Laboratorio de Datos

Primer Cuatrimestre 2024

Trabajo Práctico N° 2

El Trabajo Práctico deberá ser resuelto en grupos de dos o tres personas. No se aceptarán entregas individuales. La entrega se realizará a través del campus (pestaña Trabajos Prácticos). La fecha límite es el 02/07 a las 23:59. Deben entregar un Notebook con los nombres de les integrantes del equipo, la resolución de los ejercicios y los informes pertinentes.
Se valorará que el Notebook y el código tengan un formato prolijo: ejercicios separados por títulos (Ejercicio 1, Ejercicio 2, etc.), nombres descriptivos para las variables, comentarios, etc.

En este trabajo práctico nos sumamos a la moda de utilizar análisis de datos para la toma de decisiones en los deportes, y creamos nuestra propia empresa LDD Futbol Analytics.
Trabajaremos con el dataset `FBRef2020-21.csv`[1] que contiene datos sobre los principales torneos de clubes y selecciones de fútbol del mundo.

Preprocesamiento [1 pt.]

1. Cargar en un DataFrame los datos del archivo `FBRef2020-21.csv`.
2. Eliminar a los jugadores que jugaron menos de 500 minutos en la temporada (columna `Min`).
3. Eliminar los datos faltantes. Por ejemplo, eliminar columnas con más de 100 datos faltantes y luego las filas con datos faltantes, o convertir los datos faltantes a algún valor apropiado.
4. Al finalizar la limpieza de datos, resetear los índices.
5. Definir el DataFrame `data_num` que solo contenga las variables númericas, a partir de la columna `Ast/90`, inclusive. Para clustering y clasificación no vamos a utilizar las variables categóricas ni edad ni minutos jugados.

Clustering [4 pts.]

6. Nuestro primer objetivo es realizar algún agrupamiento de jugadores con características similares.
   - **(a)** Seleccionar dos variables cualesquiera de los datos y realizar un gráfico de dispersión de una variable en función de la otra para el total de las observaciones. ¿Pueden encontrar fácilmente grupos distintos?
   - **(b)** Escalar los datos y realizar un análisis de componentes principales, quedándose solo con las dos primeras componentes. Realizar un gráfico como el del punto anterior. ¿Cuántos clusters puede distinguir en el gráfico? ¿A qué características de los jugadores pueden corresponder los clusters? ¿Cómo pueden verificar su conjetura? (realizar una visualización o algún cálculo)
   - **(c)** Para la cantidad de clusters observados en el ítem anterior, realizar un agrupamiento por k-medias, y colorear los puntos según las etiquetas obtenidas. ¿Coinciden las etiquetas con lo esperado?

[1] Fuente: https://fbref.com/en/

(d) Repetir el agrupamiento utilizando DBSCAN. ¿Cómo eligirían en este caso un valor de ε apropiado? Sugerencia: consultar la sección "Selección del hiperparámetro eps" del Notebook de la clase de DBSCAN (o utilizar cualquier otra técnica que consideren apropiada)

(e) Utilizar DBSCAN para realizar agrupamiento utilizando como datos todas las variables originales en vez de solo las dos componentes principales, modificando los valores de `epsilon` y `minPts` convenientemente. ¿Con cuál de las dos opciones obtienen mejores resultados?

Clasificación [3 pts.]

7. Ahora queremos poder predecir la posición en la que juega cada jugador según sus datos estadísticos utilizando KNN. En la columna `Pos` encontramos la posición de los jugadores. Para la mayoría de los jugadores se indica una única posición pero algunos jugadores tienen dos posiciones. Para simplificar el análisis vamos a considerar una única posición por jugador.

(a) Definir la variable `Pos_filt` que es la columna `Pos`, pero donde los jugadores deben tener una sola posición (pueden quedarse sólo con la primera posición de cada jugador o eliminar los jugadores con dos posiciones, lo que consideren más conveniente).

(b) Construir el DataFrame `data_clasif` que resulta de agregarle la columna de `Pos_filt` al DataFrame `data_num`. Dividir `data_num` en un 80% para entrenamiento y un 20% para testeo.

(c) Aplicar un esquema de validación en el conjunto de entrenamiento para seleccionar el valor óptimo de K. (Esto puede demorar mucho si prueban muchos valores de K, pueden hacerlo hasta un valor máximo de $K = 20$.)

Sugerencia: puede resultar de ayuda `KNeighborsClassifier` de `sklearn` e investigar esa librería para aplicar esquemas de validación.

(d) Para el valor de K obtenido, ¿cuál es el porcentaje de aciertos en el conjunto de testeo?

(e) Repetir el procedimiento utilizando la primeras dos componentes principales en vez de todas las variables. Indicar si se obtienen mejores resultados.

8. Repetir el mismo método de clasificación con el dataset de jugadoras de la liga inglesa femenina `superleague2023.csv`. Para esto, quedarse con la columna `Pos` y las columnas a partir de `MP`, inclusive. ¿Qué porcentaje de aciertos obtienen en este caso? ¿Puede modificarse el parámetro para obtener un porcentaje mayor?

Recomendaciones de jugadores [2 pts.]

9. Trabajamos ahora con el dataset `transfermarkt_fbref_201920.csv` que incluye la valuación de los jugadores. Una de las aplicaciones más comunes de análisis de datos en el fútbol es para obtener recomendaciones de jugadores a comprar.

Si al leer el `.csv` salta un error, intentar con:

```
data=pd.read_csv('transfermarkt_fbref_201920.csv', delimiter=';')
```

(a) En 2021 Messi fue transferido del Barcelona al PSG. Baśandose en los datos disponibles, recomendarle a Barcelona un jugador de características similares a Messi pero de menor valor.

(b) Queremos elaborar un modelo para detectar jugadores "baratos", es decir cuya valuación en el mercado (columna `value`) sea inferior a que la que nosotros estimemos. Para esto, quisiéramos ajustar el valor de mercado a partir de los datos de los jugadores (para estimar la valuación pueden incorporar la edad entre las variables explicativas). El modelo que desarrollen puede basarse en redes neuronales o en otro modelo que consideren adecuado.

(c) Según el modelo desarrollado, entre los jugadores con un valor de mercado mayor a $100000, ¿quién es el más sobrevalorado? Es decir, el jugador con mayor diferencia entre el valor de mercado y el valor predicho por el modelo. ¿Y el más infravalorado?

(d) El PSG quiere vender a Mbappé y reemplazarlo por otro jugador más barato. Hacer un listado de los 10 jugadores más parecidos a Mbappé según el criterio que elijan. De esos 10 jugadores, según el modelo que desarrollaron en el ítem anterior, ¿a qué jugador recomendarían teniendo en cuenta la valuación del mercado y la predicción del modelo? Para ese jugador, averiguar la valuación actual del jugador. ¿Hicieron una buena recomendación?

Apéndice: descripción de algunas columnas

Los datos están tomados del sitio web FBRef, en la sección "Big 5 European Leagues History". En ese sitio pueden ver el detalle de qué significa cada columna. Por ejemplo, pueden ingresar a este enlace `https://fbref.com/en/comps/Big5/stats/players/Big-5-European-Leagues-Stats` y entrar al enlace "Glossary". Algunos nombres están levemente modificados.
En la siguiente tabla incluimos la descripción de las columnas principales.

Columna	**Descripción**
Rk	*Rango* Conteo de filas de arriba hacia abajo. Se recalcula al ordenar una columna.
Nation	*Nacionalidad del jugador* Primero, verificamos nuestros registros en el juego internacional a nivel sénior. Luego, a nivel juvenil. Luego, la ciudadanía presentada en Wikipedia. Finalmente, usamos su lugar de nacimiento cuando está disponible.
Pos	*Posición*

	Posición más comúnmente jugada por el jugador: GK - Porteros DF - Defensores MF - Mediocampistas FW - Delanteros FB - Laterales LB - Laterales Izquierdos RB - Laterales Derechos CB - Defensas Centrales DM - Mediocampistas Defensivos CM - Mediocampistas Centrales LM - Mediocampistas Izquierdos RM - Mediocampistas Derechos WM - Mediocampistas Exteriores LW - Extremos Izquierdos RW - Extremos Derechos AM - Mediocampistas Ofensivos
Comp	*Competencia* La competencia. El número junto a la competencia indica qué nivel ocupa esta liga en la pirámide de ligas del país.
Age	*Edad al inicio de la temporada* Dada el 1 de agosto para ligas de invierno y el 1 de febrero para ligas de verano.
Born	*Año de nacimiento* Año de nacimiento del jugador.
MP	*Partidos Jugados* Partidos jugados por el jugador o equipo.
Starts	*Titularidades* Juego o juegos comenzados por el jugador.
Min	*Minutos* Minutos jugados.
90s	*Partidos completos jugados* Partidos completos jugados (minutos jugados divididos por 90).
Gls	*Goles* Goles marcados o permitidos.
Ast	*Asistencias* Asistencias.
G+A	*Goles + Asistencias* Goles y Asistencias.
G-PK	*Goles sin Penaltis* Goles sin incluir penaltis.
PK	*Penaltis Marcados* Penaltis marcados.
PKatt	*Penaltis Intentados* Penaltis intentados.
CrdY	*Tarjetas Amarillas* Tarjetas Amarillas.
CrdR	*Tarjetas Rojas*

	Tarjetas Rojas.
xG	*xG: Goles Esperados* Goles Esperados. Incluyen penaltis pero no tandas de penaltis (a menos que se indique lo contrario). Proporcionado por Opta. Un subrayado indica que falta información de un partido, pero se actualizará cuando esté disponible.
npxG	*npxG: Goles Esperados sin Penaltis* Goles Esperados sin Penaltis. Proporcionado por Opta. Un subrayado indica que falta información de un partido, pero se actualizará cuando esté disponible.
xAG	*xAG: Asistencias de Goles Esperados* Asistencias de Goles Esperados. xG que sigue a un pase que asiste a un tiro. Proporcionado por Opta. Un subrayado indica que falta información de un partido, pero se actualizará cuando esté disponible.
npxG + xAG	*npxG + xAG* Goles Esperados sin Penaltis más Asistencias de Goles Esperados. Incluyen penaltis pero no tandas de penaltis (a menos que se indique lo contrario). Proporcionado por Opta. Un subrayado indica que falta información de un partido, pero se actualizará cuando esté disponible. Mínimo de 30 minutos jugados por partido de equipo para calificar como líder.
PrgC	*Avances Progresivos* Conducciones que mueven el balón hacia la línea de gol del oponente al menos 10 yardas desde su punto más lejano en los últimos seis pases, o cualquier conducción dentro del área de penalti. Excluye conducciones que terminan en el 50% defensivo del campo.
PrgP	*Pases Progresivos* Pases completados que mueven el balón hacia la línea de gol del oponente al menos 10 yardas desde su punto más lejano en los últimos seis pases, o cualquier pase completado dentro del área de penalti. Excluye pases desde el 40% defensivo del campo.
PrgR	*Pases Progresivos Recibidos* Pases completados que mueven el balón hacia la línea de gol del oponente al menos 10 yardas desde su punto más lejano en los últimos seis pases, o cualquier pase completado dentro del área de penalti. Excluye pases desde el 40% defensivo del campo.
Gls/90	*Goles por 90 minutos* Goles marcados por 90 minutos. Mínimo de 30 minutos jugados por partido de equipo para calificar como líder.
Ast/90	*Asistencias por 90 minutos* Asistencias por 90 minutos. Mínimo de 30 minutos jugados por partido de equipo para calificar como líder.
G+A/90	*Goles + Asistencias por 90 minutos* Goles y Asistencias por 90 minutos. Mínimo de 30 minutos jugados por partido de equipo para calificar como líder.
G-PK/90	*Goles sin Penaltis por 90 minutos* Goles menos penaltis marcados por 90 minutos. Mínimo de 30 minutos jugados por partido de equipo para calificar como líder.
G+A-PK/90	*Goles + Asistencias sin Penaltis por 90 minutos*

	Goles más asistencias menos penaltis marcados por 90 minutos. Mínimo de 30 minutos jugados por partido de equipo para calificar como líder.
xG/90	*Goles Esperados por 90 minutos* Goles Esperados por 90 minutos. Incluyen penaltis pero no tandas de penaltis (a menos que se indique lo contrario). Proporcionado por Opta. Un subrayado indica que falta información de un partido, pero se actualizará cuando esté disponible. Mínimo de 30 minutos jugados por partido de equipo para calificar como líder.
xAG/90	*Asistencias de Goles Esperados por 90 minutos* Asistencias de Goles Esperados por 90 minutos. Proporcionado por Opta. Un subrayado indica que falta información de un partido, pero se actualizará cuando esté disponible. Mínimo de 30 minutos jugados por partido de equipo para calificar como líder.
xG+xAG/90	*Goles Esperados + Asistencias de Goles Esperados por 90 minutos* Goles Esperados más Asistencias de Goles Esperados por 90 minutos. Incluyen penaltis pero no tandas de penaltis (a menos que se indique lo contrario). Proporcionado por Opta. Un subrayado indica que falta información de un partido, pero se actualizará cuando esté disponible. Mínimo de 30 minutos jugados por partido de equipo para calificar como líder.
npxG/90	*Goles Esperados sin Penaltis por 90 minutos* Goles Esperados sin Penaltis por 90 minutos. Proporcionado por Opta. Un subrayado indica que falta información de un partido, pero se actualizará cuando esté disponible. Mínimo de 30 minutos jugados por partido de equipo para calificar como líder.
npxG+xAG/90	*Goles Esperados sin Penaltis + Asistencias de Goles Esperados por 90 minutos* Goles Esperados sin Penaltis más Asistencias de Goles Esperados por 90 minutos. Proporcionado por Opta. Un subrayado indica que falta información de un partido, pero se actualizará cuando esté disponible. Mínimo de 30 minutos jugados por partido de equipo para calificar como líder.