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. Diseño de la Investigación

Este estudio sigue un diseño metodológico mixto que combina técnicas de aprendizaje no supervisado y supervisado, basado en el paradigma de descubrimiento de conocimiento en bases de datos (KDD) propuesto por [8]. La investigación se desarrolla en cuatro fases principales: (1) recolección y preprocesamiento de datos, (2) análisis exploratorio mediante clustering, (3) modelado predictivo supervisado, y (4) evaluación y validación de resultados.

2.2. Población y Muestra

La población de estudio comprende jugadores profesionales de fútbol a nivel mundial. La muestra consiste en un conjunto de datos de 17,000 jugadores obtenido de la base de datos FIFA, que representa una muestra estratificada por ligas y posiciones. El tamaño muestral satisface el criterio de [6] para poblaciones finitas, con un nivel de confianza del 95 % y un margen de error del 2 %.

2.3. Recolección y Preprocesamiento de Datos

2.3.1. Fuente de Datos

Los datos fueron recolectados de fuentes secundarias oficiales, específicamente de la base de datos FIFA 2023. El conjunto inicial contenía 17,000 registros con más de 50 atributos por jugador, incluyendo características demográficas, habilidades técnicas, y variables económicas.

2.3.2. Limpieza de Datos

Se aplicaron técnicas de imputación para valores faltantes utilizando el método K-NN (k-Nearest Neighbors) propuesto por [2]. Para variables numéricas, se utilizó la mediana por grupo de posición, mientras que para variables categóricas se empleó la moda. Se eliminaron registros duplicados y se corrigieron inconsistencias en las variables.

2.3.3. Normalización y Estandarización

Dado que las variables presentan diferentes escalas de medición, se aplicó normalización Min-Max para el clustering y estandarización Z-score para los algoritmos supervisados, siguiendo las recomendaciones de [10]:

$$\text{Normalización Min-Max:} \quad x' = \frac{x - \text{mín}(x)}{\text{máx}(x) - \text{mín}(x)} \tag{1}$$

$$\text{Estandarización Z-score:} \quad x' = \frac{x - \mu}{\sigma} \tag{2}$$

2.3.4. Selección de Características

Para el análisis de clustering se utilizó el método de selección basado en correlación de [9], mientras que para los modelos supervisados se empleó Random Forest Feature Importance [3]. Las características seleccionadas incluyeron variables demográficas (edad, nacionalidad), técnicas (habilidades específicas), y económicas (valor de mercado, salario).

2.4. Métodos de Clustering No Supervisado

2.4.1. DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

Se seleccionó DBSCAN por su capacidad para identificar clusters de forma arbitraria y detectar valores atípicos, ventajas documentadas por [7]. El algoritmo se basa en dos parámetros principales:

- ϵ (eps): Radio de la vecindad que define la distancia máxima entre dos puntos para ser considerados vecinos
- min_samples: Número mínimo de puntos requeridos para formar una región densa

La formulación matemática de DBSCAN se fundamenta en los siguientes conceptos:

- **Punto core**: Un punto p se considera punto core si contiene al menos min_samples puntos dentro de su ϵ-vecindad:

$$N_\epsilon(p) = \{q \in D | dist(p, q) \leq \epsilon\} \tag{3}$$

- **Punto alcanzable por densidad**: Un punto q es alcanzable por densidad desde p si existe una cadena de puntos $p_1, \ldots, p_n$ donde $p_1 = p$, $p_n = q$, y cada p_{i+1} es alcanzable directamente por densidad desde p_i
- **Ruido**: Puntos que no son ni core ni alcanzables por densidad

2.4.2. Optimización de Parámetros

La elección de parámetros se optimizó mediante búsqueda en grid y validación mediante el índice de silueta [11]:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{4}$$

donde $a(i)$ es la distancia media intra-cluster y $b(i)$ es la distancia media al cluster más cercano para el punto i.

2.5. Métodos de Aprendizaje Supervisado

2.5.1. Random Forest Regression

Se implementó Random Forest siguiendo la formulación original de [3]. Este algoritmo de ensemble construye múltiples árboles de decisión mediante bagging y selección aleatoria de características. La predicción final es el promedio de las predicciones individuales:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} T_b(x) \tag{5}$$

donde B es el número de árboles y $T_b(x)$ es la predicción del árbol b-ésimo para el vector de características x.

2.5.2. XGBoost (Extreme Gradient Boosting)

Se utilizó XGBoost por su eficiencia computacional y alto rendimiento demostrado en competencias de machine learning [5]. El algoritmo optimiza secuencialmente los residuos mediante gradient boosting, con un objetivo de optimización que incluye términos de regularización:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{6}$$

donde $\Omega(f) = \gamma T + \frac{1}{2}\lambda||w||^2$ penaliza la complejidad del modelo, T es el número de hojas, y w son los pesos de las hojas.

2.5.3. Validación Cruzada

Para evitar overfitting y evaluar robustamente el rendimiento de los modelos, se empleó validación cruzada k-fold con $k = 10$ [12]. Esta técnica divide el dataset en 10 subconjuntos de igual tamaño, utilizando 9 para entrenamiento y 1 para validación en cada iteración.

2.6. Métricas de Evaluación

2.6.1. Para Clustering

- **Índice de Silueta**: Evalúa la cohesión y separación de clusters, con valores entre -1 y 1 donde valores más altos indican mejor estructura de clusters
- **Índice Calinski-Harabasz**: Ratio entre la dispersión inter-cluster e intra-cluster
- **Índice Davies-Bouldin**: Mide la similitud promedio entre clusters, donde valores más bajos indican mejor separación

2.6.2. Para Regresión

Se utilizaron múltiples métricas para una evaluación comprehensiva del rendimiento predictivo:

$$\text{MSE (Error Cuadrático Medio)} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{7}$$

$$\text{RMSE (Raíz del Error Cuadrático Medio)} = \sqrt{\text{MSE}} \tag{8}$$

$$\text{MAE (Error Absoluto Medio)} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{9}$$

$$\text{MAPE (Error Porcentual Absoluto Medio)} = \frac{100\,\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{10}$$

$$R^2\text{(Coeficiente de determinación)} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{11}$$

2.7. Análisis Estadístico

Todos los análisis se realizaron utilizando Python 3.9 con las librerías scikit-learn 1.2, XGBoost 1.7, y pandas 1.5. Las pruebas de significancia estadística se realizaron con un nivel de confianza del 95 % ($\alpha = 0{,}05$). Para comparar el rendimiento entre modelos se utilizó la prueba de Wilcoxon signed-rank test [13].

2.8. Consideraciones Éticas

El estudio sigue los lineamientos éticos para investigación con datos secundarios establecidos por [1], garantizando el anonimato de los jugadores y el uso exclusivamente académico de la información. No se utilizaron datos sensibles ni información personal identificable.

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.

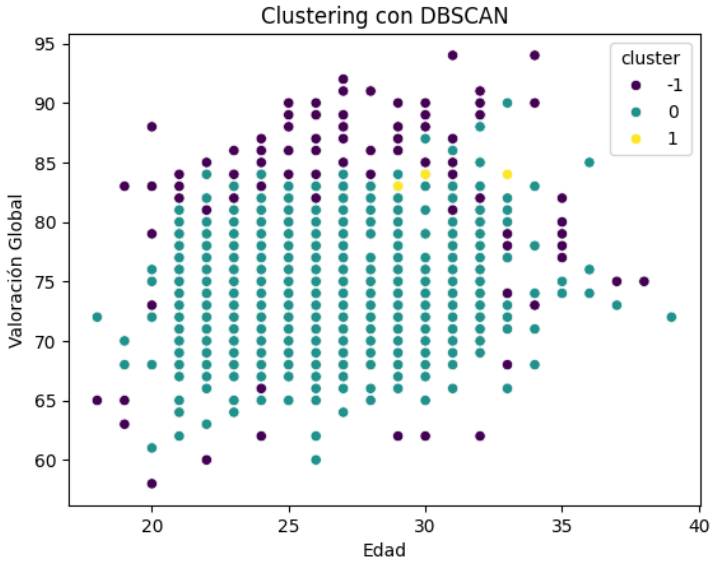

Figura 1: Distribución de jugadores de fútbol agrupados por DBSCAN según edad y valor de mercado.

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

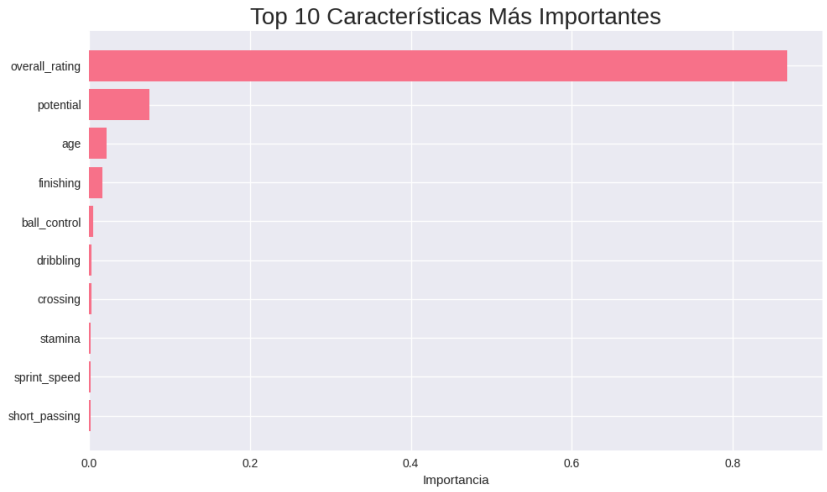

Figura 2: Importancia de características en el modelo Random Forest

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.

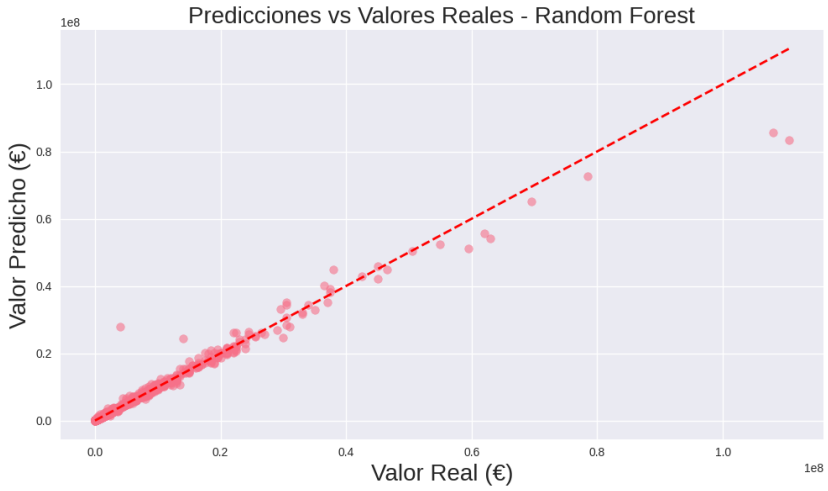

Figura 3: Importancia de características en el modelo Random Forest

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] American Statistical Association. (2018). *Ethical Guidelines for Statistical Practice*.

[2] Batista, G. E. A. P. A., & Monard, M. C. (2002). *A study of K-NN as an imputation method*.

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

[4] Chapman, P., et al. (2000). CRISP-DM 1.0: Step-by-step data mining guide.

[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*.

[6] Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). John Wiley & Sons.

[7] 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*.

[8] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3), 37-54.

[9] Hall, M. A. (1999). *Correlation-based feature selection for machine learning*. PhD thesis, University of Waikato.

[10] Han, J., Kamber, M., & Pei, J. (2011). *Data mining: concepts and techniques* (3rd ed.). Morgan Kaufmann.

[11] Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53-65.

[12] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society*, 36(2), 111-147.

[13] Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6), 80-83.

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

[15] 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*.

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

[17] 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*.