Universidad de La Sabana

Facultad de Ingeniería

Maestría en Analítica Aplicada (Coterminal)

Proyecto Final – Herramientas de Big Data

Predicción del Valor de Mercado de Futbolistas

Autores:

Esteban Bernal - C´od. 271930 Juan Montes - C´od. 272113 Juan Gomez - C´od. 286774z

Asignatura: Herramientas de Big Data (Coterminal)

Profesor: Hugo Franco, Ph.D.

Chía, Colombia

Septiembre 5, 2025

Índice

1. Resumen Ejecutivo (Brief)

Este proyecto presenta el diseño e implementación de un pipeline de datos bajo el paradigma ETL (Extract, Transform, Load) para la predicción del valor de mercado de futbolistas profesionales. Se integraron fuentes de datos como Transfermarkt (valores históricos) y FBref (métricas de rendimiento), procesados mediante Python y orquestados en un entorno reproducible con PostgreSQL y Docker. El objetivo fue entrenar modelos de Machine Learning (Regresión Lineal, Random Forest y XGBoost), demostrando que es posible obtener estimaciones objetivas y transparentes frente a los valores de referencia del mercado.

2. Abstract

This project presents the design and implementation of a data pipeline under the ETL (Extract, Transform, Load) paradigm for predicting the market value of professional football players. We integrated multiple data sources, such as Transfermarkt (historical values) and FBref (performance metrics), processed with Python and orchestrated in a reproducible environment using PostgreSQL and Docker. The goal was to train Machine Learning models (Linear Regression, Random Forest, and XGBoost), showing that it is possible to generate objective and transparent estimates compared to market reference values.

3. Introducción

3.1. Formulación del problema

La valoración de jugadores de fútbol profesional es clave para clubes, agentes y patrocinadores. Sin embargo, los valores reportados en plataformas como Transfermarkt carecen de transparencia metodológica. Esto genera incertidumbre y riesgos en decisiones de inversión deportiva.

3.2. Marco conceptual

- **ETL Pipeline**: extracción, transformación y carga de datos.

- **Métricas Avanzadas**: indicadores como goles esperados (xG), asistencias esperadas (xA).
- **EDA**: análisis exploratorio de datos para identificar patrones.
- **Modelos**: comparación de regresión lineal, Random Forest y XGBoost.

3.3. Antecedentes

Diversos estudios muestran correlación entre rendimiento ofensivo y valor de mercado, aunque suelen limitarse a ligas específicas o datasets pequeños. Nuestro enfoque amplía la escala y aplica mejores prácticas de reproducibilidad.

3.4. Objetivo

Desarrollar un pipeline reproducible que integre datos de mercado y rendimiento para entrenar modelos de predicción robustos del valor de futbolistas.

4. Datos empleados

4.1. Fuentes de datos

1. **FBref.com**: vía web scraping para estadísticas de rendimiento (goles, asistencias, xG, xA, tackles).
2. **Transfermarkt (Kaggle)**: dataset histórico de valores de mercado, datos contractuales y demográficos.

4.2. Contenido y variables

- **Dependiente:** Valor de mercado en euros.
- **Independientes:** Edad, posición, nacionalidad, goles/90, asistencias/90, xG, xA, tackles, club, liga.

5. Materiales y Métodos

5.1. Pipeline ETL

1. **Extracción:** Scraping de FBref y carga de Transfermarkt desde Kaggle.
2. **Transformación:** Limpieza, unificación y normalización de datos.
3. **Carga:** Generación de un dataset consolidado (`dataset_ready.csv`) almacenado en PostgreSQL vía Docker.

5.2. Análisis Exploratorio (EDA)

Distribuciones, correlaciones de Pearson, boxplots por posición y outliers identificados.

5.3. Modelado

Comparación de regresión lineal, Random Forest y XGBoost con métricas MAE, RMSE y R^2.

6. Resultados

- XGBoost obtuvo el mejor desempeño predictivo.
- Se identificaron jugadores sobrevalorados e infravalorados respecto a Transfermarkt.
- Predicciones para 2025/2026 muestran consolidación de estrellas jóvenes como Bellingham y Yamal.

7. Discusión y Conclusiones

7.1. Discusión

Los resultados validan que el mercado tiende a sobrevalorar talento ofensivo y jóvenes promesas, mientras que los jugadores veteranos muestran ajustes a la baja. El modelo se

centra en métricas objetivas, excluyendo factores externos (marketing, popularidad).

7.2. Conclusiones

- Se construyó un pipeline reproducible en Python con PostgreSQL y Docker.
- El modelo XGBoost ofrece estimaciones más objetivas que Transfermarkt.
- El sistema puede ampliarse con datos de lesiones o redes sociales para futuras versiones.

8. Bibliografía

1. Franco, H. (2025). *Material de clase – Herramientas de Big Data*. Universidad de La Sabana.
2. McKinney, W. (2017). *Python for Data Analysis*. O'Reilly Media.
3. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer.

9. Anexos

9.1. Código Fuente

El código fuente completo (pipelines ETL, notebooks de modelado y consultas) está disponible en el repositorio de GitHub: `https://github.com/usuario/bigdata-futbol-valor-mercado`

9.2. Datos empleados

Los datos crudos y el dataset procesado final (`dataset_ready.csv`) se encuentran en el directorio `/data` del repositorio.