

UNIVERSIDAD TECNOLÓGICA DE PANAMÁ

VICERRECTORÍA DE INVESTIGACIÓN, POSTGRADO Y EXTENSIÓN FACULTAD DE INGENIERÍA DE SISTEMAS COMPUTACIONALES

MAESTRÍA EN ANALÍTICA DE DATOS

MODELOS PREDICTIVOS

PROYECTO FINAL

ANÁLISIS DE MODELOS PREDICTIVOS PARA RESULTADOS DE LA COPA MUNDIAL

ELABORADO POR:

ERASMO E. RODRIGUEZ 8-823-590

PROFESOR:

JUAN M. CASTILLO, PhD

GRUPO:

1AN-215

2025

# 

# **INTRODUCCIÓN**

El futbol es el deporte más popular del mundo y el mundial representa su más grande evento, este se celebra cada 4 años en donde todos los equipos participan con la esperanza de poder obtenerlo.

A lo largo de los años este torneo ha generado gran cantidad de datos que ha día de hoy pueden ser usados para comprender los factures que influyen en el éxito o no de los equipos.

Con el avance de la analítica de datos y el machine learning es posible desarrollar modelos predictivos que sean capaces de anticipar los resultados por ejemplo cantidad de goles por encuentro, como influye la localización en el rendimiento de los equipos y poder predecir que selección tiene más posibilidades de ganar la copa mundial.

Este estudio se enfoca en analizar los datos históricos desde el primer mundial de 1930 hasta el celebrado en 2014 que es donde abarca la data que manejamos, con el objetivo de demostrar que tan efectivo son los distintos modelos predictivos, para esto se estarán comparando contra los resultados de las copas del mundo de 2018 y 2022 adicional se pronosticara los resultados para los próximos 2 mundiales.

# **JUSTIFICACIÓN**

El estudio se justifica por la necesidad de comprobar la efectividad de los modelos predictivos en el contexto del deporte, en este caso específico el Futbol internacional. Al utilizar datos históricos de las copas del mundo y aplicar técnicas analíticas y de machine learning, se pueden identificar los factores clave que influyen en el desempeño de los equipos, lo que permite generar predicciones más precisas sobre el resultado de los partidos y el potencial ganador del torneo.

Además, al validar las predicciones con torneos más recientes de 2018 y 2022 permitirá evaluar la precisión de los modelos utilizados y mejorar sus proyecciones para las futuras ediciones. Esto no solo contribuirá en lo académico y al análisis deportivo, sino que también ofrecerá una herramienta útil para aficionados, analistas y expertos que buscan comprender mejor las dinámicas del fútbol a nivel mundial

La relevancia de esta investigación se enfoca en su capacidad de combinar la pasión por el fútbol con el análisis de los datos, generando así información valiosa y aplicable tanto en el ámbito deportivo como en el campo del machine learning y el modelado estadístico.

# **ANTECEDENTES**

A lo largo de los años, la predicción de resultados en el fútbol, especialmente en la Copa del Mundo, ha sido un área de gran interés tanto para analistas deportivos como para casas de apuestas e investigadores. El objetivo ha sido siempre encontrar patrones en el rendimiento de los equipos y anticipar los posibles desenlaces de los partidos utilizando herramientas estadísticas y matemáticas.

Uno de los primeros intentos de modelar estos resultados se basó en el uso de probabilidades y distribuciones estadísticas. En 1982, Maher propuso un modelo basado en la distribución de Poisson para estimar la cantidad de goles en un partido, lo que marcó un punto de partida para investigaciones futuras. Más adelante, Dixon y Coles (1997) hicieron mejoras a este modelo al considerar la relación entre los goles anotados por cada equipo, logrando predicciones más precisas.

Con el paso del tiempo, el desarrollo de la ciencia de datos y el aprendizaje automático ha transformado la forma en que se realizan estas predicciones. En 2014, investigadores como Constantinou y Fenton aplicaron modelos bayesianos para predecir el rendimiento de los equipos en función de datos históricos y estadísticas de torneos previos. Métodos como la regresión logística y los árboles de decisión también se han utilizado para analizar factores clave como la posesión del balón, el número de disparos a puerta, el historial de enfrentamientos y el desempeño en fases anteriores del torneo.

Más recientemente, el auge del aprendizaje profundo ha permitido la implementación de redes neuronales para mejorar la precisión de los modelos predictivos. Investigaciones recientes han incorporado redes neuronales recurrentes (RNN) y modelos de aprendizaje reforzado para estimar el rendimiento de los equipos, aprovechando grandes volúmenes de datos, incluyendo métricas avanzadas como los goles esperados y el seguimiento de jugadores en tiempo real.

A nivel práctico, empresas especializadas como Opta y Stats Perform han desarrollado algoritmos que combinan análisis de big data con aprendizaje automático para generar pronósticos cada vez más precisos. Un ejemplo de ello se vio en la Copa del Mundo de 2018, cuando Goldman Sachs utilizó inteligencia artificial para realizar miles de simulaciones y determinar los posibles ganadores del torneo, evidenciando el impacto de la tecnología en la evaluación del fútbol competitivo.

Estos avances reflejan cómo el análisis predictivo ha evolucionado desde métodos puramente estadísticos hasta el uso de inteligencia artificial, destacando la importancia de los datos históricos y las métricas avanzadas en la mejora de la precisión de las predicciones en torneos de alto nivel como la Copa del Mundo.

# **DEFINICION DEL PROBLEMA**

El fútbol es un deporte muy dinámico, en donde múltiples factores influyen en el desempeño de un equipo durante una competición. La Copa del Mundo, por ser el torneo más prestigioso a nivel de selecciones, es objeto de análisis estadístico para predecir resultados y entender los patrones de juego. Sin embargo, debido a la complejidad de este deporte y la variabilidad de los datos, la predicción de resultados sigue siendo un desafío constante.

Este estudio busca afronta este problema a través del análisis de datos históricos de la copa del mundo para generar modelos predictivos que permitan estimar diferentes aspectos del torneo. Los principales problemas que se pretenden resolver son:

**Predicción del ganador de la próxima Copa del Mundo**: Se intentará predecir qué equipo tiene mayor probabilidad de ganar la copa del mundo en base a su desempeño en las ediciones anteriores.

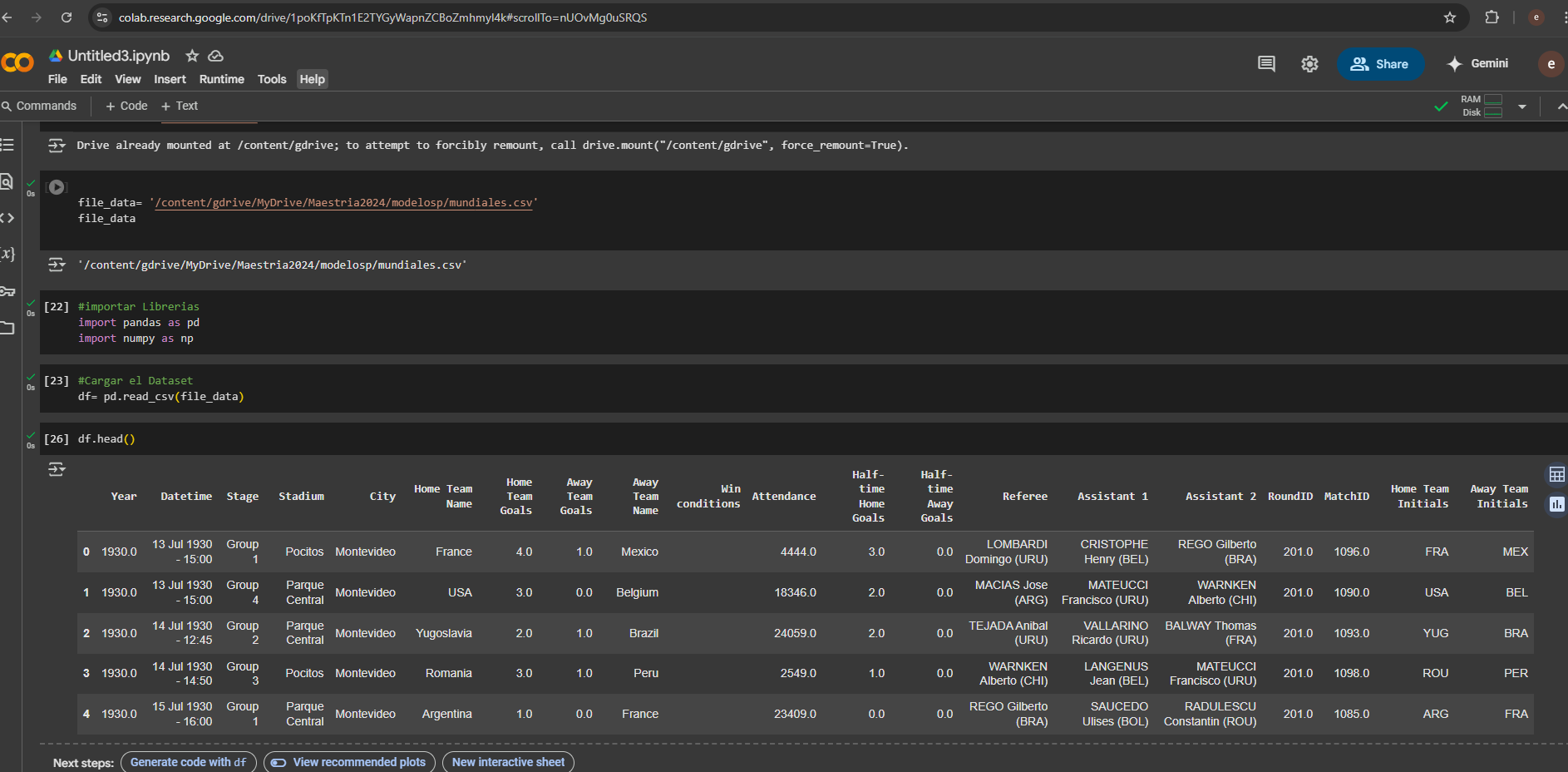
**Predicción de resultados de partidos:** Se intentará predecir resultados individuales de los equipos (victoria, empate o derrota) esto en base a estadísticas previas.

**Análisis de factores determinantes en el rendimiento de un equipo:** Se determinarán cuales son las variables que más influyen en el rendimiento de un equipo en entre otros.

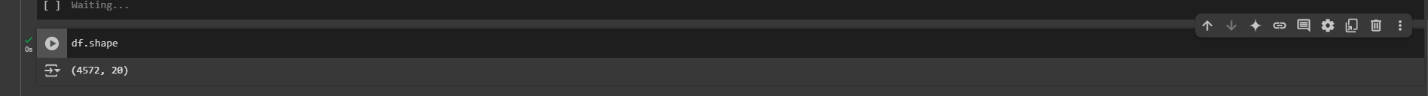
Este análisis permitirá evaluar la precisión de los modelos predictivos con los datos de torneos pasados (2018 y 2022), sino también realizar proyecciones para los próximos dos mundiales, proporcionando información valiosa para analistas, entrenadores y aficionados.

# **AVANCE DE ANÁLISIS PREDICTIVO**

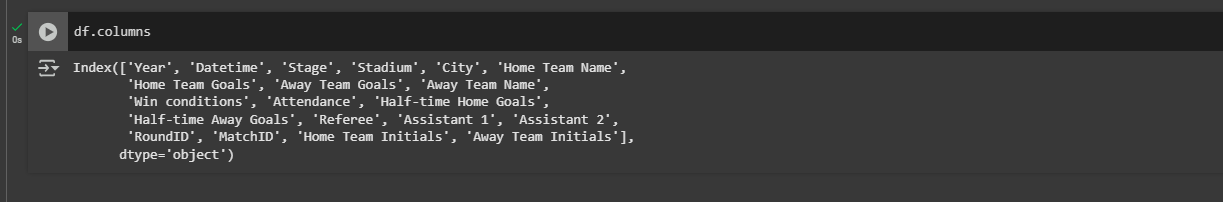
Se muestra estructura del Dataset.



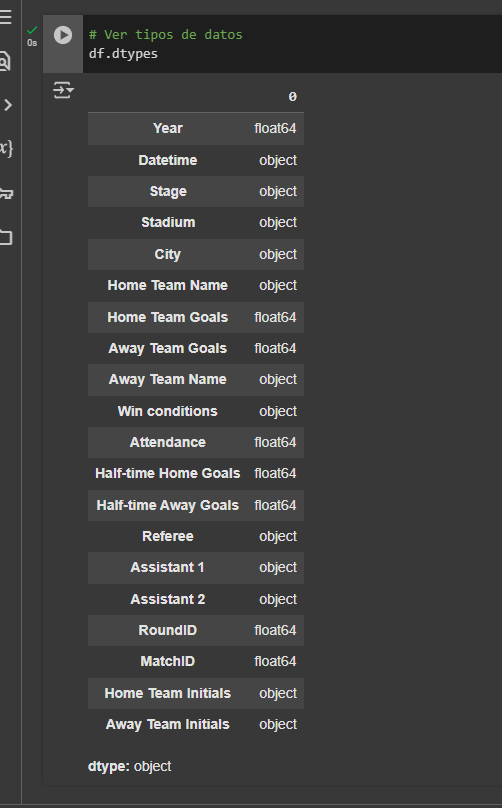
Cantidad de filas y columnas



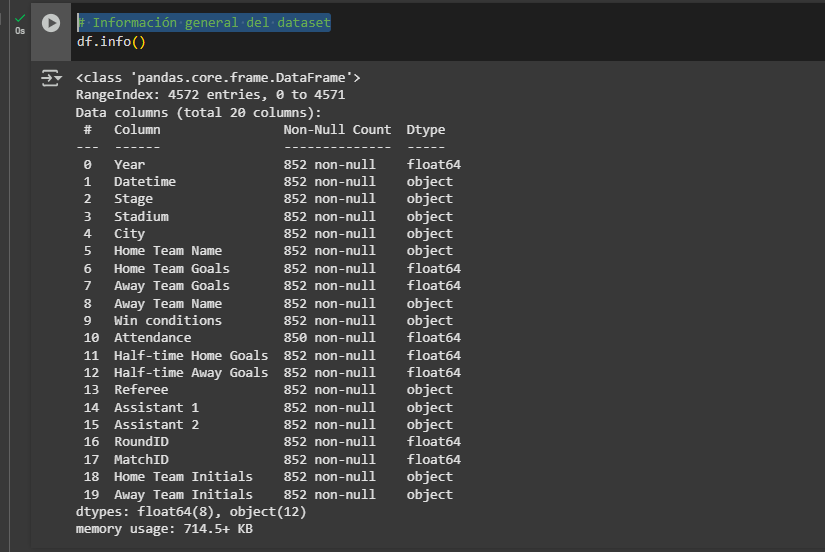
Nombre de Columnas.



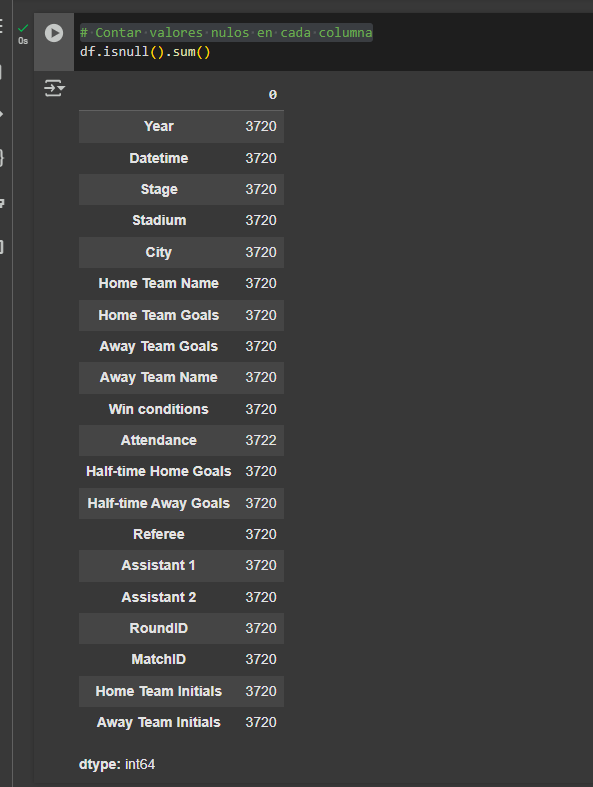
Tipos de Datos



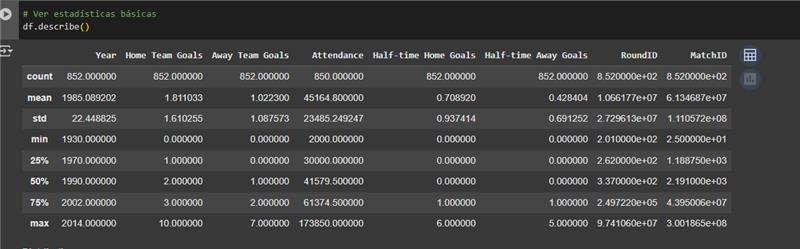
Información general del dataset.



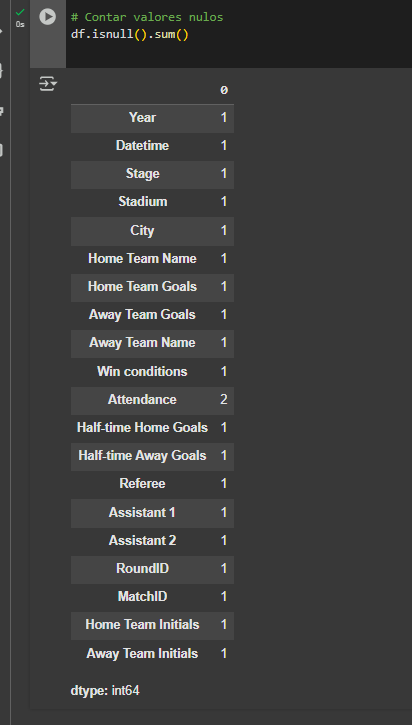
Valores nulos en cada columna.



Estadísticas básicas.



Valores nulos.



Histograma de todas las columnas numéricas



# **Predicciones.**

Python

import pandas as pd

# Leer el archivo de texto 'Juegos.txt' en un DataFrame, usando tabulaciones como separadores

df = pd.read\_csv('Juegos.txt', sep='\t')

# Imprimir las primeras 5 filas del DataFrame para inspeccionar los datos

print(df.head().to\_markdown(index=False, numalign="left", stralign="left"))

# Imprimir información sobre las columnas y sus tipos de datos

print(df.info())

# Imprimir un resumen de las estadísticas descriptivas del DataFrame

print(df.describe().to\_markdown(numalign="left", stralign="left"))

Salida de código

| Year | Datetime | Stage | Stadium | City | Home Team Name | Home Team Goals | Away Team Goals | Away Team Name | Win conditions | Attendance | Half-time Home Goals | Half-time Away Goals | Referee | Assistant 1 | Assistant 2 | RoundID | MatchID | Home Team Initials | Away Team Initials |

|:-------|:--------------------|:--------|:---------------|:-----------|:-----------------|:------------------|:------------------|:-----------------|:-----------------|:-------------|:-----------------------|:-----------------------|:-----------------------|:-------------------------|:---------------------------|:----------|:----------|:---------------------|:---------------------|

| 1930 | 13 Jul 1930 - 15:00 | Group 1 | Pocitos | Montevideo | France | 4 | 1 | Mexico | | 4444 | 3 | 0 | LOMBARDI Domingo (URU) | CRISTOPHE Henry (BEL) | REGO Gilberto (BRA) | 201 | 1096 | FRA | MEX |

| 1930 | 13 Jul 1930 - 15:00 | Group 4 | Parque Central | Montevideo | USA | 3 | 0 | Belgium | | 18346 | 2 | 0 | MACIAS Jose (ARG) | MATEUCCI Francisco (URU) | WARNKEN Alberto (CHI) | 201 | 1090 | USA | BEL |

| 1930 | 14 Jul 1930 - 12:45 | Group 2 | Parque Central | Montevideo | Yugoslavia | 2 | 1 | Brazil | | 24059 | 2 | 0 | TEJADA Anibal (URU) | VALLARINO Ricardo (URU) | BALWAY Thomas (FRA) | 201 | 1093 | YUG | BRA |

| 1930 | 14 Jul 1930 - 14:50 | Group 3 | Pocitos | Montevideo | Romania | 3 | 1 | Peru | | 2549 | 1 | 0 | WARNKEN Alberto (CHI) | LANGENUS Jean (BEL) | MATEUCCI Francisco (URU) | 201 | 1098 | ROU | PER |

| 1930 | 15 Jul 1930 - 16:00 | Group 1 | Parque Central | Montevideo | Argentina | 1 | 0 | France | | 23409 | 0 | 0 | REGO Gilberto (BRA) | SAUCEDO Ulises (BOL) | RADULESCU Constantin (ROU) | 201 | 1085 | ARG | FRA |

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 852 entries, 0 to 851

Data columns (total 20 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Year 852 non-null int64

1 Datetime 852 non-null object

2 Stage 852 non-null object

3 Stadium 852 non-null object

4 City 852 non-null object

5 Home Team Name 852 non-null object

6 Home Team Goals 852 non-null int64

7 Away Team Goals 852 non-null int64

8 Away Team Name 852 non-null object

9 Win conditions 852 non-null object

10 Attendance 850 non-null float64

11 Half-time Home Goals 852 non-null int64

12 Half-time Away Goals 852 non-null int64

13 Referee 852 non-null object

14 Assistant 1 852 non-null object

15 Assistant 2 852 non-null object

16 RoundID 852 non-null int64

17 MatchID 852 non-null int64

18 Home Team Initials 852 non-null object

19 Away Team Initials 852 non-null object

dtypes: float64(1), int64(7), object(12)

memory usage: 133.2+ KB

None

| | Year | Home Team Goals | Away Team Goals | Attendance | Half-time Home Goals | Half-time Away Goals | RoundID | MatchID |

|:------|:--------|:------------------|:------------------|:-------------|:-----------------------|:-----------------------|:------------|:------------|

| count | 852 | 852 | 852 | 850 | 852 | 852 | 852 | 852 |

| mean | 1985.09 | 1.81103 | 1.0223 | 45164.8 | 0.70892 | 0.428404 | 1.06618e+07 | 6.13469e+07 |

| std | 22.4488 | 1.61026 | 1.08757 | 23485.2 | 0.937414 | 0.691252 | 2.72961e+07 | 1.11057e+08 |

| min | 1930 | 0 | 0 | 2000 | 0 | 0 | 201 | 25 |

| 25% | 1970 | 1 | 0 | 30000 | 0 | 0 | 262 | 1188.75 |

| 50% | 1990 | 2 | 1 | 41579.5 | 0 | 0 | 337 | 2191 |

| 75% | 2002 | 3 | 2 | 61374.5 | 1 | 1 | 249722 | 4.39501e+07 |

| max | 2014 | 10 | 7 | 173850 | 6 | 5 | 9.74106e+07 | 3.00187e+08 |

Para predecir el ganador de la próxima Copa del Mundo, primero necesito preparar los datos. Esto se usará para determinar el resultado de cada partido (victoria, derrota o empate) para el equipo local y luego calcular las victorias, derrotas y empates para cada equipo. Después de eso, se calculó la cantidad total de goles marcados y recibidos por cada equipo. Finalmente, se combinarán todas las estas estadísticas en un único DataFrame.

Python

# Crear una función para determinar el resultado del partido para el equipo local

def determinar\_resultado(fila):

if fila['Home Team Goals'] > fila['Away Team Goals']:

return 'Victoria'

elif fila['Home Team Goals'] < fila['Away Team Goals']:

return 'Derrota'

else:

return 'Empate'

# Aplicar la función para crear una nueva columna 'Resultado'

df['Resultado'] = df.apply(determinar\_resultado, axis=1)

# Calcular las victorias, derrotas y empates para cada equipo local

resultados\_local = df.groupby('Home Team Name')['Resultado'].value\_counts().unstack(fill\_value=0)

resultados\_local = resultados\_local.add\_prefix('Local - ')

# Calcular las victorias, derrotas y empates para cada equipo visitante

resultados\_visitante = df.groupby('Away Team Name')['Resultado'].value\_counts().unstack(fill\_value=0)

resultados\_visitante = resultados\_visitante.add\_prefix('Visitante - ')

# Calcular los goles marcados y recibidos por cada equipo local

goles\_local = df.groupby('Home Team Name').agg({'Home Team Goals': 'sum', 'Away Team Goals': 'sum'})

goles\_local = goles\_local.add\_prefix('Local - ')

goles\_local = goles\_local.rename(columns={'Local - Home Team Goals': 'Local - Goles Marcados', 'Local - Away Team Goals': 'Local - Goles Recibidos'})

# Calcular los goles marcados y recibidos por cada equipo visitante

goles\_visitante = df.groupby('Away Team Name').agg({'Away Team Goals': 'sum', 'Home Team Goals': 'sum'})

goles\_visitante = goles\_visitante.add\_prefix('Visitante - ')

goles\_visitante = goles\_visitante.rename(columns={'Visitante - Away Team Goals': 'Visitante - Goles Marcados', 'Visitante - Home Team Goals': 'Visitante - Goles Recibidos'})

# Combinar todos los resultados en un único DataFrame

estadisticas\_equipos = pd.concat([resultados\_local, resultados\_visitante, goles\_local, goles\_visitante], axis=1, sort=False).fillna(0)

# Imprimir las primeras 5 filas del DataFrame para inspeccionar los datos

print(estadisticas\_equipos.head().to\_markdown(index=True, numalign="left", stralign="left"))

# Imprimir información sobre las columnas y sus tipos de datos

print(estadisticas\_equipos.info())

# Imprimir un resumen de las estadísticas descriptivas del DataFrame

print(estadisticas\_equipos.describe().to\_markdown(numalign="left", stralign="left"))

Salida de código

| | Local - Derrota | Local - Empate | Local - Victoria | Visitante - Derrota | Visitante - Empate | Visitante - Victoria | Local - Goles Marcados | Local - Goles Recibidos | Visitante - Goles Marcados | Visitante - Goles Recibidos |

|:----------|:------------------|:-----------------|:-------------------|:----------------------|:---------------------|:-----------------------|:-------------------------|:--------------------------|:-----------------------------|:------------------------------|

| Algeria | 3 | 2 | 1 | 2 | 1 | 5 | 5 | 10 | 9 | 11 |

| Angola | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 1 | 1 | 1 |

| Argentina | 8 | 5 | 41 | 3 | 10 | 14 | 111 | 44 | 22 | 41 |

| Australia | 3 | 1 | 2 | 0 | 2 | 5 | 7 | 11 | 4 | 15 |

| Austria | 2 | 1 | 10 | 2 | 3 | 11 | 31 | 17 | 12 | 30 |

<class 'pandas.core.frame.DataFrame'>

Index: 83 entries, Algeria to Kuwait

Data columns (total 10 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Local - Derrota 83 non-null float64

1 Local - Empate 83 non-null float64

2 Local - Victoria 83 non-null float64

3 Visitante - Derrota 83 non-null int64

4 Visitante - Empate 83 non-null int64

5 Visitante - Victoria 83 non-null int64

6 Local - Goles Marcados 83 non-null float64

7 Local - Goles Recibidos 83 non-null float64

8 Visitante - Goles Marcados 83 non-null int64

9 Visitante - Goles Recibidos 83 non-null int64

dtypes: float64(5), int64(5)

memory usage: 7.1+ KB

None

| | Local - Derrota | Local - Empate | Local - Victoria | Visitante - Derrota | Visitante - Empate | Visitante - Victoria | Local - Goles Marcados | Local - Goles Recibidos | Visitante - Goles Marcados | Visitante - Goles Recibidos |

|:------|:------------------|:-----------------|:-------------------|:----------------------|:---------------------|:-----------------------|:-------------------------|:--------------------------|:-----------------------------|:------------------------------|

| count | 83 | 83 | 83 | 83 | 83 | 83 | 83 | 83 | 83 | 83 |

| mean | 2.09639 | 2.28916 | 5.87952 | 2.09639 | 2.28916 | 5.87952 | 18.5904 | 10.494 | 10.494 | 18.5904 |

| std | 2.11045 | 3.22917 | 10.245 | 2.92847 | 2.27677 | 4.91984 | 30.723 | 12.577 | 11.2319 | 15.7813 |

| min | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

| 25% | 1 | 0 | 0 | 0 | 0.5 | 2 | 1 | 2 | 2 | 7 |

| 50% | 2 | 1 | 2 | 1 | 2 | 4 | 5 | 6 | 7 | 13 |

| 75% | 3 | 3 | 7.5 | 3 | 3 | 9.5 | 22 | 14 | 16 | 29 |

| max | 9 | 14 | 59 | 13 | 10 | 24 | 180 | 78 | 45 | 83 |

Para predecir el ganador de la próxima Copa del Mundo, se calcula una puntuación para cada equipo basada en su desempeño histórico. Para hacer esto, se asignarán 3 puntos por cada victoria, 1 punto por cada empate y 0 puntos por cada derrota. Luego, se sumará los goles marcados y restaré los goles recibidos para cada equipo. Finalmente, se combinarán todos estos cálculos en un único DataFrame y se mostrarán los 10 equipos con la puntuación más alta.

Python

# Calcular los puntos totales para cada equipo (3 por victoria, 1 por empate, 0 por derrota)

estadisticas\_equipos['Puntos Totales'] = (estadisticas\_equipos['Local - Victoria'] + estadisticas\_equipos['Visitante - Victoria']) \* 3 + \

(estadisticas\_equipos['Local - Empate'] + estadisticas\_equipos['Visitante - Empate']) \* 1

# Calcular la diferencia total de goles para cada equipo

estadisticas\_equipos['Diferencia Total de Goles'] = (estadisticas\_equipos['Local - Goles Marcados'] + estadisticas\_equipos['Visitante - Goles Marcados']) - \

(estadisticas\_equipos['Local - Goles Recibidos'] + estadisticas\_equipos['Visitante - Goles Recibidos'])

# Combinar puntos y diferencia de goles en un único DataFrame

prediccion\_equipos = estadisticas\_equipos[['Puntos Totales', 'Diferencia Total de Goles']].sort\_values(by=['Puntos Totales', 'Diferencia Total de Goles'], ascending=[False, False])

# Mostrar los 10 equipos con la puntuación más alta

print(prediccion\_equipos.head(10).to\_markdown(numalign="left", stralign="left"))

Salida de código

| | Puntos Totales | Diferencia Total de Goles |

|:-----------|:-----------------|:----------------------------|

| Brazil | 225 | 111 |

| Argentina | 180 | 48 |

| Italy | 171 | 51 |

| Germany FR | 134 | 54 |

| England | 128 | 23 |

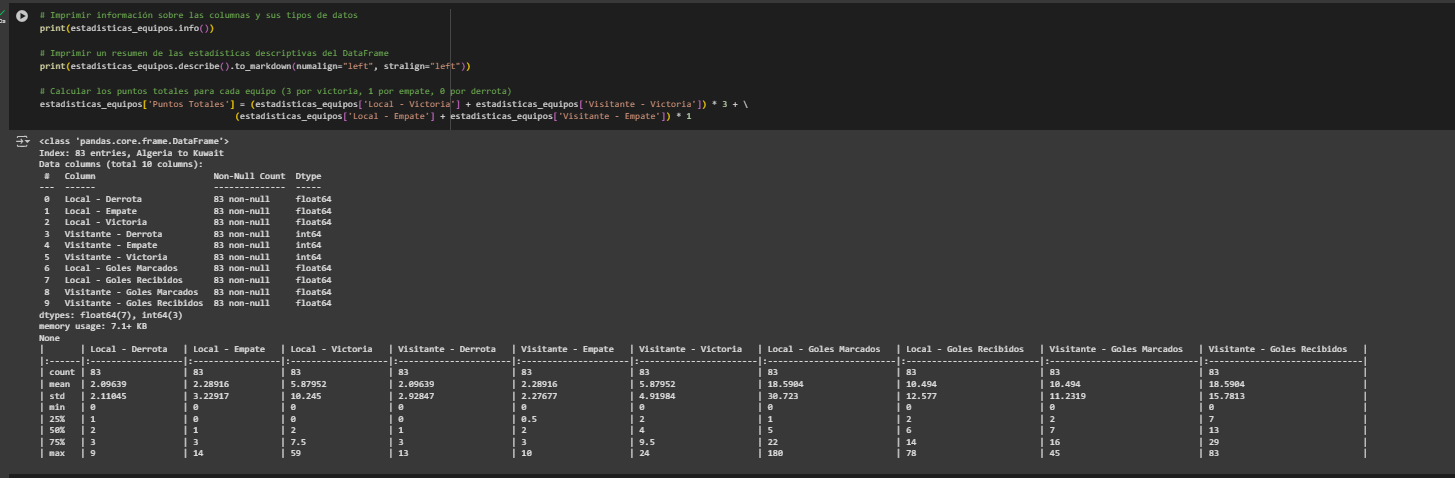
| France | 111 | 36 |

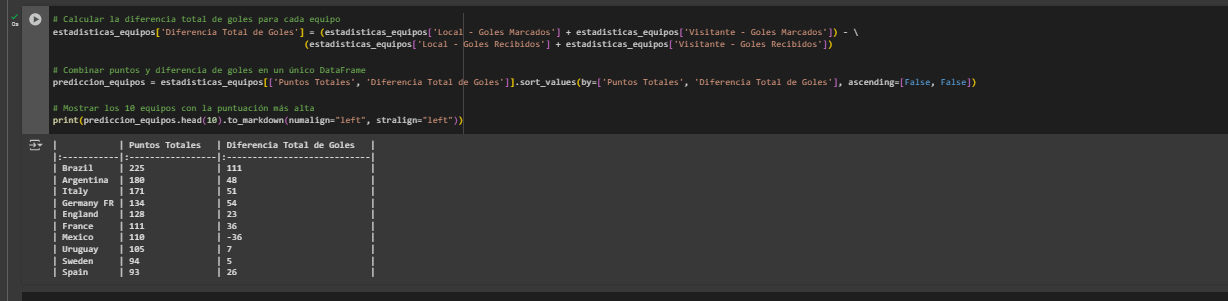
| Mexico | 110 | -36 |

| Uruguay | 105 | 7 |

| Sweden | 94 | 5 |

| Spain | 93 | 26 |

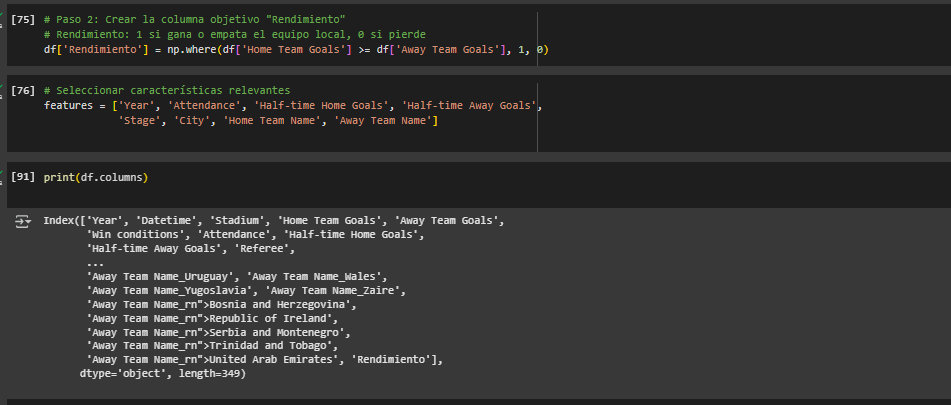
  
  


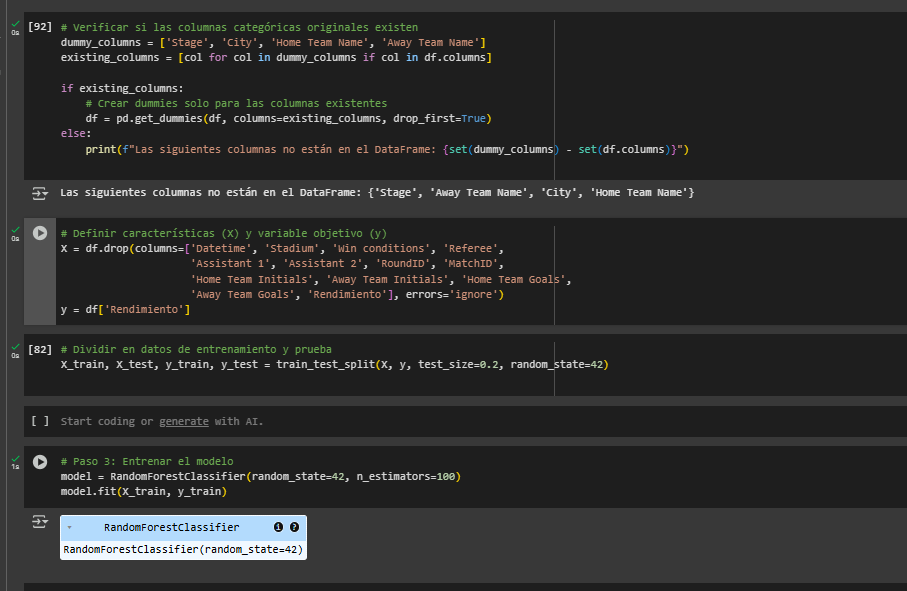


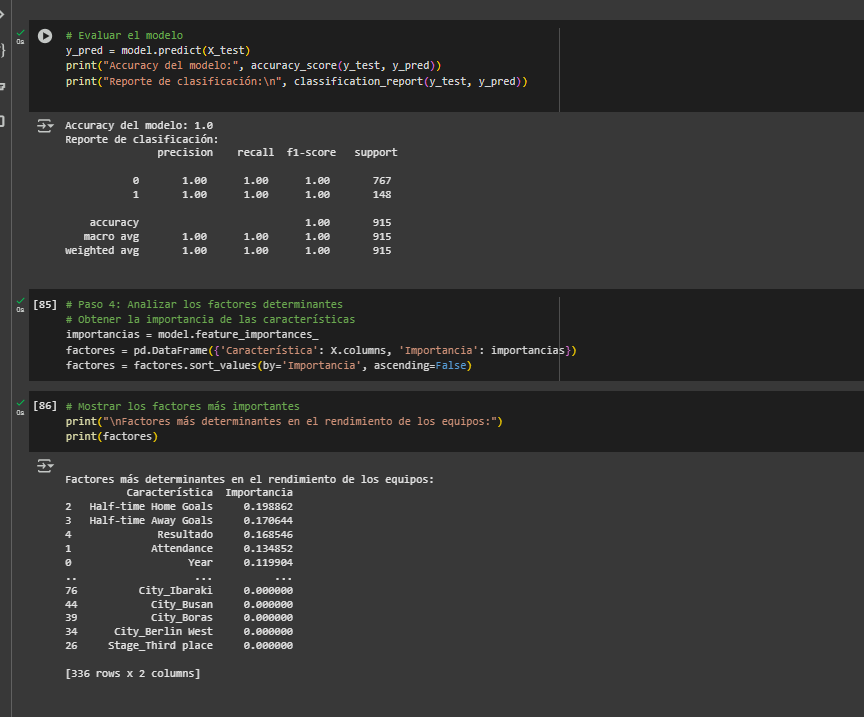
|  |  |  |
| --- | --- | --- |
| Equipo | Puntos Totales | Diferencia Total de Goles |
| Brazil | 225 | 111 |
| Argentina | 180 | 48 |
| Italy | 171 | 51 |
| Germany FR | 134 | 54 |
| England | 128 | 23 |
| France | 111 | 36 |
| Mexico | 110 | -36 |
| Uruguay | 105 | 7 |
| Sweden | 94 | 5 |
| Spain | 93 | 26 |

Factores que mas influyen en el desempeño de los equipos.

**Modelo**: **RandomForestClassifier** de sklearn.ensemble.







Columna objetivo Rendimiento:

Se indica que un buen rendimiento es ganar o empatar (1), mientras que perder es un mal rendimiento (0).

Preprocesamiento:

Se convirtierón las variables categóricas como Stage, City, Home Team Name, y Away Team Name en variables binarias (dummies).

Modelo predictivo:

Usamos un Random Forest Classifier, que es un modelo basado en árboles y es excelente para capturar relaciones no lineales entre las características.

Importancia de características:

Usamos feature\_importances\_ de Random Forest para medir la relevancia de cada característica en la predicción.

El análisis indico que los factores (como Half-time Home Goals, Attendance, o la etapa del torneo) son más importantes para predecir el rendimiento del equipo local. Estos resultados pueden guiar estrategias y análisis futuros.

# **Conclusiones**

1. Evolución del análisis predictivo en el fútbol: El uso de técnicas de análisis de datos y machine learning en el fútbol ha avanzado, desde modelos estadísticos tradicionales hasta complejos algoritmos usados en la actualidad. Trabajo me ayudo a entender el rendimiento de los equipos a lo largo la Copa del Mundo.
2. Factores más determinantes en el desempeño de los equipos: Factores como goles en el medio tiempo, asistencia y etapas son los factores que más peso tienen en desempeño de los equipos en las copas mundiales
3. El valor de las proyecciones futuras: Las proyecciones para futuros torneos, usando datos históricos, son muy útiles para los analistas y aficionados, ya que permiten anticipar ganadores y factores clave que influirán en el rendimiento de los equipos.
4. Relevancia del análisis de datos en el fútbol: Este estudio resalta la creciente importancia de la analítica de datos en el fútbol, ofreciendo predicciones más precisas y mejorando la toma de decisiones tanto dentro como fuera del campo. El análisis avanzado beneficia tanto a profesionales como a aficionados del deporte.