Instituto Tecnológico y de Estudios Superiores de Monterrey

Inteligencia Artificial Avanzada para la Ciencia de Datos (Gpo 101)

Limpieza del Conjunto de Datos

Equipo 5

Jorge Eduardo De León Reyna - A00829759
David Esquer Ramos - A01114940
Francisco Mestizo Hernández - A01731549
Adrián Emmanuel Faz Mercado - A01570770

Agosto 12 , 2023

El desastre del Titanic ha sido uno de los eventos más impactantes en la historia de la humanidad, pues alrededor de 1,500 personas perdieron la vida en el hundimiento del barco. En este reto, se nos presenta un conjunto de datos de las personas que abordaron el barco, incluyendo información detallada de cada tripulante, como la edad, el sexo, la clase en la que viajaba, el precio que se pagó, entre otros. Además, se presenta una columna que indica si el pasajero sobrevivió o no a la catástrofe.

Lo que se busca es generar un modelo que permita predecir si un pasajero sobrevivió o no al evento en base a la información que se proponga del mismo. Se busca entender la correlación que puede existir entre las variables de estudio para saber si una persona sobrevivió o no.

Primeramente, es necesario realizar un análisis de los datos que se nos proporcionan, para así poder comenzar a identificar los posibles factores que pudieran ser significativos para saber si una persona sobrevivió o no. Las variables que se nos proporcionaron en los archivos iniciales fueron las siguientes:

- *Passenger ID*
- *Survival*
- Ticket class (PClass)
- *Sex*
- *Age*
- *Number of siblings / spouses aboard the Titanic (sibsp)*
- *Number of parents / children aboard the Titanic (parch)*
- *Ticket number*
- *Passenger fare*
- *Cabin number*
- *Port of Embarkation*

Una vez entendido el significado de cada variable, analizamos cuáles eran las que conisderabamos más relevantes para definir si una persona sobrevivió o no al desastre. A continuación presentamos las columnas que creemos más importantes y las cuales nos brindan información valiosa en este reto:

1. **Sexo (*Sex)*:** Consideramos que el sexo de una persona puede ser crucial para definir si una persona sobrevive o no, especialmente por el hecho de que se pudo tomar la creencia de que “las mujeres y los niños van primero”, y por esto mismo, puede que

las mujeres tengan una mayor probabilidad de sobrevivir. Sin embargo, al estar investigando, también nos encontramos con una noticia que indica lo contrario, pues de acuerdo a BBC, en la mayoría de los desastres marítimos, las mujeres han tenido un menor porcentaje de supervivencia que los hombres (BBC, 2012). Nos interesa ver cuál es la tendencia en este caso y creemos que sí puede ser un factor determinante para obtener una buena predicción. Los datos de esta columna están completos en todas las filas, lo cual nos facilita también su análisis.

2. **Ticket class (Pclass):** Para los pasajeros, existían tres clases en las que podían viajar. Estas se encuentran identificadas en 3 clases diferentes: Primera, segunda y tercera clase. Creemos que la clase en la que viajaban los habitantes representa también un factor importante, principalmente por el hecho de que las personas que viajaban en primera clase pudieron quizás tener algún tipo de prioridad o privilegio al momento de estar realizando los protocolos de rescate. De igual forma, al estar investigando, pudimos encontrar que las personas de tercera clase se encontraban en una parte más escondida del barco, lo cual podría significar que sería más difícil para ellos salir del barco y encontrar una manera de salvarse. Por otro lado, las habitaciones de primera clase estaban en la parte superior y central del barco, lo cual también pudo facilitar su sobrevivencia. Dentro de los datos estos se identifican con su equivalente numérico, no existe ningún registro sin información de la clase y todas tienen valores correctos.

3. **Edad (Age):** Al igual que con la justificación de sexo, creemos que la edad pudiera ser un factor importante a considerar por el hecho de que quizás se le pudo dar prioridad a los niños para salvarse, o de igual forma, quizás para las personas de las tercera edad pudo ser más difícil la movilidad al momento de intentar salvarse. En el caso de la edad, identificamos que existen varios registros sin edad, pero más adelante se explicará la manera en la que los manejamos.

4. **Cantidad de hermanos/parejas (sibsp):** La cantidad hermanos y parejas que se encontraban a bordo, nos podría indicar si hubo personas que prefirieron que sus familiares se salven en lugar de ellos. Creemos que el hecho de estar juntos en familia les pudo haber permitido tener más posibilidades de sobrevivir o por el contrario, por el hecho de estar juntos perder agilidad y no sobrevivir. Creemos que esto sí pudo tener un impacto en si las personas sobrevivieron o no, y nos gustaría analizarlo a

detalle. Los datos de esta columna se encuentran completos, son consistentes y no identificamos alguna anomalía.

5. **Cantidad de padres/hijos(parch):** Al igual que la variable anterior, creemos que esta variable si pudo haber afectado en si las personas sobreviven o no. Es muy probable que los padres del barco hayan hecho todo lo posible por salvar a sus hijos o incluso sacrificar su vida por ellos. De igual forma, es posible que si no tenían a nadie a quien quisieran salvar hayan tenido una mayor facilidad para salvarse. Estos datos también están completos y no se identificó ninguna anomalía.

6. **Cuota (Fare):** Este dato está de cierta forma ligado con la clase en la que viajaba la persona. Consideramos que el hecho de que alguien haya pagado una mayor cuota quizás podría significar que tenga más privilegios o beneficios al momento de que sucediera una emergencia y que esto incremente sus probabilidades de sobrevivir. Quizás podría significar que tiene acceso inmediato a salvavidas o a abordar los barcos de rescate, pero sigue siendo una variable que queremos analizar a detalle. Nos dimos cuenta que los datos de esta columna estaban completos, pero al analizar los datos con una gráfica, existía una pequeña anomalía pues identificamos tres filas en donde la cuota que se pagó sobrepasaba los 500 dólares, mientras que el promedio de los datos era de 32 dólares. Más adelante se explica lo que se hizo con esto.

7. **Survival:** En este caso, la variable de “Survival” representa nuestra variable dependiente y definitivamente es necesaria y muy relevante, pues es la que nos ayudará a entrenar al modelo y quien nos indica si el pasajero sobrevivió o no. Esta es la variable que queremos llegar a calcular en nuestro modelo, pero es sumamente necesaria en los datos de nuestro entrenamiento, pues nos permitirá identificar patrones y relaciones para definir el modelo más preciso posible.

8. **Embarked:** Esta es una variable que mantenemos ya que puede influenciar en el camarote que se le asignó a cada pasajero. Por ejemplo, las personas que se embarcaron en los primeros puestos tuvieron mayor oportunidad de elegir donde querían estar ya que la mayoría de cuartos se encontraban vacíos. Por el contrario los

que embarcaron en otros puertos tuvieron menos posibilidades de elegir ya que el barco se encontraba más lleno.

Las variables que no consideramos muy relevantes para generar el modelo de predicción fueron el número de ticket, el número de cabina y el id del pasajero. A continuación presentamos las razones por las que decidimos descartar esas columnas:

1. **Ticket:** Al observar la estructura de los números de ticket, nos dimos cuenta que la mayoría de los registros consisten solamente en un número, y este solamente representa un identificador que se le asigna a cada ticket, por lo que no nos es muy útil para encontrar algún patrón, tendencia, o para entrenar al modelo. También identificamos que algunos de los tickets tienen al inicio unas letras en mayúsculas, pero como no se sabe con claridad lo que estas letras significan y son muy pocos los boletos que las tienen, decidimos de igual forma descartar esta variable.
2. **Número de cabina:** Si bien el número de cabina podría ser un buen indicador para saber si la zona en la que se encontraba cada cabina les podría facilitar la salida a los pasajeros o bien que fuera más difícil sobrevivir si la cabina estaba en una zona escondida, decidimos eliminar esta columna por la cantidad de valores que tenía el conjunto de datos. De los 981 registros que había en el archivo, solamente 294 tenían el campo de “Cabina” lleno, lo que significa que más del 70% de los datos no tenían un número de cabina. Debido a que son tan pocos los datos que tenemos de esta columna, consideramos que lo mejor fue no incluirla en el análisis y no representaría correctamente al conjunto de datos.
3. **Id del pasajero:** Elegimos descartar el id del pasajero porque este simplemente representa un identificador de cada registro, cada persona tenía un id distinto y en realidad no nos da ninguna información valiosa para identificar alguna tendencia o correlación con otro dato. Este dato no nos sirve más que para identificar a cada pasajero.
4. **Nombre:** Al igual que el id, el nombre de cada pasajero es único, y este no tiene ninguna correlación con la columna de si sobrevivió la persona o no. Sin embargo, esta columna si nos sirvió para generar una nueva que sí podría tener una importancia significativa en el análisis. Si bien el nombre de la persona no nos indica nada, quizás el título de la persona sí nos pueda decir algo acerca de si es más probable que la

persona sobreviva. Por ejemplo, si alguna persona tiene algún titulo de la realeza, es probable que tuvo una mayor protección o preferencia al momento de iniciar el proceso de rescate. Más adelante se explicará cómo extraímos cierta información de esta columna para generar una nueva y finalmente eliminamos esta variable.

Transformaciones realizadas

Una vez seleccionadas las variables más relevantes para generar nuestro modelo, comenzamos con la limpieza del conjunto de datos.

Cuenta de campos vacíos

Como primer paso, decidimos revisar si existían columnas que tenían datos vacíos y cuántos registros eran. Analizar estos datos vacíos es de suma importancia para saber si existen columnas que no cuentan con suficientes datos para darnos información valiosa o si debíamos aplicar alguna técnica para rellenar los campos vacíos.

Con el método isnull(), revisamos cuántos valores de cada columna estaban vacíos, y nos dimos cuenta que casi todas estaban completas, excepto las columnas de “Cabin”, “Age” y “Embarked”.

Mapeo de títulos

Anteriormente mencionamos que consideramos que la columna de nombres no era relevante para el análisis que realizaremos y para generar el modelo, sin embargo, pudimos darnos cuenta que todos los registros de nombres tenían un título dependiendo del puesto que tenían en el barco o simplemente de acuerdo a su estatus. Algunos de estos títulos incluían *Mr. , Mrs. , Don. , Major. , Lady, Master*, entre otros. Consideramos que estos títulos sí podrían tener una relación importante en cuanto a las personas que sobrevivieron en el barco, por lo que decidimos solamente mantener estos títulos para continuar con el análisis y para utilizar en el modelo. Sin embargo, aún así existía una gran cantidad de títulos en los diferentes

nombres, por lo que decidimos que una mejor idea sería agrupar estos títulos en categorías más amplias, para así poder acortar los objetos de estudio. Agrupamos todos los títulos disponibles en 6 categorías:

- Officer
- Royal
- Mrs.
- Mr.
- Miss
- Master

Antes de eliminar la columna del nombre, creamos una nueva columna llamada “Title”, en la cual fue rellenada con los títulos correspondientes a cada nombre. Los datos resultantes de esta columna nos pueden acercar a una mejor predicción de la supervivencia de cada pasajero ya que los cuartos asignados a las primeras clases se encontraban más arriba en el barco y por esto tuvieron un lugar más “privilegiado” al momento del hundimiento.

Eliminación de columnas que no utilizaremos

Para continuar con la limpieza de los datos, finalmente eliminamos las columnas que no consideramos relevantes para el análisis y generación del modelo. Para eliminar estas columnas, utilizamos el método “.drop()” y así eliminamos las columnas de “PassengerId”, “Name”, “Cabin” y “Ticket” completamente.

Las razones por las que decidimos eliminar estas columnas se mencionaron previamente en el documento, pero principalmente es porque algunas de ellas solamente se utilizan como identificadores de los datos y no nos proveen información que pueda ser utilizada para saber si alguien sobrevivió o no, y en el caso de “Cabin”, la cantidad de valores vacíos superaba el 70% de sus registros, por lo que no nos iba a poder dar resultados que representen a todo el conjunto.

Eliminación de registros con muy pocos campos vacíos

En cuestión de la columna de “Embarked”, pudimos darnos cuenta que solamente eran 2 los registros que estaban vacíos en esta columna. Debido a que solamente eran 2 registros, y esto no representaba algún riesgo para nuestro modelo o para nuestro análisis, decidimos

simplemente eliminar los registros de estas 2 personas que no contaban con ese dato. Creemos que dado que solo dos registros están incompletos en comparación con el tamaño total del conjunto de datos, eliminarlos tiene un impacto mínimo en el tamaño de nuestra muestra y en la potencia estadística del análisis. Además, dado que se trata de una variable categórica, no es tan sencillo el poder generar una media de los datos y simplemente ponersela a todos los faltantes, por lo que lo mejor fue eliminar estos 2 registros y mantener la integridad de los datos.

Ajuste de campos vacíos de "Edad"

Existe una gran cantidad de registros que no tienen marcada su edad, pero al ser tantos, no es conveniente retirarlos del dataset. Es por esto que para calcular la edad de los pasajeros que no la tienen, lo que hicimos fue sacar un promedio de la edad que tienen los demás pasajeros con el mismo título, clase y sexo. Estos tres datos nos pueden dar una buena idea de la edad de cada pasajero ya que los estamos agrupando en grupos específicos que los representan. Por ejemplo, las "mrs" son mayores que las "miss". O los miembros del staff del barco tienen que tener una edad parecida ya que se encontraban ahí por cuestiones de trabajo.

Transformación de variable categórica "Sexo" a numérica

Otra de las transformaciones que realizamos fue cambiar los valores de "Sexo" de ser strings a enteros. La variable 'Sex' es categórica y toma dos valores distintos: 'female' y 'male'. Sabemos que para entrenar el modelo, lo mejor es trabajar con variables numéricas, por lo que decidimos reemplazar todos los valores de 'female' a 0, y 'male' a 1. Dado que solo tiene dos posibles valores, no fue necesario utilizar ninguna técnica compleja para convertir los valores categóricos o agregar más columnas, simplemente reemplazamos los 2 valores posibles por 0 y 1.

One hot encoding para variable "Embarked"

"Embarked" es otra de las variables categóricas que están presentes en nuestros datos, y existen 3 posibles resultados para ella, dependiendo en cuál es el puerto donde los tripulantes

se subieron al barco. Utilizamos One-Hot Encoding en la columna "Embarked" del conjunto de datos del Titanic porque representa puertos de embarque, que son categorías sin un orden intrínseco. Al convertir cada puerto en una columna binaria separada, evitamos que el modelo asuma una jerarquía entre los puertos y permitimos que capture patrones específicos relacionados con cada puerto individual, facilitando la interpretación y optimizando el rendimiento del algoritmo.

Eliminación de outliers de la columna “Fare”

Como parte del análisis para la limpieza de datos, decidimos crear un diagrama de caja y bigotes para analizar la dispersión de nuestros datos. Al realizar estos diagramas, pudimos darnos cuenta que en la columna de “Fare”, habían 3 datos que consideramos como “outliers”, estos 3 registros presentaban valores de tarifa significativamente más altos que el resto, situándose en un rango de 532 pesos mientras que el promedio estaba alrededor de 32. Dado que solamente eran 3 registros y se encontraban muy alejados de la media e incluso de los datos más altos dentro del diagrama, optamos por **eliminar** estos 3 registros.

Decidimos eliminarlos porque estos outliers podrían distorsionan nuestras estadísticas clave o llegar a sesgar nuestros modelos predictivos. Además, representan solo un 0.3% de los datos, por lo que su eliminación no afecta significativamente la representatividad de nuestro conjunto.

Referencias

1. BBC Mundo, (2012). *Se hunde el mito de "mujeres y niños primero"* BBC Mundo: Sin país. Recuperado el 10 de Agosto del 2023, de:

https://www.bbc.com/mundo/noticias/2012/04/120413_mujeres_ninos_primero_mito_adz

2. InfoBAE (2022). El día que zarpó el Titanic: el lujo, las diferencias entre clases, la pileta climatizada y la escasez de baños. Recuperado el 12 de Agosto del 2023, de: https://www.infobae.com/historias/2022/04/10/el-dia-que-zarpo-el-titanic-el-lujo-las-diferencias-entre-clases-la-pileta-climatizada-y-la-escasez-de-banos/