Predicción Ganador UEFA Champions League

Lidia Velicia Ruiz
Ana García Saiz

Inteligencia Artificial
Grado Ingenería Matemática
Universidad Alfonso X el Sabio

Índice

1. Introducción

El objetivo de este proyecto era simple, haciendo uso de técnicas de Análisis de Datos y Machine Learning, predecir el equipo ganador del torneo de la UEFA Champions League (UCL) 2023/24. Con total libertad para elegir el planteamiento del análisis, los datos, los modelos a utilizar...

2. Metodología

2.1. Recopilación de datos

El vasto mundo del fútbol y las páginas de estadísticas futbolísticas es complicado. Hay muchas webs con mucha información y recopilar datos en grandes cantidades no es tarea fácil.

2.1.1. Extracción

Por suerte, dimos con la librería ScraperFC de **@ouseymour**, para el scrappeo de datos futbolísticos. Entre las diferentes webs de las que permite obtener datos, elegimos FBRef, ya que nos pareció la más completa en términos de las estadísticas que ofrece y la gran variedad de ligas internacionales que maneja.

Con el fin de predecir el ganador del torneo de la UCL 23/24, hemos tomado los siguientes datos:

- Estadísticas de la UCL 2022/2023
- Estadísticas de la fase de grupos de la UCL 2022/2023
- Estadísticas de las ligas nacionales 2022/2023
- Estadísticas de las ligas nacionales 2023/2024 (hasta la fecha)

Decidimos coger los datos de las 2 últimas temporadas porque, en general, los equipos suelen mantener su plantilla relativamente constante durante este períodos.

2.1.2. Estadísticas equipos

Para obtener una visión completa de cada equipo, seleccionamos variables que cubren tanto aspectos de ataque como de defensa, así como otras variables relevantes. Las variables recopiladas para el análisis y su significado son las siguientes:

Ataque

- **Gls90**: Goles por 90 minutos.
- **Ast90**: Asistencias por 90 minutos.
- **G+A90**: Goles y asistencias combinados por 90 minutos.
- **G-PK90**: Goles no provenientes de penaltis por 90 minutos.
- **G+A-PK90**: Goles y asistencias no provenientes de penaltis por 90 minutos.
- **SCA90**: Acciones que conducen a una oportunidad de gol por 90 minutos.
- **GCA90**: Acciones que conducen a un gol por 90 minutos.
- **Sh90**: Disparos por 90 minutos.
- **SoT90**: Disparos a puerta por 90 minutos.
- **SoT %**: Porcentaje de disparos a puerta.
- **G/Sh**: Goles por disparo.
- **G/SoT**: Goles por disparo a puerta.

Defensa

- **GA90**: Goles concedidos por 90 minutos.
- **Save %**: Porcentaje de paradas realizadas por el portero.
- **SoTA90**: Disparos a puerta enfrentados por 90 minutos.

2.2. Preprocesamiento

2.2.1. Ponderaciones Ligas Nacionales

Dado que cada liga europea presenta un nivel competitivo diferente, creímos necesario ponderar las estadísticas de los equipos en sus ligas nacionales, según el "nivel futbolístico"de la liga de procedencia:

- Las cinco principales ligas europeas Premiere League (Inglaterra), La Liga (España), Ligue 1 (Francia), Serie A (Italia) y Bundesliga (Alemania) se ponderaron con un factor de 1.
- Las ligas "de 2º categoría"serían la Eredivisie (Países Bajos), Superliga (Dinamarca) y Primeira Liga (Portugal), consideradas de un nivel ligeramente inferior, se ponderaron con un factor de 0.67.

2.2.2. Ponderaciones UCL

Para reflejar de manera precisa las diferencias de nivel entre los partidos de fase de grupos de la UCL y los enfrentamientos en fases más avanzadas del torneo, es esencial ponderar las estadísticas de los equipos. La mayoría de los partidos en las fases iniciales de la UCL son contra equipos significativamente inferiores, lo cual puede inflar las estadísticas de los equipos más fuertes. Para contrarrestar este efecto y lograr una representación más equilibrada, es necesario ajustar las estadísticas de la UCL añadiendo las estadísticas de las ligas nacionales de cada equipo.

La fórmula de ponderación utilizada para calcular las estadísticas combinadas es la siguiente:

$$\text{Ponderación Final} = \frac{(\text{Est}_{\text{UCL}} + 0{,}5 \cdot \text{Est}_{\text{NAC}})}{2}$$

Esta fórmula permite integrar las estadísticas nacionales, dándoles un peso moderado al multiplicarlas por 0.5, antes de combinarlas con las estadísticas de la UCL. Finalmente, el resultado se divide por 2 para obtener una ponderación equilibrada entre ambos conjuntos de datos. Esta metodología asegura que las estadísticas utilizadas en el análisis sean representativas del rendimiento real de los equipos, considerando tanto su desempeño en la UCL como en sus respectivas ligas nacionales.

2.2.3. Dataset partidos

Nuestro enfoque para predecir quién ganará el campeonato será simular los enfrentamientos, prediciendo los goles marcados por cada equipo, con ello, según quien gane, llegaremos al ganador de la UCL 23/24.

Para ello, entrenaremos los modelos con los partidos de la temporada completa 22/23 y la fase de grupos de la 23/24. Junto con las variables relativas al enfrentamiento, incluiremos las estadísticas de cada equipo, para que el modelo pueda aprender de ellas para predecir los goles.

Las variables del enfrentamiento son:

- **Nombre_Eq1**: equipo 2
- **Nombre_Eq2**: equipo 1
- **Gol_Match_Eq1**: goles quipo 1
- **Gol_Match_Eq2**: goles equipo 2
- **GANADOR**: nombre del equipo ganador

De esta forma el dataset nos queda con las variables:

Match_ID	Nombre_Eq1	Nombre_Eq2	Gol_Match_Eq1	Gol_Match_Eq2	GANADOR	Gls90_Eq1	Ast90_Eq1	G+A90_Eq1	G-PK90_Eq1	G+A-PK90_Eq1
1	hr Dinamo Zagreb	eng Chelsea	1	0	hr Dinamo Zagreb	0.67	0.33	1.0	0.67	1.0

SCA90_Eq1	GCA90_Eq1	Sh90_Eq1	SoT90_Eq1	SoT%_Eq1	G/Sh_Eq1	G/SoT_Eq1	GA90_Eq1	Save%_Eq1	SoTA90_Eq1	Gls90_Eq2	Ast90_Eq2	G+A90_Eq2	G-PK90_Eq2
16.67	1.33	9.17	3.33	0.36	0.07	0.2	1.83	0.77	5.83	1.2	0.7	1.9	1.0

SCA90_Eq1	GCA90_Eq1	Sh90_Eq1	SoT90_Eq1	SoT%_Eq1	G/Sh_Eq1	G/SoT_Eq1	GA90_Eq1	Save%_Eq1	SoTA90_Eq1	Gls90_Eq2	Ast90_Eq2	G+A90_Eq2	G-PK90_Eq2
16.67	1.33	9.17	3.33	0.36	0.07	0.2	1.83	0.77	5.83	1.2	0.7	1.9	1.0

3. Análisis

Nuestro objetivo es predecir el número de goles que marca cada equipo en cada partido en funcion de sus estadisticas y las del adversario. Por lo tanto, al estar prediciendo una variable numérica, nuestros modelos serán de regresión.

Para simular el torneo, hemos hecho que en las fases de octavos, cuartos y semis haya ida y vuelta, para que le modelo pueda considerar el factor jugar en çasa". Es por esto que se muestran los goles acumulados, salvo en la final que es a partido único.

Los modelos que hemos utilizado son:

3.1. Modelo Regresión Lineal

Fase	Equipo 1	Gol Acum. Eq 1	Equipo 2	Gol Acum. Eq 2
Octavos	PSG	3.612	Real Sociedad	3.165
Octavos	Copenhague	3.421	Manchester City	5.581
Octavos	Barcelona	3.647	Napoli	2.308
Octavos	Atlético Madrid	3.539	Inter de Milán	3.313
Octavos	Dortmund	3.819	PSV	2.808
Octavos	Bayern Múnich	3.186	Lazio	2.166
Octavos	Arsenal	4.346	Porto	3.916
Octavos	Real Madrid	4.161	Leipzig	3.245
Cuartos	Dortmund	2.995	Atlético Madrid	3.4
Cuartos	Barcelona	3.059	PSG	2.941
Cuartos	Arsenal	2.918	Bayern Múnich	2.294
Cuartos	Real Madrid	4.014	Manchester City	4.259
Semis	Barcelona	3.131	Atlético Madrid	2.992
Semis	Arsenal	3.509	Manchester City	4.2
Final (part. único)	Barcelona	1.461	Manchester City	1.818

Cuadro 1: Resultados modelo Regresión Lineal

El modelo de regresión lineal predice que el ganador sería el Manchester City.

3.2. Modelo Red Neuronal

Hemos contruído una red neuronal (DNN) multicapa, formada por una secuencia de capas lineales intercaladas con funciones de activación ReLU. La primera capa recibe el número de variables independientes del dataset y conecta a una capa oculta de 300 neuronas. Este esquema se repite, reduciendo el número de neuronas en las capas sucesivas hasta alcanzar la capa de salida, que tiene 2 neuronas correspondientes a las dos variables objetivo (goles de cada equipo).

En el código podrá encontrar 2 modelos de red neuronal, para cada uno hemos probado distintos parámetros en la construcción de la red, y presentamos los que mejores resultados han conseguido:

3.2.1. DNN 1: estadísticas futbolísticas

En un principio creamos una red neuronal utilizando, como estadísticas de los equipos enfrentados, las variables futbolísticas que describimos anteriormente. Estos fueron los resultados de los enfrentamientos:

Fase	**Equipo 1**	**Gol Acum. Eq 1**	**Equipo 2**	**Gol Acum. Eq 2**
Octavos	PSG	1.969	Real Sociedad	1.055
Octavos	Copenhague	1.534	Manchester City	3.468
Octavos	Barcelona	3.061	Napoli	0.762
Octavos	Atlético Madrid	1.912	Inter	1.452
Octavos	Dortmund	2.379	PSV	1.124
Octavos	Bayern Múnich	2.417	Lazio	0.747
Octavos	Arsenal	1.991	Porto	1.791
Octavos	Real Madrid	2.608	Leipzig	0.982
Cuartos	Dortmund	1.538	Atlético Madrid	1.668
Cuartos	Barcelona	1.967	PSG	1.698
Cuartos	Bayern Múnich	1.586	Arsenal	1.862
Cuartos	Real Madrid	1.958	Manchester City	2.608
Semis	Barcelona	2.104	Atlético Madrid	1.457
Semis	Arsenal	1.245	Manchester City	2.741
Final (part. único)	Barcelona	0.628	Manchester City	1.371

Cuadro 2: Resultados DNN con sólo estadísticas futbolísticas

Nuestra red neuronal predice que el ganador sería el Manchester City.

3.2.2. DNN 2: estadísticas históricas

Después, nos dimos cuenta de que podíamos añadir otras variables sobre el equipo que reflejasen de alguna forma la historia del club en la competición de la Champions. Para que el modelo pudiese también tener en cuenta este dato acerca de la experiencia del equipo en las fases finales del torneo.

Las variables que creamos y que añadimos a las estadísticas de cada equipo en cada enfrentamiento son:

- **Probabilidad_SemiF**: Probabilidad de alcanzar las semifinales.
- **Probabilidad_Won**: Probabilidad de ganar el torneo.

Estos fueron los resultados que obtuvimos al entrenar la red incluyendo estas nuevas variables:

Fase	Equipo 1	Gol Acum. Eq 1	Equipo 2	Gol Acum. Eq 2
Octavos	PSG	3.581	Real Sociedad	2.767
Octavos	Copenhague	3.159	Manchester City	4.446
Octavos	Barcelona	3.622	Napoli	2.892
Octavos	Atlético Madrid	3.312	Inter de Milán	3.011
Octavos	Dortmund	3.282	PSV	3.117
Octavos	Bayern Múnich	3.006	Lazio	2.402
Octavos	Arsenal	3.294	Porto	3.456
Octavos	Real Madrid	3.954	Leipzig	3.053
Cuartos	Dortmund	2.992	Atlético Madrid	3.397
Cuartos	Barcelona	3.667	PSG	3.765
Cuartos	Bayern Múnich	3.042	Arsenal	3.288
Cuartos	Real Madrid	4.07	Manchester City	4.462
Semis	PSG	3.688	Atlético Madrid	3.502
Semis	Arsenal	3.42	Manchester City	4.433
Final (part. único)	PSG	1.915	Manchester City	2.31

Cuadro 3: Resultados DNN con más variables

Esta versión de la red neuronal también daría la victoria al Manchester City. Pero a diferencia de los modelos anteriores, que ambos daban al Barcelona como finalista, en esta pasaría el PSG hasta la final.

4. Evaluación

Como podemos observar, comparando los resultados de nuestros modelos con la realidad, los modelos predijeron perfectamente los equipos que pasaron a cuartos de final, pero los equipos que pasaron a semifinales, no coinciden con lo que pasó realmente.

Es por esto que, para ver cómo responderían nuestros modelos, hemos impuesto los equipos clasificados en Semis. Estos serían los resultados:

4.1. Regresión Lineal

Fase	**Equipo 1**	**Gol Acum. Eq 1**	**Equipo 2**	**Gol Acum. Eq 2**
Semis	PSG	3.349	Dortmund	2.923
Semis	Bayern Munich	2.353	Real Madrid	3.423
Final (part. único)	PSG	1.782	Real Madrid	2.076

Cuadro 4: Resultados regresión lineal imponiendo semis

El modelo predice bien la clasificación del Real Madrid, pero es el PSG, en lugar del Dortmund el que pasaría a la final. En la final el ganador sería el Real Madrid.

4.2. DNN

Fase	**Equipo 1**	**Gol Acum. Eq 1**	**Equipo 2**	**Gol Acum. Eq 2**
Semis	PSG	1.854	Dortmund	1.986
Semis	Bayern Munich	1.556	Real Madrid	2.366
Final (part. único)	Dortmund	0.759	Real Madrid	1.102

Cuadro 5: Resultados red neuronal imponiendo semis

El modelo predice exactamente lo sucedido, tanto Dortmund como Real Madrid pasan a la final. El ganador, ante esta final, sería el Real Madrid.

5. Conclusiones

El uso de técnicas de Machine Learning para predecir resultados de la UEFA Champions League puede ofrecer predicciones valiosas, aunque no infalibles, como consecuencia de las variaciones en el rendimiento de los equipos y la naturaleza impredecible del fútbol. Hemos decidido considerar variables estrictamente futbolísticas, pero existen otros factores no deportivos que también pueden llegar a afectar a los resultados, factores económicos, factores de salud (lesiones de jugadores), decisiones arbitrales... Asímismo, también se pueden haber perdido detalles al considerar los equipos en su conjunto, un planteamiento con jugadores individuales podría permitir mayores niveles de precisión. Además, ayudaría a tener una mayor cantidad de datos, lo que beneficiaría el entrenamiento del modelo.

Aún así, nuestros modelos, aunque mejorables, han demostrado una buena capacidad para identificar patrones y tendencias en los enfrentamientos entre equipos. Por lo tanto, tras el análisis, conociendo ya quiénes son los equipos en la final y teniendo en cuenta los resultados de nuestros modelos, podemos dar respuesta a la cuestión planteada como objetivo:

El ganador de la Champions 2023/24 será el Real Madrid.