Informe final. Ejercicio feedback

Iker de Cabo González

January 31, 2025

En este trabajo se ha realizado un análisis exhaustivo de los resultados de selecciones de fútbol masculino en la historia documentada, desde la década de 1870 hasta el día de hoy. A continuación se van a presentar los resultados de este ejercicio.

1 Exploración Inicial y Análisis Descriptivo

Para comenzar, vamos a analizar en DataFrame con el que estamos trabajando, realizando una exploración y un análisis descriptivo minucioso del conjunto de datos.

1.1 Exploración Inicial, EDA y Limpieza de Datos

En esta sección vamos a comentar las características principales del conjunto de datos con el que estamos trabajando, además de modificaciones y limpieza de datos que le hayamos aplicado. Para empezar, el DataFrame se compone de todos los partidos jugados desde 1870 hasta hoy, incluyendo todos los partidos de selecciones documentados en campeonatos y amistosos. El DataFrame incluye las siguientes variables para cada uno de los partidos:

Variable	Tipo
date	object
home_team	object
away_team	object
home_score	int64
away_score	int64
tournament	object
city	object
country	object
neutral	bool

Table 1: Descripción de las variables del conjunto de datos

Como podemos observar, el dataframe nos da la información de la fecha, los equipos que jugaron, quién jugó como local y quién como visitante, el resultado, la ciudad y el país donde se jugó y si el campo era neutral o no. Es decir, tenemos un Dataframe con do variables numéricas, 5 categóricas, una lógica y la fecha. Los datos se encuentran bastante limpios, de hecho, no hay ni un solo valor faltante o duplicado en todo el Dataset, por lo que no hay más que añadir en este apartado.

1.2 Análisis Descriptivo Avanzado y EDA

Lo primero que vamos a analizar son los goles por partido, haciendo distinción entre los partidos jugados como local y como visitante.

	Goles como local	Goles como visitante
Campo no neutral	1.79	1.12
Campo neutral	1.67	1.36

Table 2: Promedio de goles por partido

Como cabría esperar, los equipos marcan más goles de local que como visitante, marcando de media 1,79 goles como local y 1,12 como visitante, cuando se juega en campo no neutral. Es decir, observamos una diferencia de 0,72 goles, o lo que es lo mismo, una reducción del 37,4% en los goles marcados para los visitantes. En cuanto a los partidos jugados en campo neutral, llama la atención que la diferencia de goles marcados como local en comparación con los goles marcados como visitante sigue siendo significativa, algo que a priori podriamos pensar que no debería ser así, al eliminar la ventaja de jugar en tu campo. No obstante, persiste una reducción del 18,56% en los goles marcados como visitante respecto a los del equipo local. Aunque menor que en los partidos disputados en campo no neutral, sigue siendo una diferencia estadísticamente significativa.

A continuación podemos observar un histograma con la distribución de goles en todos los partidos, sin hacer distinción entre partidos jugados en campo neutral o no neutral.

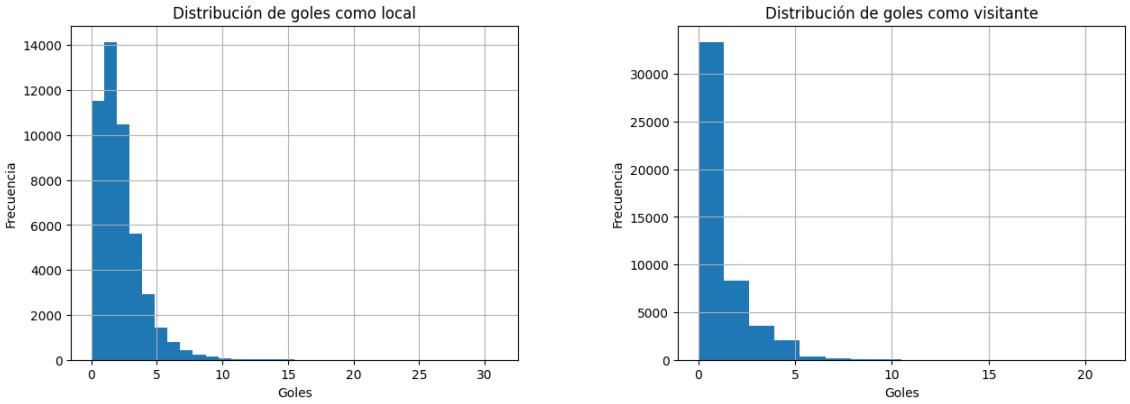

Figure 1: Histograma de goles marcados como local y como visitante

Debido a los outliers este histograma resulta algo complicado de interpretar. Por lo tanto, vamos a aplicar un filtro, de tal forma que podamos hacer una interpretación más sencilla y representativa de la distribución de goles. Vamos a eliminar los registros con más de 7 goles marcados por un mismo equipo, ya que en el 98,36% de los partidos ninguno de los dos equipos marca más de 7 goles, por lo que la muestra resulta representativa.

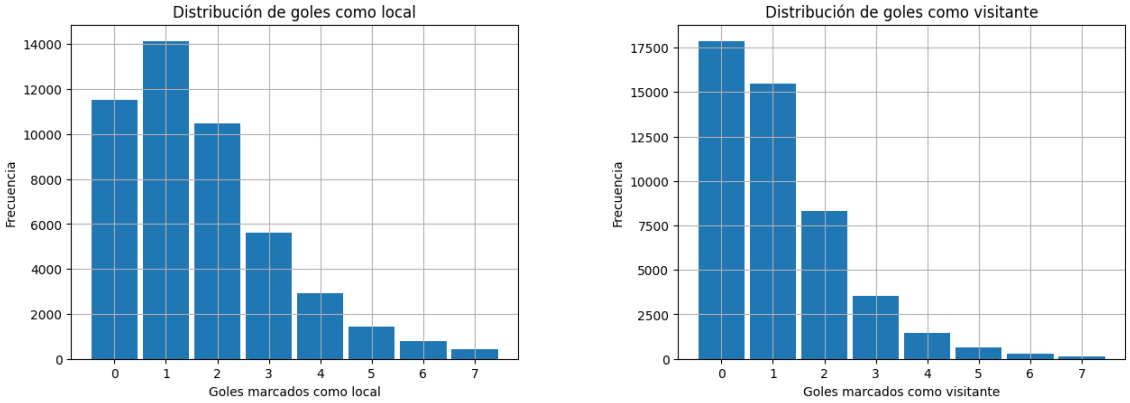

Figure 2: Histograma de goles marcados como local y como visitante filtrados

Se puede observar que el resultado más común cuando se juega en casa es marcar un gol, mientras que cuando se juega como visitante es no marcar ningún gol. Finalmente, en la figura (3) vemos la evolución de goles totales marcados por partido a lo largo de las décadas.

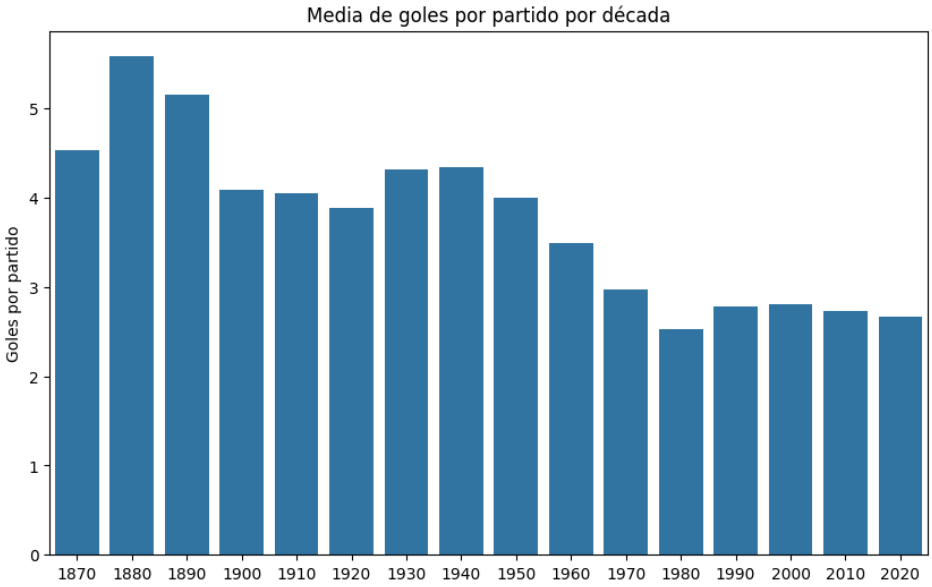

Figure 3: Goles por partido a lo largo de las décadas

En esta gráfica se aprecia como el número de goles ha ido disminuyendo con el paso de las décadas. Inicialmente se marcaban más de 4 goles por partido, y el número de goles ha ido disminuyendo paulatinamente, hasta estabilizarse alrededor de 2,75 para las últimas 4 décadas.

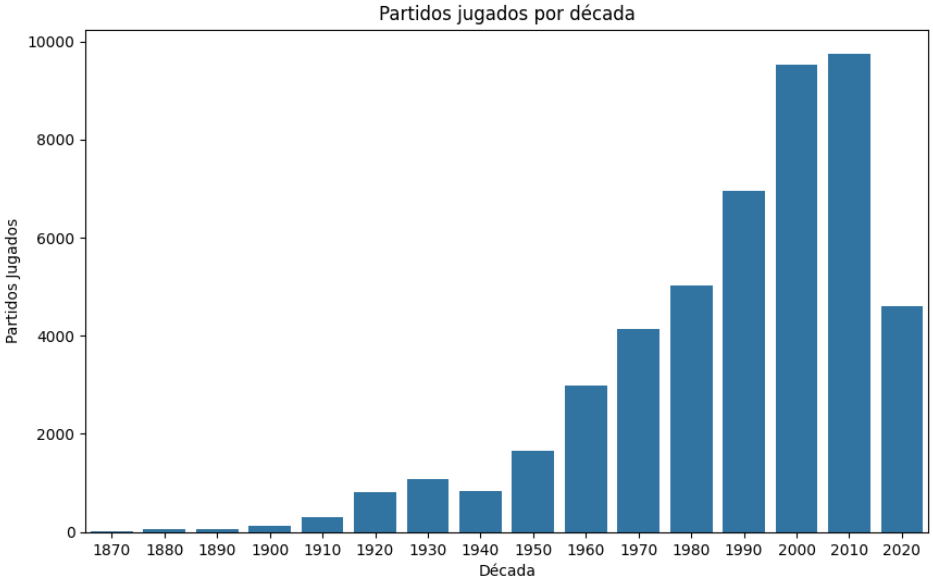

Figure 4: Partidos totales jugados por década

Podemos ver que la tendencia de partidos jugados por década ha ido en aumento a lo largo de los años, con solo dos excepciones, la década de 1940 y la de 2020. La década de 1940 tiene como explicación la Segunda Guerra Mundial, periodo en el que por motivos bélicos el número de partidos jugados disminuyó de manera considerable. La década de 2020 no ha terminado, lo que explica que el número de partidos sea menor que en la década anterior, además de que en 2020, debido a la pandemia, se jugaron menos partidos de lo habitual (347 en 2020 frente a 1115 en 2021). Sin embargo, a pesar de esto, lo normal es que la década termine al menos con un número similar de partidos a la década de 2010, ya que el número de partidos por año (más de 1000 en esta década) ha seguido aumentando

comparado con la década anterior.

En cuanto al formato de los partidos, como podemos observar en la figura [5], los partidos amistosos son los más comunes con una gran diferencia, algo lógico teniendo en cuenta que se juegan amistosos como preparación para todos los torneos.

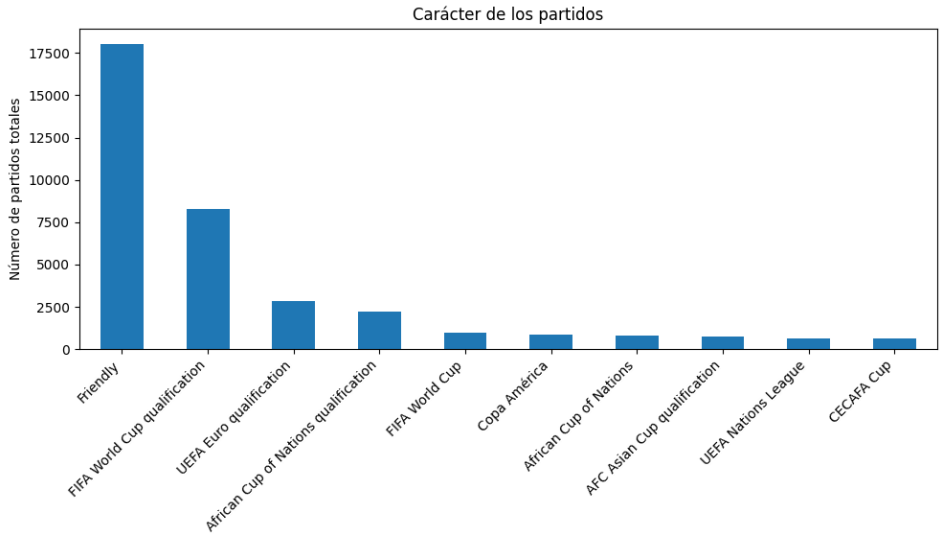

Figure 5: Competiciones más populares

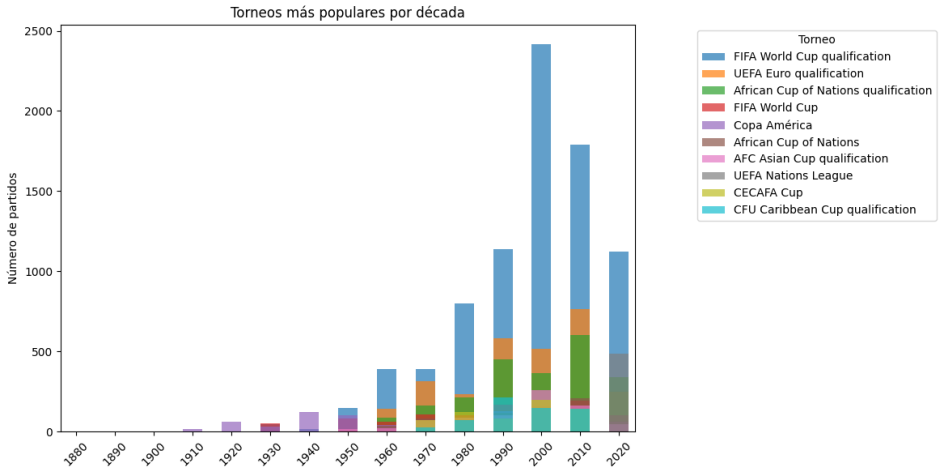

Figure 6: Competiciones más populares por década

En cuanto a torneos oficiales, desde la década de 1960 el más jugado son las clasificatorias para el mundial, mientras que antes de esto los más jugados eran competiciones contientales, como la Copa Asiática o la copa América. Cabe mencionar que en esta gráfica, dado la grandísima cantidad de torneos distintos que existen, se han cogido solo los 10 más jugados, para poder visualizar mejor los resultados. Es por esto que en las décadas de 1880, 1890 y 1900 no aparece ningún partido, ya que los torneos más jugados históricamente aún no existían en aquella época.

En las figuras 7 y 8 tenemos representado el porcentage de victorias para los partidos que se juegan en campo nuetral y los que se juegan en campo no neutral. Se ve claramente como el hecho de jugar como local tiene una gran correlación con la probabilidad de ganar. Sin embargo, al igual que en el caso de los goles en campo neutral, se observa que el equipo designado como local tiene una mayor probabilidad de ganar que el designado como visitante (44% vs 33%), que aunque no es tan grande como para el caso en campo no neutral, sigue siendo significativo teniendo en cuenta que en campo neutral ninguno de los dos equipos debería verse beneficiado a priori. Por lo tanto llama la atención que ante un dataset tan grande, con todos los partidos oficiales desde 1880, exista esta disparidad tan grande en las victorias del equipo 'local' para los partidos jugados en campo neutral.

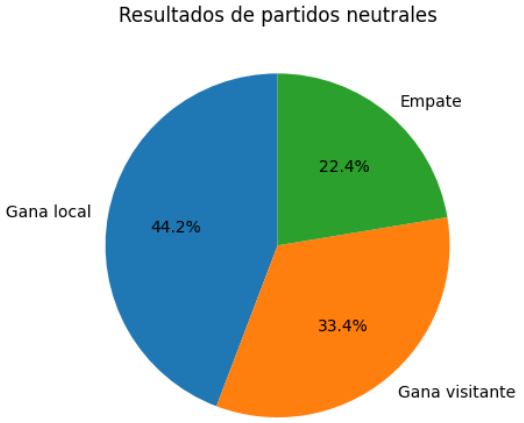

Figure 7: Resultado de los partidos jugados en campo neutral

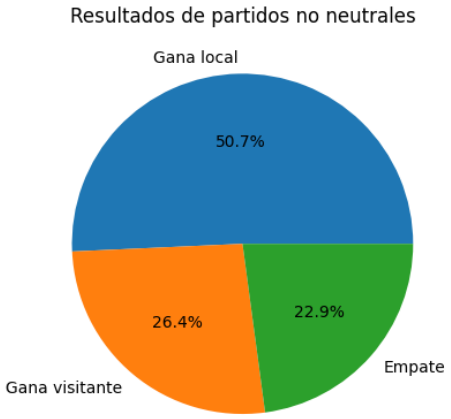

Figure 8: Resultado de los partidos jugados en campo no neutral

En la siguiente tabla podemos observar los equipos que tienen mejor rendimiento como locales. Dado que había equipos con muy pocos partidos (un solo partido en algunos casos), lo que generaba porcentajes de victoria artificialmente altos debido al escaso número de encuentros disputados, hemos aplicado un filtro para tener en cuenta solo equipos que hayan jugado al menos un total de 50 partidos.

Equipo	Win rate	partidos totales
Brazil	0.712397	605
Jersey	0.700787	127
Spain	0.687657	397
Argentina	0.670588	595
Czech Republic	0.660819	171
Guernsey	0.659091	132
Egypt	0.657407	432
Ivory Coast	0.648562	313
Iran	0.641176	340
New Caledonia	0.627586	145

Table 3: Win rate de los mejores equipos como locales

Los mejores equipos jugando como locales resultan ser los 10 expuestos en la tabla 4. En la gráfica 9 está expuesta la evolución del winrate como local para estos 10 equipos. Los mejores equipos en casa presentan porcentajes de victoria de más del 60%, con Brasil y Jersey como únicos equipos con más de un 70% de winrate. Esto supone un porcentaje considerablemente mayor que el caso general de $50,7\%$, sobre todo teniendo el cuenta la cantidad de partidos jugados por alguno de estos equipos (por ejemplo, Brasil y Argentina tienen alrededor de 600 partidos jugados como local).

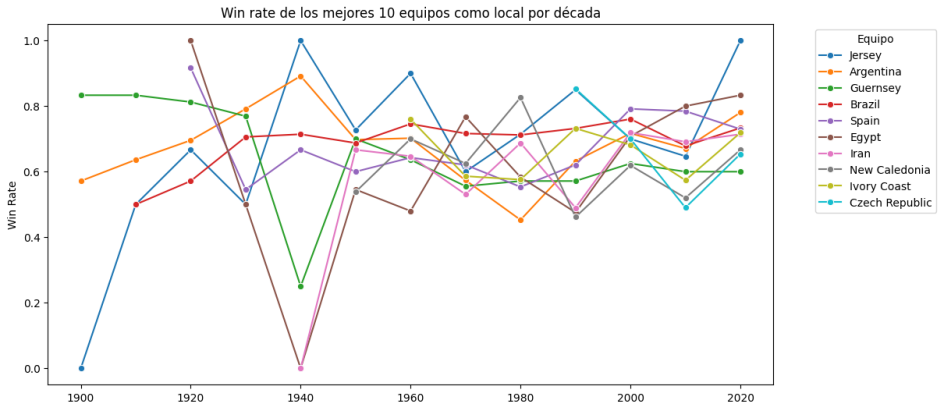

Figure 9: Winrate de los mejores equipos como locales por décadas

1.3 Reducción de Dimensionalidad

En este apartado vamos a analizar la correlación existente entre variables derivadas que hemos añadido al DataFrame y el resultado final del partido, para lo que hemos realizado una matriz de correlación. Las variables que hemos añadido son las siguientes: décadas, goles marcados y recibidos, tanto como local como visitante, carácter del partido (amistoso o de competición) número total de goles, diferencia de goles, y la media de gooles marcados y recibidos por ambos equipos en los últimos dos años. A continuación exponemos los resultados que hemos obtenido.

- En primer lugar, se observa que la década no tiene relación con el resultado del partido (gana local, gana visitante, empate), si no que simplemente tiene una correlación negativa con el número total de goles, como ya habiamos comentado antes. Aunque debemos decir que existe una mínima correlación negativa entre la década y ganar como local, y una mínima correlación positiva entre la década y empate o ganar como visitante.

- También vemos que el número total de goles y la diferencia de goles si tienen una correlación considerable con el resultado del partido. Este resultado era esperable, ya que estas dos variables reflejan información clave sobre el desenlace del partido. Se observa que el número total de goles tiene una correlación inversa con las probabilidades de empate, una leve correlación con la victoria visitante y una fuerte correlación con la victoria local. Es decir, los partidos con muchos goles tienden a beneficiar al equipo local, que suele marcar más. Esto último explica la ligera correlación entre la década y los resultados de los partidos, debido a la tendencia a la baja en el número total de goles a lo largo de las décadas.

- Se encuentra que el hecho de que el partido sea amistoso o de competición oficial no tiene una gran relación con el resultado del partido, favoreciéndose ligeramente la posibilidad de empate en comparación con victoria local o visitante, aunque esta correlación es prácticamente despreciable.

- También observamos que la media de goles marcados o recibidos a lo largo del último año sí que tiene una correlación importante con el resultado del partido, siendo más probable la victoria local si el equipo tiene una media de goles alta el último año, y una probabilidad más baja de victoria local en caso contrario. Igualmente se encuentra la misma relación entre goles marcados o recibidos como visitante y las probabilidades de victoria visitante. Además, estas 4 variables tienen una correlación negativa con la posibilidad de empate.

- Finalmente, encontramos cierta relación entre la media de goles marcados o recibidos en los dos últimos años, siendo más probable una victoria local si la media de goles marcados por el quipo local es mayor o si la media de goles recibidos por el equipo visitante es mayor, y viceversa para el equipo visitante. La relación de estas variables con la posibilidad de empate es menor, aunque resulta más improbable para un mayor número de goles marcados o recibidos.

Se va a presentar un mapa de calor de la matriz de correlación para ilustrar los resultados recién expuestos, sin embargo, no se van a añadir las décadas, ya que hace que la matriz sea demasiado grande,

dificultando considerablemente su explicabilidad, además de tener una correlación muy pequeña con el resultado del partido. Además, también eliminaremos las variables 'goles totales' y 'diferencia de goles' tanto de la matriz de correlación como del modelo que desarrollaremos, ya que esta información no está disponible antes del partido y, por lo tanto, no tiene sentido utilizarla para predecir su resultado.

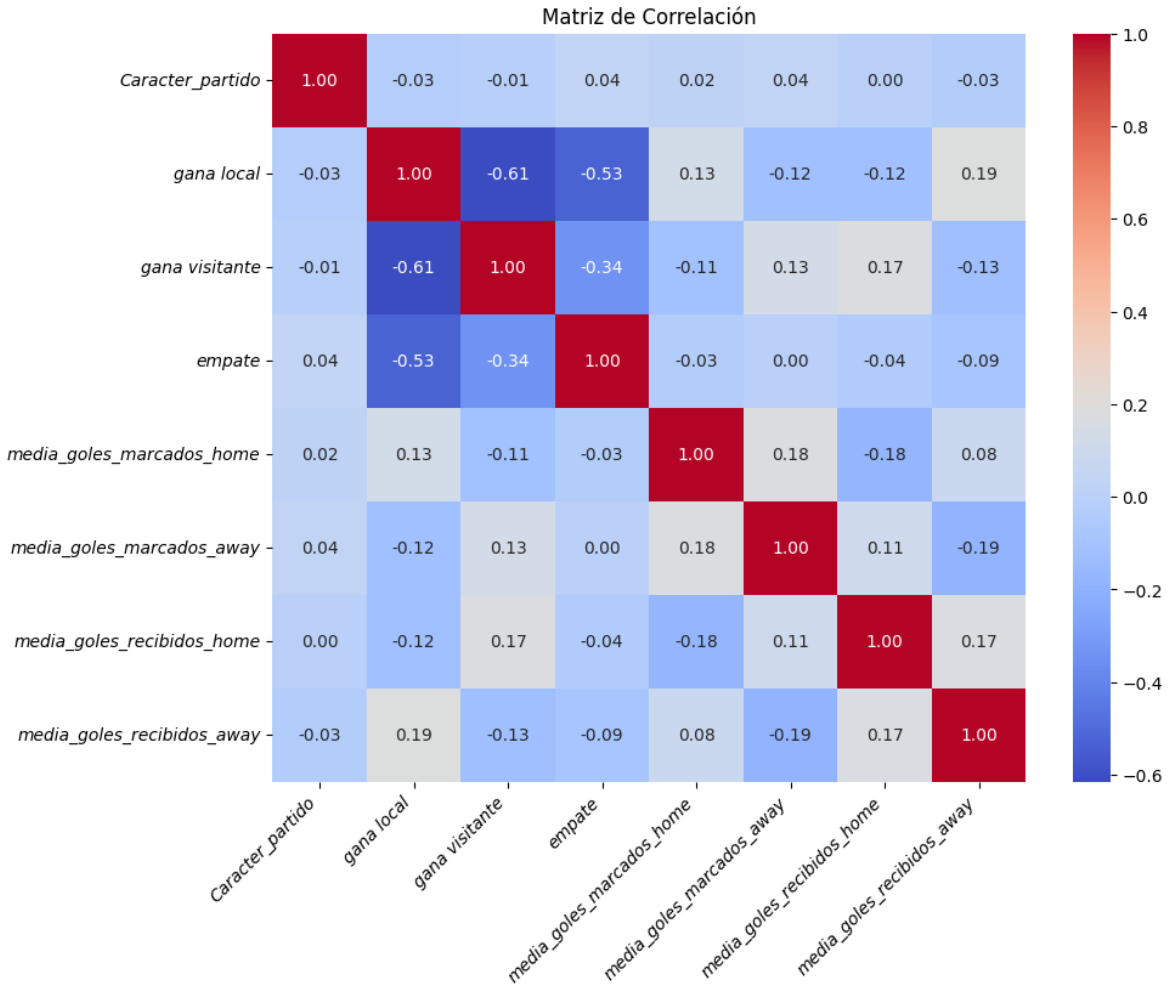

Figure 10: Matriz de correlación

A continuación vamos a exponer los resultados correspondientes al análisis de componentes principales. Para este análisis vamos a usar las mismas variables que en el caso de la matriz de correlación. No se van a añadir las décadas por que de nuevo harían la interpretación de los resultados considerablemente más complicada, y como más adelante usaremos las componentes principales para modelos predictivos, no se van a añadir las variables que puedan darle información del resultado. Es decir, solo vamos a añadir las variables que se van a encontrar presentes en el análisis de componentes principales que utilizaremos más adelante en los modelos predictivos.

Componente principal 1	Componente principal 2	Componente principal 3
0.18142	0.16165	0.15283

Table 4: Varianza explicada por cada componente principal

En la figura 11 se encuentran representados los pesos de los componentes principales. Se observa que las variables con mayor peso son las medias de goles marcados en los dos últimos años. Esto nos indica que la variable más importante sería el momento de forma del equipo. Después de estas

variables, la siguiente más importante son los equipos involucrados en el partido, tanto el local como el visitante, y finalmente el carácter del aprtido, es decir, si es amistoso o de competición.

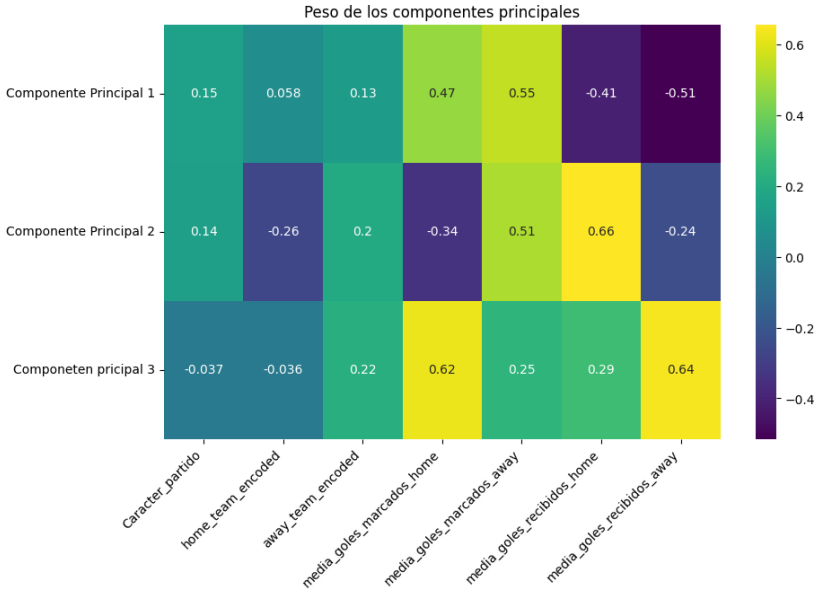

Figure 11: Peso de los componentes principales

Finalmente tenemos representados los partidos clasificados por resultado frente a la componente principal 1 y 2. Esta gráfica resulta complicada de interpretar, pero podemos observar que los empates tienden a estar menos dispersos, concentrandose alrededor de valores bajos de las componetnes. Esto tiene sentido, ya que como hemos comentado antes, el número de goles está inversamente relacionado con la posibilidad de empate, y las principales variables de estas componentes son la media de goles marcados. También podemos observar que las vistorias visitantes se encuentras más dispersas que las victorias como local, lo que puede indicar ue valroes grandes de los componetnes principales, y consecuentemente la tendencia goleadora de los equipos, está correlacionada con las posibilidades de victoria visitante.

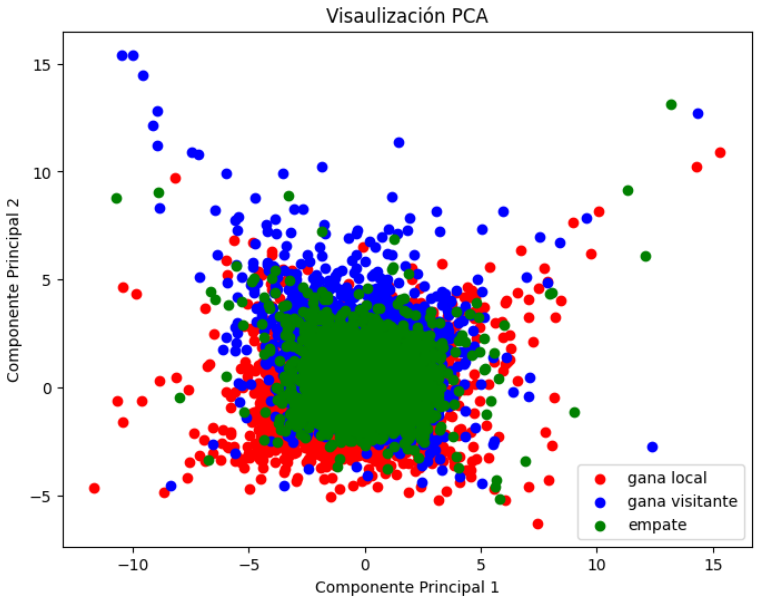

Figure 12: Visualización de las componentes principales

2 Análisis Avanzado y Modelos Predictivos

2.1 Ventaja de Localía y Probabilidades Condicionales

Ya hemos calculado previamente las probabilidades ganar, empatar o perder para partidos neutrales y no neutrales usando el método frecuentista. Ahora vamos a hacer para la probabilidad de ganar como local y la probabilidad de empatar en campo neutral utilizando el método de Bayes. Para ello, en vez de simplemente contar las frecuencias y dividir por el número total de partidos, vamos a utilizar el teorema de Bayes:

$$P(H|D) = \frac{P(D|H) * P(H)}{P(D)} \tag{1}$$

donde $P(H|D)$ es la probabilidad a posteriori, $P(D|H)$ es la verosimilitud, P(D) es la probabilidad marginal y P(H) la probbilidad a priori.

- Probabilidad de ganar jugando como local

  Para este caso, donde queremos calcular la probabilidad de ganar jugando como local, $P(D|H)$ representa la probabilidad de jugar como local dado que se gana el partido, P(D) es la probabilidad de jugar como local y P(H) la probabilidad de ganar. Utilizando este método obtenemos que la probabilidad de ganar como locl es de un $51,4\%$, la cual es bastante cercana a la probabilidad frecuentista.

- Probabilidad de empate jugando en campo neutral

  Para la probabilidad de empate jugando en campo neutral, $P(D|H)$ representa la probabilidad de jugar en campo neutral dado que el resultado es empate, P(D) es la probabilidad de jugar en campo neutral y P(H) es la probabilidad de empate. En este caso también obtenemos una probabilidad prácticamente idéntica al caso frecuentista, con un $22,4\%$.

	Probabilidad de ganar como local	Probabilidad de empate campo neutral
Frecuentista	$50,7\%$	$22,4\%$
Bayesiana	$51,4\%$	$22,4\%$

Table 5: Comparación de probabilidades entre el enfoque frecuentista y el bayesiano

2.2 Modelos de Clasificación

En este apartado vamos a exponer los modelos para la predicción del resultado de los partidos. En primer lugar, hemos eliminado cualquier variable que le pudiese proporcionar informaciónm al modelo acerca del resultado, es decir, hemos eliminado las variables 'home_score' y 'away_score', 'diferencia de goles' y 'goles totales'. Si estas se incluyeran, le estariamos diciendo el resultado al modelo, y los resultados no serían indicativos de su capacidad de predicción real.

2.2.1 GaussianNB

El primer modelo que hemos realizado ha sido un GaussianNB con los componentes principales. Con esto, obtenemos una precisión de alrededor del $51,83\%$. Cuando se ahn utilizado las variables originales y derivadas, es decir, sin hacer la reducción de dimensionalidad, obtenemos un resultado muy parecido, con una precisión del $51,81\%$. Si analizamos como clasifica cada clase, observamos que en el caso de las componetnes principales no clasifica ningúin partido como empate, mientras que en el caso de las variables normales, es capaz de clasificarlos con un 28% de precisión.

2.2.2 Regresión Logística

Utilizando la regresión logística obtenemos los mismos resultados que en el apartado anterior, con una precisión del $52,98\%$ para las variables normales y un $51,66\%$ para el caso de las componentes principales. Además, en este caso ninguno de los dos clasifica partidos como empate.

2.2.3 CategoticalNB

Para este caso, no es posible realizar el análisis de componentes principales, ya que este modelo no admite valores negativos, y el análisis de componentes principales inserta valores negativos al realizar la reducción de dimensionalidad. En este caso, el modelo es capaz de asignar el resultado empate a un partido, aunque lo hace con una precisión bastante mala, de tan solo un 34%. Con esto, este modelo es capaz de predecir el resultado del partido con un $55,4\%$ de precisión.

2.2.4 RandomForest

En el caso de Random Forest, obtenemos una precisión del $51,35\%$ para las variables originales y derivadas, mientras que la precisión empeora hasta un $44,56\%$ para las componentes principales. En este caso si obtenemos predicciones de empate, pero con una precisión menor que para el CategoricalNB, con aproximadamente un 30% de precisión para ambos casos.

Podemos concluir por lo tanto que los modelos, sin añadir ninguna variable que le de algún tipo de información acerca del resultado del partido, no son capaces de predecir adecuadamente el resultado del partido, más allá de precisiones de alrededor del 50%, siendo el mejor el NaiveBayes categórico con las variables originales y derivadas, con una precisión del $55,4\%$. Si añadimos algún dato del resultado, como la diferencia de goles, los modelos mejoran considerablemente, sin embargo, esta información no estaría disponible si quisieramos hacer una predicción de un partido que fuese a jugarse mañana, por lo que hemos concluido que no tiene sentido añadirlo a pesar de la mejora en precisión. Si que se podrían añadir otro tipo de estadísticas más avanzadas al modelo, como porcentaje de posesión, estilo de juego o lesiones para darle una imagen más completa de los equipos, sin embargo no tenemos disponibles esas estadísticas con el dataset con el que estamos trabajando. También debemos comentar que los modelos han tenido problemas serios para clasificar correctamente los empates, por lo que tal vez se podria mejorar los modelos balanceando las clases para que el modelo sea mejor reconociendo los empates. En el notebookasociado al informe se puede ver un análisis más detallado de las métricas de precisión, con la precisión, recall y f-1 score para cada método.

Model	Variables originales	Componentes principales
GaussianNB	$51,81\%$	$51,83\%$
Regresión Logística	$52,98\%$	$51,66\%$
CategoricalNB	Nan	$55,32\%$
RandomForest	$51,35\%$	$44,56\%$

Table 6: Precisión de los distintos modelos desarrollados

3 Redes y Técnicas Avanzadas

Vamos a presentar un grafo donde cada nodo representa a los distintos equipos, y cada arista representa partidos jugados entre los equipos que unen. Debido al grandísimo número de partidos, hacer un grafo de todo el Dataset no tiene mucho sentido, ya que es imposible hacer una interpretación del mismo. Por lo tanto vamos a ir aplicando filtros y mostrando distintos grafos, calculando métricas de centralidad, grado y componentes conectados para identificar equipos clave. Lo primero que vamos a hacer es eliminar todos los partidos que no sean de competición, es decir, aquellos que están marcados como 'friendly'. Aún así el dataset sigue siendo muy grande, con lo que vamos a imponer que los equipos hayan jugado al menos un total de 250 partidos. Con esto obtenemos el siguiente grafo o centralidad.

Tras este filtrado nos encontramos con un grafo complejo pero donde se pueden empezar a observar patrones. Se puede ver como existe una cierta agrupación por continentes, existiendo cuatro grupos diferenciados, siendo estos los correspondientes a África (arriba), Asia (derecha), Europa (izquierda) y América (abajo). Este es un resultado lógico ya que es de esperar que los equipos próximos jueguen más a menudo. Este grafo está compuesto por 100 nodos y 1819 aristas, es decir, tenemos 100 equipos y 1819 combinaciones de partidos entre ellos. En la tabla 7 podemos ver los equipos más relevantes según su grado de conexión en el grafo. Finalmente el resultado de estudiar los componentes conectados nos

da un resultado de 1, es decir, tenemos un grafo altamente interconectado, donde no existen subgrupos bien definidos, si no una red altamente conectada.

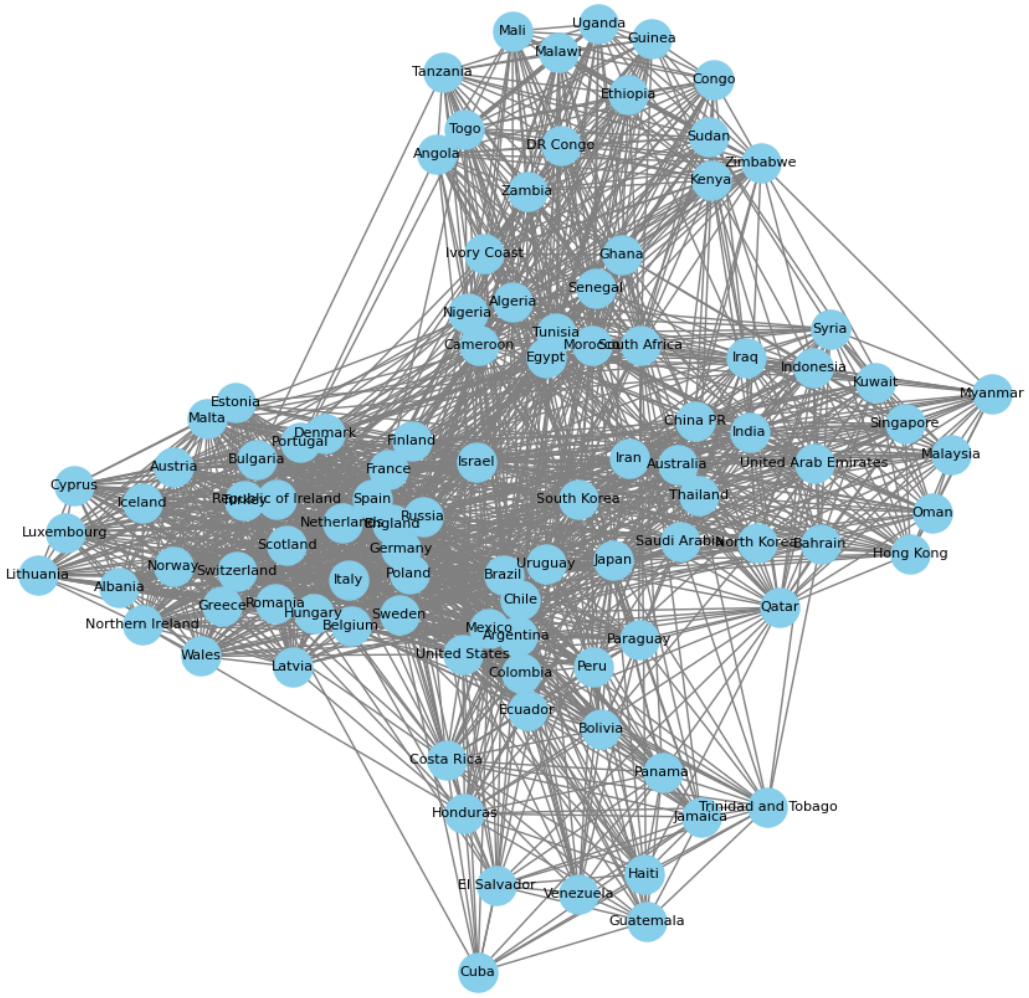

Figure 13: Grafo de partidos de todas las competiciones

Equipo	Centralidad	Grado
Japan	0.636	63
South Korea	0.596	59
Brazil	0.576	57
Germany	0.556	55
Sweden	0.545	54
Egypt	0.545	54
Spain	0.545	54
Russia	0.545	54
England	0.535	53
France	0.535	53

Table 7: Grado y centralidad de todas las competiciones

A continuación vamos a realizar el mismo análisis pero para los grafos de competiciones continentales. De nuevo, obtenemos redes muy conectadas, y casi sin subgrupos o equipos poco conectados. Si que cabe mencionar que vemos diferencias significativas entre las redes de los distintos continentes. En primer lugar, podemos observar que el grafo de las competiciones europeas es considerablemente más complejo que el resto, debiéndose esto al mayor númerod e países europeos representados en comparación con el resto de continentes, y seguramente también a la mayor tradición de fútbol existente en este país, lo que hace que se hayan jugado históricamente más partidos en estas competiciones, comparadas con sus homólogas en otros países.

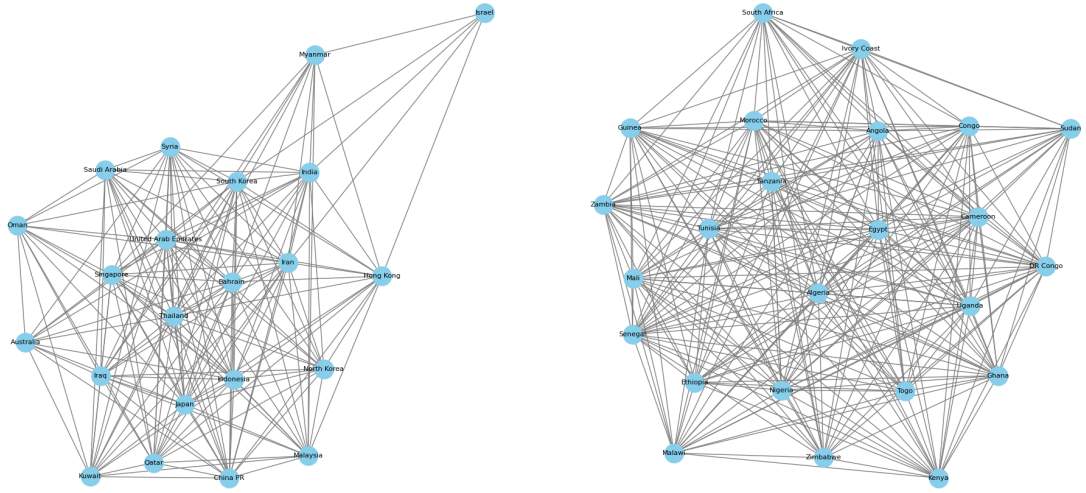

Figure 14: Grafo de competiciones asiáticas (izquierda) y competiciones africanas (derecha)

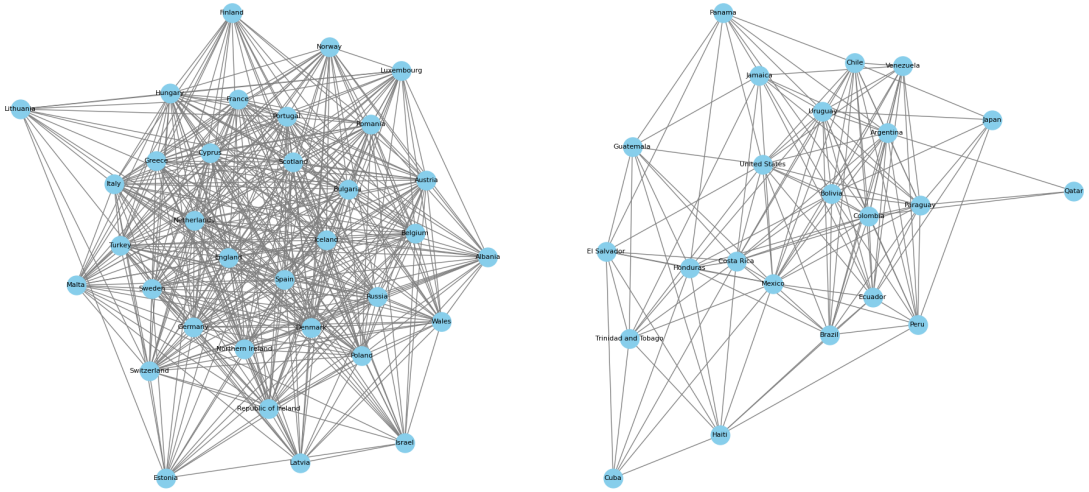

Figure 15: Grafo de competiciones europeas (izquierda) y competiciones americanas (derecha)

Además, como podemos observar en la tabla 8, estos países se encuentran muy bien conectados, con métricas de centralidad altas. En contraste, el grafo de las competiciones americanas muestra

una menor complejidad y presenta nodos con centralidades significativamente más bajas que en otros continentes. Igualemente todos los grafos están muy conectados, teniendo todos un solo componente conectado, es decir, se puede alcanzar cualquier nodo desde cualquier otro nodo directa o indirectamente.

Continente	Top Equipos	Grado	Centralidad
América	Mexico	19	0.86
	Uruguay	16	0.73
	Costa Rica	16	0.73
	Argentina	15	0.68
	Paraguay	15	0.68
África	Egypt	23	1.00
	Nigeria	23	1.00
	Uganda	23	1.00
	Tunisia	23	1.00
	Morocco	23	1.00
Asia	Iran	20	0.95
	South Korea	20	0.95
	Thailand	20	0.95
	Japan	19	0.90
	Indonesia	19	0.90
Europa	Portugal	31	0.94
	Germany	31	0.94
	Spain	30	0.91
	Denmark	30	0.91
	Netherlands	30	0.91

Table 8: Grado y centralidad por continentes