**Tarea PF2**

Alejandro Hernández Beneito, Javier Montaner de Fez y Salvador Gisbert Sempere

**2. Resumen**

En nuestro grupo hemos decidido aplicar los conocimientos que hemos adquirido en la asignatura de Modelos Lineales al dataset "2022-2023 Football Player Stats.csv" y "2021-2022 Football Player Stats.csv" extraídos de <https://www.kaggle.com/datasets/vivovinco/20222023-football-player-stats?resource=download> y <https://www.kaggle.com/datasets/vivovinco/20212022-football-player-stats?resource=download> respectivamente.

Es por ello que trabajaremos con el fin último de predecir los máximos goleadores y asistentes de las mejores ligas de fútbol europeas.

**3. Introducción**

Nuestros datasets originales cuentan con 124 variables y +2500 observaciones y 143 variables y +2900 observaciones respectvamente. Cabe destacar que no trabajaremos con los datasets originales sino que los juntaremos y haremos con uno con menos variables y observaciones que se acople al nivel del análisis que realizaremos. Para reducirlo hemos tenido en cuenta sólo algunas variables que consideramos son las más influyentes en cuanto a goles y asistencias.

A medida que hemos avanzado con el preprocesado de los datos, nos hemos dado cuenta de que algunas variables no eran del tipo que les tocaba para poder ser analizadas correctamente. Por ejemplo la variable SoT (Shots on Target) originalmente era de tipo character. Para seguir con nuestro análisis hemos cambiado el tipo de algunas variables.

Otro problema con el que nos hemos encontrado es que algunas variables estaban registradas en variable/partido. Para obtener las variables totales hemos tenido que modificar nuestro dataset.

Por último, hemos eliminado todos aquellos jugadores con menos de 90 minutos jugados, hemos eliminado porteros (no suelen marcar ni asistir y no sería coherente predecir que uno de ellos sea máximo goleador o asistente). Finalmente hemos eliminado jugadores sin goles ni asistencias.

Así, trabajamos con un dataset de 27 variables y 2152 observaciones. A continuación se muestran los nombres de las variables junto con algunas observaciones para poder tener una idea del dataset con el que trabajaremos:

Tabla

Descripción generada automáticamente

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Las variables que contiene nuestro dataset son:

|  |  |
| --- | --- |
| Player | Nombre del futbolista |
| Age | Minutos jugados |
| Goals | Goles totales |
| Shots | Tiros por partido |
| SoT | Tiros a puerta por partido |
| PassTotCmp | Pases totales por partido |
| Assists | Asistencias por partido |
| PasAss | Pases clave por partido |
| Total\_Assists | Asistencias totales |
| Total\_Shots | Tiros totales |
| Total\_SoT | Tiros a puerta totales |
| Total\_Passes | Pases totales |
| Total\_Key\_Passes | Pases clave totales |

Debemos mencionar que al haber 2 temporadas recogidas con datos en nuestro dataset hemos duplicado las variables que se mencionan anteriormente de manera que tendremos Goals\_2021\_2022 y Goals\_2022\_2023…

**4. Resultados**

**4.1. Recta mínimos cuadrados**

Analizamos nuestros datos. Esta es la recta de mínimos cuadrados que mejor ajusta la relación entre Tiros Totales y Goles en la temporada 2022-2023 (Para a temporada 2021-2022, ver rmd). Vemos como puede existir una relación lineal entre ambas variables.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Se observa como a más tiros, más goles. ¿Ocurrirá lo mismo con las asistencias?

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Vemos como no necesariamente los que más pases hacen son los que más asisten. Esta distribución tiene forma de campana y los que realizan entre los 500 y los 1000 pases son los que más asisten (siempre hay excepciones). Esto se puede ver reflejado en los defensas, por ejemplo. Con equipos que juegan desde atrás, como el FC Barcelona, los centrales por ejemplo realizan muchos pases y sin embargo pocas asistencias.

Calendario

Descripción generada automáticamente

Imagen que contiene Tabla

Descripción generada automáticamente

En efecto, esto se confirma.

**4.2 Diagnóstico del modelo**

Hagamos un diagnóstico de estos modelos: ¿cumplen con la linealidad, normalidad y homocedasticidad?

Para el modelo de goles frente a tiros:

Gráfico, Gráfico de dispersión

Descripción generada automáticamenteGráfico, Gráfico de líneas

Descripción generada automáticamente

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Podemos afirmar que la relación tiros-goles es lineal (si no tenemos en cuenta los outliers). Las observaciones se distribuyen alrededor de la recta de mínimos cuadrados y los residuos entorno a 0.

Podemos ver que además cumple con la normalidad. Aceptamos también la heterocedasticidad, no cumple con la homocedasticidad pues para que la haya no debe haber patrones en la distribución de los residuos. En los nuestros claramente se ve que se organizan en una recta descendente.

Para el modelo de asistencias frente a pases:

Gráfico, Gráfico de dispersión

Descripción generada automáticamenteGráfico, Gráfico de líneas

Descripción generada automáticamente

Calendario

Descripción generada automáticamente

Utilizando los mismos razonamientos que en el modelo de los goles, afirmamos que:

No podemos saber si es lineal, pues no sólo no se distribuyen los datos en forma de recta sino que además los residuos no se distribuyen exactamente entorno al 0 sino que está un poco desplazado.

No cumple con la normalidad y tampoco con la homocedasticidad. Aceptamos la heterocedasticidad.

**4.3 Correlación**

Para empezar a modelar en función de nuestro objetivo, veamos las correlaciones entre las distintas variables:

Gráfico, Gráfico de burbujas

Descripción generada automáticamente

Vemos que para predecir goles tendría sentido tomar en cuenta los, los tiros y los tiros a portería. Así se cumple en los máximos goleadores, que suelen ser los que más tiran a portería. Así se puede ver en el dataframe. En cuanto a las asistencias vemos que están muy relacionadas con los pases clave. También podemos ver esto en los máximos asistentes.

A continuación mostramos las dos temporadas en conjunto y vemos como se cumple también esto.

Gráfico de dispersión

Descripción generada automáticamente

**4.4 Modelaje**

Comenzamos con un modelo en que la variable respuesta son los Goles y las variables predictoras son todas las demás. Seguidamente, usando el método step(), elegiremos el mejor modelo para predecir nuestra variable.

Para predecir la variable respuesta, necesitaremos que las variables predictoras no estén vacías (no tendría sentido predecir los goles en una temporada si no sabemos los datos de esa misma temporada). Por ello, tomaremos un dataset que simulará los datos de la temporada 2023-2024, que hemos sacado de la media aritmética entre las dos temporadas de las que sí que tenemos datos. No tendría sentido que un jugador que ha jugado 0 minutos una temporada y 300 otra, de repente jugase 3000, por ello pensamos que la media aritmética sería un buen estimador. Con los modelos construidos y el dataset de datos nuevos, obtenemos:

Tabla

Descripción generada automáticamente Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Viendo los datos de las últimas temporadas, podemos decir que la predicción es correcta. Los goles y asistencias predichos son coherentes con el rendimiento de los jugadores en las últimas temporadas.

**5. Conclusión**

A medida que hemos ido avanzando nos hemos dado cuenta de varias cosas:

* Aunque parezca que no, es importante también trabajar con datasets arreglados, cuando se manejan muchos datos puede haber confusiones.
* A más datos mejor predicción. En principio trabajábamos sólo con 1 dataset de 1 temporada. Tuvimos que añadir uno de otra temporada, para así mejorar el modelo y predecir con un poco más exactitud.
* Tener un modelo u otro hace variar mucho las predicciones. Hemos probado con varios modelos, algunos prediciendo barbaridades, y nos hemos dado cuenta de la importancia que tiene la correlación, de que las covariables sean buenas predictoras de la variable respuesta.
* Es importante contrastar siempre los resultados que obtenemos para que sean coherentes y no obtengamos datos erróneos.