Pronóstico del rendimiento de jugadores de la NBA mediante Random Forest Regressor

Autor: Salomón García López

18 de noviembre de 2025

Resumen

Este trabajo presenta un estudio sobre la aplicación de un algoritmo de aprendizaje supervisado, **Random Forest Regressor**, con el fin de realizar un pronóstico del rendimiento de jugadores de la NBA. Se utiliza como variable objetivo la métrica *Game Score* (GmSc), la cual resume el impacto de un jugador en un partido individual. A partir de diversas estadísticas de juego —minutos, tiros de campo, triples, rebotes, asistencias y robos— se entrena un modelo para predecir dicho valor. Se discute el modelo matemático, las métricas de error empleadas (MAE, MSE, RMSE, MAPE) y los resultados obtenidos. Los modelos basados en árboles, como Random Forest, demostraron ser adecuados para capturar relaciones no lineales en los datos deportivos.

1. Introducción

El análisis predictivo en el deporte profesional se ha convertido en una herramienta fundamental para la evaluación de talento, estrategias de rotación y proyecciones de desempeño. En la NBA, donde cada partido genera un conjunto amplio de estadísticas, los métodos supervisados permiten construir modelos que anticipen métricas claves como el rendimiento global del jugador.

Este trabajo aborda la predicción de la métrica **Game Score (GmSc)**, una representación numérica del impacto del jugador en un partido, mediante un modelo supervisado basado en Random Forest. A diferencia de métodos lineales tradicionales, Random Forest permite capturar patrones no lineales y relaciones complejas entre variables.

2. Modelo matemático: Random Forest Regressor

Random Forest [1] es un método basado en *ensambles* de árboles de decisión. Para un conjunto de entrenamiento

$$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\},$$

el algoritmo construye múltiples árboles $T_1, T_2, \ldots, T_M$ mediante:

1. **Bootstrap:** muestreo con reemplazo para cada árbol.
2. **Selección aleatoria de variables:** en cada partición del árbol se evalúa solo un subconjunto aleatorio de características.

Cada árbol genera una predicción $T_m(x)$ y la predicción final del bosque es:

$$\hat{y} = \frac{1}{M} \sum_{m=1}^{M} T_m(x)$$

Este esquema disminuye la varianza respecto a un solo árbol y mejora la capacidad de generalización.

3. Marco teórico sobre métricas de desempeño y diseño experimental

En problemas de regresión, como la predicción del rendimiento de jugadores de la NBA mediante la métrica Game Score (GmSc), es fundamental contar con indicadores cuantitativos que permitan evaluar la calidad de los modelos. La literatura de aprendizaje automático y estadística aplicada recomienda el uso de diversas métricas de error, tales como el Error Absoluto Medio (MAE), el Error Cuadrático Medio (MSE), su raíz cuadrada (RMSE) y el Error Porcentual Absoluto Medio (MAPE), además del coeficiente de determinación R^2 [**?**, **?**].

3.1. Métricas de desempeño

Sea y_i el valor real de la variable objetivo (GmSc) para la observación i, y $\hat{y}_i$ la predicción del modelo. Las métricas empleadas se definen como:

- **Error Absoluto Medio (MAE):**

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

  Mide el error promedio en unidades de la variable original y es robusto ante valores atípicos moderados.

- **Error Cuadrático Medio (MSE):**

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

  Penaliza con mayor fuerza los errores grandes debido al cuadrado de la diferencia.

- **Raíz del Error Cuadrático Medio (RMSE):**

  $$RMSE = \sqrt{MSE}$$

  Es la métrica más utilizada en muchos estudios por mantenerse en la misma escala que y_i y reflejar fuertemente los errores grandes.

- **Error Porcentual Absoluto Medio (MAPE):**

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

  Expresa el error en términos porcentuales, lo que facilita su interpretación en algunos contextos, aunque puede ser inestable si existen valores de y_i muy cercanos a cero.

- **Coeficiente de determinación (R^2):**

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

  Indica la proporción de variabilidad de la variable objetivo explicada por el modelo [**?**].

En este trabajo se emplean principalmente MAE y RMSE como métricas de referencia para comparar configuraciones de modelos supervisados, mientras que R^2 se utiliza como indicador adicional de ajuste global.

3.2. Diseño de experimentos para modelos de pronóstico

Además de seleccionar métricas de desempeño, es necesario definir un *diseño de experimentos* que permita comparar distintas configuraciones del modelo y cuantificar su efecto sobre el error de predicción. Siguiendo la literatura de diseño factorial en modelos de ensamble [**?**], se definieron los siguientes factores para el algoritmo Random Forest Regressor:

- **Factor A:** número de árboles en el bosque ($n\_estimators$), con niveles $\{50, 100, 200\}$.
- **Factor B:** profundidad máxima de los árboles (max_depth), con niveles $\{\text{None}, 5, 10\}$.

La combinación de estos factores da lugar a un diseño factorial 3×3, es decir, nueve tratamientos diferentes. Para cada tratamiento se entrenó un modelo Random Forest con la configuración correspondiente y se evaluó su desempeño sobre un conjunto de prueba utilizando las métricas MAE, MSE, RMSE, MAPE y R^2.

El diseño experimental permite:

1. Identificar la configuración que minimiza el error (por ejemplo, RMSE).
2. Observar la sensibilidad del modelo a los cambios en el número de árboles y la profundidad máxima.
3. Comparar el modelo seleccionado con baselines más simples, como la regresión lineal, dentro del mismo esquema de evaluación.

Este enfoque sistemático, apoyado en métricas estandarizadas y un diseño factorial claro, proporciona una base sólida para seleccionar el modelo supervisado más adecuado para el pronóstico del rendimiento de jugadores NBA.

4. Metodología

La metodología utilizada en este estudio combina técnicas de aprendizaje no supervisado y supervisado para analizar y pronosticar el rendimiento de jugadores de la NBA. El flujo de trabajo se desarrolló en cuatro etapas principales: (1) selección y preparación del conjunto de datos, (2) preprocesamiento y estandarización de variables, (3) modelado mediante algoritmos sustentados en la literatura especializada y (4) evaluación cuantitativa y visual de los resultados.

4.1. Conjunto de datos

Se empleó un conjunto de datos correspondiente a partidos de la temporada 2024–2025 de la NBA, que contiene estadísticas individuales de desempeño por jugador. Entre las variables consideradas se incluyen minutos jugados (MP), puntos anotados (PTS), eficiencia de tiro (FG %, 3P %, FT %), rebotes (ORB, DRB, TRB), asistencias (AST), robos (STL), tapones (BLK), pérdidas de balón (TOV) y la métrica *Game Score* (GmSc), propuesta por Hollinger [3].

4.2. Preprocesamiento

Se realizaron las siguientes transformaciones sobre los datos:

- **Eliminación de valores nulos:** Se descartaron registros con información incompleta para garantizar consistencia estadística.
- **Estandarización:** Todas las variables se escalaron mediante *StandardScaler* para obtener media cero y desviación estándar uno. Esta normalización es esencial en algoritmos que dependen de distancias o combinaciones no lineales de características, como DBSCAN y Random Forest [2].
- **Reducción de dimensionalidad:** Se aplicó Análisis de Componentes Principales (PCA) [**?**] con dos componentes exclusivamente para propósitos de visualización, manteniendo el espacio original para el entrenamiento de los modelos.

4.3. Algoritmo no supervisado: DBSCAN

Para la exploración inicial de estructuras subyacentes se utilizó **DBSCAN** (Density-Based Spatial Clustering of Applications with Noise), propuesto por Ester et al. [**?**]. Este método identifica regiones de alta densidad en el espacio de características mediante dos parámetros: el radio de vecindad ε y el número mínimo de puntos *minPts* requeridos para formar un clúster.

DBSCAN es especialmente útil en datos deportivos, donde existen partidos atípicos (muy buenos o muy malos) que pueden comportarse como ruido. La elección de ε se realizó mediante el método de la *k-distancia*, graficando las distancias al k-ésimo vecino más cercano hasta localizar el punto de inflexión o "codo" de la curva.

4.4. Algoritmo supervisado: Random Forest Regressor

Para el pronóstico del rendimiento se utilizó el algoritmo supervisado **Random Forest Regressor**, un método de ensamble basado en múltiples árboles de decisión, propuesto por Breiman [1]. Su estructura permite modelar relaciones no lineales entre las variables y reduce el sobreajuste mediante:

1. Muestreo con reemplazo (*bootstrap*) para generar múltiples subconjuntos de datos.
2. Selección aleatoria de características en cada nodo del árbol.
3. Promedio de las predicciones individuales de los árboles:

$$\hat{y} = \frac{1}{M} \sum_{m=1}^{M} T_m(x)$$

Random Forest resulta especialmente adecuado para datos NBA debido a la complejidad de las interacciones entre estadísticas como puntos, eficiencia, rebotes y minutos de juego.

4.5. Evaluación del modelo

El conjunto de datos se dividió en entrenamiento (75 %) y prueba (25 %). Para evaluar el desempeño del modelo se utilizaron métricas ampliamente aceptadas en la literatura para regresión:

- **Error Absoluto Medio (MAE)**
- **Error Cuadrático Medio (MSE)**
- **Raíz del Error Cuadrático Medio (RMSE)**
- **Error Porcentual Absoluto Medio (MAPE)**

Además, se generaron visualizaciones como:

- Gráfica de valores reales vs predichos.
- Residuales (error = real – predicción).
- Importancia de características según Random Forest.
- Serie temporal ordenada de valores reales y predichos.

Estas representaciones permiten analizar tanto la calidad del ajuste como la estructura de error y los factores estadísticos que más influyen en la predicción del rendimiento.

5. Resultados

5.1. Predicción del Game Score

La figura 1 muestra la relación entre valores reales y predichos. Los puntos se agrupan alrededor de la línea de identidad, lo que indica un desempeño adecuado del modelo.

5.2. Métricas de error

El desempeño cuantitativo del modelo fue el siguiente:

- MAE = 0.115
- MSE = 0.028
- RMSE = 0.167
- MAPE = 64.362

5.3. Residuales

La distribución de residuales no muestra patrones evidentes, sugiriendo que el modelo no presenta sesgos sistemáticos.

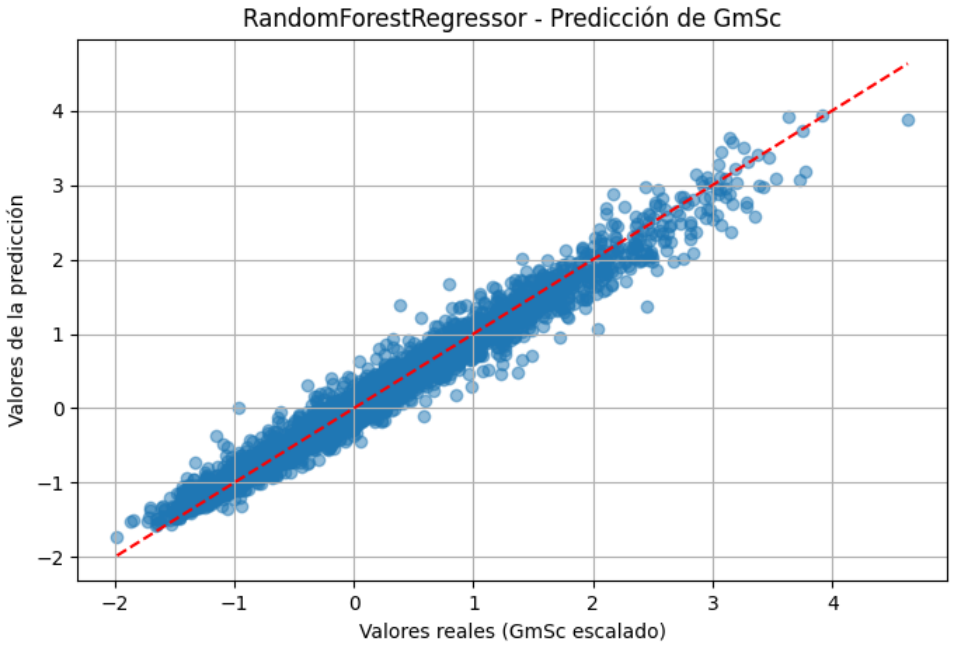

Figura 1: Valores reales vs predicción del modelo Random Forest para GmSc (escalado).

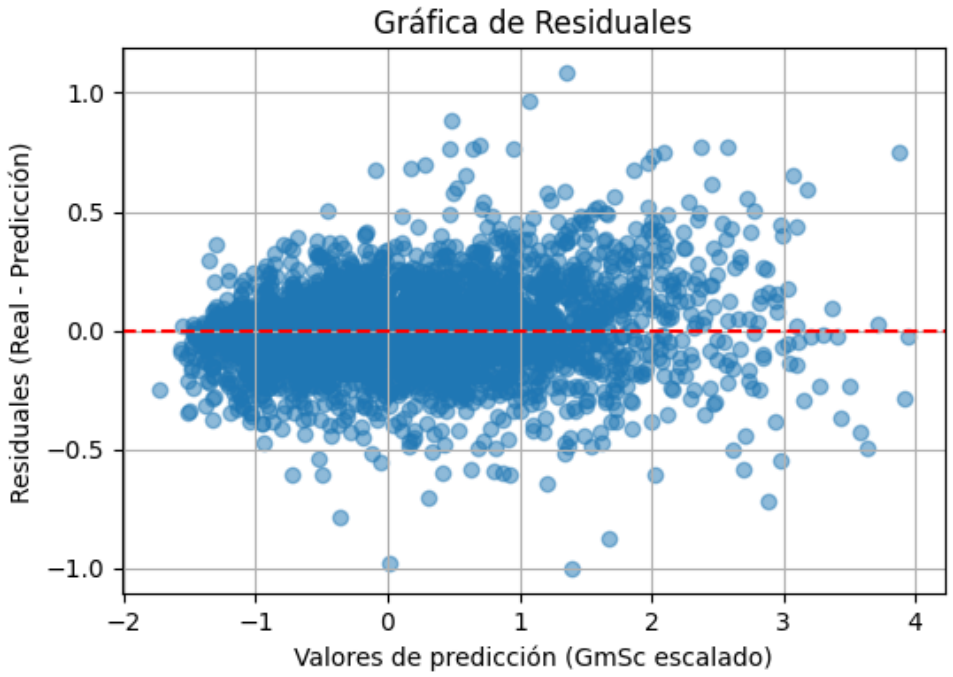

Figura 2: Gráfica de residuales Real - Predicción.

Figura 3: Importancia de las variables según Random Forest.

5.4. Importancia de características

La figura 3 muestra la contribución de cada variable según su importancia en el modelo.

5.5. Serie Real vs Predicción

6. Discusión

Los resultados sugieren que los modelos basados en árboles son adecuados para datos NBA, donde las relaciones entre estadísticas son inherentemente no lineales. La pequeña diferencia entre valores reales y predichos indica que Random Forest captura correctamente patrones de rendimiento.

El análisis de importancia de variables confirma que los puntos anotados, el tiempo en cancha y la eficiencia de tiro son los factores más determinantes del impacto global del jugador.

7. Conclusiones

El uso de Random Forest permitió predecir de forma razonablemente precisa el Game Score de jugadores NBA. Este tipo de modelos puede integrarse en sistemas de análisis deportivo para ofrecer métricas esperadas de rendimiento, apoyar decisiones tácticas o incluso estimar eficiencia proyectada.

Como trabajo futuro se propone comparar el desempeño del modelo con regresión lineal, Gradient Boosting y Redes Neuronales.

Referencias

[1] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

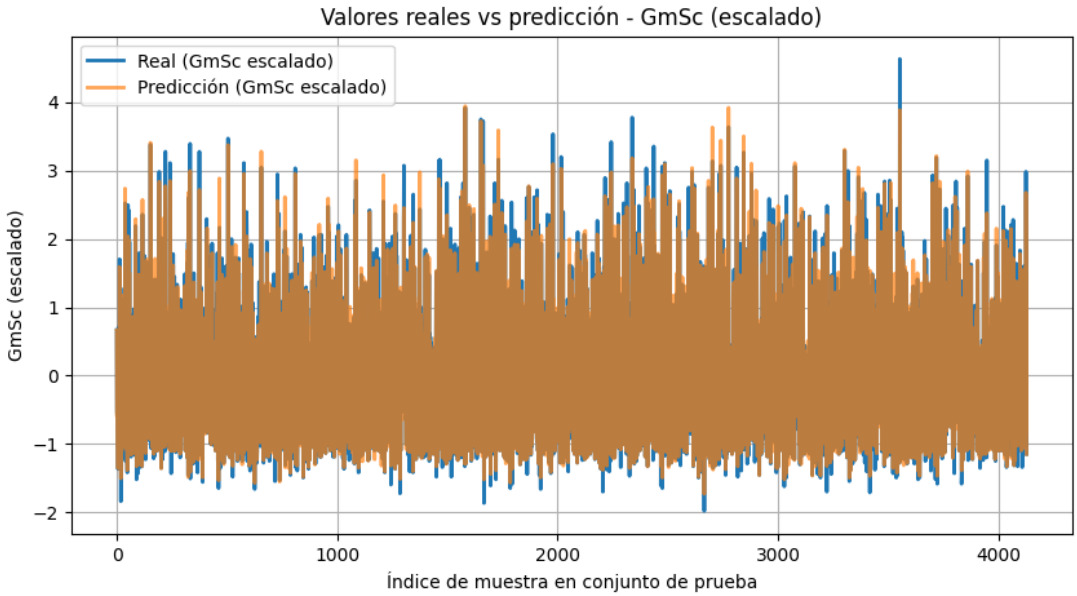

Figura 4: Comparación entre valores reales y predichos de GmSc.

[2] Scikit-learn developers, "RandomForestRegressor Documentation," 2024. Disponible en: `https://scikit-learn.org/stable/`

[3] J. Hollinger, "Game Score Metric," *ESPN Basketball Analytics*, 2011.