Memoria proyecto

Contenido

Índice de Ilustraciones

Índice de tablas

1. Introducción

"El fútbol es la cosa más importante de las cosas menos importante", Arriago Sacchi. [1]

El fútbol es más que un simple deporte, es la pasión que late en el corazón de millones de personas en todo el mundo. Como bien dijo Arriago Sacchi*, "El fútbol es la cosa más importante de las cosas menos importante"*, y es que, para muchos aficionados, este deporte es una forma de vida, una razón para seguir adelante, un escape de la rutina y los problemas cotidianos.

Cada fin de semana, los estadios se llenan de almas apasionadas, que animan, sufren y disfrutan del espectáculo del fútbol. Y aunque algunos lo miren con desdén y tristeza, lo cierto es que el fútbol es uno de los mayores negocios en nuestra sociedad actual, moviendo cantidades ingentes de dinero.

Pero hay una competición que sobresale por encima de todas las demás, la Copa del Mundo de la FIFA, que cada cuatro años reúne a los mejores equipos y jugadores del mundo en una fiesta del deporte y la cultura. En 2024 se elegirá la sede del próximo mundial, el de 2030, y en este proyecto de *datamining* se ha investigado por qué España debería presentar su candidatura y con qué estadios debería hacerlo.

Más allá de la emoción y el espectáculo que supone acoger un evento de tal magnitud, la organización de un Mundial puede tener un impacto económico y social significativo para un país. Por eso, en este proyecto se ha llevado a cabo una predicción de la variación del GDP de España tras la organización de este mundial, para demostrar que no solo es una cuestión de pasión y emoción, sino también de desarrollo y progreso para nuestro país.

En definitiva, el fútbol puede ser considerado por algunos como una simple afición, pero para muchos es mucho más que eso. Es un motor de cambio, una fuente de inspiración y una herramienta para unir a las personas más allá de las fronteras y diferencias culturales. Por todo esto, España debería presentar su candidatura al Mundial de 2030 y demostrar al mundo que estamos preparados para acoger uno de los mayores eventos deportivos y culturales del planeta.

2. Motivación

La Copa del Mundo de la FIFA es el evento deportivo más importante y esperado en todo el mundo, y España ha sido una de las grandes potencias futbolísticas en la historia de este deporte. Es por ello por lo que presentar la candidatura para organizar el mundial de 2030 es un reto que no podemos dejar pasar. Un evento de tal magnitud no solo traería un gran beneficio económico al país, sino que también reforzaría el espíritu deportivo y uniría a toda la nación detrás de un objetivo común. Además, la elección de los estadios adecuados es crucial para garantizar una experiencia única e inolvidable tanto para los equipos participantes como para los aficionados que asistirían a los partidos. Por todo ello, en este trabajo se ha realizado un análisis exhaustivo de las razones por las que España debería presentar su candidatura para el mundial de 2030 y de los estadios más adecuados para acoger el evento.

3. Objetivos

Analizar el impacto económico que la organización del Mundial de Fútbol de 2030 podría tener en España, a través de la realización de un estudio estadístico y de predicción del crecimiento del Producto Interno Bruto (PIB) del país durante los años siguientes a su realización.

Estudiar la mejor distribución de los estadios de fútbol españoles para presentar como sedes mundialistas, analizando sus estadísticas en la última temporada.

Identificar y analizar las principales características de los equipos y jugadores que participarían en el Mundial de Fútbol de 2030, con el objetivo de determinar cómo afectaría la actuación de estos en la competición en la economía del país.

Realizar un análisis de los beneficios y riesgos asociados a la organización del Mundial de Fútbol de 2030, en términos económicos, con el objetivo de determinar si es una iniciativa rentable.

4. Implementación

En esta parte de la memoria se va a explicar cada una de las técnicas de *datamining* utilizadas para el desarrollo del proyecto, los datos utilizados y se mostrarán los resultados del estudio.

Para desarrollar el proyecto se ha hecho uso del lenguaje de programación Python y el IDE *Visual Studio Code,* en el IDE se hace uso de la extensión *jupyter* para poder trabajar en el notebook donde ha sido desarrollada la solución.

4.1. Datasets

Los *datasets* utilizados son uno con varios *csv* de datos históricos mundialistas y otro con los GDP de gran parte de los países del mundo.

4.1.1. Dataset Histórico de la FIFA

En el *dataset* Histórico de la FIFA, se muestran diversos datos históricos de los mundiales de la FIFA. Se han usado *FIFA-Winners.csv* [2], en el que se recogen los datos de, campeón, subcampeón, tercero y cuarto puesto, goles, partidos, equipos, año y organizador de cada mundial desde el 1960, primer mundial de la historia. El csv *young_player.csv* muestra los jugadores que obtuvieron el premio a mejor jugador joven en cada mundial, junto con la edad del jugador, la nacionalidad y el continente. El *csv goldenSilverBronzeBallbyplayer.csv* mostraba los jugadores galardonados con el balón de oro, balón de bronce y plata en cada mundial, pero se ha decidido usar tan solo el balón de oro porque es el más relevante. El *csv FIFA_Attendance.csv* muestra las estadísticas de asistencia a los distintos mundiales. Todos estos *csv* se han almacenado en una estructura de tipo *dataframe* de Python para poder manipular los datos y trabajar mejor con ello.

4.1.2. Dataset GDP de cada país

En este *dataset* se tienen datos de los distintos países con su GDP desde 1960 hasta 2021. Para poder hacer un uso correcto de este, se ha decidido almacenar estos datos en un *dataframe.* Este *dataframe,* contiene los datos que se encuentran en el *csv countries_gdp_hist.csv* [3]. Este *csv* contiene: código de país, nombre de la región, nombre de la subregión, región intermedia, nombre del país, grupo, año, GDP total, GDP total en millones de euros, variación del GDP.

Estos datos se han combinado para poder sacar conclusiones de cuan rentable es organizar un mundial y cuáles son las mejores condiciones para organizarlo.

4.2. ¿Cómo varía la organización de un mundial en el GDP del país anfitrión?

Para saber si es rentable y cómo de rentable es organizar un mundial, se ha decidido analizar como varía el GDP en los 3 años antes de organizar el mundial contra los 3 años posteriores de organizarlo. Para ello se ha realizado una labor de preprocesamiento de los datos y un diagrama de barras para analizar mejor los resultados obtenidos.

Para analizar la variación del GDP, se han combinado los datos de los *csv FIFA_Attendance.csv* y *countries_gdp_hist.csv.*

4.2.3. Preprocesamiento

Para poder hacer el estudio de cómo varía el GDP del país organizador del mundial en los años posteriores y anteriores a la cita mundialista en primer lugar se hace un preprocesamiento de los datos del *dataset*.

4.2.3.1. Limpieza de los datos

El primer paso en el preprocesamiento de los datos es la limpieza del *dataset*. En primer lugar, se ha limpiado el *dataframe* de los mundiales históricos. Se han mantenido tan solo las columnas: año, campeón, subcampeón, tercer puesto y país organizador. Como se tiene datos del GDP de los países desde 1960 hasta 2021, se han eliminado los mundiales anteriores a 1960 y los posteriores a 2022. Ambos inclusive.

Para el *dataframe* de los datos del GDP tan solo se mantienen las columnas: nombre del país, GDP total en millones de euros y la variación del GDP. Esta limpieza se ha realizado ya que se considera que los demás datos que aparecen en el CSV son irrelevantes para el estudio.

Por último, se han unido los *dataframes* limpiados uniéndolos por el nombre del país organizador. De esta manera, el resultado del *dataframe* serán los años desde 1960 hasta 2021 para cada país con el GDP en millones y la variación del GDP para cada año. Además, se han eliminado del *dataframe* todos aquellos países que no hayan organizado un mundial.

Además, el mundial de Corea y Japón ha sido eliminado del análisis pues fueron dos países los que organizaron el mundial.

ID	Country_name	year	total_gdp_million	gdp_variation
372	Argentina	1960	0.000000	0.000000
373	Argentina	1961	0.000000	5.427843
374	Argentina	1962	24450.604878	-0.852022
375	Argentina	1963	18272.123664	-5.308197

Tabla 1 Modelo de Dataframe para analizar los datos y después del Mundial

4.2.3.2. Generación de datos sintéticos

Se han tenido que generar datos sintéticos para aquellos países en los que no tenemos suficientes datos de GDP para evaluar, estos son Chile, que organizó el mundial en 1962 y los datos de 1958, 1959 y 1960 están sesgados o no existen. Para ello, se han utilizado los datos de 1961, 1962 y 1963 para generar datos sintéticos para esos años.

Creación de datos sintéticos de Chile

```python
# Get data for Chile
df_chile = datasetGDP[datasetGDP['country_name'] == 'Chile']

# Get the value of GDP in 1961
gdp_1959 = df_chile[df_chile['year'] == 1963]['total_gdp_million'].values[0]

# Calculate the value of GDP in 1962 if the value in 1961 is greater than zero
if gdp_1959 > 0:
    gdp_1958 = gdp_1959 / (1 + df_chile[df_chile['year'] == 1962]['gdp_variation'].values[0])
else:
    gdp_1958 = 0

# Calculate the value of GDP in 1963 if the value in 1962 is greater than zero
if df_chile[df_chile['year'] == 1963]['total_gdp_million'].values[0] > 0:
    gdp_1960 = gdp_1959 * (1 + df_chile[df_chile['year'] == 1963]['gdp_variation'].values[0])
else:
    gdp_1960 = 0

# Create the synthetic data
df_sintetico = pd.DataFrame({
    'country_name': 'Chile',
    'year': [1958, 1959, 1960],
    'total_gdp_million': [gdp_1958, gdp_1959, gdp_1960],
    'gdp_variation': [0, df_chile[df_chile['year'] == 1963]['gdp_variation'].values[0], 0]
})

# Concatenate the synthetic data with the original dataset
df_nuevo = pd.concat([datasetGDP, df_sintetico])

# Sort the dataset by country name and year
df_ordenado = df_nuevo.sort_values(by=['country_name', 'year'], ascending=[True, True])

# Save the new dataset in a CSV file
df_ordenado.to_csv('PIB-GDP/datos_sinteticos.csv', index=False)
```

Python

Ilustración 1 Código de Creación de datos Sintéticos de Chile

Para Rusia, se han tenido que generar datos sintéticos para 2022 puesto que en el CSV no existen, para poder generarlos, se han utilizado los datos del GDP de los 5 años anteriores a 2022 y se ha hecho una tasa de crecimiento promedio para este año.

Creación de datos Sintéticos de Rusia

```python
# seleccionar los datos de Rusia y los últimos 5 años
rus_df = datasetGDP[(datasetGDP['country_name'] == 'Russia') & (datasetGDP['year'] >= 2017)]

# calcular la tasa de crecimiento promedio del PIB
avg_gdp_variation = rus_df['gdp_variation'].mean()

# obtener el valor del PIB para 2021
gdp_2021_rus = datasetGDP[(datasetGDP['country_name'] == 'Russia') & (datasetGDP['year'] == 2021)]['total_gdp_million'].values[0]

# aplicar la tasa de crecimiento promedio al valor del PIB de 2021 para obtener una estimación del PIB para 2022
gdp_2022_rus = gdp_2021_rus * (1 + avg_gdp_variation)

# crear un nuevo dataframe con los datos sintéticos para 2022
new_row = {
    'country_name': 'Russia',
    'year': 2022,
    'total_gdp_million': gdp_2022_rus,
    'gdp_variation': avg_gdp_variation
}
DataSetSinteicos = datasetGDP.append(new_row, ignore_index=True)

# guardar el nuevo dataframe en un archivo CSV
DataSetSinteicos.to_csv('PIB-GDP/datos_sinteticos.csv', index=False)

datasetGDP = DataSetSinteicos
```

Ilustración 2 Código de creación de los Datos Sintéticos de Rusia

4.2.4. Análisis de los resultados

Tras haber realizado el preprocesamiento de los datos, los ilustramos en un gráfico de barras que permita apreciar la evolución del PIB de cada país en los años anteriores y posteriores, adicionalmente incluimos una tabla con los datos específicos.

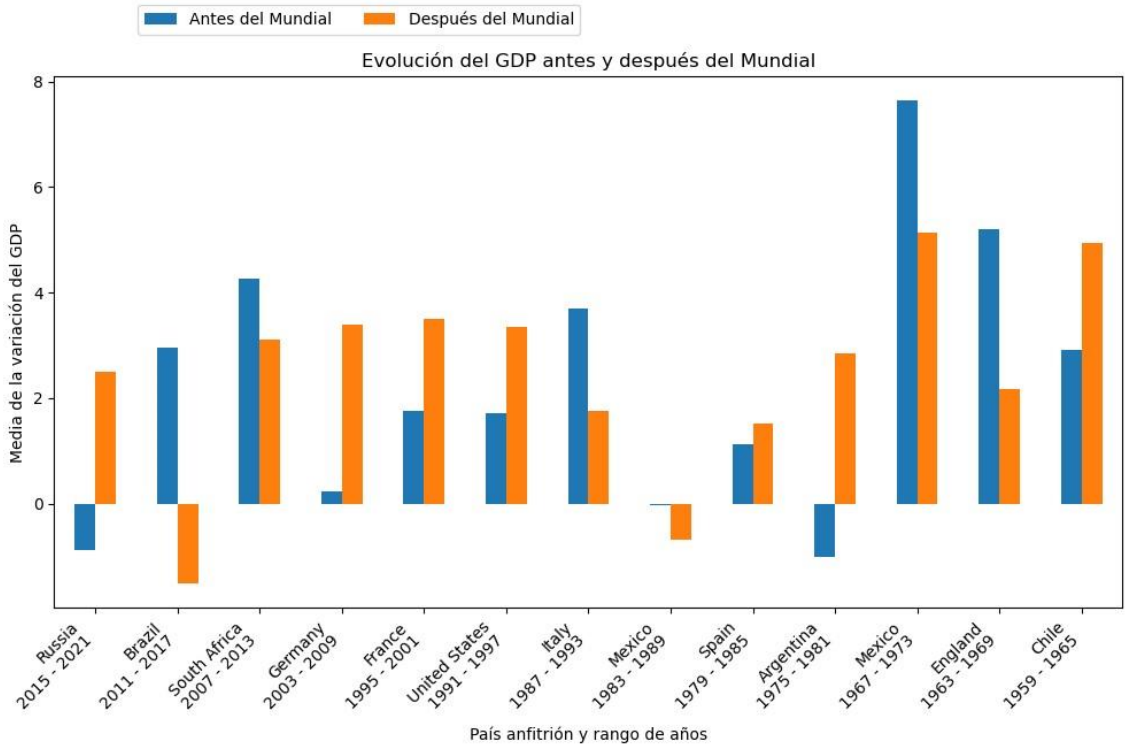

Ilustración 3 Diagrama de barras de GDP

En la tabla se pueden apreciar países en diferentes categorías y con diferentes peculiaridades:

- Países los cuales ser sede del mundial tuvo un impacto positivo, es decir, se incrementó el PIB tras haber sido anfitrión del mundial respecto a cómo se encontraba con anterioridad al mundial, aquí se podría encontrar países como:

    - Rusia cuyo PIB aumentó de -0.88 a 2.50 ○ Alemania cuyo PIB aumentó de 0.23 a 3.39
    - Francia cuyo PIB aumentó de 1.75 a 3.50

- Países los cuales ser sede del mundial tuvo un impacto negativo, es decir, se decrementó su PIB tras ser el anfitrión del mundial, respecto a cómo se encontraba con anterioridad al mundial, aquí se podría encontrar países como: ○ Brasil cuyo PIB descendió de 2.94 a -1.52. ○ Sudáfrica cuyo PIB descendió de 4.27 a 3.10
    - México cuyo PIB descendió de -0.037 a -0.67

- En algunos países, como Estados Unidos y Francia, el impacto económico positivo del Mundial duró solo unos años después del evento. En cambio, en otros países, como Sudáfrica, el impacto económico se mantuvo durante varios años después del evento. Esto es debido a que hay varios factores que influyen en cuánto tiempo puede durar el impacto económico positivo. Uno de los principales factores es la inversión en infraestructura realizada para albergar el evento, como la construcción de estadios, la mejora de la red de transporte público y la ampliación de la capacidad hotelera. Si esta infraestructura se utiliza de manera efectiva después del evento para atraer más turismo y actividades económicas, el impacto económico puede durar varios años. Por otro lado, si la infraestructura no se utiliza de manera efectiva después del evento, el impacto económico puede desaparecer rápidamente.

- Existen diferencias regionales en el impacto económico del Mundial. Por ejemplo, los países de América Latina experimentaron un aumento significativo en su PIB después de albergar un Mundial, mientras que los países europeos experimentaron un aumento más modesto y existen varias razones por las que los países de América Latina pueden experimentar un mayor aumento en su PIB después de albergar el evento en comparación con los países europeos. Uno de los factores clave es el tamaño de la economía del país anfitrión en relación con el evento. Si el país anfitrión es un mercado importante para los productos y servicios relacionados con el fútbol, es más probable que experimente un aumento significativo en su PIB. Además, en algunos países de América Latina, el turismo y las actividades económicas relacionadas con el fútbol pueden tener una mayor importancia en su economía en comparación con los países europeos.

country_name	Antes del Mundial	Después del Mundial	Rango de años
Russia	-0.889515	2.502661	2015 - 2021
Brazil	2.947800	-1.520904	2011 - 2017
South Africa	4.275759	3.104145	2007 - 2013
Germany	0.237486	3.396449	2003 - 2009
France	1.759844	3.505017	1995 - 2001
United States	1.707088	3.356505	1991 - 1997
Italy	3.693169	1.762111	1987 - 1993
Mexico	-0.037804	-0.678258	1983 - 1989
Spain	1.125137	1.508289	1979 - 1985
Argentina	-1.023332	2.858319	1975 - 1981
Mexico	7.639102	5.132476	1967 - 1973
England	5.204022	2.179788	1963 - 1969
Chile	2.920070	4.933413	1959 - 1965

Ilustración 4 Tabla de variacion GDP mundiales

4.3. ¿Qué variables afectan más en la rentabilidad positiva de organizar un mundial?

4.3.5. Combinación de datasets

Para responder a la pregunta sobre qué variables tienen mayor impacto en la rentabilidad positiva de organizar un mundial, se ha llevado a cabo la combinación de los diferentes conjuntos de datos en un único *dataset* que recoge las diversas características de cada mundial [4]. Esto nos ha permitido contar con suficientes datos para identificar las variables más relevantes en la obtención de una rentabilidad positiva.

El código que se muestra a continuación es el proceso de combinación de datos, en el cual se agregan al dataset principal (df_gdp_host) los demás datasets que contienen información relevante sobre los Mundiales, como los goles por partido, la asistencia total y media en los estadios, los ganadores del Balón de Oro y del jugador joven, y los datos económicos de los países anfitriones antes y después de la celebración del Mundial.

Al realizar esta combinación de datos, se ha podido contar con suficientes datos para identificar las variables más relevantes en la obtención de una rentabilidad positiva. El resultado final es el dataset df_mundiales, que contiene toda la información necesaria para poder llevar a cabo el análisis de las variables más influyentes en la rentabilidad positiva de organizar un Mundial.

```
df_mundiales = df_goals_match
df_mundiales = pd.merge(df_mundiales, df_attendance[['Year', 'Hosts', 'Venues', 'Totalattendance †', 'Averageattendance']], on='Year', how='outer')
df_mundiales = pd.merge(df_mundiales, df_goldenBall[['Continent', 'Year']], on='Year', how='outer')
df_mundiales = df_mundiales.rename(columns={'Continent': 'Continent_GB_Winner'})
df_mundiales = pd.merge(df_mundiales, df_young_player[['Continent', 'Year']], on='Year', how='outer')
df_mundiales = df_mundiales.rename(columns={'Continent': 'Continent_GB_Young_winner'})
df_gdp_host = df_gdp_host.rename(columns={'country_name': 'Hosts'})
df_mundiales = pd.merge(df_mundiales, df_gdp_host[['Antes del Mundial', 'Después del Mundial', 'Rango de años', 'Hosts']], on='Hosts', how='outer')
df_mundiales['Diferencia'] = df_mundiales['Después del Mundial'] - df_mundiales['Antes del Mundial']
rango_de_anios = df_mundiales.pop('Rango de años')
df_mundiales['Rango de años'] = rango_de_anios
df_mundiales = df_mundiales.drop(df_mundiales[df_mundiales['Year'] == 2002].index)
df_mundiales = df_mundiales.drop(df_mundiales[df_mundiales['Year'] == 1974].index)

df_mundiales = df_mundiales.sort_values(by='Year', ascending=False)
df_mundiales1 = df_mundiales
```

Ilustración 5 Codigo combinación datasets

Con este código nos queda un dataset con la siguiente forma:

	Year	Champion	goals	matches	goals_per_match	Hosts	Venues	Totalattendance †	Averageattendance	Continent_GB_Winner	Continent_GB_Young_winner	Antes del Mundial	Después del Mundial	Diferencia	Rango de años
0	2018	France	169	64	2.640625	Russia	12	3,031,768	47,371	Europe	Europe	-0.889515	2.502661	3.392175	2015 - 2021
1	2014	Germany	171	64	2.671875	Brazil	12	3,429,873	53,592	South America	Europe	2.947800	-1.520904	-4.468703	2011 - 2017
2	2010	Spain	145	64	2.265625	South Africa	10	3,178,856	49,670	South America	Europe	4.275759	3.104145	-1.171614	2007 - 2013
3	2006	Italy	147	64	2.296875	Germany	12	3,359,439	52,491	Europe	Europe	0.237486	3.396449	3.158963	2003 - 2009
5	1998	France	171	64	2.671875	France	10	2,785,100	43,517	South America	Europe	1.759844	3.505017	1.745172	1995 - 2001
6	1994	Brazil	141	52	2.711538	United States	9	3,587,538	68,991	South America	Europe	1.707088	3.356505	1.649417	1991 - 1997
7	1990	Germany	115	52	2.211538	Italy	12	2,516,215	48,389	Europe	Europe	3.693169	1.762111	-1.931058	1987 - 1993
8	1986	Argentina	132	52	2.538462	Mexico	12	2,394,031	46,039	South America	Europe	-0.037804	-0.678258	-0.640454	1983 - 1989
9	1986	Argentina	132	52	2.538462	Mexico	12	2,394,031	46,039	South America	Europe	7.639102	5.132476	-2.506626	1967 - 1973
12	1982	Italy	146	52	2.807692	Spain	17	2,109,723	40,572	Europe	Europe	1.125137	1.508289	0.383151	1979 - 1985
13	1978	Argentina	102	38	2.684211	Argentina	6	1,545,791	40,679	South America	Europe	-1.023332	2.858319	3.881651	1975 - 1981
10	1970	Brazil	95	32	2.968750	Mexico	5	1,603,975	50,124	South America	South America	-0.037804	-0.678258	-0.640454	1983 - 1989
11	1970	Brazil	95	32	2.968750	Mexico	5	1,603,975	50,124	South America	South America	7.639102	5.132476	-2.506626	1967 - 1973
15	1966	England	89	32	2.781250	England	8	1,563,135	48,848	Europe	Europe	5.204022	2.179788	-3.024234	1963 - 1969
16	1962	Brazil	89	32	2.781250	Chile	4	893,172	27,912	Europe	Europe	2.920070	4.933413	2.013343	1959 - 1965

Ilustración 6 Dataset Final

4.3.6. Preprocesamiento de datos

En este caso particular, se está trabajando con datos sobre los países que han sido anfitriones de la Copa del Mundo, los campeones de cada mundial, el país del mejor jugador del mundial y el país del mejor jugador joven. Para poder transformar esta información en valores numéricos, se necesita primero listar los países potenciales y el continente al que pertenecen, de manera que se pueda aplicar una función que permita categorizarlos por continente y crear una columna para cada uno de ellos en el dataset.

En este proceso, se utilizará la función get_dummies, que, en este caso, se aplicará a la lista de países potenciales para categorizarlos por continente y crear columnas para cada uno de ellos en el dataset. Esto permitirá trabajar con variables numéricas que pueden ser utilizadas en un modelo de aprendizaje automático.

4.3.7. Aplicación PCA

Con estos datos ya se puede realizar el análisis de los componentes principales (PCA) para poder identificar las variables más influyentes en la rentabilidad de un Mundial para el país anfitrión. Para este algoritmo se obtiene el siguiente resultado:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15
goals	-0.2786	0.2561	-0.0347	-0.0728	-0.3782	-0.2046	0.1615	0.3238	-0.2040	-0.0847	-0.5953	0.3624	-0.0000	-0.0000	-0.0000
matches	-0.3140	0.2461	-0.1186	-0.0832	-0.1857	0.0717	0.0590	0.2869	-0.2529	-0.0173	0.1499	-0.7796	0.0000	0.0000	0.0000
Venues	-0.2991	0.0539	-0.0318	0.0319	-0.3426	0.2251	0.5701	-0.6028	0.1713	0.1256	0.0615	0.0161	-0.0000	-0.0000	-0.0000
goals_per_match	0.2544	-0.0902	0.2843	-0.0402	-0.3195	-0.6692	0.2651	0.0513	0.2183	-0.3059	0.2516	-0.1320	0.0000	0.0000	0.0000
Totalattendance †	-0.2635	0.3622	0.1254	-0.0216	-0.0464	0.1720	-0.0263	0.2933	0.0551	-0.1130	0.6683	0.4497	-0.0000	-0.0000	-0.0000
Averageattendance	-0.0612	0.3685	0.4819	-0.0137	0.1358	0.2501	-0.1712	-0.1231	0.4097	-0.4531	-0.3109	-0.1811	0.0000	0.0000	0.0000
Continent_GB_Winner_Europe	-0.1751	-0.4299	0.1484	0.1395	0.1056	0.1892	0.2693	0.3188	0.1051	-0.0930	-0.0454	-0.0135	0.0091	-0.4140	-0.5042
Continent_GB_Winner_South America	0.1751	0.4299	-0.1484	-0.1395	-0.1056	-0.1892	-0.2693	-0.3188	-0.1051	0.0930	0.0454	0.0135	0.0091	-0.4140	-0.5042
Continent_GB_Young_winner_Europe	-0.2706	-0.0011	-0.2839	0.4660	0.0454	-0.2097	-0.1609	-0.0743	0.1501	-0.1696	0.0114	-0.0102	-0.3028	-0.1595	-0.1923
Continent_GB_Young_winner_South America	0.2706	0.0011	0.2839	-0.4660	-0.0454	0.2097	0.1609	0.0743	-0.1501	0.1696	-0.0114	0.0102	-0.3028	-0.1595	-0.1923
Champion_Europe	-0.3304	-0.1144	0.0506	-0.3075	0.0155	-0.1874	-0.1980	0.0247	0.3330	0.3061	-0.0173	-0.0229	0.1553	0.5128	-0.4561
Champion_South America	0.3304	0.1144	-0.0506	0.3075	-0.0155	0.1874	0.1980	-0.0247	-0.3330	-0.3061	0.0173	0.0229	0.1553	0.5128	-0.4561
Hosts_Africa	-0.0833	0.1753	-0.3118	-0.3384	0.6213	-0.2029	0.4129	-0.0236	0.0224	-0.2227	-0.0136	0.0234	-0.2792	0.0903	-0.0125
Hosts_America	0.0079	0.2916	0.4276	0.4461	0.2344	-0.1952	0.1737	0.0681	-0.0719	0.5420	-0.0420	-0.0561	-0.2792	0.0903	-0.0125
Hosts_Europe	-0.2688	-0.2863	0.2590	-0.0872	-0.1097	-0.0304	-0.2603	-0.2631	-0.4108	-0.2338	0.0666	0.0459	-0.5483	0.1773	-0.0245
Hosts_South America	0.3017	0.0477	-0.3122	0.0318	-0.3201	0.2289	-0.0377	0.2361	0.4281	0.0699	-0.0376	-0.0287	-0.5583	0.1805	-0.0249

Correlación con la variable "Diferencia":

PC1	-0.303770
PC2	-0.372945
PC3	0.462634
PC4	0.246877
PC5	0.399142
PC6	0.150396
PC7	-0.309717
PC8	0.142589
PC9	-0.324689
PC10	-0.431071
PC11	-0.083348
PC12	-0.059140
PC13	0.130564
PC14	-0.352597
PC15	0.013182

Ilustración 7 Resultado PCA

Para analizar el análisis de PCA es importante entender que este método se utiliza para reducir la dimensionalidad de un conjunto de datos. En este caso, se tienen datos de los mundiales de fútbol, donde se evalúa la rentabilidad del país que organiza el mundial en términos de la variación del PIB en el año en que se llevó a cabo el mundial.

Los componentes principales (PC) representan combinaciones lineales de las variables originales, que buscan maximizar la varianza en los datos. En este caso, se tienen 15 componentes, cada uno con un peso diferente para cada una de las variables originales.

Observando los pesos de los diferentes componentes, podemos inferir cuáles son las variables que más influyen en la rentabilidad del mundial. En particular, se observa que PC1, que tiene el mayor peso absoluto en la mayoría de las variables, está dominado por las variables "matches", "venues" y "totalattendance", lo que indica que la cantidad de partidos, la cantidad de estadios y la asistencia total son los factores más importantes para determinar la rentabilidad del mundial.

Por otro lado, PC4, que tiene un peso positivo en la variable "goals_per_match" y un peso negativo en la variable "goals", parece estar capturando el impacto del entretenimiento del torneo en la rentabilidad.

Cabe destacar que las variables "Continent_GB_Winner_Europe", "Continent_GB_Winner_South America", "Continent_GB_Young_winner_Europe" y "Continent_GB_Young_winner_South America" tienen pesos significativos en PC7, PC8, PC9 y PC10, lo que indica que la región de los equipos ganadores y la edad promedio de los jugadores también pueden influir en la rentabilidad del mundial.

En resumen, para analizar los datos de España y su posible correlación con la rentabilidad de los mundiales, se deberían considerar principalmente las variables de cantidad de partidos, estadios y asistencia total, así como también el entretenimiento del torneo y la región de los equipos ganadores y la edad promedio de los jugadores.

Seguido a esto, se ha calculado el Porcentaje de Varianza Explicada y Acumulada por Componente Principal, con la siguiente tabla como resultado:

Porcentaje de Varianza Explicada y Acumulada por Componente Principal:

Componente Principal	Porcentaje de Varianza Explicada	Porcentaje de Varianza Acumulada
1.0000	0.4286	0.4286
2.0000	0.2158	0.6443
3.0000	0.1047	0.7490
4.0000	0.0923	0.8413
5.0000	0.0704	0.9117
6.0000	0.0330	0.9447
7.0000	0.0203	0.9650
8.0000	0.0173	0.9823
9.0000	0.0151	0.9974
10.0000	0.0025	0.9999
11.0000	0.0001	1.0000
12.0000	0.0000	1.0000
13.0000	0.0000	1.0000
14.0000	0.0000	1.0000
15.0000	0.0000	1.0000

Ilustración 8 Varianzas PCA

Se puede observar que las componentes más influyentes y que más datos abarcan son las 3 primeras, consiguiendo representar el 65% de los datos con dos componentes y el 75% de los datos con tres componentes. Estos datos nos vendrán muy bien para analizar posteriormente la agrupación de individuos según sus características.

4.4. Agrupación de los individuos según sus características

Para poder responder a la pregunta de cuál es la combinación de características para tener un mundial rentable se ha decidido realizar un proceso de *clustering y* así poder agrupar en rentable los mundiales históricamente organizados según sus características. Para saber cuál es el número óptimo de *clusters* se ha hecho uso del método del codo. Este método consiste en graficar el número de *clusters* en el eje x y la suma de los cuadrados de las distancias de cada punto a su centroide más cercano (en inglés*, Within-Cluster sum of Squares o WCSS*) en el eje y. Luego, se busca el punto en la gráfica donde la disminución en la *WCSS* empieza a aplanarse significativamente, lo que se asemeja a un codo en la curva. Este punto indica que añadir más *clusters* no mejorará significativamente la calidad del *clustering*, por lo que se toma el número de *clusters* correspondiente a ese punto como el número óptimo de *clusters* para el análisis.

En el caso de estudio la gráfica resultante es la siguiente:

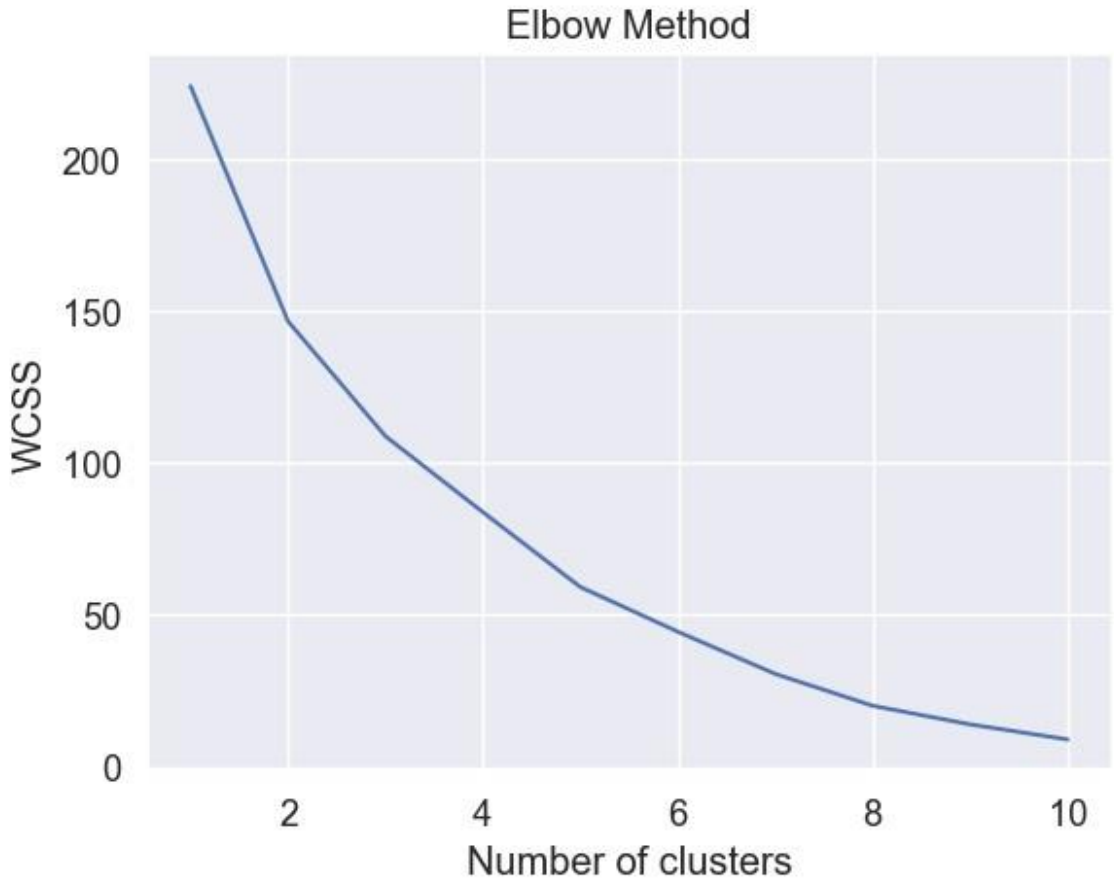

Ilustración 9 Gráfica del Método del Codo

Como se puede observar en la gráfica, el número óptimo de *clusters* para este problema es 2 *clusters.*

Para poder analizar correctamente la agrupación de características se han utilizado dos *clusters* y el resultado es el siguiente:

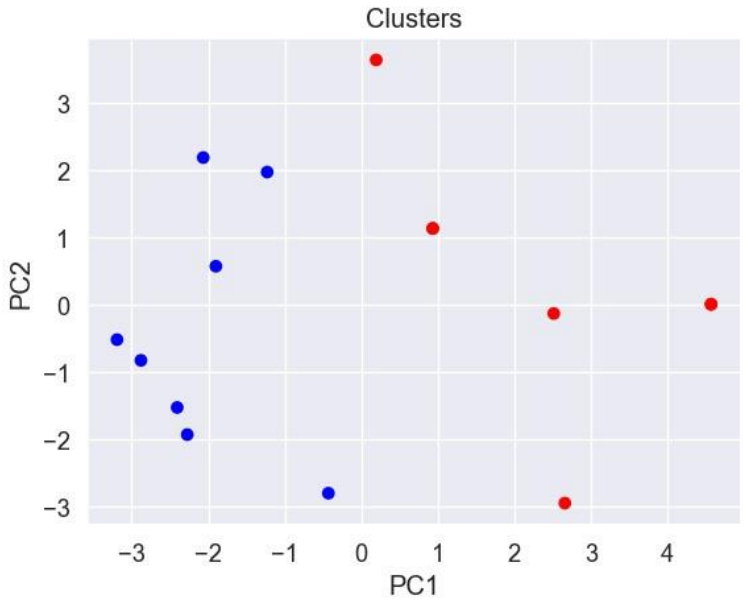

Ilustración 10 Clustering con dos componentes

Con este *clustering* se representa el 64,43% de los datos que se tienen a partir del PCA. Aún eso, se entiende que no se están representando suficientes datos y se ha decidido realizarlo con 3

componentes. Utilizando 3 componentes podemos mostrar el 75% de los datos y son reconocibles los resultados y analizables. Como se observa en la ilustración, se observan los grupos claramente distintos, mundial rentable y no rentable. Se puede observar cómo en estos grupos bien diferenciados, uno de ellos comparte muchas características en común, observando que siguen un patrón para la rentabilidad del mundial y el otro tiene sus características más dispersas.

Como se observa, con 3 componentes podemos analizar mejor y se representan mejor los grupos entre rentable y no rentable ya que tenemos más cantidad de datos a representar.

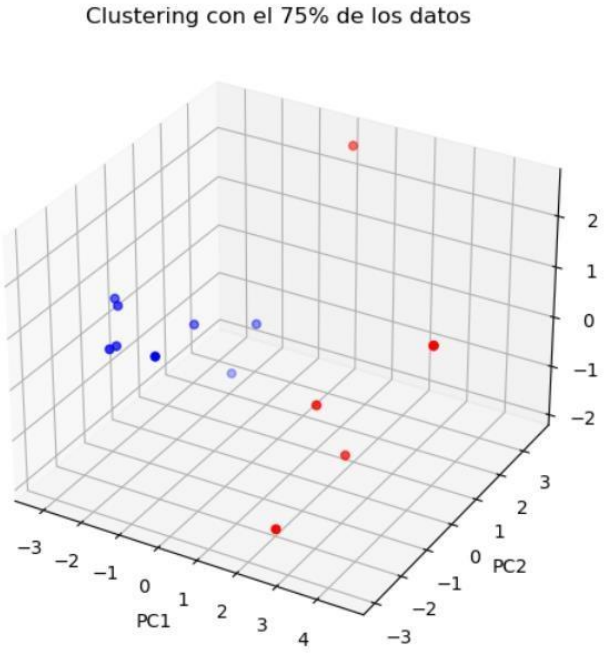

Ilustración 11 Clustering con 3 componentes

4.5. ¿Será rentable para España organizar el Mundial de 2030?

Para responder a esta pregunta se ha decidido hacer uso de técnicas de predicción gracias a la regresión lineal.

Lo primero que se ha realizado es el cálculo de la media de goles y la media de goles por partido de los equipos de los 15 mundiales que se han utilizado. Como se observa en el PCA, comentado anteriormente, los goles es de las características que más afectan a un mundial para que el país organizador acabe con un balance positivo o negativo del GDP. En aras de hacer un mejor análisis se ha usado los datos de asistencia de la primera división española en la temporada 21/22 con los estadios de mayor asistencia de esta competición en la última temporada para la predicción.

Los estadios que mayor asistencia tuvieron en la temporada 21/22 fueron el Camp Nou y el Cívitas Metropolitano. Para hacer la predicción de la asistencia total se han sumado las de los dos estadios.

4.5.1. Predicción

La predicción ha sido realizada gracias a un modelo de regresión lineal, para ello, se han usado los datos de los mundiales anteriores y se han creado nuevos para la candidatura de España en 2030.

Los valores usados para la predicción son:

- **Media goles:** media de los goles marcados en los anteriores mundiales.
- **Partidos:** Número de partidos del mundial. El mundial de 2030 será un mundial atípico pues la FIFA quiere introducir 48 participantes generando que se celebren 80 partidos.

- **Número de estadios:** Para establecer el número de estadios se ha tenido como base los estadios que España ha propuesto para la candidatura de 2030.
- **Media de goles por partido:** Media de goles por partido en los anteriores mundiales.
- **Asistencia total:** Sumatorio de la asistencia total de los dos estadios con más asistencia de la Liga Española en la temporada 21/22.
- **Media de asistencia:** Asistencia media de la temporada 21/22 en la Liga Española.
- **Continente campeón del mundial:** Se ha supuesto el peor caso, que un equipo europeo gane el mundial. Según el análisis de componentes realizado anteriormente, que el campeón del mundial sea europeo es la peor de las situaciones para que mejore la variación del GDP del país.
- **Continente del ganador del Balón de Oro del Mundial:** Al igual que para el continente campeón se ha supuesto el peor caso, que el campeón del balón de oro sea sudamericano. Es lo que peor influye para el GDP del país.
- **Continente del ganador del Balón de Oro joven del Mundial:** Al igual que para el continente campeón se ha supuesto el peor caso, que el campeón del balón de oro sea europeo. Es lo que peor influye para el GDP del país.

Con todo esto, con una predicción conservadora, podemos afirmar que la diferencia del GDP español tras la organización del mundial de 2030 será de 18.96 millones. Como vemos, la predicción es más que próspera y, en términos económicos, que España organice el mundial provocará un crecimiento económico notable.

4.5.1.1. ¿Por qué la predicción del GDP es tan alto?

Para responder esta pregunta se ha analizado el dato que más afecta a la variación del GDP de un país en un mundial, los goles. Para ello se han realizado dos líneas temporales de la variación de goles en los mundiales y los goles por partidos en los mundiales.

Ilustración 13 Histórico de goles en los mundiales

Ilustración 12 Histórico de goles por partido en los mundiales

Como podemos observar, el número de goles en los mundiales tiene un crecimiento al alza, cómo se juegan más partidos la probabilidad de celebrar o llorar un gol en un estadio es mucho más grande. En lo que atañe a los goles por partido, se ve un crecimiento en los últimos años teniendo sus peores datos en el mundial del 90 de Italia con el momento de mayor esplendor del *Catenaccio* [5], defensa en bloque bajo, pocos ataques y esperar para salir a la contra. Esta filosofía de juego tuvo su mayor auge en los años 90, coincidiendo con la organización del mundial de los creadores de esta filosofía de juego. Es por esto por lo que el mundial de 1990 es el que menos goles por partido tiene.

5. Conclusiones

El fútbol es más que un deporte, es una pasión que une a pueblos, culturas y naciones en un mismo sentimiento: la emoción de la victoria y la grandeza del juego.

Como bien se menciona, uno de los principales motivos por los que España debería albergar el Mundial de Fútbol de 2030 es el impacto económico positivo que generaría. Un evento de esta magnitud atraería a millones de turistas, lo que aumentaría el consumo en distintos sectores, como el turístico, hotelero, gastronómico, entre otros. Esto, a su vez, tendría un efecto multiplicador en la economía, ya que se generarían empleos temporales y se dinamizaría la actividad económica en general.

Además, España cuenta con una infraestructura deportiva de alta calidad, que le permitiría albergar un evento de esta envergadura sin mayores inconvenientes. El país tiene una larga trayectoria organizando eventos deportivos de alto nivel, como los Juegos Olímpicos de Barcelona 92, la Copa del Mundo de Fútbol de 1982 o la Final de la Champions League de 2019.

Otro punto a favor de España como sede del Mundial de Fútbol de 2030 es la gran asistencia de público que tiene la Liga Española. Los estadios españoles son conocidos por ser lugares llenos de pasión y colorido, lo que les confiere un ambiente único que los aficionados al fútbol adoran. Esta atmósfera de celebración podría ser trasladada al Mundial de Fútbol de 2030, lo que lo convertiría en un evento aún más emocionante.

Por último, es importante destacar la importancia que tienen los goles en la variación del Producto Interno Bruto (PIB) de un país durante un Mundial. Según algunos estudios, un aumento en el número de goles de los equipos de un país anfitrión puede tener un efecto positivo en la economía de ese país, debido a que se incrementa la felicidad y el optimismo de la población. Como se menciona en el texto, la selección española cuenta con equipos competitivos y atractivos para el público, lo que podría aumentar las posibilidades de éxito deportivo y, por ende, económico.

En definitiva, España cuenta con las características necesarias para ser una excelente sede del Mundial de Fútbol de 2030. Desde su infraestructura deportiva hasta su afición al fútbol, pasando por su potencial turístico y la capacidad de generar impacto económico positivo, España sería una excelente opción para acoger este evento deportivo de talla mundial.

6. Bibliografía

[1] A. Sacchi, Fútbol Total, Roca, 2016.

[2] PRANAV941, «Kaggle.com,» Kaggle, 27 04 2023. [En línea]. Available: https://www.kaggle.com/datasets/pranav941/historical-fifa-world-cups-10-awards. [Último acceso: 20 12 2023].

[3] F. A. S. SANCHEZ, «Kaggle.com,» Kaggle, 14 02 2023. [En línea]. Available: https://www.kaggle.com/datasets/fredericksalazar/pib-gdp-global-by-countries-since1960-to-2021. [Último acceso: 30 04 2023].

[4] A. A. RIMI, «Kaggle.com,» Kaggle, 01 12 2022. [En línea]. Available: https://www.kaggle.com/datasets/aklimarimi/fifa-world-cup-stadium-audience. [Último acceso: 29 04 2023].

[5] A. SEGURA, «panenka.org,» panenka, 10 11 2017. [En línea]. Available: https://www.panenka.org/pasaportes/serie-a/mago-encumbro-catenaccio/. [Último acceso: 31 04 2023].