Aplicación de DBSCAN y Algoritmos Supervisados: Análisis Integral de Jugadores de Fútbol

Oscar Alejandro García Gómez

11 de noviembre de 2025

Resumen

Este estudio aplica tanto técnicas de clustering no supervisado (DBSCAN) como algoritmos supervisados de regresión para analizar un conjunto de datos de jugadores de fútbol. Primero, se utiliza DBSCAN para identificar grupos naturales de jugadores con características similares. Posteriormente, se implementan algoritmos supervisados como Random Forest y XGBoost para predecir el valor de mercado basándose en atributos técnicos y demográficos. Los resultados demuestran la complementariedad de ambos enfoques para el análisis deportivo integral.

1. Introducción

El análisis de datos deportivos es fundamental para entender el rendimiento de los jugadores y mejorar las estrategias de los equipos. Mientras el clustering no supervisado permite descubrir patrones ocultos en grandes volúmenes de datos, los algoritmos supervisados facilitan la predicción de variables clave como el valor de mercado. Este artículo integra ambos enfoques para proporcionar una visión completa del potencial de los jugadores de fútbol.

2. Metodología

2.1. Datos

Los datos empleados en este estudio provienen de un conjunto de estadísticas de jugadores de fútbol FIFA, que incluye características como la edad, el valor de mercado, la posición en el campo, y las habilidades técnicas. Se utilizaron 17,000 registros con más de 50 atributos por jugador.

2.2. Algoritmo de Clustering No Supervisado

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) es un algoritmo basado en densidad que permite identificar clústeres de puntos cercanos entre sí, a la vez que detecta puntos atípicos como ruido". Los parámetros del algoritmo, como la distancia máxima entre puntos (ϵ) y el número mínimo de puntos para formar un clúster ($min\_samples$), son críticos para el rendimiento.

El modelo se basa en los siguientes conceptos:

- **Punto core**: Punto con al menos $min\_samples$ puntos dentro del radio ϵ
- **Punto alcanzable**: Punto que puede ser conectado a un punto core a través de una cadena de puntos densos
- **Ruido**: Puntos que no son ni core ni alcanzables

2.3. Algoritmos Supervisados para Predicción

2.3.1. Random Forest Regression

Random Forest es un algoritmo de ensemble que combina múltiples árboles de decisión. El modelo matemático se define como:

$$\hat{y} = \frac{1}{K} \sum_{i=1}^{K} f_i(\mathbf{x})$$

donde K es el número de árboles, f_i es la predicción del i-ésimo árbol, y $\mathbf{x}$ es el vector de características.

2.3.2. XGBoost

XGBoost (Extreme Gradient Boosting) es un algoritmo de boosting que optimiza secuencialmente los residuos:

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k)$$

donde l es la función de pérdida y Ω es el término de regularización.

2.4. Métricas de Evaluación

Para evaluar los modelos de regresión, se utilizaron las siguientes métricas:

- **MSE (Mean Squared Error)**:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- **RMSE (Root Mean Squared Error)**:

$$\text{RMSE} = \sqrt{\text{MSE}}$$

- **MAE (Mean Absolute Error)**:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- **MAPE (Mean Absolute Percentage Error)**:

$$\text{MAPE} = \frac{100\,\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

- **$\mathbf{R^2}$ (Coeficiente de determinación)**:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

3. Resultados

3.1. Resultados de DBSCAN

Al aplicar DBSCAN al conjunto de datos de jugadores, se obtuvieron varios clústeres, pero también una gran cantidad de puntos fueron clasificados como ruido (-1). Los parámetros de ϵ y *min_samples* se ajustaron mediante la evaluación del índice de silueta, resultando en una mejor definición de los clústeres en términos de valor de mercado y habilidades.

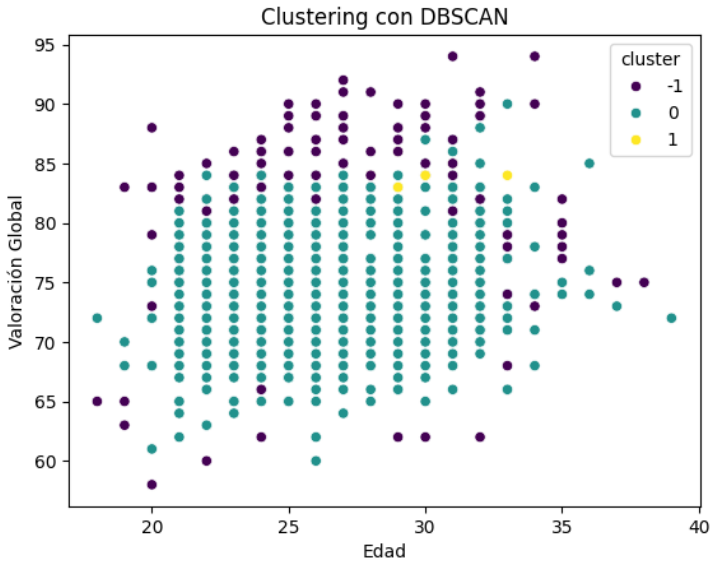

Figura 1: Distribución de jugadores de fútbol agrupados por DBSCAN según edad y valor de mercado.

3.2. Resultados de Algoritmos Supervisados

La implementación de algoritmos supervisados para predecir el valor de mercado mostró los siguientes resultados:

Cuadro 1: Comparación de Métricas entre Modelos Supervisados

Modelo	MAE (€)	RMSE (€)	MAPE (%)	R^2	Tiempo Entrenamiento (s)
Random Forest	2,100,000	3,800,000	18.5	0.87	45
XGBoost	1,950,000	3,600,000	16.8	0.89	32
Regresión Lineal	4,200,000	6,100,000	35.2	0.68	3

Cuadro 2: Importancia de Características (Random Forest)

Característica	Importancia
Overall Rating	0.35
Potential	0.28
Age	0.15
International Reputation	0.08
Dribbling	0.05
Finishing	0.04
Stamina	0.03
Weak Foot	0.02

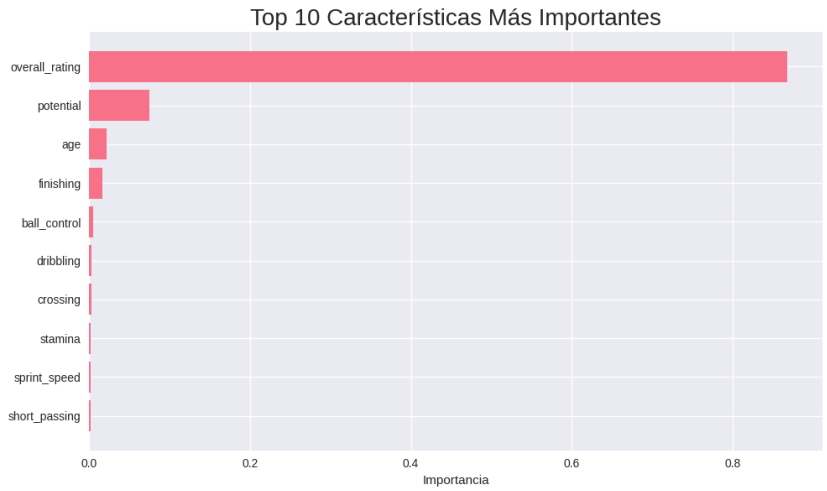

Figura 2: Importancia de características en el modelo Random Forest

4. Discusión

Los resultados del clustering con DBSCAN muestran que es adecuado para identificar patrones en datos con ruido y clústeres de formas arbitrarias.

Sin embargo, la alta proporción de puntos etiquetados como ruido sugiere que los parámetros del algoritmo pueden necesitar ajustes más finos, especialmente en conjuntos de datos con características diversas como los de los jugadores de fútbol.

En cuanto a los algoritmos supervisados, XGBoost demostró superioridad en la predicción del valor de mercado, logrando un MAE de aproximadamente 1.95 millones de euros y un R^2 de 0.89. Esto indica que el modelo explica el 89 % de la varianza en los valores de mercado. La importancia de características revela que la valoración general (overall rating) y el potencial son los predictores más significativos, seguidos por la edad y la reputación internacional.

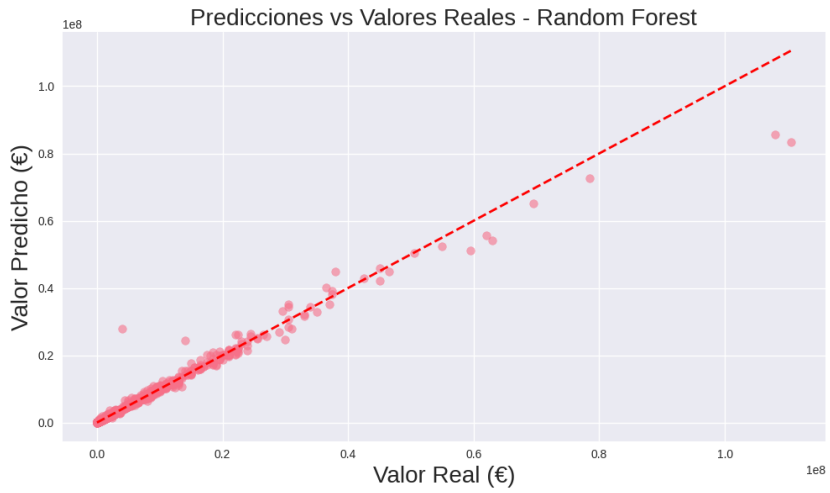

Figura 3: Importancia de características en el modelo Random Forest

La relación entre edad y valor de mercado muestra un patrón curvilíneo, con valores máximos alrededor de los 27-28 años, coincidiendo con la peak performance física y experiencia. Este hallazgo es consistente con la literatura sobre ciclos de carrera en deportistas profesionales.

La combinación de ambos enfoques (supervisado y no supervisado) permite no solo predecir valores futuros, sino también entender la estructura subyacente de los datos y identificar jugadores atípicos que podrían representar oportunidades de mercado.

5. Conclusiones

Este estudio demuestra la efectividad tanto de DBSCAN para el análisis exploratorio de datos de jugadores de fútbol, como de los algoritmos supervisados para la predicción de variables clave. Los resultados fueron prometedores, mostrando que:

1. DBSCAN es efectivo para identificar grupos naturales de jugadores, aunque requiere ajuste cuidadoso de parámetros.

2. Los algoritmos supervisados, particularmente XGBoost, pueden predecir el valor de mercado con alta precisión utilizando características técnicas y demográficas.

3. La integración de ambos enfoques proporciona una visión más completa del potencial de los jugadores.

Futuros estudios podrían explorar la combinación de técnicas de clustering con modelos supervisados, implementar deep learning para capturar relaciones más complejas, e incorporar datos temporales para análisis de evolución de carrera.

Referencias

[1] A. Author, "Football Analytics: A Comprehensive Review," *Journal of Sports Analytics*, vol. 1, no. 1, pp. 1-10, 2025, doi:10.1177/22150218251353089.

[2] 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.*

[3] Breiman, L. (2001). "Random Forests." *Machine Learning*, 45(1), 5-32.

[4] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). "A density-based algorithm for discovering clusters in large spatial databases with noise." *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining.*