PRA2: ¿Cómo realizar la limpieza y análisis de datos?

Autores: Andrés Merino Toapanta y Mario Cueva Almeida

1. Descripción del dataset

El conjunto de datos a analizar se ha recuperado de la plataforma Kaggle (en este enlace) y contiene todos los resultados de los partidos de fútbol disputados en las copas mundiales de la FIFA desde 1930 hasta 2022.

El conjunto de datos contiene 964 partidos (filas) y 44 variables (columnas). Las variables son:

- `home_team` , `away_team` : Equipos local y visitante, respectivamente.
- `home_score` , `away_score` : Marcador final del partido para cada equipo.
- `home_xg` , `away_xg` : Goles esperados para cada equipo.
- `home_penalty` , `away_penalty` : Número de penales a favor de cada equipo, en el desempate.
- `home_manager` , `away_manager` : Nombre de los DT de cada equipo.
- `home_captain` , `away_captain` : Nombre de los capitanes de cada equipo.
- `home_goal` , `away_goal` : Nombre de los anotadores de goles de cada equipo.
- `home_goal_long` , `away_goal_long` : Detalles de los anotadores de goles (minuto, asistente, etc.) de cada equipo.
- `home_own_goal` , `away_own_goal` : Nombre de los anotadores de auto-goles de cada equipo.
- `home_penalty_goal` , `away_penalty_goal` : Nombre de los anotadores de goles con penal de cada equipo.
- `home_penalty_miss_long` , `away_penalty_miss_long` : Detalles de penales que fallaron (nombre de quién cobró, minuto, razón, etc.) de cada equipo.
- `home_penalty_shootout_miss_long` , `away_penalty_shootout_miss_long` : Detalles de penales, en el desempate, que fallaron (nombre de quién cobró, marcador, razón, etc.).
- `home_penalty_shootout_goal_long` , `away_penalty_shootout_goal_long` : Detalles de penales, en el desempate, que acertaron (nombre de quién cobró, marcador, etc.).
- `home_red_card` , `away_red_card` : Detalles de tarjetas rojas dadas directamente a cada equipo.

- `home_yellow_red_card` , `away_yellow_red_card` : Detalles de tarjetas rojas dadas por acumulación de tarjetas amarillas a cada equipo.
- `home_yellow_card_long` , `away_yellow_card_long` : Detalles de tarjetas amarillas dadas a cada equipo.
- `home_substitute_in_long` , `away_substitute_in_long` : Detalles de las sustituciones realizadas por cada equipo.
- `Attendance` : Asistencia al estadio.
- `Venue` : Estadio donde se jugó el partido.
- `Officials` : Árbitros del partido.
- `Round` : Ronda en la que se jugó el partido.
- `Date` : Fecha del partido.
- `Score` : Marcador final del partido.
- `Referee` : Árbitro principal.
- `Notes` : Notas adicionales del partido.
- `Host` : País anfitrión.
- `Year` : Año en el que se jugó el partido.

Con este conjunto de datos se pueden abordar diferentes problemáticas como las siguientes:

- ¿La métrica de goles esperados (xG) es una buena métrica para evaluar el rendimiento de un equipo?
- ¿Cuáles son la variables más correlacionadas con el resultado del partido?
- ¿El número de tarjetas amarillas o rojas es mayor en los partidos de la fase eliminatoria que en la de los grupos?
- ¿Se puede predecir el resultado del desempate con los datos del partido?

Las respuestas a estas preguntas pueden, por un lado, orientar a los directores técnicos de cada equipo para plantear estrategias que les favorezca, dependiendo del escenario en el que se encuentre; por otro lado, podría ser utilizado por las personas que realizan apuestas, o casas de apuestas, para mejorar sus decisiones.

2. Integración y selección de los datos de interés a analizar.

Puede ser el resultado de adicionar diferentes datasets o una subselección útil de los datos originales, en base al objetivo que se quiera conseguir.

Dado que en el análisis no es relevante datos como los nombres de directores o jugadores, tomaremos únicamente las siguientes variables de las cuales podremos obtener la información necesaria:

- `home_score`
- `home_xg`
- `home_penalty`
- `away_score`
- `away_xg`

- `away_penalty`
- `Round`
- `home_penalty_goal`
- `away_penalty_goal`
- `home_penalty_miss_long`
- `away_penalty_miss_long`
- `home_red_card`
- `away_red_card`
- `home_yellow_card_long`
- `away_yellow_card_long`
- `home_substitute_in_long`
- `away_substitute_in_long`

Además, varias de estas variables tienen un nivel de detalle que no se involucrará en el análisis, por ejemplo, el `home_penalty_goal` posee la información de quién anotó el penal y el minuto; de esto solo nos interesa cuántos penales se anotaron, por lo cual, dividimos la lista por el separador y contamos la cantidad de datos. Esto lo guardamos con el mismo nombre de columna.

Por otro lado, la variable `Round` la refactorizamos para unicamente contar con la información de si es un partido de la etapa de grupos o de la eliminatoria.

Finalmente, generamos una variable que nos indique cuál fue el resultado, en la cual colocamos 1 si el equipo local ganó, -1 si perdió y 0 si empató.

Out[ ]:

	home_score	home_xg	home_penalty	away_score	away_xg	away_penalty	home_penalty_goal
0	3	3.3	4.0	3	2.2	2.0	1.0
1	2	0.7	NaN	1	1.2	NaN	NaN
2	2	2.0	NaN	0	0.9	NaN	NaN
3	3	2.3	NaN	0	0.5	NaN	1.0
4	1	1.4	NaN	0	0.9	NaN	NaN

3. Limpieza de los datos

3.1 Gestión de ceros o elementos vacíos

Revisamos la información de los datos:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 964 entries, 0 to 963
Data columns (total 18 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   home_score               964 non-null    int64
 1   home_xg                  128 non-null    float64
 2   home_penalty             35 non-null     float64
 3   away_score               964 non-null    int64
 4   away_xg                  128 non-null    float64
 5   away_penalty             35 non-null     float64
 6   home_penalty_goal        116 non-null    float64
 7   away_penalty_goal        84 non-null     float64
 8   home_penalty_miss_long   6 non-null      float64
 9   away_penalty_miss_long   9 non-null      float64
 10  home_red_card            51 non-null     float64
 11  away_red_card            54 non-null     float64
 12  home_yellow_card_long    621 non-null    float64
 13  away_yellow_card_long    627 non-null    float64
 14  home_substitute_in_long  740 non-null    float64
 15  away_substitute_in_long  747 non-null    float64
 16  round                    964 non-null    int32
 17  result                   964 non-null    int32
dtypes: float64(14), int32(2), int64(2)
memory usage: 128.2 KB
```

Podemos apreciar que la mayoría de variables tienen datos nulos, los cuales corresponden, por ejemplo, en la variable `home_penalty`, a que no se anotaron penales en ese partido. Por esta razón, en todas las variables de este estilo, imputamos el valor nulo por 0.

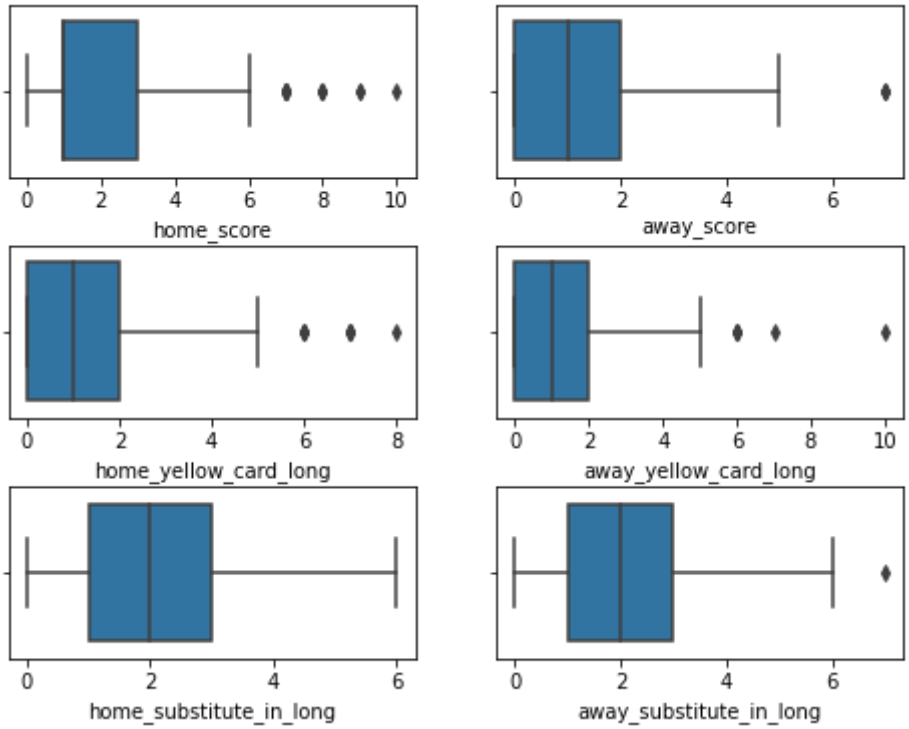
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 964 entries, 0 to 963
Data columns (total 18 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   home_score               964 non-null    int64
 1   home_xg                  128 non-null    float64
 2   home_penalty             964 non-null    float64
 3   away_score               964 non-null    int64
 4   away_xg                  128 non-null    float64
 5   away_penalty             964 non-null    float64
 6   home_penalty_goal        964 non-null    float64
 7   away_penalty_goal        964 non-null    float64
 8   home_penalty_miss_long   964 non-null    float64
 9   away_penalty_miss_long   964 non-null    float64
 10  home_red_card            964 non-null    float64
 11  away_red_card            964 non-null    float64
 12  home_yellow_card_long    964 non-null    float64
 13  away_yellow_card_long    964 non-null    float64
 14  home_substitute_in_long  964 non-null    float64
 15  away_substitute_in_long  964 non-null    float64
 16  round                    964 non-null    int32
 17  result                   964 non-null    int32
dtypes: float64(14), int32(2), int64(2)
memory usage: 128.2 KB
```

Podemos apreciar que continuamos con valores nulos en las variables relacionadas con la métrica de goles esperados. Esto se debe a que, antes de 2018, no se contaba con esta medida. Por esta razón, lo dejaremos como valor perdido y tendremos en cuenta cuando se realicen los análisis correspondientes.

3.2 Identificar y gestionar los valores extremos

Dado que algunas de las variables tienen valores igual a 0 en la mayoría de registros pues, por ejemplo, en la mayoría de partidos se tiene 0 goles de penal, vamos a dividir el análisis de valores extremos en dos etapas.

Primero analizaremos los valores extremos de las variables que suelen tomar valores diferentes de 0:



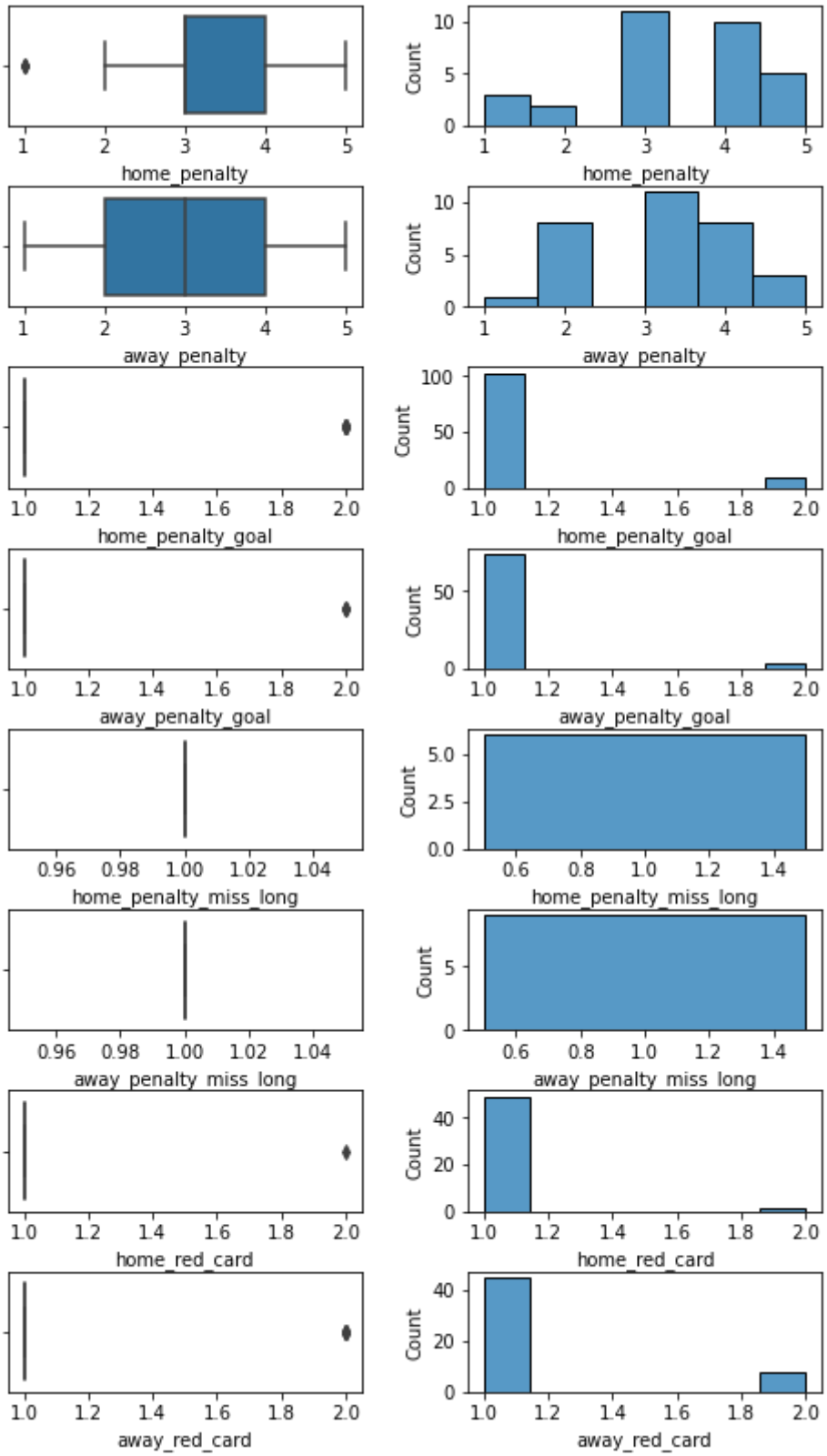
A pesar de que todos los valores extremos son válidos, pueden ocasionar problemas al momento de realizar un análisis de relación o de predicción. Procedemos a contabilizar la cantidad de estos valores:

```
Cantidad de valores extremos en home_score: 15
Cantidad de valores extremos en away_score: 3
Cantidad de valores extremos en home_yellow_card_long: 10
Cantidad de valores extremos en away_yellow_card_long: 6
Cantidad de valores extremos en home_substitute_in_long: 0
Cantidad de valores extremos en away_substitute_in_long: 1
Total de registros con valores extremos: 33
```

Dado que los 33 registros no son significativos comparado al tamaño del conjunto de datos (3.4%), procedemos a eliminarlos.

```
Total de registros: (931, 18)
```

Ahora, realizamos los gráficos del resto de variables, pero omitimos los datos iguales a 0.



Observamos que, bajo esta consideración, no se tienen valores extremos.

Para finalizar, guardamos los datos en un archivo csv.

4. Análisis de los datos

4.1. Selección de los grupos de datos que se quieren analizar/comparar

Para la segmentación de los grupos, se usará si el partido pertenece a la fase de grupos o a la de eliminatoria. Estos grupos podrían presentar diferencias dado que los partidos de eliminatoria son más decisivos. Por otro lado, también se dividirá por el resultado del partido, para ver si hay diferencias entre los partidos ganados, perdidos o empatados.

4.2. Comprobación de la normalidad y homogeneidad de la varianza

Analizaremos la normalidad de los datos utilizando el test de Kolmogorov-Smirnov ya que tenemos una cantidad considerable de datos y hemos eliminado los datos extremos. Analizando si el p valor es mayor a 0.05, podemos concluir que es probable que la variable tenga una distribución normal.

```
La columna away_substitute_in_long posiblemente sigue una distribución normal.
```

Con esto, tenemos que solo una variable, `away_substitute_in_long`, probablememnte posee una distribución normal.

Ahora, para analizar la homogeneidad de la varianza, utilizaremos el test de Levene. Iniciaremos comparando los datos entre los partidos de grupos y de eliminatoria.

```
La varianza de la columna home_score en ambos dataframes es diferente.
La varianza de la columna home_penalty en ambos dataframes es diferente.
La varianza de la columna away_score en ambos dataframes es diferente.
La varianza de la columna away_penalty en ambos dataframes es diferente.
La varianza de la columna away_red_card en ambos dataframes es diferente.
La varianza de la columna away_substitute_in_long en ambos dataframes es diferent
e.
La varianza de la columna result en ambos dataframes es diferente.
```

Tenemos que 7 variables no tienen homogeneidad en la varianza.

Ahora, realizamos lo mismo entre los grupos de partidos ganados, perdidos y empatados.

```
La varianza de la columna home_score en ambos dataframes es diferente.
La varianza de la columna home_penalty en ambos dataframes es diferente.
La varianza de la columna away_score en ambos dataframes es diferente.
La varianza de la columna away_penalty en ambos dataframes es diferente.
La varianza de la columna home_penalty_goal en ambos dataframes es diferente.
La varianza de la columna away_penalty_goal en ambos dataframes es diferente.
La varianza de la columna home_red_card en ambos dataframes es diferente.
La varianza de la columna away_red_card en ambos dataframes es diferente.
La varianza de la columna home_substitute_in_long en ambos dataframes es diferent
e.
La varianza de la columna away_substitute_in_long en ambos dataframes es diferent
e.
La varianza de la columna round en ambos dataframes es diferente.
```

Como era de esperarse, tenemos más variables que presentan diferencias en la varianza.

4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos.

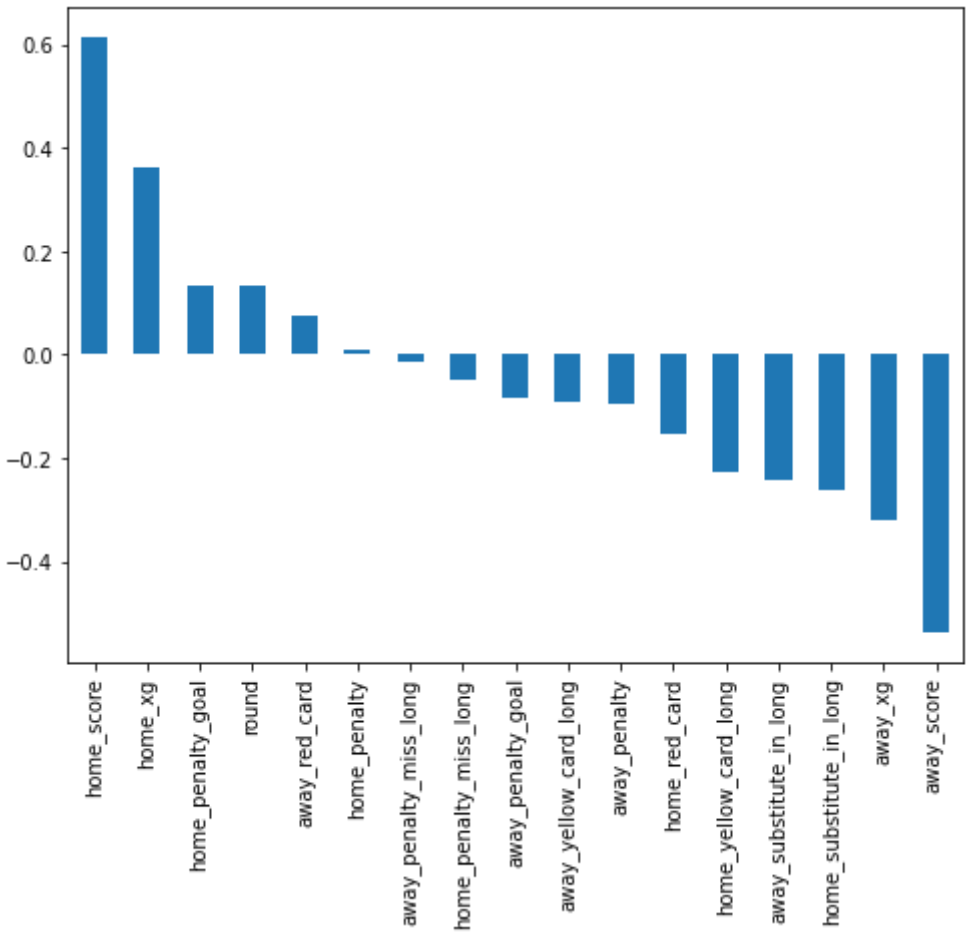
En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

Iniciamos analizando la correlación entre la cantidad de goles anotados y la cantidad de goles esperados.

```
La correlación entre col_1 y col_2 es:  0.5474207148146943
p-valor : 3.6459353590661565e-20
```

Vemos que el nivel de correlación no es alto.

Ahora, analicemos la correlación entre las variables y el resultado del partido.



Tenemos que las variables más correlacionadas son la cantidad de goles y los goles esperados, en ambos equipos.

A continuación, analicemos si en un partido de la ronda eliminatoria se tienen más tarjetas que en los partidos de la fase de grupos.

Para esto, la hipótesis nula sería que la media de tarjetas en los partidos en la fase eliminatoria es menor o igual a la media en la fase de grupos. La hipótesis alternativa sería que la media de tarjetas en la fase eliminatoria es mayor a la media en la fase de grupos.

Dado que no tenemos normalidad en las variables, utilizamos la prueba no paramétrica de Mann-Whitney U.

```
p-valor:0.03543211223321652
Rechazamos la hipótesis nula
```

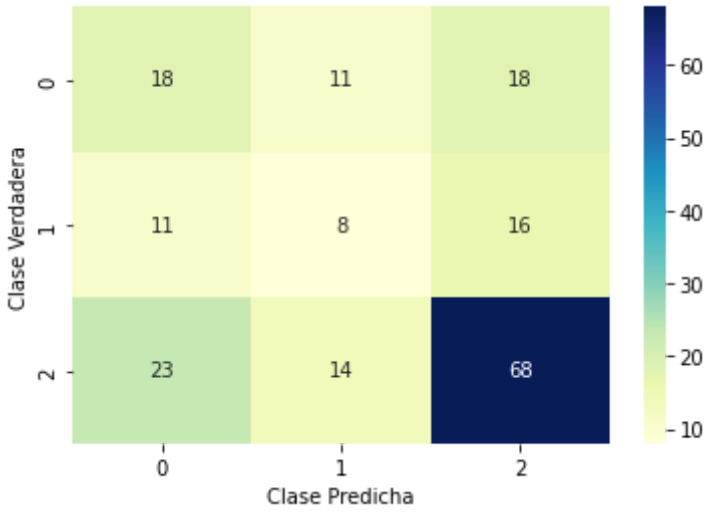
Al rechazar la hipótesis nula, tenemos que hay evidencia estadística para decir que en los partidos de fase eliminatoria se tienen más tarjetas que en los partidos de la fase de grupos.

Finalmente, para revisar si es posible predecir los resultados del partido en función de los datos del mismo (omitiendo los datos relacionados a la cantidad de goles), planteamos un modelo de Árbol de decisión y lo entrenamos con el 80% de los datos para probarlo en el 20%.

Out[ ]:
```
▾ DecisionTreeClassifier
DecisionTreeClassifier()
```

Generamos las predicciones y presentamos su exactitud y la matriz de confusión:

```
Exactitud: 0.50
```



Podemos ver que se alcanza una exactitud del 50%, es decir, el modelo no es mejor que un modelo aleatorio.

6. Resolución del problema.

A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

1. Uno de los elementos más controversiales presentados en los mundiales de fútbol, es saber con anticipación ¿quién va a ganar en cada partido?, ¿quién será el campeón?, entre otras inquietudes. Para resolver estas inquietudes, ha sido habitual recurrir a los datos estadísticos y su análisis, sin embargo, no ha habido un método adecuado que resuelva esta controversia.
2. Con el avance de la tecnología, se ha desarrollado la Ciencia de Datos, la misma que cuenta con varios algoritmos de Inteligencia Artificial, que luego de un buen entrenamiento con una cantidad considerable de datos, ayudan a predecir ciertos eventos. En este sentido, el fútbol no ha sido la excepción y tomando en cuenta que ya existen experiencias exitosas, decidimos entrenar un árbol de regresión para determinar si es posible predecir el resultado del desempate con los datos generados en el partido.

En el presente trabajo, hemos visto que esto no es posible, ya que la exactitud es apenas del 0.5, lo que equivale a usar el azhar. Esto quizá se deba a que los datos con los que se cuenta, no son sufientes. Quizá sea necersario medir el nivel de agotamiento físico y psicológico de los jugadores, entre otras variables posibles para mejorar el entrenamiento del modelo.

3. A pesar de tener 964 registros, aparentemente un buen número para entrenar un algoritmo y obtener buenos resultados, parece ser que no es suficiente, como resultó en el caso de entrenar un árbol de decisión con las datos de un solo partido. Esto posiblemente se dé porque los mundiales de fútbol se realizan cada cuatro años, tiempo en el cual, muchas de las condiciones de cada uno de los equipos cambian, más aún de partido a partido; por ejemplo: el entrenado ya no es el mismo, algunos jugadores ya envejecieron, otros cambiaron de nacionalidad, hay jugadores que ya no pueden jugar ya sea por lesiones o por acumulación de tarjetas. Es así que en el último mundial, 2022, casi todas las selecciones acudieron con al menos un jugador nacionalizado.

4. Con los resultados obtenidos, se ha podido resolver el problema en un cien porciento, aunque esto no signifique que todo es positivo, como la respuesta que se tiene a la pregunta de si se puede predecir el resultado del desempate con solo los datos obtenidos en el partido.

Contribuciones	Firma
Investigación previa	Andrés Merino Toapanta, Mario Cueva Almeida
Redacción de las respuestas	Andrés Merino Toapanta, Mario Cueva Almeida
Desarrollo del código	Andrés Merino Toapanta, Mario Cueva Almeida
Participación en el vídeo	Andrés Merino Toapanta, Mario Cueva Almeida