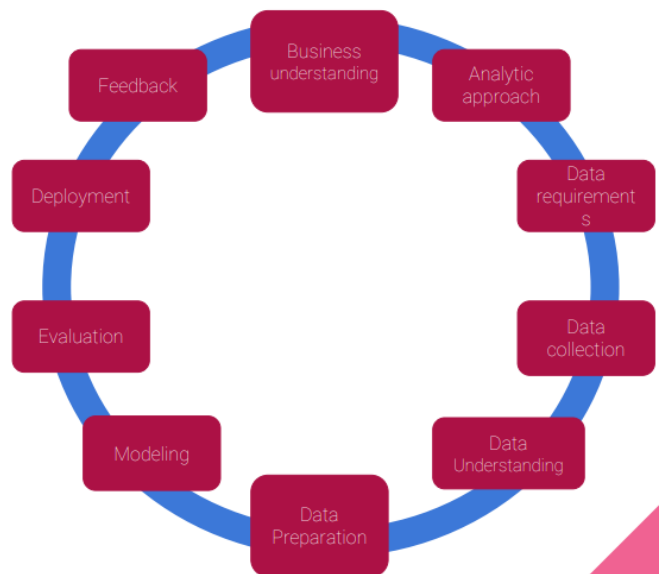


# NEOLAND

## Modelo predictivo play-offs de la NBA



Leonardo Bustamante Urbina

Juan Carlos Sanz González

Bootcamp Data Science, Abril 2021

# Índice

1. Introducción.....	4
1.1 Formato de la NBA.....	5
1.2 Objetivo .....	6
1.3 Planteamiento .....	6
1.4 Descripción del proyecto.....	6
2. Análisis exploratorio de los datos.....	8
3. Modelos.....	14
3.1 No supervisados.....	14
3.2 Supervisados.....	14
3.3 Preparación de los datos para el modelo.....	14
3.4 Elección del conjunto de prueba y entrenamiento.....	15
4. Resultados.....	16
5. Conclusiones.....	17
6. Futuras investigaciones.....	18

## **Abreviaciones de las estadísticas por jugador**

Min: Minutos

FGM: Tiros de campo anotados

FGA: Tiros de campo intentados

FG%: Porcentaje de tiros anotados de campo

3PM: Tiros de tres puntos anotados

3PA: Tiros de tres puntos intentados

3P%: Porcentaje de tiros anotados de tres puntos

FTM: Tiros libres anotados

FTA: Tiros libres intentados

FT%: Porcentaje de tiros libres anotados

OREB: Rebotes ofensivos

DREB: Rebotes defensivos

REB: Rebotes totales

AST: Asistencias dadas

STL: Robos

BLK: Tapones

PF: Faltas personales provocadas

PTS: Puntos anotados.

PLUS\_MINUS +/-: Resultado que ha hecho tu equipo mientras un jugador estaba en pista.

# 1. Introducción

El sector de los juegos de azar, específicamente el de las apuestas deportivas, ha sufrido un incremento importante, más aún con el aumento de plataformas web, que facilitan la práctica de esta actividad desde cualquier lugar.

Las páginas de apuestas deportivas ofrecen una gran variedad de posibilidades para los usuarios, de las que se podría sacar un beneficio económico. Dentro de los deportes más populares está el baloncesto, con el que se pueden hacer diferentes tipos de apuestas, desde predecir el ganador o perdedor hasta cuántos puntos puede hacer un jugador particular en determinado momento de un juego.

En el baloncesto se ha visto un creciente interés por las estadísticas de los jugadores, pues su estudio permite identificar a aquellos que aportan valor a un equipo, aunque no se les conozca como estrellas del deporte.































Basándonos en esto último, es decir, en el valor que aporta cada jugador de un equipo hemos pensado en la posibilidad de llegar a predecir el ganador de un partido de baloncesto.

Para desarrollar el proyecto se procederá primero a la obtención de los datos, de la página oficial de la NBA, a través del web scraping y de la NBA\_API. Luego se procesarán los datos para quedarnos con los relacionados con el resultado de un partido, según un análisis exploratorio, y libres de valores nulos o vacíos.

Se abordará el proyecto como un problema de clasificación de jugadores, entre ganadores y perdedores, al final el equipo con más jugadores ganadores tendrá mayor probabilidad de ganar. Se utilizará el modelo random forest, pues por su complejidad es capaz de asumir variables de tipo categórico, y variables numéricas, ambas importantes para el desenlace de un juego de baloncesto.

## 1.1 Formato de la NBA

En la NBA existen dos conferencias: Este y Oeste, donde se enfrentan entre sí los equipos de cada una. Los que pertenecen a la misma conferencia se enfrentan un total de 3 veces en la fase regular y los que pertenecen a distinta conferencia, 2 veces. Cada equipo juega un total de 72 partidos en la fase regular.

Conferencia Este			Conferencia Oeste		
CLASIF.		NOMBRE	CLASIF.		NOMBRE
01		e - Philadelphia 76ers	01		w - Utah Jazz
02		x - Brooklyn Nets	02		p - Phoenix Suns
03		c - Milwaukee Bucks	03		x - Denver Nuggets
04		x - New York Knicks	04		x - LA Clippers
05		se - Atlanta Hawks	05		sw - Dallas Mavericks
06		x - Miami Heat	06		x - Portland Trail Blazers
07		x - Boston Celtics	07		x - Los Angeles Lakers
08		x - Washington Wizards	08		x - Memphis Grizzlies
09		Indiana Pacers	09		Golden State Warriors
10		Charlotte Hornets	10		San Antonio Spurs
11		o - Chicago Bulls	11		o - New Orleans Pelicans
12		o - Toronto Raptors	12		o - Sacramento Kings
13		o - Cleveland Cavaliers	13		o - Minnesota Timberwolves
14		o - Orlando Magic	14		o - Oklahoma City Thunder
15		o - Detroit Pistons	15		o - Houston Rockets

**Figura 1.** Equipos de diferentes conferencias de la NBA.

Una vez finalizada la fase regular, los 8 primeros equipos juegan la fase de play-offs. Esta fase se juega entre los equipos de la misma conferencia, jugando el primero contra el octavo, el segundo contra el séptimo y así sucesivamente. Estos partidos son al mejor de 7 partidos, es decir, el que gane 4 pasa de ronda, por lo tanto, en play-offs se juegan como mínimo 60 partidos. En la final de la NBA se enfrentan los ganadores de cada conferencia.



Figura 2. Fase de play-offs de la NBA.

## 1.2 Objetivo.

El objetivo del presente trabajo es predecir el equipo ganador de un partido de play-offs de la NBA.

## 1.3 Planteamiento

Para alcanzar el objetivo hemos planeado estudiar a cada jugador a través de sus estadísticas entendiendo el equipo como la suma de lo que aportan los jugadores por separado. Si existen muchas partes buenas, habrá un equipo ganador.

Esto es importante por la propia naturaleza del baloncesto que permite que jugadores con mayor habilidad pueden ser decisivos en el resultado final de un partido.

## 1.4 Descripción del proyecto.

Se busca construir un modelo al que se le puedan pasar dos equipos, que se enfrentarán en un futuro partido, en tiempo de play-offs, la fecha de inicio de temporada y devuelva el equipo con mayor probabilidad de ganar, es decir con más jugadores clasificados como ganadores.

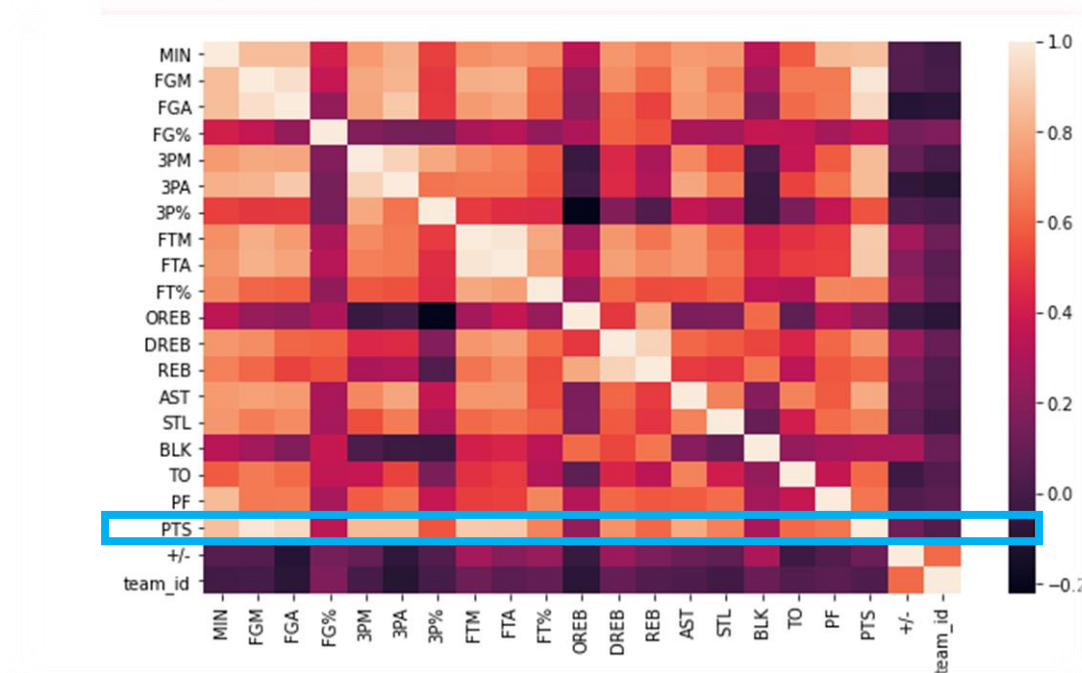
Para predecir el resultado de un partido se usarán los datos de partidos precedentes. Equipos que se enfrenten en play-offs tienen mínimo tres juegos

anteriores entre ellos, de la fase regular de la competición. Sin embargo, la final de la NBA como enfrenta a dos equipos de diferente conferencia, y solo han jugado dos veces entre ellos previamente, nuestro modelo no es capaz de predecir con tan pocos enfrentamientos.

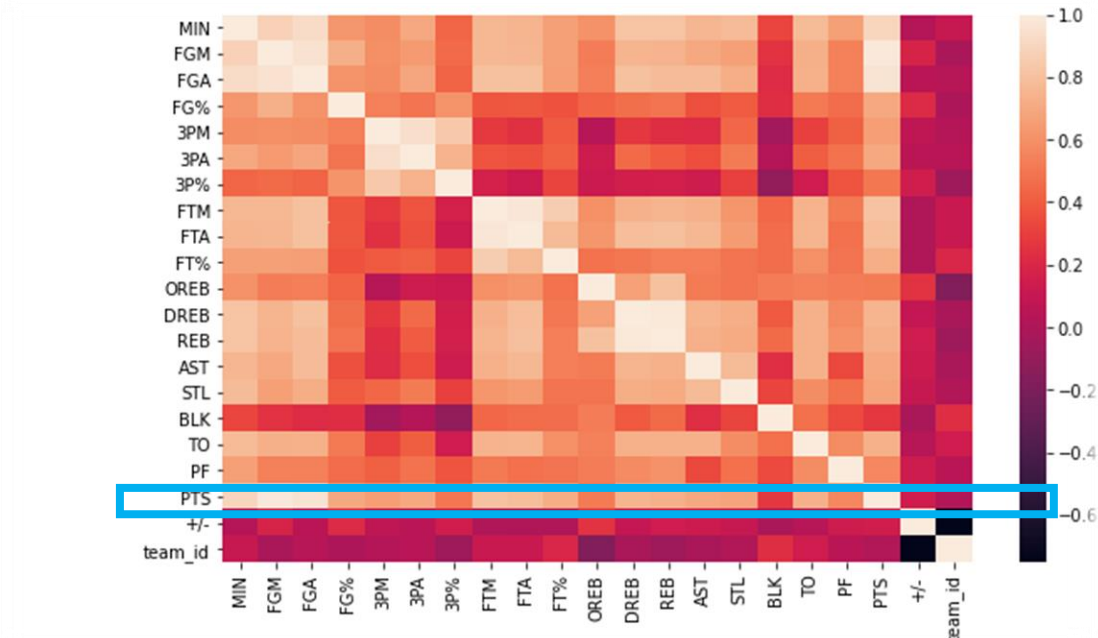
## 2. Análisis exploratorio de los datos.

Para el análisis exploratorio de los datos, se ha estudiado un buen número de encuentros de play-offs. En las siguientes gráficas se tiene el heat map de una muestra de los equipos estudiados:

MIL vs MIA



BOS vs BKN

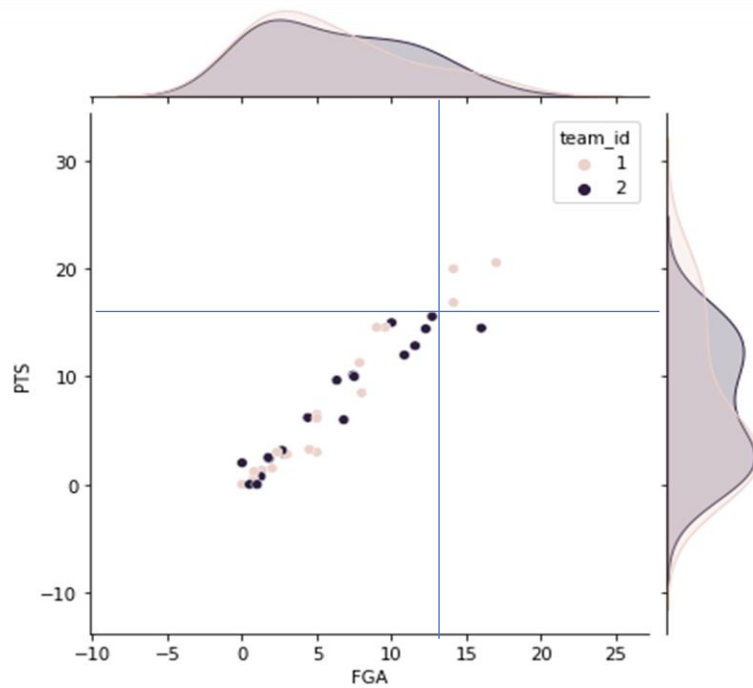


Se puede observar que en los gráficos de ambos enfrentamientos, la variable PTS, los puntos, que decide el ganador, tiene una correlación mayor con las variables: MIN, FGA, FG%, 3PA, 3P%, FTA, FT%, DREB, REB, AST, STL, TO.

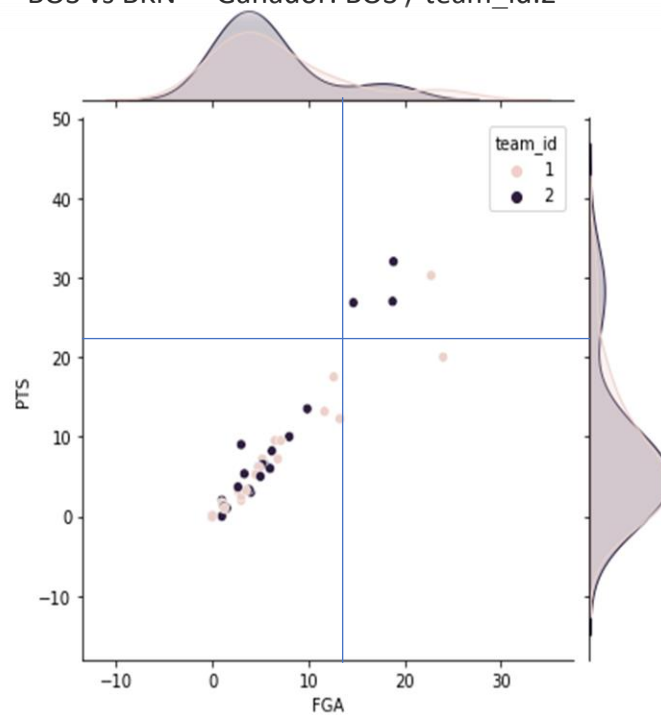


### FGA vs PTS

MIL vs MIA    Ganador: MIL / team\_id:1



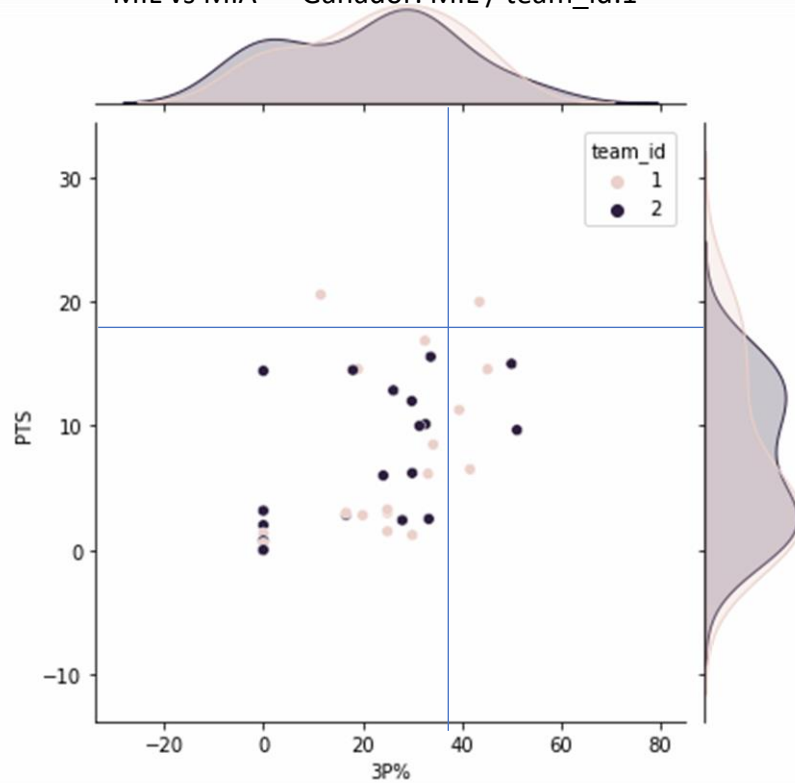
BOS vs BKN    Ganador: BOS / team\_id:2



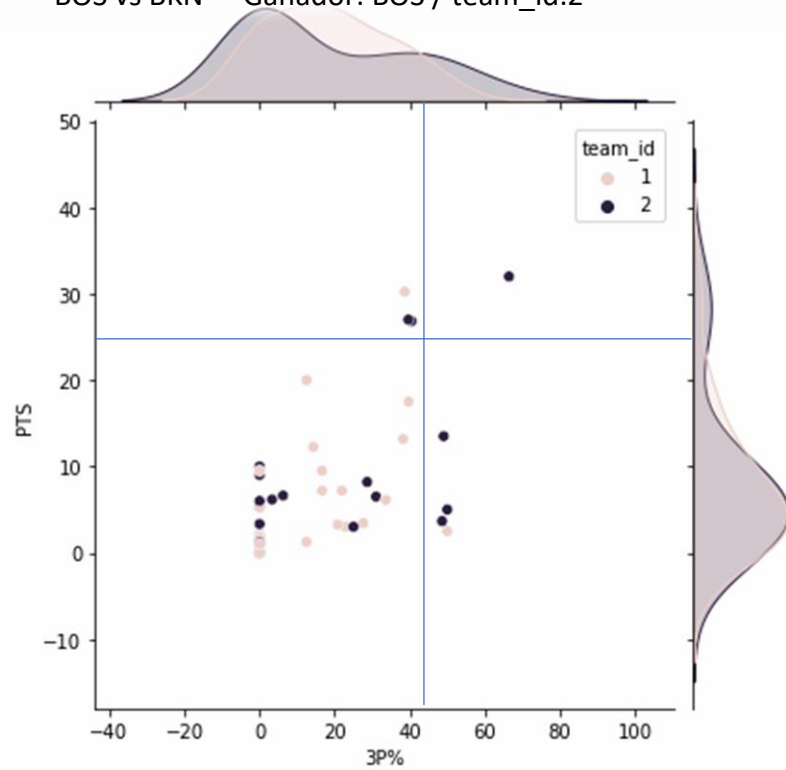
En ambos gráficos el equipo ganador tiene más jugadores con mayores tiros intentados, y a la vez más jugadores que consiguen más puntos.

### 3P% vs PTS

MIL vs MIA Ganador: MIL / team\_id:1



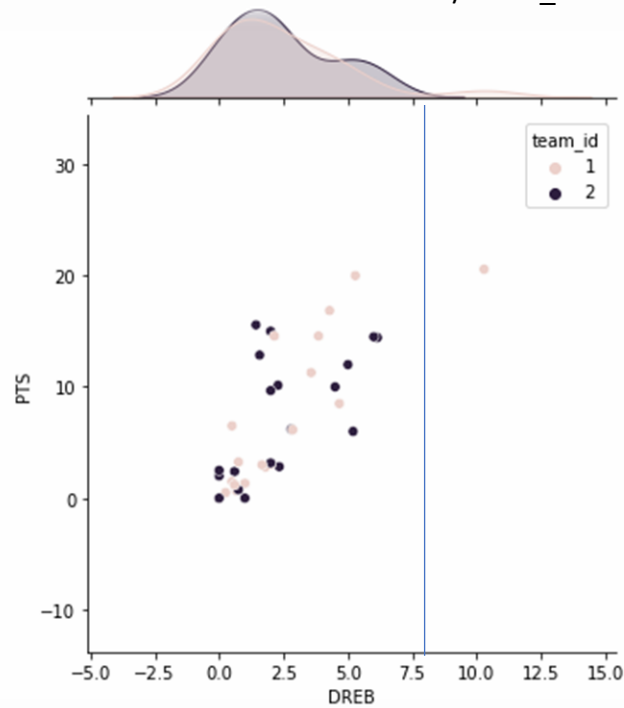
BOS vs BKN Ganador: BOS / team\_id:2



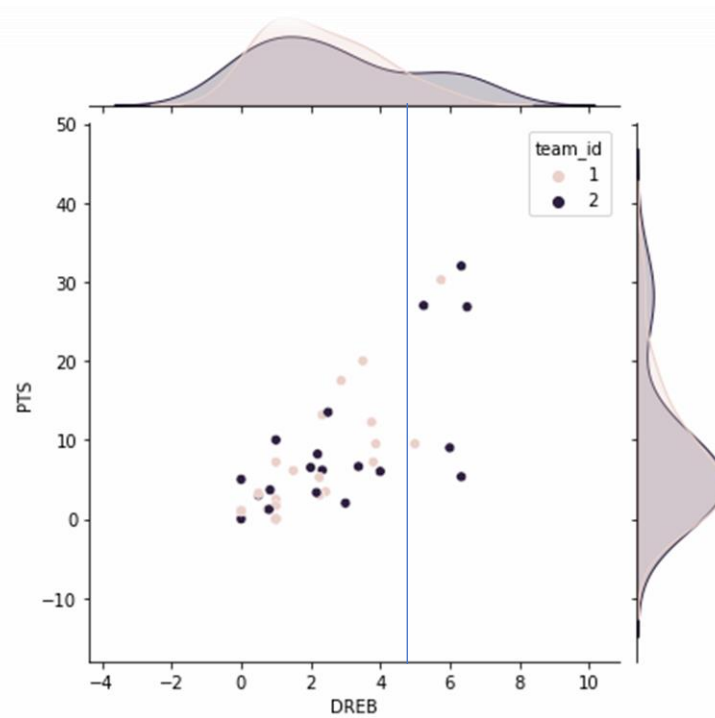
En ambos gráficos el equipo ganador tiene más jugadores con mayor efectividad de tiros de tres, y a la vez más jugadores que consiguen más puntos.

## DREB VS PTS

MIL vs MIA Ganador: MIL / team\_id:1



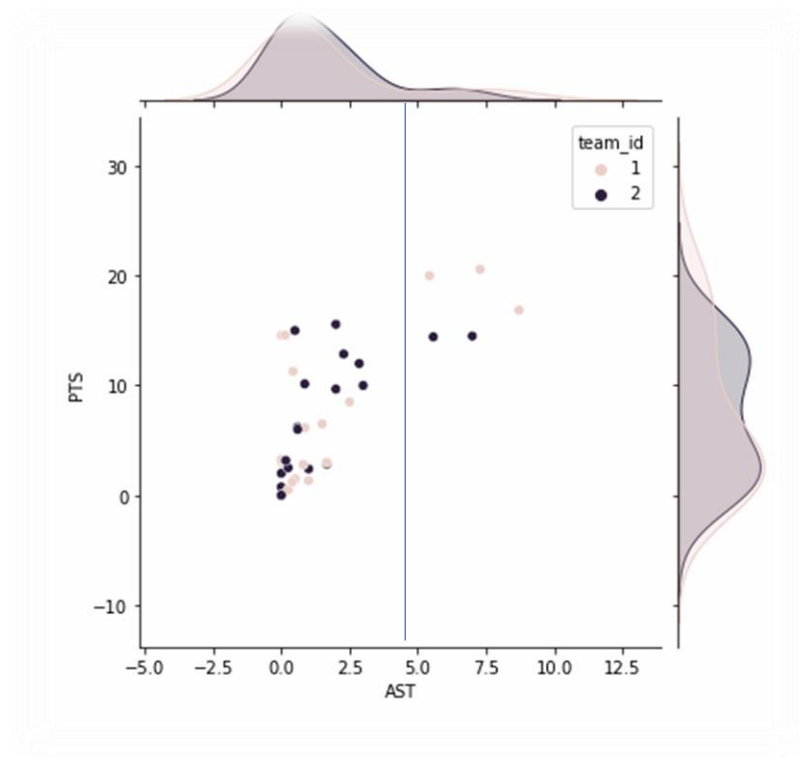
BOS vs BKN Ganador: BOS / team\_id:2



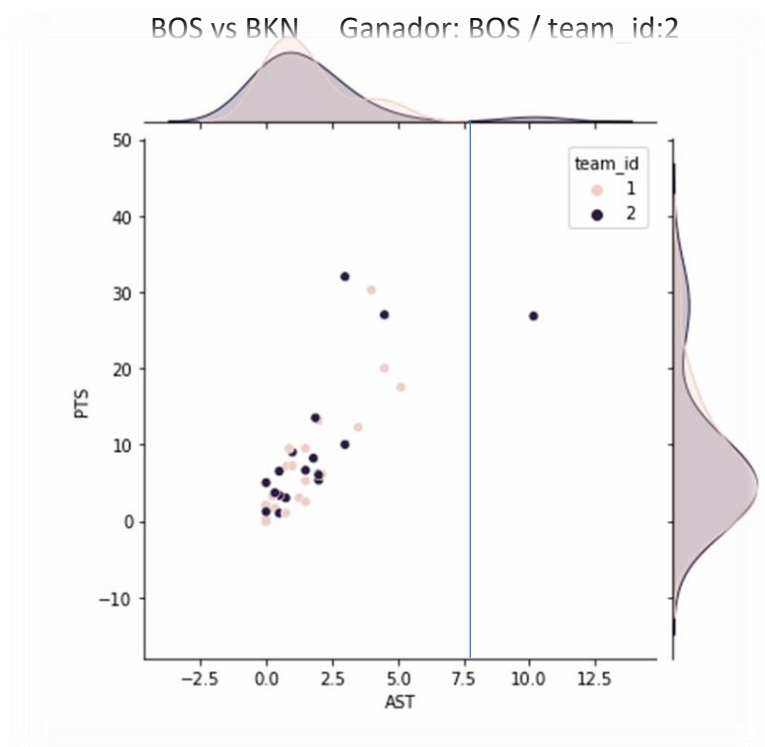
En ambos gráficos el equipo ganador tiene más jugadores que acumulan mayor número de rebotes defensivos.

## AST VS PTS

MIL vs MIA Ganador: MIL / team\_id:1



BOS vs BKN Ganador: BOS / team\_id:2



En ambos gráficos el equipo ganador tiene más jugadores que acumulan mayor número de asistencias.

Del análisis exploratorio de los datos se puede apreciar que estudiar el comportamiento de un equipo, en un enfrentamiento, por jugador, tiene sentido pues existen jugadores cuyos valores más altos de las variables estudiadas tienen mayor presencia en el equipo ganador.

### **3. Modelos**

#### **3.1 No supervisados.**

Se realizó la aplicación de modelos no supervisados a los datos, con la idea de observar si eran capaces de agrupar los jugadores, por sus números en el equipo ganador.

Para este propósito se usaron el Kmeans y el DBSCAN. En ambos casos no se encontraron agrupaciones que distinguieran entre jugadores ganadores y perdedores.

En el caso del Kmeans, donde se establece la creación de 2 clusters, se observaron agrupados en una clase jugadores con valores de sus estadísticas altos de ambos equipos, y en la otra el resto de jugadores.

Cuando se probó con el DBSCAN generó más de dos clases, lo que hace imposible la categorización ganador-perdedor de jugadores que buscamos.

#### **3.2 Modelos supervisados.**

Luego de observar los resultados con los modelos no supervisados se decidió optar por modelos de clasificación supervisados. Se eligió el Random Forest, pues es un modelo complejo de toma de decisiones, que puede trabajar con variables categóricas, presentes en los datos que se obtuvieron para construir el modelo, y así evitar pérdidas de información.

Para elegir los mejores parámetros del Random Forest se aplicaron varios GRID SEARCH.

#### **3.3 Preparación de los datos para el modelo**

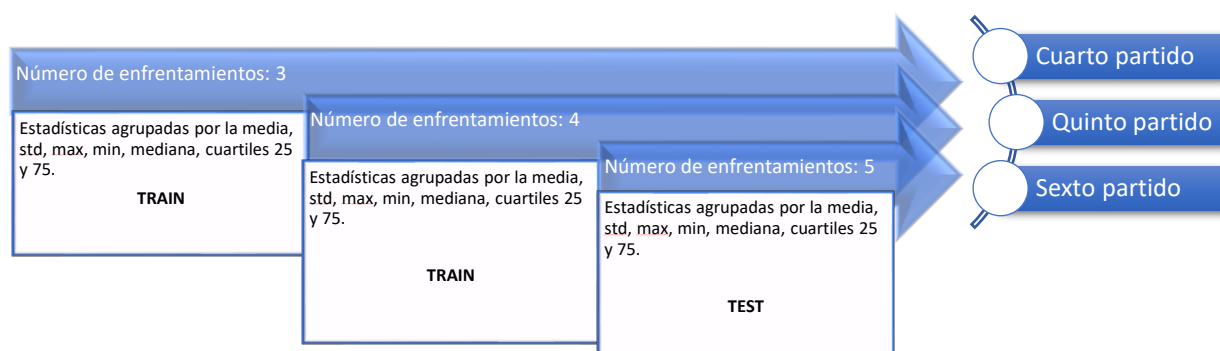
Se creó un dataset con estadísticas resumidas de los jugadores, en los partidos anteriores a un encuentro. Las estadísticas usadas son: MIN, FGA, FG%, 3PA, 3P%, FTA, FT%, DREB, REB, AST, STL, TO, agrupadas por la media, la desviación estándar, máximos,

mínimos, mediana, cuartil 25 y cuartil 75. Así se tendrá para un número de enfrentamientos previos el resumen de las estadísticas de cada jugador

Después del preprocesamiento de los datos, se procede a la elección del conjunto de prueba y entrenamiento.

### 3.4 Elección del conjunto de prueba y entrenamiento.

Se eligió como conjunto de prueba el resumen de las estadísticas que corresponde con el mayor número de enfrentamientos anteriores, que implica predecir el resultado del último partido que jugaron. Esta elección se debe al orden cronológico de los juegos, captando la evolución de los jugadores a lo largo de la temporada. Las estadísticas que corresponden con los otros números de encuentros serán el conjunto de entrenamiento.



## 4. Resultados.

Para la evaluación de los resultados, se creó un data frame que contiene una columna con los ganadores reales de los partidos del conjunto de prueba y otra columna con los ganadores que el modelo predijo. Se hace una comparación entre ambas columnas, obteniendo un True si hay coincidencia y un False si no la hay.

Será elegido como ganador el equipo con mayor número de jugadores clasificados como ganadores y que a la vez tenga el menor número de jugadores clasificados como perdedores. En el caso de que el número de ganadores sea igual, el menor número de perdedores decidirá el ganador. El mismo razonamiento se usa si empatan en el número de perdedores. Sin embargo, si hay el mismo número de jugadores clasificados en una u otra categoría el modelo no será capaz de decidir el equipo ganador.

El conjunto de prueba contiene 417 juegos, coincidiendo el ganador real con el que predice el modelo 346 veces, que implica un 83% de acierto

Mostramos un ejemplo de las variables y su importancia para la predicción del resultado de un partido, esto es cambiante según los equipos que se enfrentan:

Variable	Importancia
PLUS_MINUS_q75	10,9069 %
Número de_enfrentamientos	7,6754 %
FGA_std	6,4569 %
PLAYER_ID	6,1610 %
AST_q25	5,6829 %
DREB_max	5,5977 %



## **5. Conclusiones**

- Modelos supervisados agrupan mejor a los jugadores que los no supervisados.
- Las variables categóricas son importantes para el modelo.
- Debe reentrenarse el modelo a los 20-30 primeros partidos disputados de play-offs, para mejorar las predicciones siguientes.
- El equipo que más jugadores ganadores tiene y menos jugadores perdedores, tiene más probabilidad de ganar.

## **6. Futuras investigaciones.**

- Probar el modelo:
  1. Usando estadísticas avanzadas, que resultan de operaciones matemáticas entre las estadísticas usadas en el modelo.
  2. Combinando las estadísticas avanzadas con los datos del modelo actual.
  3. Agregando la clasificación de la fase regular y datos de los entrenadores.