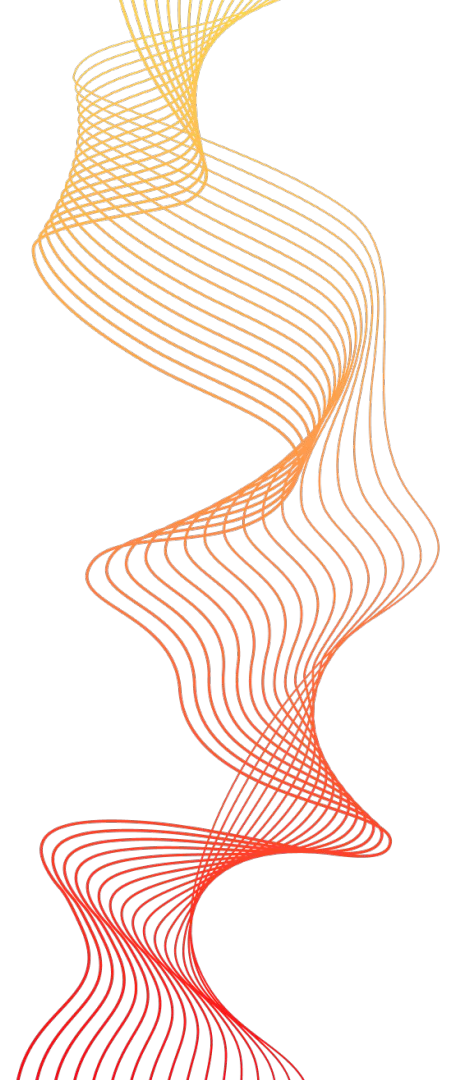




CoderHouse


# Predicción carrera Rookie NBA

Autor: Federico Gil Braile





# Análisis de Datos NBA

- Partimos de un dataset de clasificación binaria, en el que contamos con las estadísticas de jugadores jóvenes NBA
  - Nuestro objetivo es predecir si un jugador alcanzará una carrera de 5 años o más.
- 

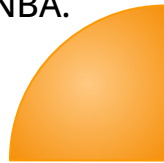


# Motivación y Audiencia

La NBA es la liga de basketball más importante e influyente del mundo, comprender los factores que influyen en la duración de la carrera de un jugador puede ser de gran interés para los entrenadores, áreas de scouting, equipos, empresarios, etc.

La capacidad de predecir si un jugador joven puede tener una carrera duradera en la NBA puede tener implicaciones significativas tanto para los jugadores como para los equipos, establece puntos de comparación, niveles de exigencia.

Esta presentación está dirigida a profesionales del basket, entrenadores, empresarios, analistas y fanáticos del deporte que estén interesados en utilizar el poder del análisis de datos para obtener una comprensión más profunda de la longevidad de los jugadores en la NBA.





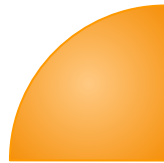
# Resumen de metadatos

El conjunto de datos analizado tiene información sobre 1340 jugadores jóvenes de la NBA y sus estadísticas de juego.

La variable de respuesta es binaria, indica si un jugador alcanzó una carrera de 5 años o más.

Contamos con un total de 19 variables entre las que destacan partidos jugados, minutos jugados, puntos y goles, entre otros.

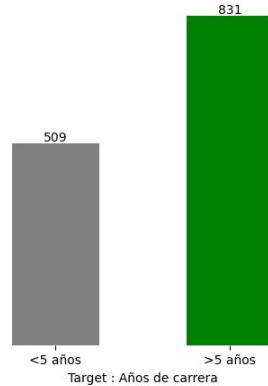
No hay datos faltantes o incorrectos y todas las variables son numéricas.



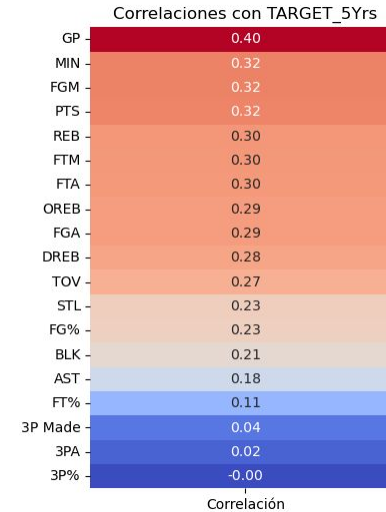
# Preguntas ,hipótesis y visualizaciones

¿Cuántos jugadores cumplieron con el target?

¿Cuántos jugadores cumplieron con el target?

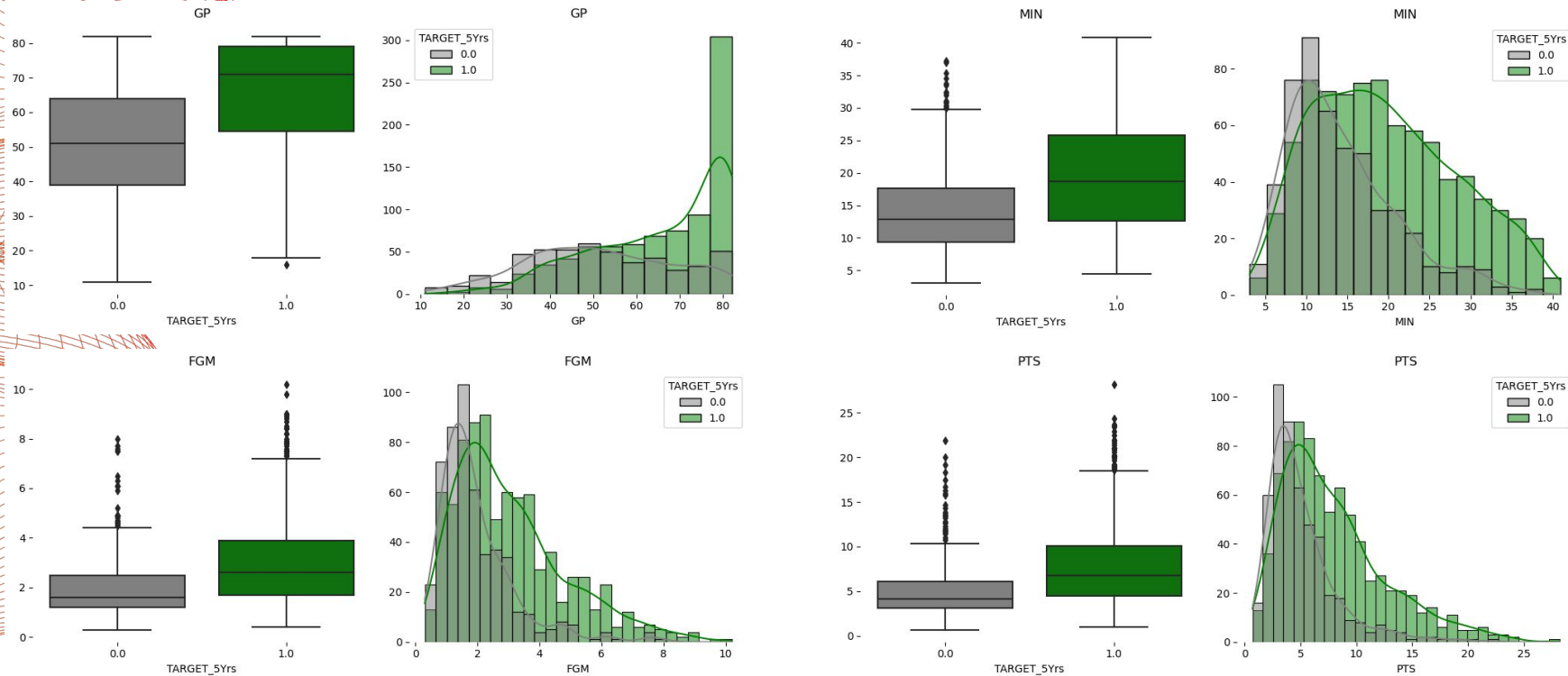


¿Cuales son las variables que mayor correlación tienen vs la variable de rpt?



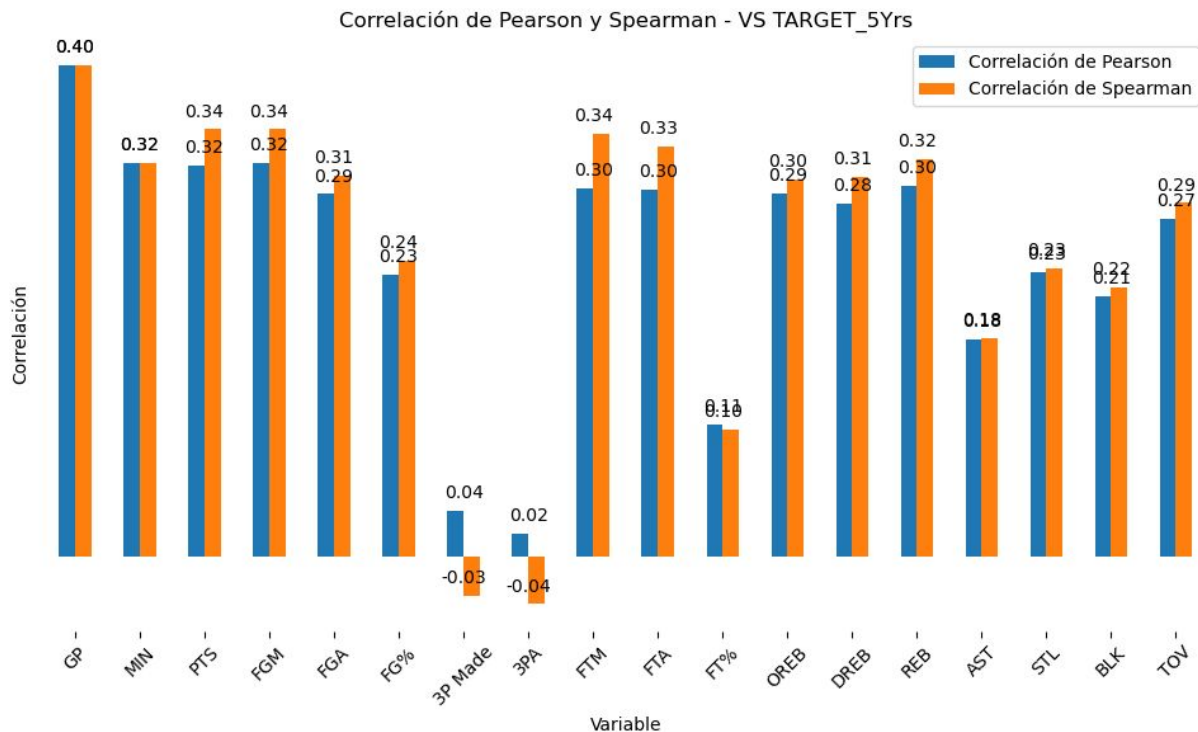
# Preguntas ,hipótesis y visualizaciones

¿Cómo se comportan las 4 variables con mayor correlación?



# Preguntas ,hipótesis y visualizaciones


Dado el comportamiento de las variables... debería considerar la correlación de Pearson o Spearman?







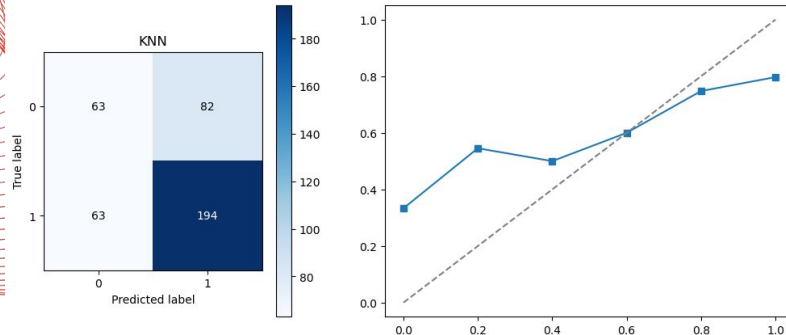
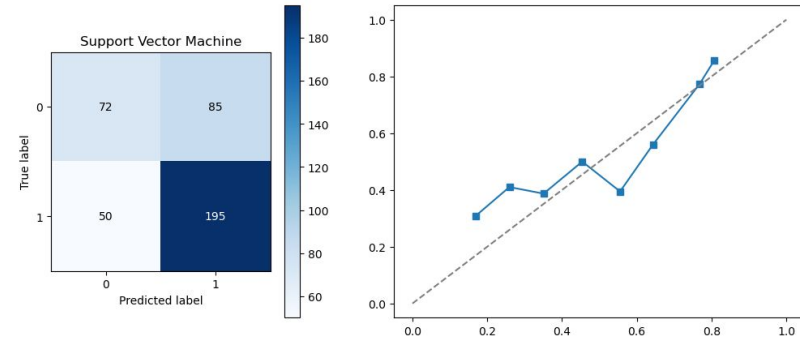
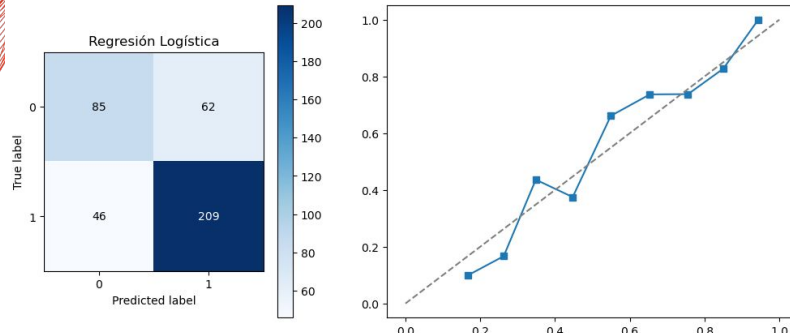
# Insights

- El 62% de los rookies de nuestro dataset, alcanzaron el objetivo de jugar  $> 5$  años en la NBA
  - Partidos, minutos jugados, goles y puntos realizados son las variables que más influyen sobre el target
    - Aquellos jugadores que cumplieron el objetivo tienden a tener mejores "valores" para cada variable vs aquellos que no cumplieron el objetivo
  - Por el contrario, las variables relacionadas con tiros de 3 puntos, no guardan relación con la variable de respuesta
  - Dado el comportamiento de nuestras variables, refiero a que todas cuentan con outliers, sobre todo al analizarlo vs target, la correlación de Spearman es la más acertada
- 



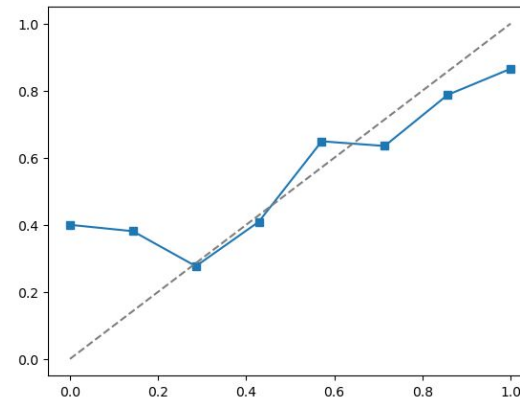
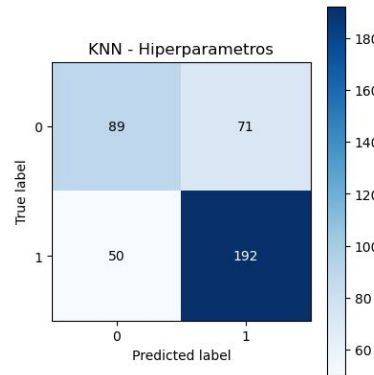
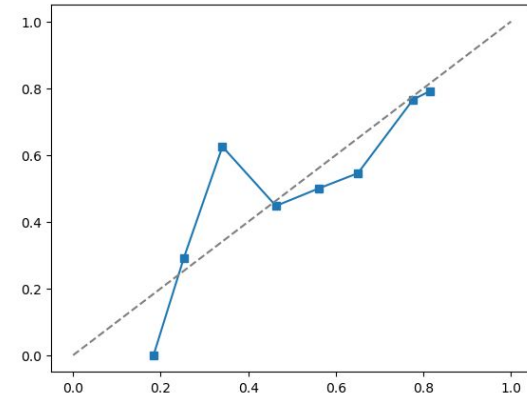
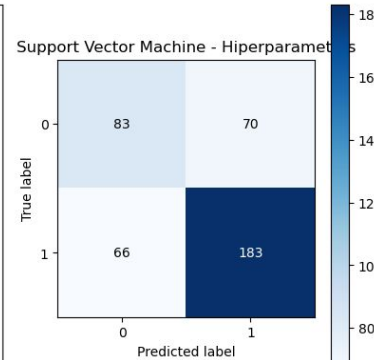
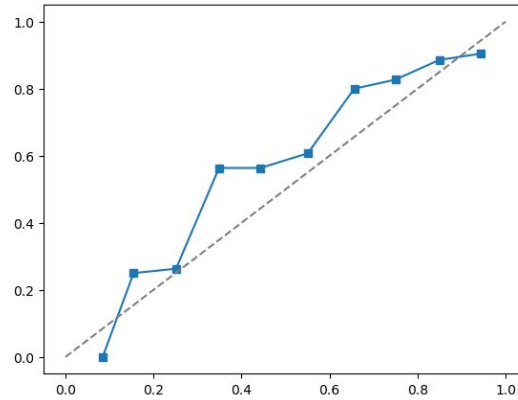
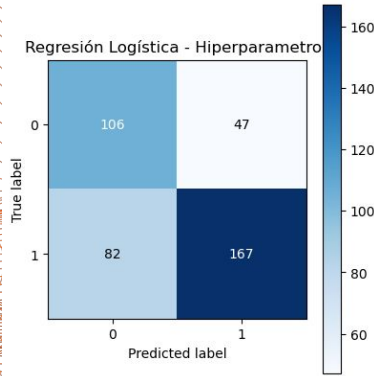
# Modelos Machine Learning

Evaluaremos 6 modelos diferentes de ML, con el objetivo de ver su performance y calibración (sin ajustar hiperparametros)



Estos 3 son los modelos que mayor calibración y métricas de performance tienen de los evaluados, por lo que procedere a ajustar los hiperparametros

# Modelos Machine Learning - Hiperparametros ajustados





# Insights modelos ML

- Se observan cambios de performance y calibración en todos los modelos evaluados, al correr el código en diferentes ocasiones
    - Esto se debe, principalmente, al tamaño del dataset utilizado.
  - A pesar de lo mencionado, Regresión logística es el modelo que tiende a tener mejor performance de predicción como así también una curva más calibrada para la resolución de esta necesidad
  - Para poder ser concluyente deberíamos contar con un mayor muestreo de datos, de todas formas considero que el mejor modelo según lo trabajado es **Regresión Logística**
- 