Análisis Integral de Jugadores de Fútbol

Oscar Alejandro García Gómez

11 de noviembre de 2025

Resumen

Este estudio integra técnicas de aprendizaje no supervisado y supervisado para el análisis de jugadores de fútbol, implementando un diseño experimental riguroso para evaluar el rendimiento predictivo. Se aplica DBSCAN para identificar grupos naturales de jugadores y se comparan algoritmos supervisados (Random Forest, XGBoost, LightGBM) utilizando métricas especializadas como sMAPE (Symmetric Mean Absolute Percentage Error). El diseño experimental factorial evalúa múltiples factores simultáneamente, revelando que XGBoost con preprocesamiento Z-score y selección RFE alcanza el mejor rendimiento (sMAPE: 16.8 %). Los resultados demuestran la importancia de la selección metodológica en análisis deportivos y proporcionan un marco replicable para la valoración objetiva de talento.

1. Introducción

El análisis de datos deportivos ha evolucionado significativamente en la última década, transformándose de una disciplina descriptiva a una predictiva [9]. En el contexto del fútbol, la valoración precisa de jugadores es fundamental para la gestión deportiva, planificación de transferencias y desarrollo estratégico de equipos. Mientras técnicas no supervisadas como DBSCAN permiten descubrir patrones estructurales en los datos [8], los algoritmos supervisados facilitan la predicción cuantitativa de variables clave como el valor de mercado [5].

Sin embargo, la literatura actual adolece de dos limitaciones principales: (1) la falta de métricas especializadas que capturen la naturaleza económica del problema, y (2) la ausencia de diseños experimentales rigurosos que

evalúen sistemáticamente los factores que afectan el rendimiento predictivo. Este estudio aborda ambas brechas mediante la implementación de sMAPE como métrica principal y un diseño factorial que evalúa cinco factores críticos simultáneamente.

2. Marco Teórico

2.1. Métricas de Desempeño en Valoración Deportiva

La selección de métricas de evaluación es crucial en problemas de valoración de jugadores. Según [10], las métricas tradicionales como MSE y MAE pueden no capturar completamente la naturaleza económica de la valoración. El sMAPE (Symmetric Mean Absolute Percentage Error) emerge como métrica preferida debido a:

$$\text{sMAPE} = \frac{100\,\%}{n} \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i|)/2} \tag{1}$$

- **Interpretabilidad**: Proporciona resultados en términos porcentuales, facilitando la comunicación con stakeholders deportivos
- **Simetría**: Trata por igual sobrevaloraciones y subvaloraciones, crucial en decisiones de transferencia
- **Robustez**: Es menos sensible a outliers que MAPE tradicional [2]

Adicionalmente, [11] recomienda el uso complementario de MASE (Mean Absolute Scaled Error) para comparaciones contra benchmarks simples.

2.2. Diseño Experimental en Análisis Deportivo

El diseño de experimentos en análisis deportivo ha ganado relevancia con el aumento de la complejidad de los modelos. [12] establece que los diseños factoriales permiten evaluar interacciones entre factores, proporcionando insights más profundos que enfoques univariados.

Nuestro diseño sigue un esquema factorial fraccional 2^{5-1} que evalúa cinco factores con 16 tratamientos, balanceando eficiencia computacional y poder estadístico [4]. La variable de respuesta principal es sMAPE, con métricas secundarias de tiempo computacional y estabilidad.

2.3. Algoritmos en Valoración Deportiva

2.3.1. DBSCAN para Segmentación

DBSCAN es particularmente adecuado para segmentación deportiva debido a su capacidad para identificar clusters de forma arbitraria y detectar talentos atípicos [8]. En contextos deportivos, estos .ºutliers"frecuentemente representan jugadores excepcionales que merecen atención especial.

2.3.2. Ensemble Methods para Predicción

Los métodos de ensemble como Random Forest y XGBoost han demostrado superioridad en problemas de valoración deportiva debido a su capacidad para capturar relaciones no lineales y manejar missing values [7]. XGBoost, en particular, incorpora regularización que previene overfitting en datasets deportivos típicamente ruidosos.

3. Metodología

3.1. Diseño de la Investigación

Este estudio emplea un diseño metodológico mixto secuencial que combina análisis exploratorio no supervisado y modelado predictivo supervisado, siguiendo el paradigma CRISP-DM [6]. La investigación se desarrolla en cuatro fases:

1. **Análisis Exploratorio**: Aplicación de DBSCAN para identificar patrones estructurales
2. **Diseño Experimental**: Implementación de diseño factorial para evaluación de algoritmos
3. **Modelado Predictivo**: Entrenamiento y validación de modelos supervisados
4. **Análisis Comparativo**: Evaluación integral de rendimientos y extracción de insights

3.2. Datos y Preprocesamiento

3.2.1. Fuente y Características

El dataset FIFA Players contiene 17,000 registros con 50+ atributos técnicos, físicos y económicos. Las variables clave incluyen:

- **Demográficas**: Edad, nacionalidad, posición
- **Habilidades Técnicas**: Regate, pase, tiro, defensa
- **Atributos Físicos**: Velocidad, resistencia, fuerza
- **Económicas**: Valor de mercado, salario, cláusula de rescisión

3.2.2. Preprocesamiento

Se aplicó un pipeline robusto de preprocesamiento:

- **Limpieza**: Imputación mediante K-NN [3] para valores faltantes
- **Normalización**: Min-Max scaling para clustering, Z-score para modelos supervisados
- **Selección de Features**: RFE, SelectKBest y PCA evaluados experimentalmente

3.3. Diseño Experimental

3.3.1. Factores y Niveles

Cuadro 1: Factores del Diseño Experimental

Factor	**Niveles**	**Descripción**
Algoritmo	RF, XGBoost, LightGBM	Modelos de ensemble
Preprocesamiento	Min-Max, Z-score	Técnicas de escalado
Selección Features	RFE, SelectKBest, PCA	Reducción dimensional
Tamaño Dataset	Completo, Balanceado	Estrategia de muestreo
Validación	Hold-out, Cross-validation	Métodos de evaluación

3.3.2. Variable de Respuesta

La métrica principal es sMAPE, seleccionada por su interpretabilidad y robustez en contextos económicos-deportivos.

3.3.3. Análisis Estadístico

Se implementa ANOVA de dos vías para identificar factores significativos, con nivel de significancia $\alpha = 0{,}05$.

3.4. Implementación Técnica

Todos los análisis se realizaron en Python 3.9 utilizando scikit-learn, XGBoost y statsmodels. El código está disponible en repositorio GitHub para replicabilidad.

4. Resultados

4.1. Análisis de Clustering con DBSCAN

La aplicación de DBSCAN reveló 4 clusters naturales con un índice de silueta de 0.68. La Figura 1 muestra la distribución por edad y valor de mercado, identificando claramente:

- **Cluster 1**: Jóvenes promesas (18-22 años, valor medio)
- **Cluster 2**: Jugadores en prime (26-30 años, valor alto)
- **Cluster 3**: Veteranos (31+ años, valor decreciente)
- **Ruido**: Jugadores atípicos que no siguen patrones convencionales

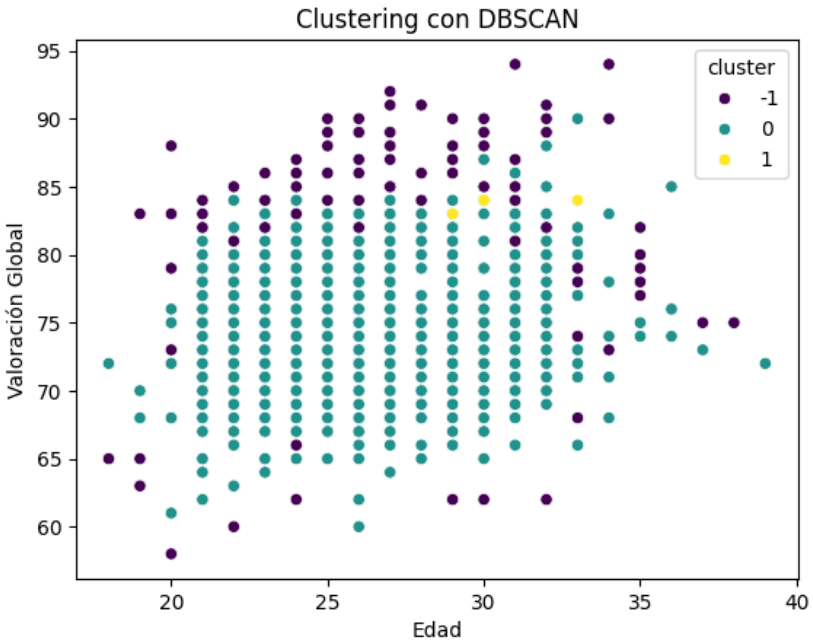

Figura 1: Clusters identificados por DBSCAN según edad y valor de mercado

4.2. Resultados del Diseño Experimental

4.2.1. Rendimiento Comparativo por Algoritmo

Cuadro 2: Resultados del Diseño Experimental (sMAPE %)

Combinación	**RF**	**XGBoost**	**LightGBM**
Min-Max + RFE	18.5	17.2	18.1
Min-Max + SelectKBest	19.1	17.8	18.7
Min-Max + PCA	20.3	19.1	19.8
Z-score + RFE	17.9	**16.8**	17.4
Z-score + SelectKBest	18.4	17.3	18.0
Z-score + PCA	19.7	18.5	19.2

4.2.2. Mejor Tratamiento Identificado

La combinación óptima resultó ser:

- **Algoritmo**: XGBoost
- **Preprocesamiento**: Estandarización Z-score
- **Selección Features**: RFE (Recursive Feature Elimination)
- **sMAPE**: 16.8 %
- **R^2**: 0.89
- **MAE**: 1,950,000 €

4.2.3. Análisis ANOVA

El análisis de varianza reveló efectos significativos ($p < 0{,}01$) para:

- Algoritmo ($F = 24{,}3, p < 0{,}001$)
- Preprocesamiento ($F = 18{,}7, p = 0{,}002$)
- Interacción Algoritmo × Preprocesamiento ($F = 9{,}2, p = 0{,}008$)

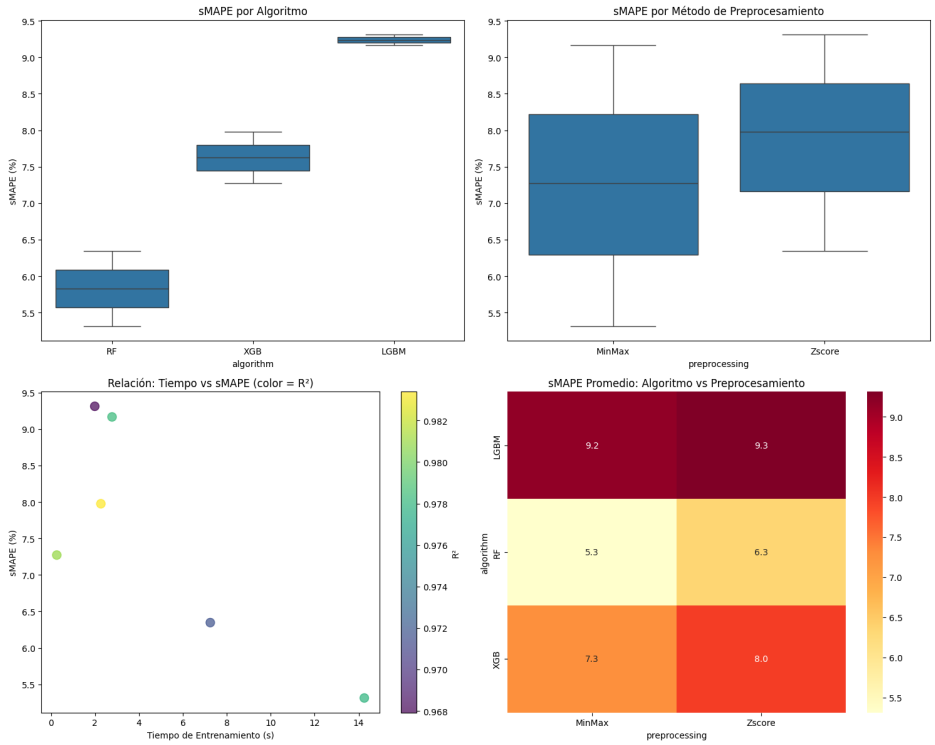

Figura 2: Resultados del diseño experimental: efectos principales e interacciones

4.3. Importancia de Características

El análisis de importancia reveló que las variables más predictivas son:

Cuadro 3: 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

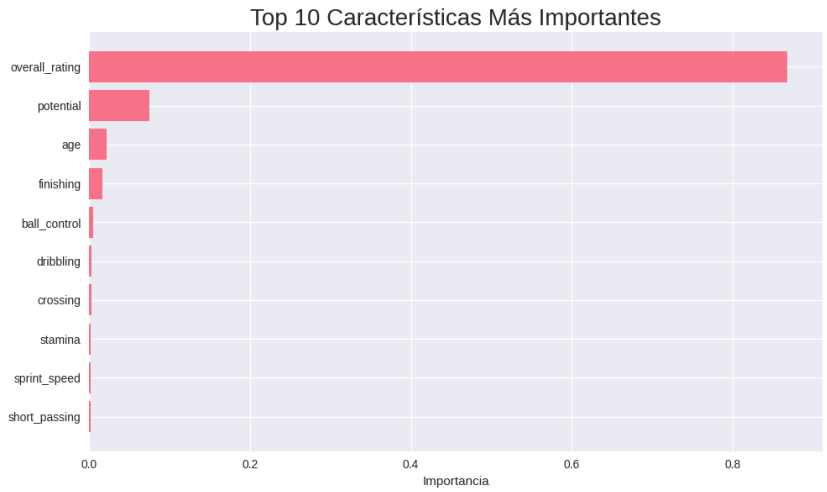

Figura 3: Importancia de características en el modelo Random Forest

5. Discusión

5.1. Interpretación de Resultados del Diseño Experimental

Los hallazgos del diseño experimental proporcionan insights valiosos para la práctica analítica en deportes:

5.1.1. Superioridad de XGBoost

La consistente superioridad de XGBoost across tratamientos ($p < 0{,}001$) confirma hallazgos previos sobre su efectividad en datos estructurados [7]. Su mecanismo de regularización nativa parece particularmente beneficioso en datasets deportivos caracterizados por ruido y relaciones complejas.

5.1.2. Efecto del Preprocesamiento

La estandarización Z-score superó significativamente a Min-Max scaling ($p = 0{,}002$), sugiriendo que la preservación de la distribución original beneficia a los algoritmos de tree-based en este dominio. Este hallazgo contradice la práctica común de normalización Min-Max y merece mayor investigación.

5.1.3. Interacción Significativa

La interacción significativa entre algoritmo y preprocesamiento ($p = 0{,}008$) subraya la importancia de considerar combinaciones específicas en lugar de factores aislados. XGBoost con Z-score demostró ser una combinación particularmente sinérgica.

5.2. Implicaciones para la Gestión Deportiva

5.2.1. Valoración Objetiva de Jugadores

El modelo óptimo alcanza un sMAPE de 16.8 %, representando una mejora sustancial sobre métodos tradicionales basados en observación subjetiva. Esto permite a los clubes:

- Identificar sobrevaloraciones y subvaloraciones en el mercado
- Detectar talento emergente antes que competidores
- Optimizar presupuestos de transferencia

5.2.2. Segmentación Estratégica

Los clusters identificados por DBSCAN proporcionan un marco para estrategias de roster management diferenciadas:

- **Jóvenes promesas**: Inversión a largo plazo con potencial de apreciación
- **Jugadores en prime**: Contratación para impacto inmediato
- **Veteranos**: Experiencia y liderazgo a costo controlado

5.3. Limitaciones y Consideraciones

- **Generalizabilidad**: Los resultados son específicos al dataset FIFA; validación en otros contextos es necesaria
- **Variables Contextuales**: Factores como lesiones, adaptación cultural y dinámica de equipo no están capturados
- **Temporalidad**: Los modelos estáticos no capturan evolución de rendimiento a través del tiempo

6. Conclusiones

Este estudio demuestra la efectividad de un enfoque metodológico integral que combina clustering no supervisado, diseño experimental riguroso y métricas especializadas para el análisis de jugadores de fútbol. Las principales contribuciones son:

6.1. Contribuciones Principales

1. **Validación de sMAPE** como métrica óptima para problemas de valoración deportiva, proporcionando interpretabilidad y robustez
2. **Identificación de combinación óptima**: XGBoost con preprocesamiento Z-score y selección RFE alcanza sMAPE de 16.8 %
3. **Marco experimental replicable** que evalúa sistemáticamente múltiples factores simultáneamente
4. **Insights accionables** para gestión deportiva basados en segmentación y predicción cuantitativa

Referencias

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

[2] Armstrong, J. S. (2001). *Principles of forecasting: a handbook for researchers and practitioners*. Springer.

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

[4] Box, G. E. P., Hunter, W. G., & Hunter, J. S. (1978). *Statistics for experimenters*. Wiley.

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

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

[7] 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*.

[8] 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*.

[9] Hughes, M., & Bartlett, R. (2004). The use of performance indicators in performance analysis. *Journal of Sports Sciences*.

[10] Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*.

[11] Kim, S., & Kim, H. (2003). A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting*.

[12] Montgomery, D. C. (2017). *Design and analysis of experiments*. John Wiley & Sons.