conjunto de entrenamiento.
- **Métricas de Evaluación:** El rendimiento del modelo se evaluó en el conjunto de prueba utilizando el Error Absoluto Medio (MAE), la Raíz del Error Cuadrático Medio (RMSE) y el Coeficiente de Determinación ($R^2$).

# 6. Resultados

## 6.1. Resultados del Pipeline de Datos

El resultado principal de la fase de ingeniería es un pipeline ETL completamente funcional y reproducible. El sistema es capaz de generar, a partir de las fuentes brutas, un dataset analítico final de **más de 25,000 registros únicos (jugador-temporada)** y **158 variables**, constituyendo un activo de datos robusto y de alto valor para la investigación en analítica deportiva.

## 6.2. Resultados del Análisis Exploratorio de Datos

El análisis de correlación fue particularmente revelador. Se encontró una fuerte asociación positiva entre el logaritmo del valor de mercado y las métricas de contribución ofensiva. Las variables con mayor coeficiente de correlación de Pearson fueron:

- Contribuciones a gol por 90 min (`goal_contributions_per_90`)
- Goles esperados sin penales por 90 min (`non_penalty_xg_per_90`)
- Tiros a puerta por 90 min (`shots_on_target_per_90`)

En contraste, las métricas defensivas mostraron correlaciones cercanas a cero o ligeramente negativas, validando la hipótesis de que el mercado prioriza el impacto ofensivo.

## 6.3. Resultados del Modelo Predictivo

El modelo XGBoost entrenado demostró una notable capacidad para predecir el valor de mercado de los jugadores. La evaluación sobre el conjunto de prueba arrojó los siguientes resultados concretos:

| **Métrica de Evaluación** | **Valor** |
|---|---|
| Coeficiente de Determinación ($R^2$) | 0,85 |
| Error Absoluto Medio (MAE) | 3,627,133,45 |
| Raíz del Error Cuadrático Medio (RMSE) | 7,949,603,95 |

Cuadro 2: Métricas de rendimiento del modelo XGBoost en el conjunto de prueba.

Un valor de $\mathbf{R^2}$ **de 0.854** es un resultado excelente, indicando que el modelo es capaz de explicar el **85.4 %** de la varianza en el logaritmo del valor de mercado de los jugadores, basándose únicamente en las variables de rendimiento y demográficas seleccionadas. El MAE indica que, en promedio, las predicciones del modelo se desvían en aproximadamente 3.6 millones de euros del valor real, una cifra competitiva dado el rango de valores en el mercado.

## 6.4. Importancia de las Características

El análisis de la importancia de las características del modelo entrenado confirma y cuantifica los hallazgos del EDA. Las variables que más contribuyeron a la capacidad predictiva del modelo fueron, en orden descendente:

| Ranking | Característica (Feature) |
|---|---|
| 1 | Contribuciones a Gol por 90 min (`goal_contributions_per_90`) |
| 2 | Goles Esperados sin Penales por 90 min (`non_penalty_xg_per_90`) |
| 3 | Edad (`age`) |
| 4 | Asistencias por 90 min (`assists_per_90`) |
| 5 | Tiros a Puerta por 90 min (`shots_on_target_per_90`) |

Cuadro 3: Top 5 características más importantes según el modelo XGBoost.

# 7. Discusión y Conclusiones

## 7.1. Discusión

El proyecto valida de manera contundente la hipótesis de que el valor de mercado en el fútbol está fuertemente anclado al rendimiento cuantificable en el campo. El alto coeficiente de determinación ($\mathbf{R^2 = 0.854}$) obtenido por el modelo XGBoost es un indicador poderoso de que, a pesar de la subjetividad inherente al mercado, la gran mayoría de la valoración puede ser explicada y predicha mediante un enfoque *data-driven*.

La preponderancia de las métricas ofensivas en la importancia de características del modelo refleja una realidad económica del deporte: el talento que genera goles es el recurso más escaso y, por lo tanto, el más valorado. Es notable que las métricas de çalidad"(como `non_penalty_xg_per_90`) sean casi tan importantes como las de çantidad"(contribuciones a gol), lo que demuestra la madurez del análisis de datos en el fútbol. Este hallazgo tiene implicaciones directas para los departamentos de scouting, que pueden utilizar modelos como este para identificar jugadores infravalorados cuyo rendimiento subyacente (xG) es superior a su producción real.

El pipeline de datos construido es el activo más valioso del proyecto. Su naturaleza modular y automatizada permite no solo la actualización de los datos con nuevas temporadas, sino también su extensión. Se podrían incorporar nuevas fuentes de datos, como métricas de tracking físico o datos de popularidad en redes sociales, para intentar explicar el 14.6 % de la varianza que el modelo actual no captura.

**Limitaciones y Trabajo Futuro:**

- **Factores Exógenos:** El modelo no captura factores no relacionados con el rendimiento en el campo, como la duración del contrato, el potencial de marketing o el club de procedencia, que se sabe influyen en el valor.
- **Orquestación:** Para una implementación en un entorno de producción, el pipeline se beneficiaría enormemente de un orquestador de flujos de trabajo como **Prefect** (discutido en la Sesión 5). Esto permitiría programar ejecuciones automáticas, gestionar dependencias de manera explícita y manejar reintentos en caso de fallos.
- **Despliegue:** Un paso futuro sería desplegar el modelo entrenado como una API REST, permitiendo a los usuarios obtener predicciones de valor de mercado en tiempo real para un jugador específico.

## 7.2. Conclusiones

Este proyecto ha cumplido exitosamente su objetivo de desarrollar una solución de extremo a extremo para la predicción del valor de mercado de futbolistas. Se ha construido un pipeline de datos ETL robusto y reproducible que transforma datos brutos de la web en un dataset analítico de alta calidad. El modelo de Machine Learning resultante no solo posee una alta precisión predictiva, sino que también ofrece insights interpretables sobre los factores que impulsan el valor en el fútbol moderno.

El trabajo realizado demuestra una aplicación integral de los conceptos clave de la asignatura, desde la ingeniería de datos y la automatización de flujos de trabajo hasta el análisis estadístico y el modelado predictivo, encapsulando el ciclo de vida completo de un proyecto de Big Data.

# 8. Bibliografía

1. Franco, H. (2025). *Session 4: Exploratory Analysis: Descriptive Statistics and Visualization.* [Material de clase]. Herramientas de Big Data, Universidad de La Sabana.

2. Franco, H. (2025). *Session 5: Data Pipelines - Prefect. Introduction to Cloud Computing.* [Material de clase]. Herramientas de Big Data, Universidad de La Sabana.

3. McKinney, W. (2017). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython.* O'Reilly Media.

4. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning.* Springer.

5. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

# 9. Anexos

## 9.1. Código Fuente y Estructura del Repositorio

El código fuente completo del proyecto, que garantiza la total reproducibilidad de este informe, se encuentra en el repositorio de GitHub adjunto a la entrega. La estructura de archivos es la siguiente:

```
/Bigdata-SoccerMarketAnalytics
|-- .venv/
|-- data/
|   |-- raw/
|   |-- processed/
|   |   |-- dataset_ready.csv
|-- notebooks/
|   |-- 01_EDA.ipynb
|-- .gitignore
|-- clean_transfermark.py
|-- fbref_scraper_all.py
|-- merge_quick.py
|-- merge_transfermarkt.py
|-- prepare_dataset.py
|-- requirements.txt
```

## 9.2. Datos Empleados

Tanto los datos brutos extraídos como el dataset procesado final (`dataset_ready.csv`) están disponibles en la carpeta `/data` del repositorio, permitiendo la replicación completa de los análisis y resultados aquí presentados.