NBA SEASON DATA

ANÁLISIS DEL RENDIMIENTO DE LOS JUGADORES DE LA NBA POR TEMPORADA DE 1978-2016

09 de enero del 2024

Elaborado por
Fábio Marques

Resumen

Este proyecto de análisis de datos se enfoca en explorar las dinámicas del rendimiento de los jugadores de la NBA a lo largo de cuatro décadas (1978-2016), con el propósito de comprender la influencia de factores como la edad, el true shooting y características físicas en el desempeño individual y colectivo de los equipos.

INTRODUCCIÓN

La NBA ha sido un escenario de cambios y transformaciones constantes a lo largo de su historia. Este proyecto tiene como objetivo arrojar luz sobre la relación entre el rendimiento de los jugadores y factores clave como la edad, el true shooting y otras características físicas y de experiencia.

CONTEXTO COMERCIAL

La National Basketball Association (NBA) es una de las ligas deportivas más populares y competitivas del mundo, con millones de seguidores y un mercado en constante crecimiento. La capacidad de evaluar y comprender el rendimiento de los jugadores es fundamental para los equipos, entrenadores y directores generales, ya que les permite tomar decisiones informadas para mejorar el desempeño del equipo, tomar decisiones de contratación y maximizar el valor de los jugadores.

PROBLEMA COMERCIAL

Evaluar el rendimiento de los jugadores tanto indivudualmente como dentro del equipo, identificar fortalezas y debilidades, y tomar decisiones estratégicas sobre la alineación y la gestión de minutos.

Para la gestión de personal y la toma de decisiones en el draft, es importante evaluar el potencial de los jugadores. Los equipos pueden utilizar datos históricos y estadísticas avanzadas para identificar talentos emergentes y tomar decisiones de contratación informadas.

OBJETIVOS

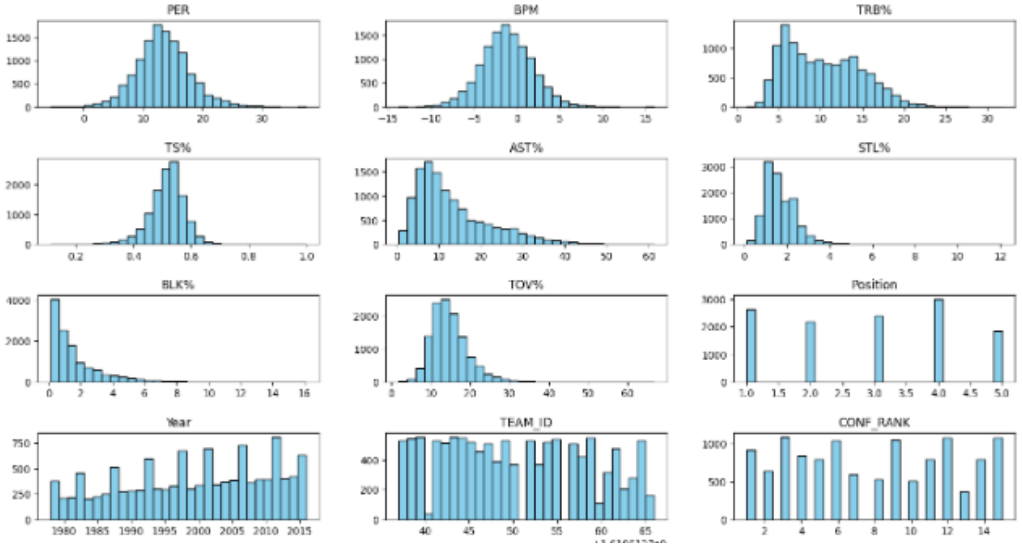
Predecir el rendimiento de un jugador con base a la posición en la que juega. Las variables de rendimiento que se usaran son tanto globales como el PER (Player Efficiency Rating) y el BPM (Box Plus-Minus) como individuales como ts(True Shooting Percentage) , trb (Total Rebound Percentage), ast (Assist Percentage), stl (Steal Percentage), blk (Block Percentage), tov (Turnover Percentage) del jugador.

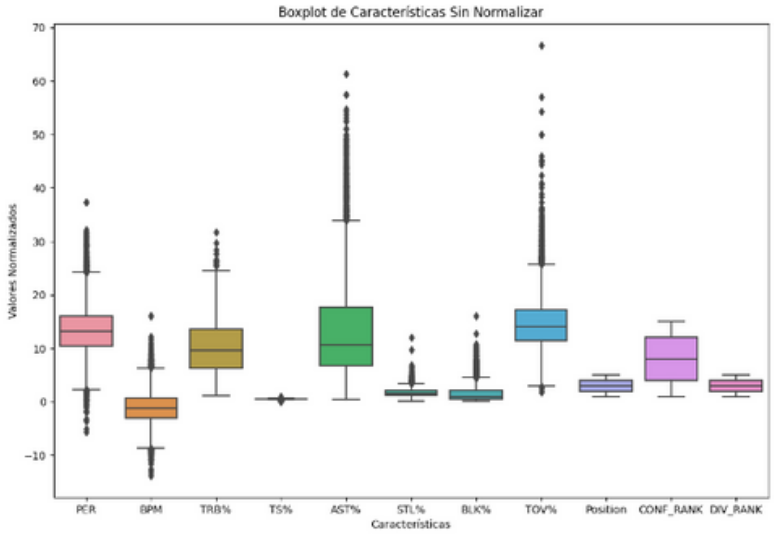
HIPÓTESIS

Hipótesis Nula (H0): La posición que un jugador ocupa en el campo (por ejemplo, base, escolta, alero, ala-pívot, pívot) no tiene un impacto significativo en su rendimiento en términos de Player Efficiency Rating (PER). En otras palabras, la posición no influye en el PER promedio de los jugadores en la NBA.

Hipótesis Alternativa (H1): La posición que un jugador ocupa en el campo (por ejemplo, base, escolta, alero, ala-pívot, pívot) tiene un impacto significativo en su rendimiento en términos de Player Efficiency Rating (PER). En otras palabras, la posición influye en el PER promedio de los jugadores en la NBA.

Análisis Univariado



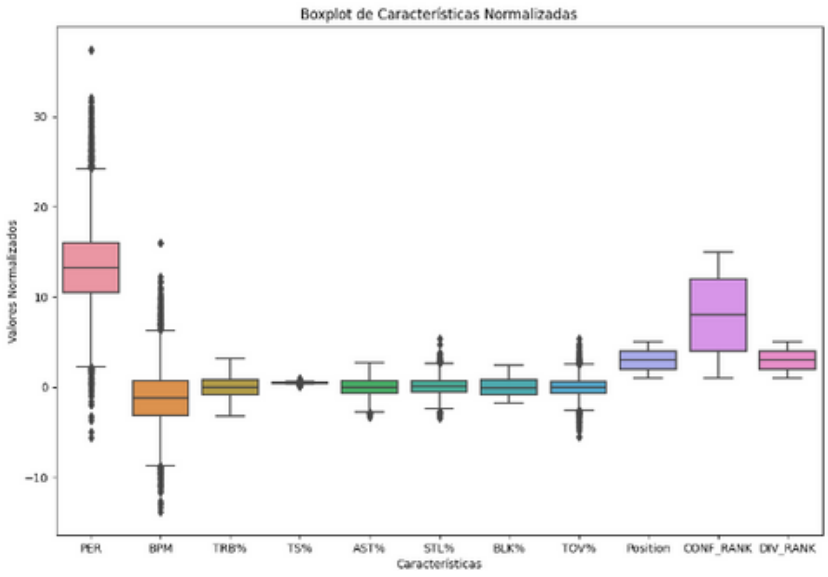


En este resumen descriptivo podermos ver que alunos de nuestros datos no están normalizados

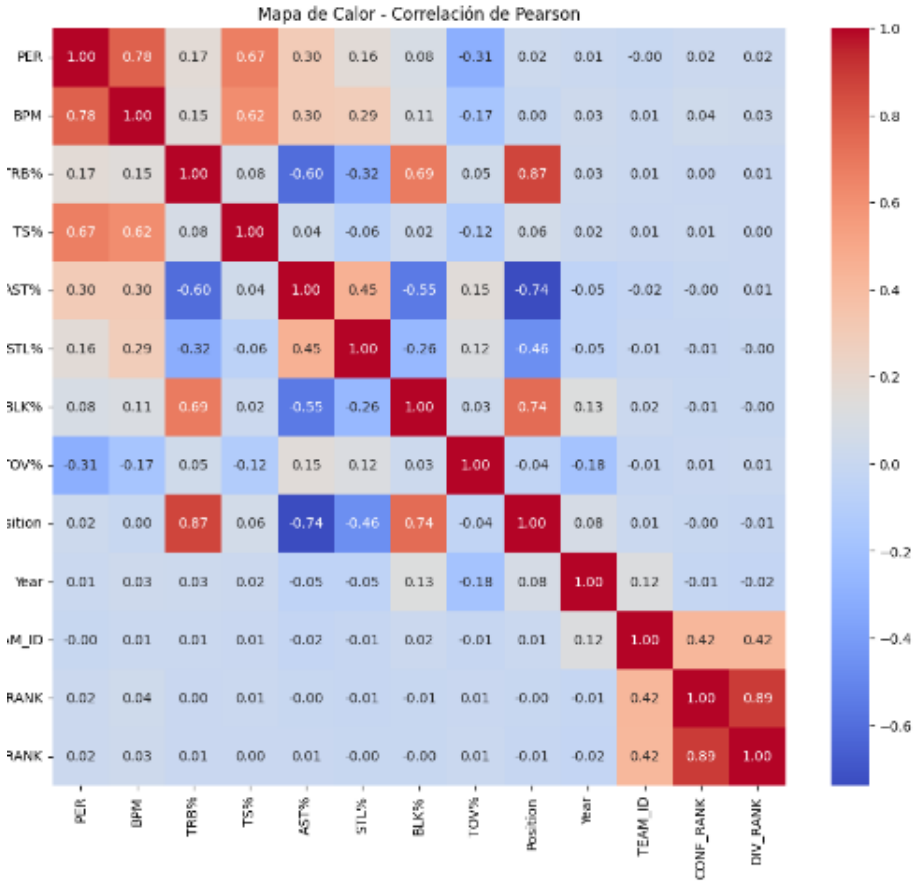
1. Media: La media de cada variable no parece ser cercana a cero. En un conjunto de datos normalizado, la media debería estar cerca de cero.
2. Desviación Estándar (Std):La desviación estándar no parece cercana a uno para todas las variables. En datos normalizados, la desviación estándar debería ser cercana a uno.
3. Rango (Min, Max): El rango de cada variable varía significativamente. En datos normalizados, los valores deberían estar en una escala similar.

Así que el siguiente paso es normalizar los datos para poder usar nuestros modelos.

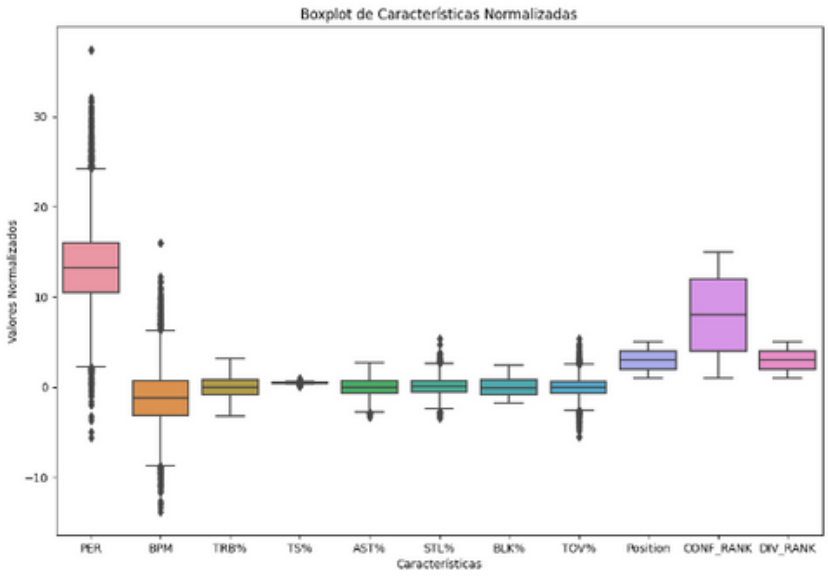
Datos después de la normalización



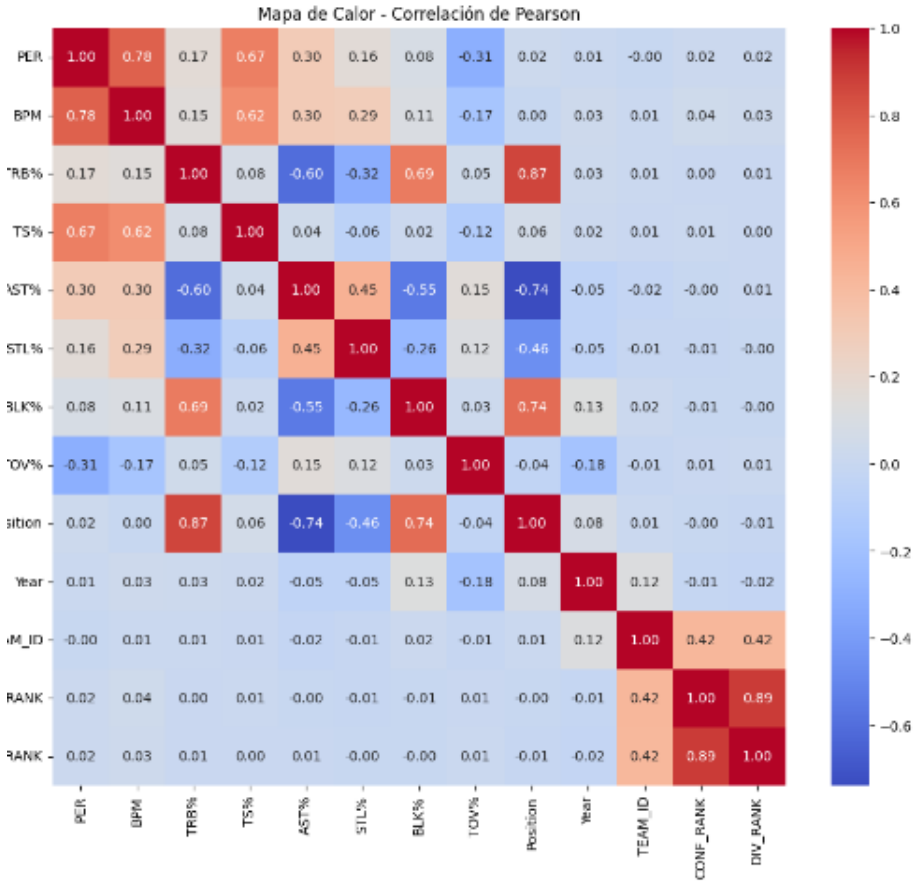
Correlación de Pearson



Datos después de la normalización



Correlación de Pearson



Modelos de Aprendizaje Usados

REGRESIÓN LINEAL
KNN
XGBOOST CON KFOLD VALIDATION

Conclusión

En este estudio, se evaluaron tres modelos diferentes para predecir el rendimiento (PER y BPM) de los jugadores de NBA en función de diversas métricas. Los modelos considerados fueron Regresión Lineal, K Vecinos Más Cercanos (KNN) y XGBoost.

Primero que nada se observó que los datos no estaban normalizados, así que se realizó una normalización por el método Power Transformer. Una vez hecha la transforamción de los datos se procedión a realizar una elección de variables a usar por medio de un Feature Selection, en específico se utilizó el método de Forward Selection. Después de aplicar este método usamos una matriz de correlación para poder evaluar las características que más correlación tenían entre sí.

Con estos datos en mano, se procedió a realizar los Modelos de Aprendizaje.

En cuanto a la Regresión Lineal, los resultados indican que el modelo tiene un rendimiento modesto. El Mean Squared Error (MSE) promedio fue de aproximadamente 0.727, lo que sugiere que hay una variabilidad considerable en las predicciones. Además, el coeficiente de determinación R2 promedio fue de alrededor del 27%, indicando que el modelo explica aproximadamente el 27% de la variabilidad en los datos que es un valor significativamente muy bajo.

El modelo KNN mostró un rendimiento similar, con un MSE promedio de aproximadamente 0.845 y un R2 promedio de alrededor del 15% (aún más bajo que en la regresión lineal). Esto sugiere que el modelo KNN tiene dificultades para capturar la complejidad de los datos y realizar predicciones precisas.

Por último, el modelo XGBoost presentó resultados intermedios con un MSE promedio de alrededor de 0.749 y un R2 promedio de aproximadamente el 25%. Aunque muestra una mejora respecto a los modelos anteriores, aún hay margen para optimizar su rendimiento.

En resumen, los modelos evaluados proporcionan una estimación del rendimiento de los jugadores, pero existe la necesidad de mejorar la precisión de las predicciones. Esto podría lograrse mediante la reconsideración de las características usadas, la inclusión de características adicionales o la exploración de modelos más avanzados.