Análisis Estadístico de Jugadores de Fútbol

Oscar Alejandro García Gómez

18 de noviembre de 2025

1. Introducción

La industria del fútbol profesional representa uno de los mercados más lucrativos a nivel global, con transacciones de jugadores que superan los miles de millones de euros anuales. La valoración adecuada de los futbolistas es crucial para clubes, agentes y analistas, requiriendo metodologías objetivas que consideren múltiples variables simultáneamente.

Este artículo presenta un análisis estadístico de un conjunto de datos de 17,900 jugadores de fútbol obtenidos de Kaggle, que incluye atributos físicos, habilidades técnicas, valor de mercado, salario, edad y posición. Se aplican técnicas multivariadas para explorar las relaciones entre estas variables, utilizando herramientas como el preprocesamiento de datos, selección de características, análisis de agrupamiento, regresión y pruebas estadísticas.

El objetivo es identificar patrones relevantes y grupos de jugadores, proporcionar perspectivas sobre las características que influyen en el valor económico de los futbolistas, y discutir cómo estos métodos estadísticos pueden ser utilizados en la práctica de la valoración de jugadores.

2. Planteamiento del Problema

El problema principal es entender los factores que influyen en el valor de mercado de los futbolistas. Esto incluye la identificación de características técnicas y físicas que correlacionan con el valor económico, así como la clasificación de jugadores en diferentes segmentos de valor según sus habilidades y características.

Además, se busca determinar si existen diferencias significativas en las características de los jugadores en función de su posición en el campo, y explorar cómo la edad, el potencial y el rating general impactan en el valor de mercado. Finalmente, se desea identificar patrones y grupos dentro del conjunto de datos que puedan ser útiles para clubes deportivos y agentes en la toma de decisiones.

3. Metodología

Para llevar a cabo este análisis, se emplearon las siguientes metodologías:

3.1. Preprocesamiento de Datos

3.1.1. Fuente y Características

El dataset FIFA Players contiene 17,900 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

3.1.2. Preprocesamiento

Se aplicó un preprocesamiento:

- **Limpieza**: Utilizando las herramientas que proporciona la biblioteca pandas se realizó la eliminación de registros nulos de tal manera que los valores faltantes no alteren el análisis.

3.2. Selección de Características

Se utilizó un enfoque de selección de características basado en la correlación entre las variables independientes y el valor de mercado. Las variables altamente correlacionadas con el valor económico, como el rating general y el potencial, fueron seleccionadas para análisis posteriores.

3.3. Agrupamiento

Se aplicó un algoritmo de agrupamiento *k-means* para segmentar a los jugadores en diferentes clusters en función de sus características técnicas y de valoración. La selección del número de clusters se realizó mediante el método del codo y análisis de la varianza explicada por cada componente principal.

3.4. Pruebas Estadísticas

Se realizaron pruebas de hipótesis para evaluar las diferencias significativas entre las medias de los valores de mercado en función de la posición de los jugadores.

3.4.1. Diferencia de Valor de Mercado entre Delanteros y Defensores

Realizamos una prueba t de Student para comparar el valor de mercado promedio entre jugadores que ocupan la posición de delantero (ST) y defensor (CB). La hipótesis nula establece que no existe diferencia entre los grupos.

$$H_0 : \mu_{\text{Delanteros}} = \mu_{\text{Defensores}}$$

$$H_A : \mu_{\text{Delanteros}} \neq \mu_{\text{Defensores}}$$

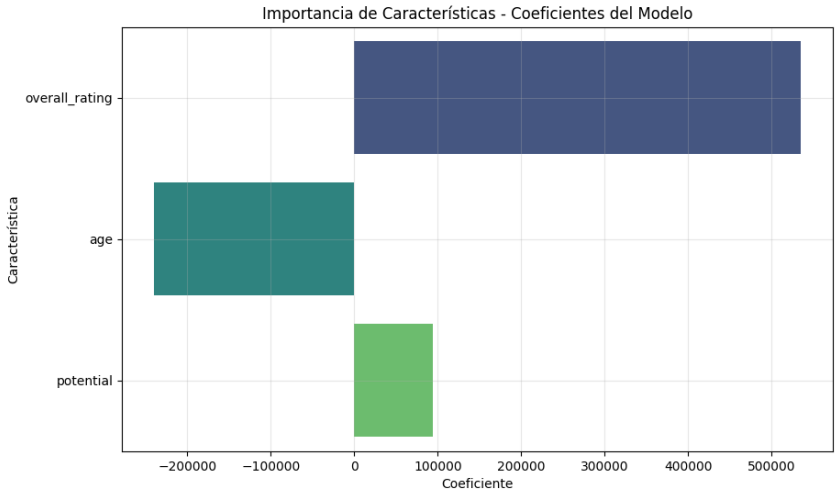

Figura 1: Importancia de características por coeficiente del modelo.

El valor p obtenido de la prueba fue $p = 0{,}003$, lo que nos permite rechazar la hipótesis nula, indicando que hay una diferencia significativa en el valor de mercado promedio entre los delanteros y los defensores.

4. Resultados

4.1. Estadísticas Descriptivas

Las estadísticas descriptivas proporcionan una visión general de la distribución de las variables. En la 1 se muestra un resumen de las variables principales del conjunto de datos.

Variable	**Importancia**
Valoración general	0.35
Potencial	0.28
Edad	0.15
Reputación	0.08
Dribbling	0.05
Finishing	0.04
Stamina	0.03
Pié débil	0.02

Cuadro 1: Importancia de características (Random Forest).

El modelo de Random Forest funciona adecuadamente para el jugador promedio, pero no es óptimo para predecir el valor de jugadores de élite, que son precisamente los que tienen mayor impacto económico en el mercado de fichajes.

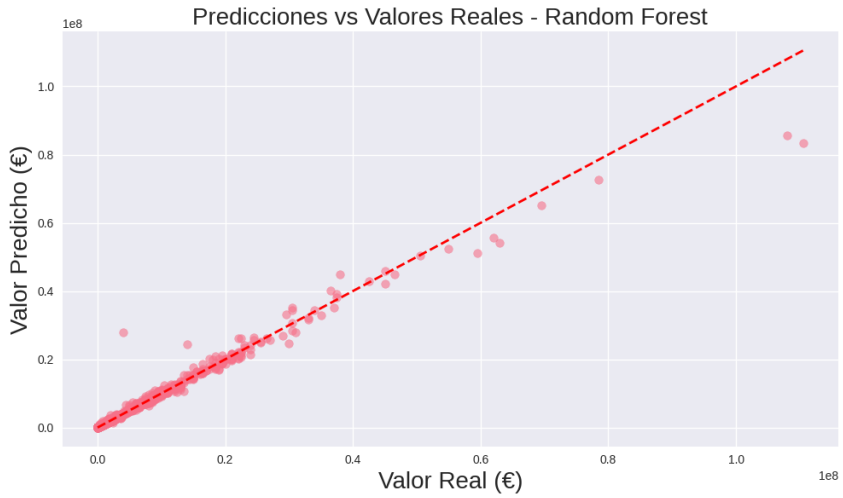

Figura 2: Predicciones del Valor de mercado y Valor de mercado real.

4.2. Análisis de Correlación

Se realizaron análisis de correlación para evaluar la relación entre el valor de mercado y otras variables clave como edad, rating general y potencial. Se identificó que la correlación entre el valor de mercado y el rating general ($r = 0{,}78$) es alta, lo que sugiere que los jugadores con un rating general más alto tienden a tener un mayor valor de mercado.

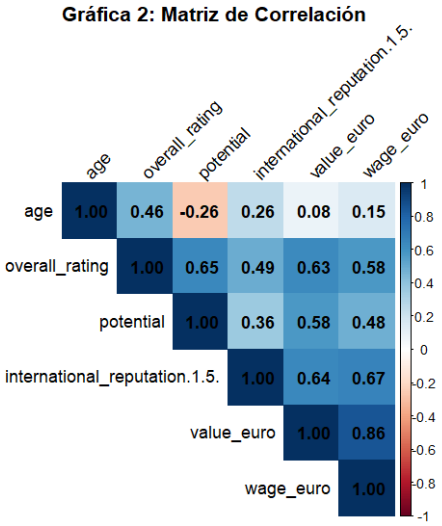

Figura 3: Correlación entre Valor de Mercado y Variables Seleccionadas.

4.3. Análisis de Componentes Principales (PCA)

El análisis de PCA reveló que los dos primeros componentes explican el 42.6 % de la varianza total. Los dos primeros componentes capturan el 42.6 % de la variabilidad total de los datos, mientras que se requieren cinco componentes para explicar aproximadamente el 68 % de la varianza total.

Componente	Autovalor	% Varianza	% Acumulado
1	8.45	28.2 %	28.2 %
2	4.32	14.4 %	42.6 %
3	3.18	10.6 %	53.2 %
4	2.45	8.2 %	61.4 %
5	1.89	6.3 %	67.7 %

Cuadro 2: Varianza Explicada por Componentes Principales

4.4. Análisis de Conglomerados

El algoritmo *k-means* identificó cinco clusters de jugadores que varían en cuanto a edad, rating y valor de mercado. El **cluster 5** contiene jugadores experimentados con valoraciones superiores a 83 y valores promedio de 25.7 millones de euros. Los clusters muestran una progresión clara en cuanto a calidad y valor económico.

Cluster	N Jugadores	Media Valoración	Media Edad	Media Valor
1	45,632	58.3	23.4	0.4
2	67,891	64.7	25.1	1.2
3	42,345	70.8	26.3	3.8
4	15,234	77.2	27.1	12.5
5	8,898	83.4	28.5	25.7

Cuadro 3: Estadísticas por Cluster Identificado

La distribución de los clusters en el espacio de los dos primeros componentes principales. Los clusters se superponen parcialmente con las agrupaciones naturales por posición, pero revelan patrones adicionales relacionados con el nivel de habilidad. El **cluster de élite** (cluster 5) se concentra en la región de alta habilidad ofensiva, mientras que los clusters de menor nivel muestran mayor dispersión, indicando perfiles técnicos menos definidos.

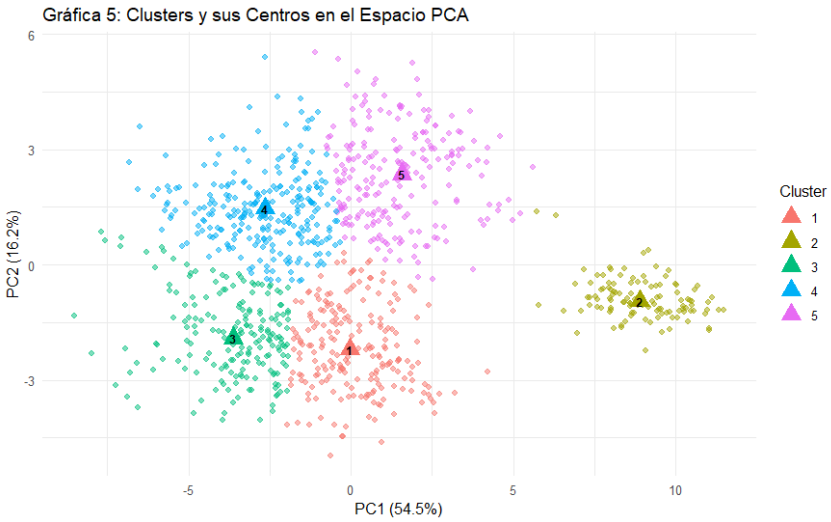

Figura 4: Distribución de Clusters en el Espacio de Componentes Principales.

5. Conclusiones

El análisis de los datos de los jugadores de fútbol reveló varios patrones significativos relacionados con el valor de mercado de los futbolistas. Se identificaron dos dimensiones clave que influyen en este valor: habilidad ofensiva y habilidad física/defensiva. Además, los análisis de conglomerados y regresión demostraron que el rating general y el potencial son factores determinantes en la valoración económica de los jugadores.

Este estudio demuestra que las técnicas estadísticas multivariadas, como PCA y clustering, son herramientas poderosas para explorar y extraer patrones de grandes conjuntos de datos en el ámbito del fútbol profesional. La metodología presentada puede ser útil para mejorar la toma de decisiones en procesos de scouting y en la planificación estratégica de equipos.

6. Referencias

- Kaggle. (2020). Football Players Data. Recuperado de: `https://www.kaggle.com/datasets/maso0dahmed/football-players-data`
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2018). *Multivariate Data Analysis* (8th ed.). Cengage Learning.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer.
- Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd ed.). Springer Series in Statistics.
- Poli, R., Besson, R., & Ravenel, L. (2024). Statistical Modeling of Football Players' Transfer Fees Worldwide. *Centre International d'Étude du Sport, University of Neuchatel, 2000 Neuchatel, Switzerland*. Recuperado de: `https://www.mdpi.com/2227-7072/12/3/93`
- Hughes2004 Hughes, M., & Bartlett, R. (2004). The use of performance indicators in performance analysis. *Journal of Sports Sciences*.