

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

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

el algoritmo construye múltiples árboles T_1, T_2, \dots, 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. Metodología

3.1. Datos

Se emplearon estadísticas de jugadores NBA del periodo 2024–2025. Las variables predictoras incluyeron:

- Minutos: MP.
- Tiros: FG, FGA, FG %, 3P, 3PA, 3P %, FT, FTA, FT %.
- Rebotes: ORB, DRB, TRB.
- Asistencias (AST), robos (STL), tapones (BLK).
- Pérdidas (TOV), faltas (PF) y puntos anotados (PTS).

La variable objetivo fue **GmSc**.

3.2. Preprocesamiento

Debido a las diferencias de escala entre variables (por ejemplo, rebotes vs porcentajes), se aplicó *StandardScaler* normalizando todas las columnas a media cero y varianza uno. Posteriormente se dividieron los datos en entrenamiento (75 %) y prueba (25 %).

3.3. Entrenamiento del modelo

El modelo seleccionado fue un **RandomForestRegressor** con los hiperparámetros por defecto e inicialización fija (*random_state = 42*). El entrenamiento consistió en:

1. Ajustar el bosque al conjunto de entrenamiento.
2. Generar predicciones sobre el conjunto de prueba.
3. Evaluar el desempeño mediante métricas de error.

4. Resultados

4.1. 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.

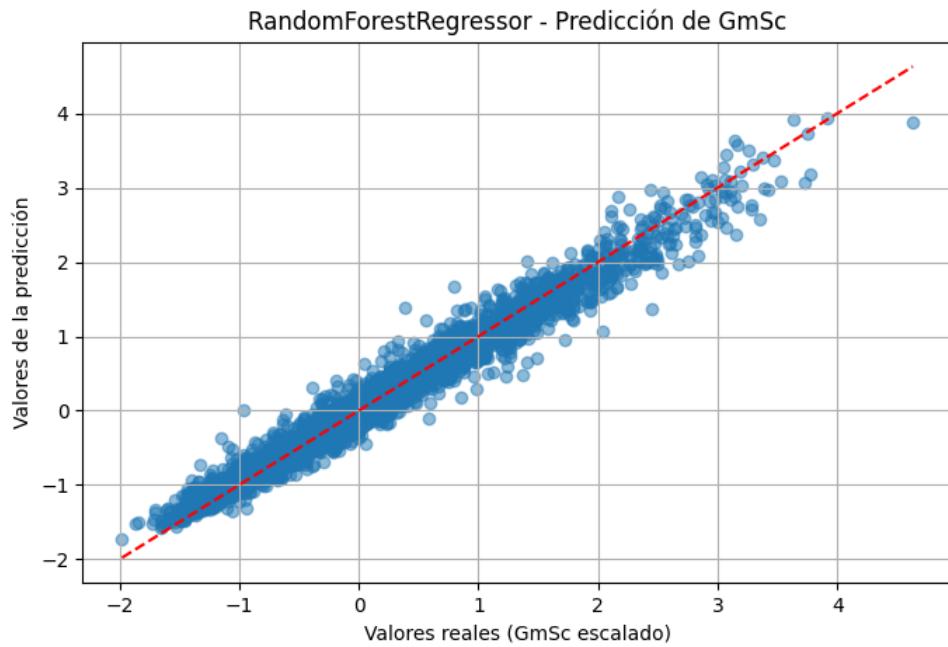


Figura 1: Valores reales vs predicción del modelo Random Forest para GmSc (escalado).

4.2. Métricas de error

El desempeño cuantitativo del modelo fue el siguiente:

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

4.3. Residuales

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

4.4. Importancia de características

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

4.5. Serie Real vs Predicción

5. 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.

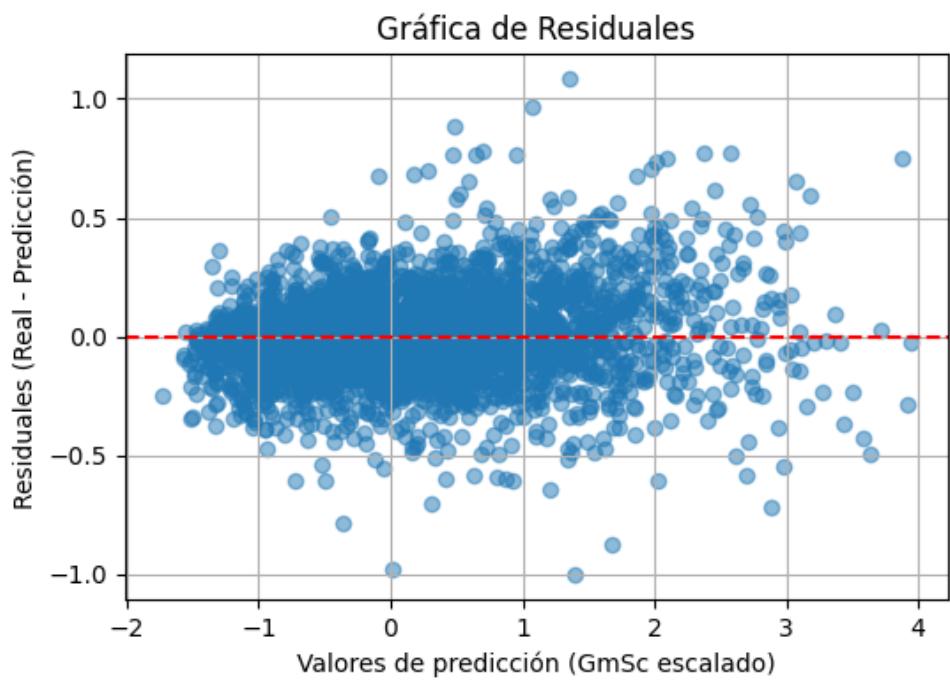


Figura 2: Gráfica de residuales Real - Predicción.

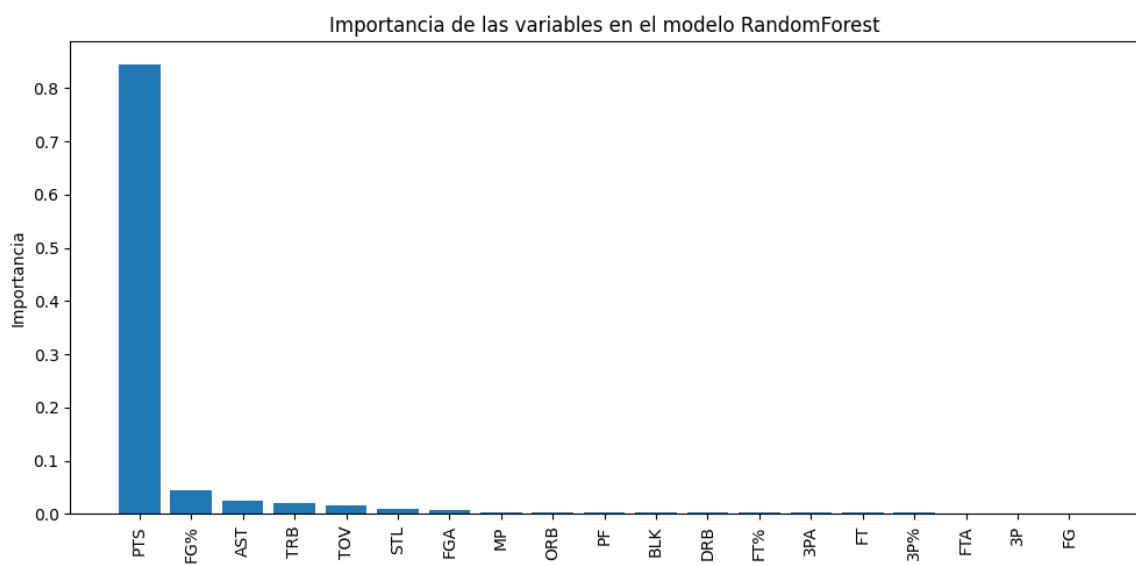


Figura 3: Importancia de las variables según Random Forest.

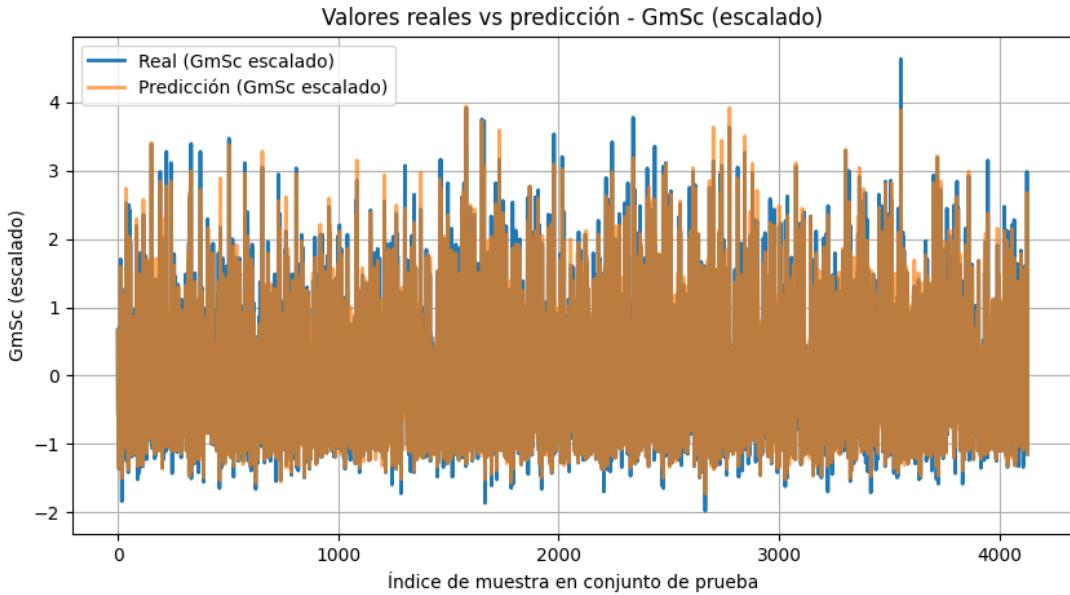


Figura 4: Comparación entre valores reales y predichos de GmSc.

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.

6. 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.
- [2] Scikit-learn developers, “RandomForestRegressor Documentation,” 2024. Disponible en: <https://scikit-learn.org/stable/>
- [3] J. Hollinger, “Game Score Metric,” *ESPN Basketball Analytics*, 2011.