Proyecto final:

Conformación de las parejas en padel

Curso Data scientist – Coder house

Fecha 22/08/23

Alumno: Lambre, Santiago

INDICE

Introducción

El padel es un deporte en pleno crecimiento, el cual se juega 2 vs 2, y cada equipo cuenta con un entrenador sentado en el banquillo, quien puede hablar con los jugadores cada vez que cambian de lado de cancha. Se juega al mejor de 3 sets, con partidos que pueden durar desde 45 minutos a 3 horas.

Al ser un deporte en crecimiento, los premios, el dinero específicamente, no son tan abundantes como en otros deportes, por lo que para todas las parejas es importante tener buenas actuaciones, cada cierto tiempo al menos

Una forma de maximizar las oportunidades de tener resultados, o cambiar ante los adversos, es con el cambio de pareja. Estos cambios se pueden llevar a cabo varias veces en el año, solo hace falta el acuerdo entre los nuevos integrantes de la pareja.

Descripción de la temática de los datos:

A partir de 2 bases de datos que contiene estadísticas de partidos de padel masculinos de la World Padel Tour (WPT), se evaluará cuáles datos pueden servir para realizar un modelo de Machine Learning

El dataset para cada jugador consta de 1128 filas y 26 columnas, siendo cada fila un set de un jugador. Los datos son del tipo numérico y categórico nominal u ordinal

El dataset de los equipos tiene 175 filas y 27 columnas, siendo cada fila las estadísticas de una pareja en un set. Los datos son del tipo numérico y categórico nominal u ordinal. Lo importante del dataset es la característica de quién ganó el set

Nivel de aplicación:

Estratégico: los resultados podrían ser de utilidad para el armado de una pareja de padel, y para el entrenador que pueda servirse de los datos para enfocar el entrenamiento en los puntos flojos o altos del equipo

Objetivos:

Se buscará generar un modelo que permita evaluar cuáles jugadores se beneficiarían jugando juntos y cuáles no. Mismo, qué podrían enfocarse en entrenar, y así mejorar, para tener mas probabilidades de ganar los sets

Hallazgos encontrados por el EDA

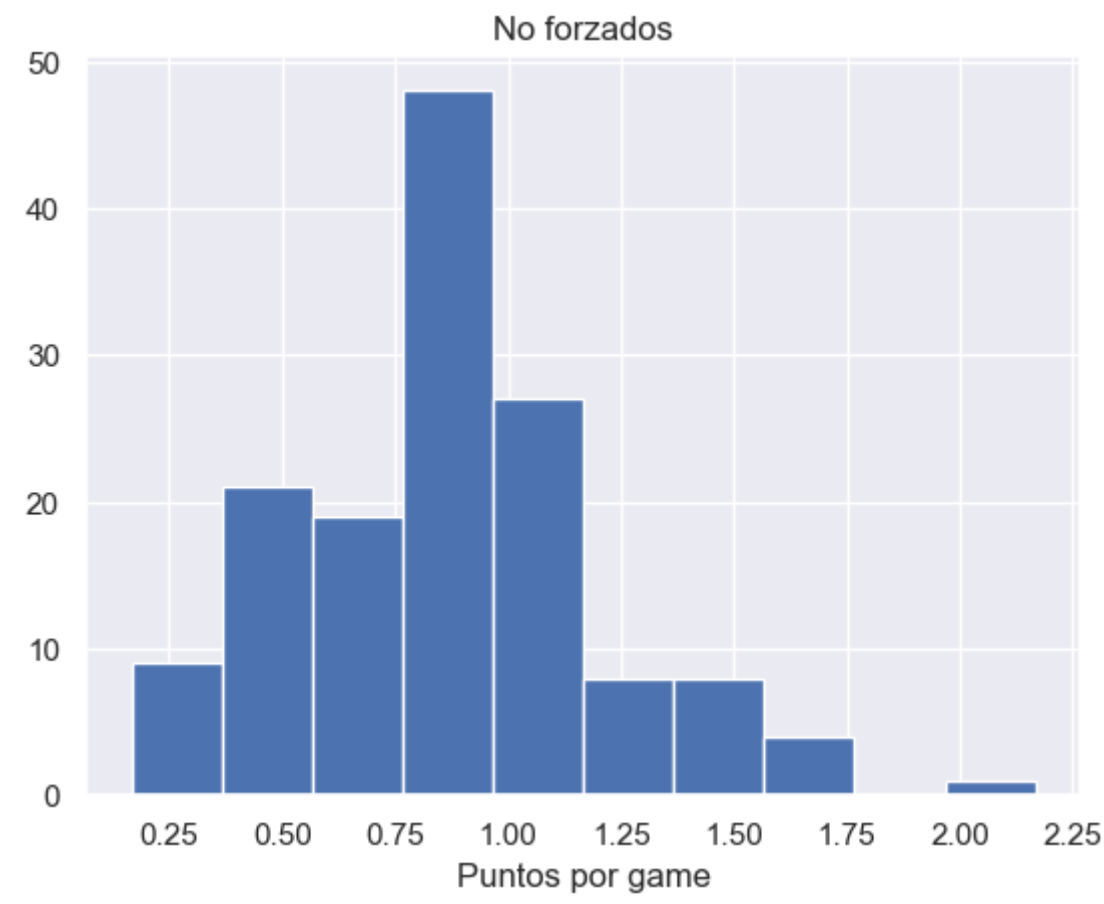
Dataset de los equipos:

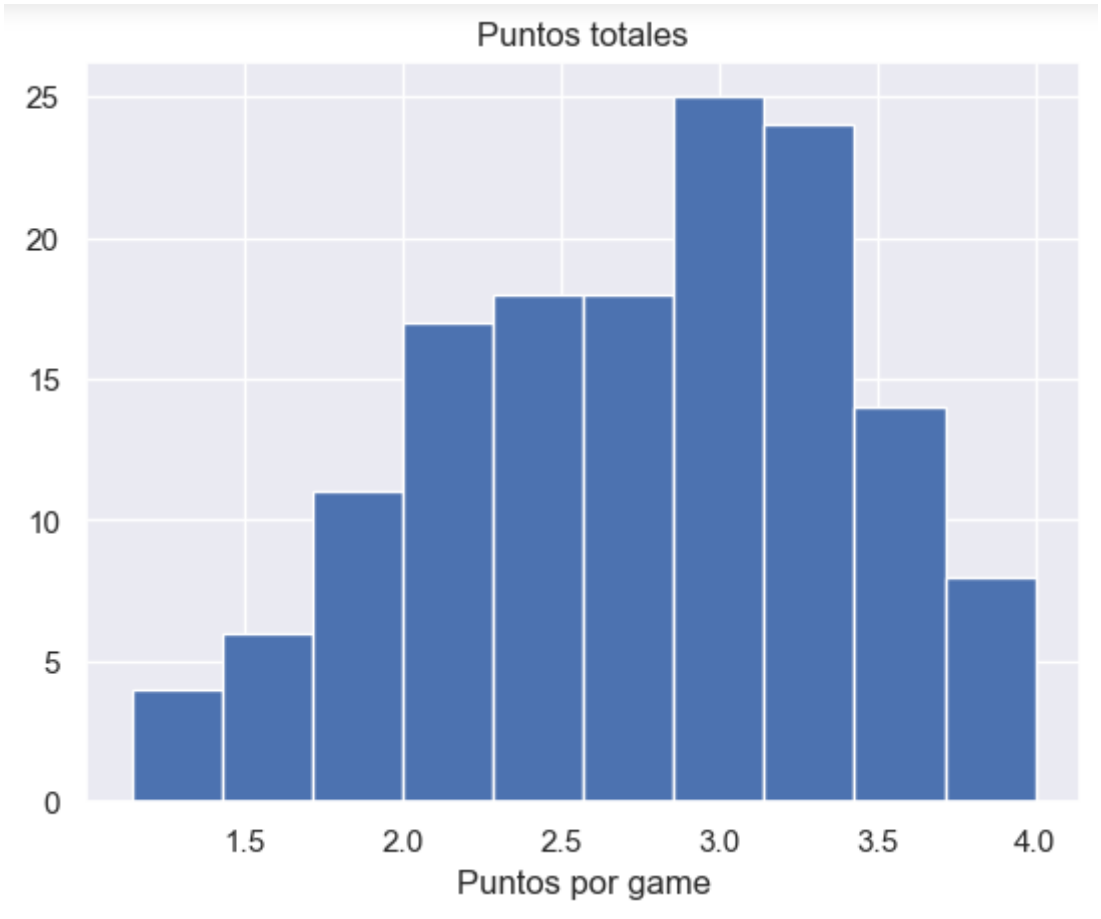
Luego de la limpieza del dataset, se utilizaron 4 columnas para predecir quienes ganan o pierden el set

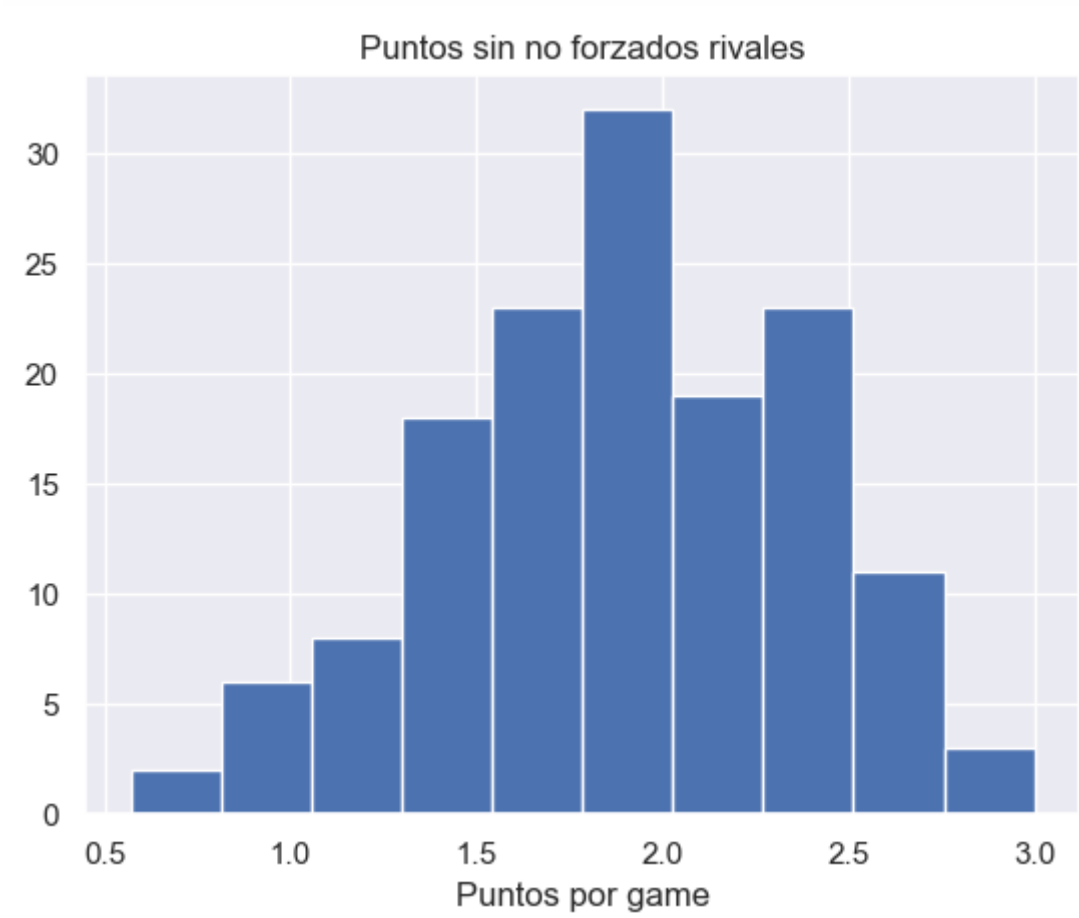
Las columnas son:

1- Errores no forzados – nf
2- Puntos totales por game – pto_tot
3- Puntos totales sin errores no forzados del rival – tot_snf
4- Resultado del set – result. Siendo 1 para quien ganó, y 0 para el perdedor

Encontramos que las 4 características tienen una distribución normal







Correlacion:

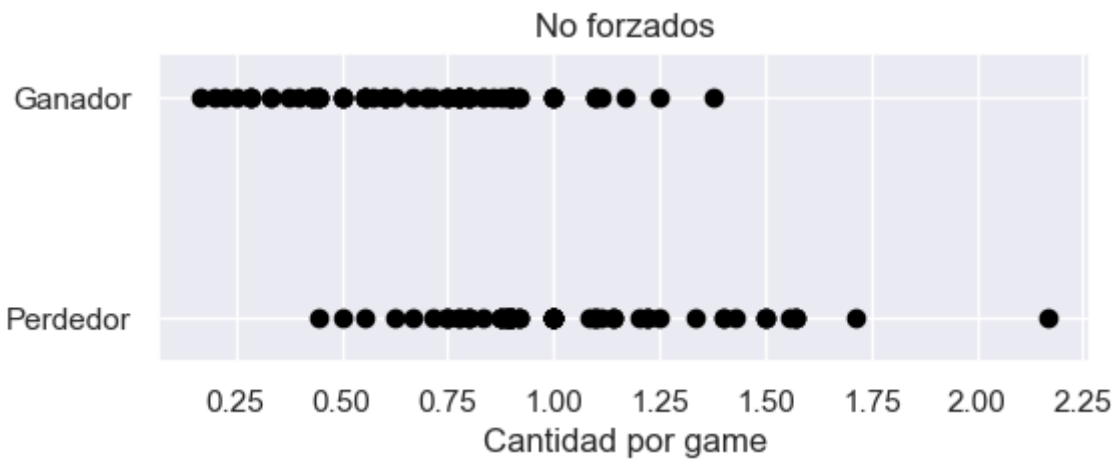
:

	result
nf	-0.522001
pto_tot	0.789150
tot_snf	0.694979
result	1.000000

Hay una fuerte correlacion con result,(ganar) y los puntos que hace una pareja y los puntos totales con los puntos sin no forzados porque uno contiene al otro. Si bien se podría eliminar una de las columnas de los puntos, sirve tener las 2 columnas para diferenciar a los jugadores que hacen muchos puntos y tambien muchos errores no forzados de los que hacen muchos puntos y pocos errores no forzados, que son el ideal del padel y que mas sets ganan

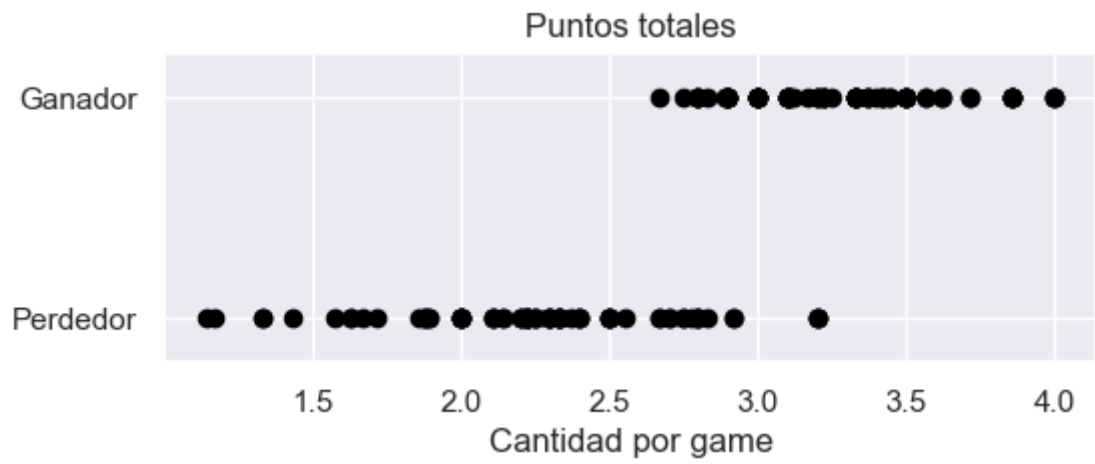
Mientras que los errores no forzados muestran correlacion inversa con ganar, pero no tan fuerte como la cantidad de puntos hechos, ya que los puntos pueden superar a los errores no forzados y ganar el partido, a pesar de tener ambos valores altos

Análisis bivariado



Se observa que hacer muy pocos errores no forzados (lo cual no es nada sencillo y suele ocurrrir cuando hay mucha diferencia en el ranking o una pareja en particular que son los numero 1 del ranking) es un buen predictor del triunfo del set.

Los errores no forzados tienen una correlacion con el resultado de: -0.522001, siendo que quien haga mas de estos tiene mas probabilidades de perder. Como se aprecia en el grafico, hay un rango de 0.47 a 1.35 puntos en los cuales se puede ganar o perder el set

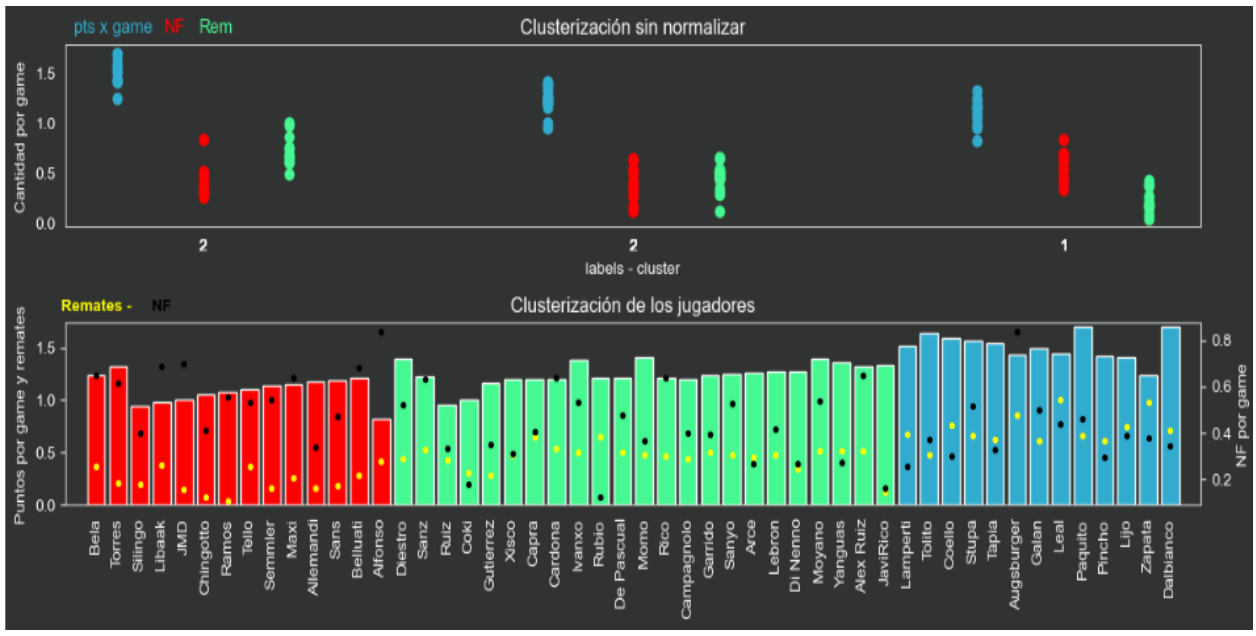


En el grafico se puede apreciar a simple vista que los ganadores del set suelen hacer mas puntos por set de promedio. Hay partidos muy parejos en los cuales los perdedores han hecho mas de 2.6 puntos por set. Se puede intuir que para ganar hay que promediar mas de 2.6 puntos por game

Teniendo en cuenta que para ganar el set, hay que promediar 2.6 puntos por game como minimo, equipos que promedien mas de 2.4 puntos sin errores no forzados, ganan el partido el 100% de las veces, ya que a los puntos que hacen hay que sumarle los puntos de los errores no forzados del rival

Modelo elegido

Para agrupar a los jugadores utilizamos un modelo no supervisado, KMEANS ya que no contamos con etiquetas para separar a cada jugador, y al ser un deporte dinámico, los jugadores van teniendo diferentes formas de jugar y roles. Si sabíamos que hay perfiles de jugadores ofensivos, defensivos y mixtos, por eso se eligió 3 clusters



En el grafico de arriba se observan los 3 clusters que agrupan a los jugadores segun sus estadisticas por game que juegan, siendo las barras los puntos que hacen, en puntos negros los errores no forzados y en puntos amarillos los remates que realizan.

Descripcion de clusters:

Cluster 1, rojo: jugadores que realizan menos remates que la media es el rasgo principal. Tambien suelen cometer mas errorez no forzados que la media, y la cantidad de puntos es variada, pero suelen hacer entre la media y menos

Cluster 2, verde: son jugadores hibridos, presentan valores intermedios entre el cluster 1 y 2

Cluster 3, azul: jugadores que realizan mas puntos que la media y mas remates. Errores no forzados es variado

Todavia no vimos modelos asi que no tengo métricas, pero si ya tengo el otro modelo, que predice quien gana y quien pierde (supervisado), pero, como tampoco lo vimos, todavía no se si esta bien asi que no lo pasé