Tecnológico de Monterrey

Primera Entrega Equipo 5

Luis Mario Lozoya Chairez A00833364

Jackeline Conant Rubalcava A01280544

Francisco Salas Porras A01177893

José Andrés Orantes Guillén A01174130

Javier Eduardo Corrales Cardoza A01742328

09 de agosto del 2024

Inteligencia artificial avanzada para la ciencia de datos I

Profesores

Antonio Carlos Bento

Alfredo Esquivel Jaramillo

Mauricio González Soto

Julio Antonio Juárez Jiménez

Frumencio Olivas Álvarez

Jesús Adrián Rodríguez Rocha

Hugo Terashima Marín

Tecnológico de Monterrey

Enumerar

Introducción

El trágico hundimiento del RMS Titanic en la noche del 14 al 15 de abril de 1912, durante su viaje inaugural desde Southampton hacia Nueva York, es uno de los desastres marítimos más infames de la historia. De las aproximadamente 2.224 personas a bordo, más de 1.500 perdieron la vida, convirtiendo este evento en una tragedia de proporciones colosales. A pesar de las innovaciones tecnológicas de la época, como la compartimentación hermética diseñada para limitar el daño en caso de una brecha en el casco, el Titanic se hundió tras chocar con un iceberg en el Atlántico Norte. Pero cuatro días después del naufragio, el transatlántico británico RMS Carpathia atracó en Nueva York con más de 700 sobrevivientes

El desastre reveló varias deficiencias en las normativas y prácticas de seguridad, incluida la insuficiente cantidad de botes salvavidas, lo que contribuyó significativamente al elevado número de víctimas. Solo había espacio en los botes para 1.178 personas, menos del 50% del total a bordo. Además, las barreras sociales y físicas, como las rejas que separaban a los pasajeros de tercera clase de los de primera y segunda, complicaron aún más la evacuación, resultando en una menor tasa de supervivencia entre los pasajeros de tercera clase. Posterior al hundimiento, una serie de investigaciones y reformas, como la adopción del Convenio SOLAS (Safety of Life at Sea) en 1914, se implementaron para corregir estas deficiencias y mejorar la seguridad en los buques mercantes.

Hoy en día, con el avance de la tecnología, es posible utilizar técnicas de Machine Learning para analizar los datos históricos de los pasajeros y predecir las probabilidades de supervivencia basadas en diversas características. Este análisis no solo tiene un valor académico y de comprensión histórica, sino que también permite explorar cómo se podrían haber tomado decisiones más informadas en el momento.

→ disponible en URL...

Data set original

El Data set proporcionado se divide en dos partes principales: el conjunto de entrenamiento (train.csv) y el conjunto de prueba (test.csv). El conjunto de entrenamiento incluye los resultados reales para cada pasajero, lo que permite construir y entrenar modelos de aprendizaje automático utilizando características como el género y la clase de los pasajeros. Por otro lado, el conjunto de prueba se utiliza para evaluar el rendimiento del modelo en datos no vistos, donde el objetivo es predecir si los pasajeros sobrevivieron al hundimiento del Titanic.

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Inglés

Tabla 1. Texto de tabla.

Tecnológico
de Monterrey

Data Dictionary

El diccionario de datos incluye variables como supervivencia, clase, sexo, edad, y detalles familiares (sibsp para hermanos/cónyuges y parch para padres/hijos). La edad puede ser fraccional o estimada. Estas variables permiten analizar las dinámicas sociales y familiares del Titanic, esenciales tanto para la creación de modelos predictivos.

Variables y Justificación de su Uso

Supervivencia (survival)

- **Definición**: Indica si un pasajero sobrevivió (1) o no (0) al desastre.
- **Justificación**: Esta es la variable objetiva que el modelo intentará predecir. Entender los factores que influyeron en la supervivencia es clave para realizar análisis históricos y predictivos.

Clase de Boleto (pclass)

- **Definición**: Clase del boleto del pasajero (1 = Primera clase, 2 = Segunda clase, 3 = Tercera clase).
- **Justificación**: La clase del boleto es un fuerte indicador del estatus socioeconómico del pasajero, lo cual tuvo un impacto directo en su acceso a los botes salvavidas y, por lo tanto, en su probabilidad de supervivencia. Los pasajeros de primera clase tenían más acceso a los botes, mientras que los de tercera clase se enfrentaban a mayores dificultades debido a las barreras físicas y la distancia a las cubiertas superiores.
- **Comentarios**: Esta variable fue transformada con la función One-Hot-Encoding a 3 variables categóricas.
  - **Pclass_1:** Primera clase.
  - **Pclass_2:** Segunda clase.
  - **Pclass_3:** Tercera clase.

Sexo (sex)

- **Definición**: Género del pasajero.
- **Justificación**: Se dio prioridad a mujeres y niños durante la evacuación, por lo que el género del pasajero es un factor crucial en la probabilidad de supervivencia. Este comportamiento estuvo influenciado por las normas sociales de la época y fue implementado rigurosamente, especialmente por oficiales como Charles Lightoller, quien insistió en llenar los botes solo con mujeres y niños. ✓

Referencia

Tecnológico de Monterrey

Edad (age)

- **Definición**: Edad del pasajero en años.
- **Justificación**: La edad es otro factor significativo, ya que los niños recibieron prioridad en la evacuación y los pasajeros más jóvenes pudieron tener más capacidad física para llegar a los botes salvavidas. Además, las personas mayores podrían haber tenido más dificultades para moverse rápidamente durante el caos.

Relatives cuál es esta variable?

- **Definición:** Representa los lazos familiares, incluyendo posiblemente parientes directos como no directos.
- **Justificación:** Adecuada para análisis más específicos, ajustándose para diferenciar tipos de relaciones familiares y permitiendo agrupaciones más finas en modelos analíticos.

Puerto de Embarque (embarked)

- **Definición**: Puerto donde el pasajero embarcó (C = Cherbourg, Q = Queenstown, S = Southampton).
- **Justificación**: El puerto de embarque podría correlacionarse con la clase socioeconómica de los pasajeros, ya que algunos puertos tenían más pasajeros de tercera clase. Este dato puede proporcionar información adicional sobre el perfil del pasajero y, en consecuencia, sobre su probabilidad de supervivencia.
- Comentarios: Esta variable fue transformada con la función One-Hot-Encoding a 3 variables categóricas.
  - **Embarked_C:** Cherbourg.
  - **Embarked_Q:** Queenstown.
  - **Embarked_S:** Southampton.

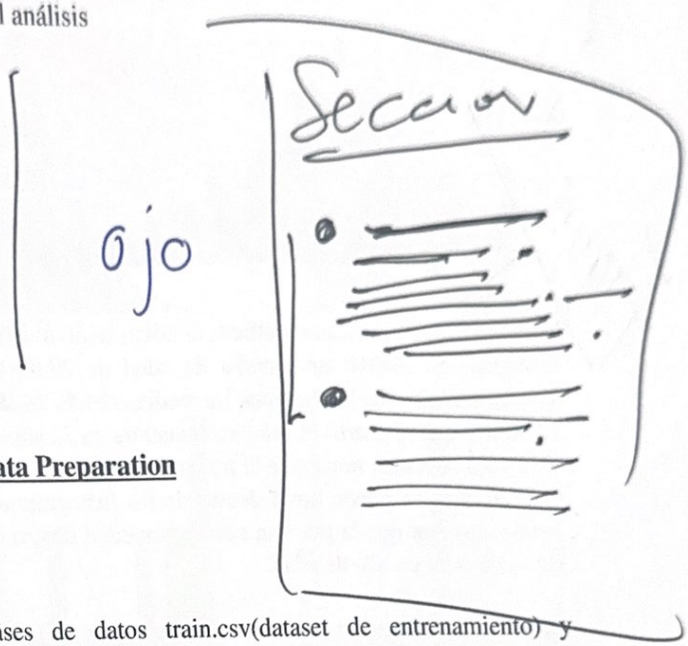
Variables descartadas

cuántos?

Se decidió descartar la variable "Cabina" debido a la gran cantidad de datos faltantes y valores nulos presentes en la base de datos. Aunque la ubicación de la cabina podría haber influido en el tiempo que un pasajero tardó en llegar a la cubierta y, por ende, en su probabilidad de acceder a un bote salvavidas, la falta de información completa impide su uso efectivo en el modelo. Sin datos suficientes y precisos sobre esta variable, su inclusión podría introducir sesgos o errores, disminuyendo la calidad del análisis.

Aunque la "Tarifa del Pasaje" podría servir como un indicador indirecto del estatus socioeconómico de los pasajeros y, por ende, de su clase de boleto, decidimos descartarla. Si bien es cierto que las personas que pagaron tarifas más altas probablemente viajaban en clases superiores, donde las tasas de supervivencia fueron mayores, ya hemos incluido otras variables que reflejan de manera más directa el estatus socioeconómico y la clase de los pasajeros. Además, utilizar esta variable podría introducir multicolinealidad en el modelo, complicando la

interpretación de los resultados. Por lo tanto, optamos por no utilizar la "Tarifa del Pasaje" para evitar redundancias y mejorar la robustez del análisis



Data Preparation

Carga de librerías y datos

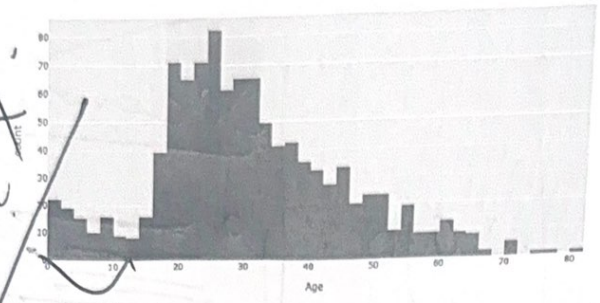
Nos encargamos de extraer las bases de datos train.csv(dataset de entrenamiento) y test.csv(dataset de evaluación) y almacenarlas dentro del repositorio, con el propósito de entrenar un modelo de predicción y su evaluación dentro de nuestro script de Jupyter Notebook. Para iniciar el proceso de limpieza de datos primero se hizo la carga de los datos que se utilizarán para el desarrollo del proyecto se procedió con el análisis exploratorio de los datos, al iniciar este notamos que había varios datos nulos dentro de las columnas de ambos datasets, por lo que se realizó la unificación de los datasets para realizar una limpieza conjunta en ambos datasets. Paralelamente, se realizó la exploración de los datos, para poder entender en que forma se realizaría la limpieza y llenado de los datos. Se encontró una columna de datos con demasiados datos nulos, en adición a esto, los datos que se encontraban contenidos en esta no eran posibles de rellenar, dado que no se contaba con algo que nos pueda ayudar a realizar el relleno, por lo que se eliminó esta columna.

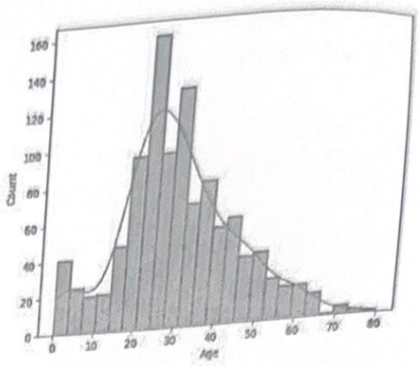
→ Será correcto!

Luego, se pudo encontrar que dentro de los datos de embarcación y tarifa se mostraba con una mínima cantidad de datos nulos. Por lo tanto, se utilizó la función de moda para poder calcular los valores más constantes y poder asignarlo a las variables faltantes. Posteriormente, fue necesario resolver el problema de los datos faltantes que existen dentro de la columna edad, donde se generó un histograma que mostrará los datos previamente a sustituirlos dentro del dataset. Además, se generó una segunda gráfica que permitiera la visualización del polígono de frecuencia de los datos en el histograma.

Se lee como narrativa de hechos, en ver de un reporte descriptivo.

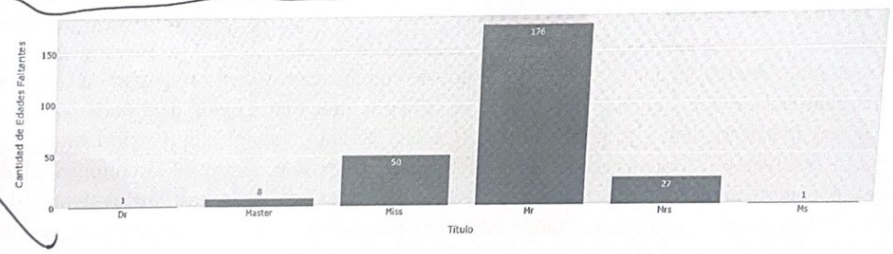
Tecnológico de Monterrey





Adicionalmente, de estos gráficos también se obtuvieron las medidas de tendencia central de esta columna. Se obtuvo una media de edad de 29.88 años, que de cierta manera se puede ver reflejado dentro del histograma. La mediana es de 28 años y la moda de 24 años. Adicional a esto se obtuvo que el cuartil inferior se encuentra en 21 años y el superior se encuentra en los 39 años. Con esto se puede notar que la mayoría de los pasajeros se encontraban entre los 21 y 39 años, esto también se puede notar dentro de los histogramas presentados anteriormente. También se puede observar que la persona con mayor edad dentro del barco contaba con 80 años y la menor con menos de un año de edad.

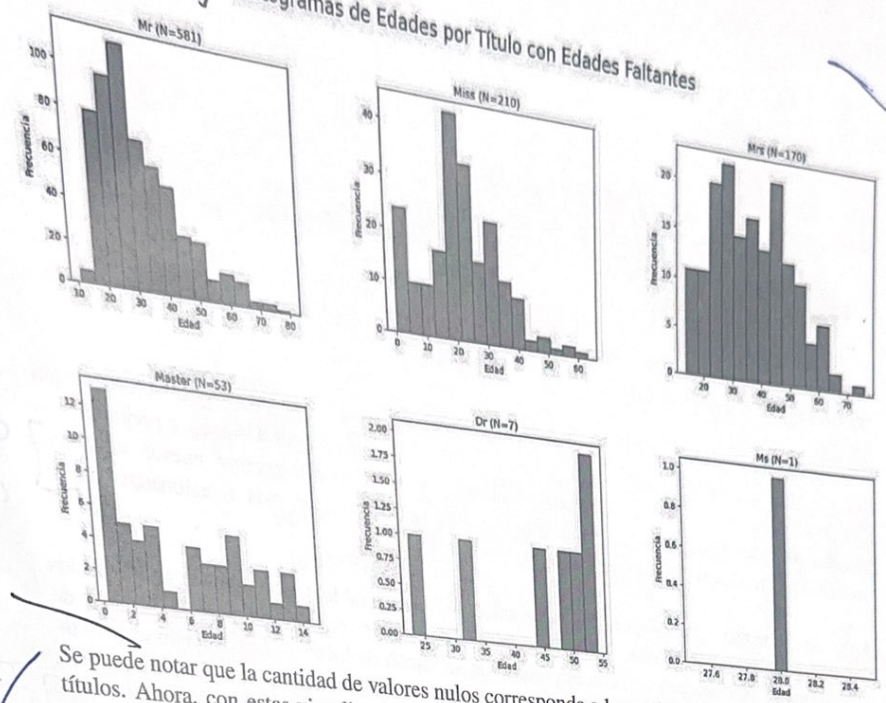
Después de obtener una visualización más clara de las edades, se notó que la mejor manera para sustituir estos datos no era simplemente escogiendo una sola medida de tendencia para el llenado de datos. Se tomó la decisión de utilizar otro de los datos que pertenecen a la base de datos para poder aproximar la edad de forma óptima con otro enfoque. Se notó que los nombres de los pasajeros cuentan con títulos, por lo que se buscó hacer uso de los títulos de las personas para darles una medida, es decir, si una persona tiene el título Mr. y no tiene edad, se usará la media móvil de las personas con ese título. Para realizar esto, primero se generó una nueva columna que contuviera únicamente los títulos de cada pasajero. Con esto visualizamos primeramente a cuantos títulos les faltaban datos.



Con esta visualización es posible observar que los pasajeros con el título *Mr.* son los que tienen mayor número de valores nulos, seguidos de los títulos *Miss*, *Mrs*, *Master*, *Ms* y *Dr*. Teniendo en cuenta esto se procedió con la visualización de la distribución de las edades de estos títulos.

Cuidar la calidad de imagen

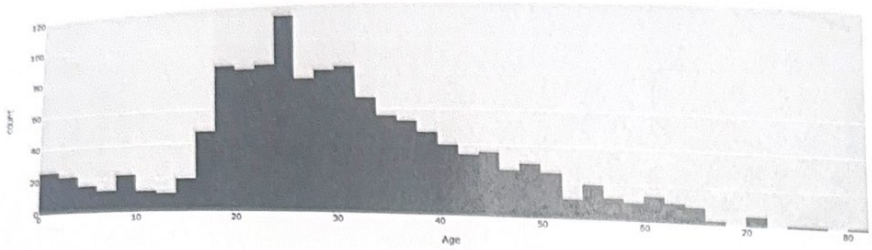
Tecnológico de Monterrey



Se puede notar que la cantidad de valores nulos corresponde a la cantidad de datos que hay en los títulos. Ahora, con estas visualizaciones, se hizo una evaluación sobre cuál medida usar para el llenado de valores faltantes. Se busca que los datos que llenen a los faltantes no alteren la forma que ya cuentan los datos que se tienen, por lo que se decidió utilizar la mediana móvil para los títulos con más de 100 datos. Lo anterior se debe a que, en estos datos, parecen haber valores que pueden sesgar la imputación de valores, por lo que la mediana, al ser robusta, permite que el sesgo no se proyecte al imputar los valores. Además, la mediana solo se utiliza en estos títulos debido a que son los únicos en los que se ha considerado que hay suficientes datos.

Para el resto de títulos se optó por el uso de la media móvil, puesto que no cuentan con suficientes datos para que se aplique la mediana, también parece que la forma de los datos no va a ser influida por la imputación con media, además la cantidad de datos que serán imputación es baja. Cabe destacar que para el título *Ms* se utilizó la moda, pero se pudo haber utilizado cualquier otra medida de tendencia central y se habría obtenido el mismo resultado. Tras haber terminado la imputación de los datos se hizo un segundo histograma de edades, que se presenta a continuación del lado derecho, mientras que en el izquierdo se presenta el histograma con valores faltantes, en el cual se puede notar que la distribución original de los datos se mantiene.

> En la Figura 7,



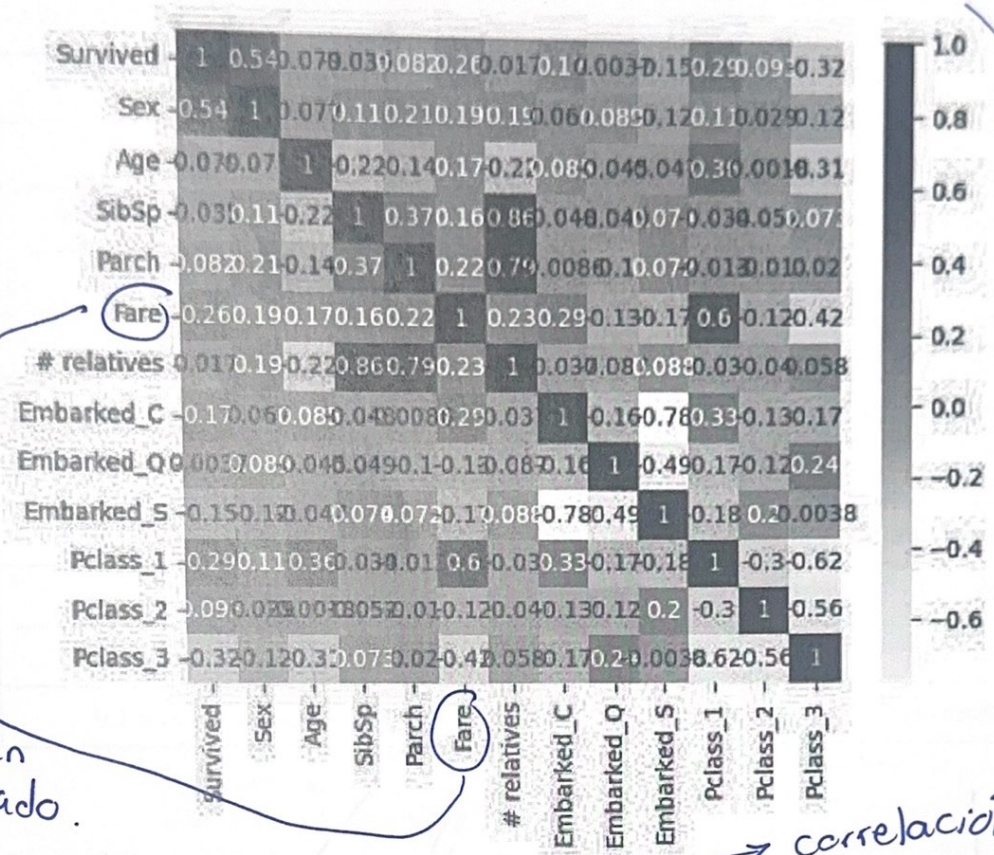
Con el llenado de los datos de la edad de los pasajeros finalmente terminado, se consideró completa la limpieza de la base de datos.

A continuación, se inició el proceso de depuración de datos para determinar qué factores serían útiles en la creación del modelo de predicción deseado y cuáles podrían resultar perjudiciales. Con los datos del Titanic previamente limpiados, se comenzaron a crear nuevas variables, basadas en investigaciones sobre la tragedia, que podrían estar más relacionadas con la probabilidad de supervivencia que buscamos obtener.

Antes de comenzar con la depuración, se planteó utilizar una tabla de correlación entre todos los factores para visualizar cuáles serían los valores de mayor importancia para la variable de supervivencia. Así, se comenzó con el proceso de selección de factores, convirtiendo el sexo de los pasajeros en valores binarios para su uso en el modelo. El sexo masculino fue reemplazado por el valor 0, y el sexo femenino por el valor 1. Posteriormente, se creó la variable de parientes relacionados por sangre o matrimonio, sumando los valores de los factores SibSp (hermanos y cónyuges) y Parch (padres e hijos). Luego, se eliminaron las variables de nombre, ticket y título de los pasajeros que se encontraban en la base de datos, ya que dejaron de ser prioritarias y no serían útiles para los siguientes pasos en la elección de hiperparámetros. Posteriormente se seleccionarán los datos categóricos de la base de datos y se utilizará la función *One-Hot-Encoder* para transformar los valores categóricos en valores numéricos. Estos factores podrán integrarse en una gráfica de calor que mostrará la correlación entre los datos de entrenamiento con la supervivencia de los pasajeros. Las variables categóricas transformadas son *Embarked* y *Pclass*, dado que estas variables parecen ser importantes, es importante señalar que *Pclass* terminaba generando una correlación negativa anteriormente, puesto que eran números y por lo mismo era tratada como variable numérica, sin embargo, se trata de una variable categórica. El mapa de calor que presenta la correlación entre las variables se encuentra a continuación.

Tecnológico de Monterrey

cuidar la calidad de imagen



La habían descartado.

correlación con qué?

Se puede notar que las variables que tienen mayor correlación son las *Embarked*, las cuales se obtuvieron después del One-Hot Encoding. Por lo que su relación se esperaba. La correlación que se quiere observar es entre la variable *Survived* y el resto. Las mejores correlaciones de la variable objetivo, tanto positivas como negativas, fueron: *Sex*, *Pclass* y *Fare*. Sin embargo, estos datos no son suficientes, por lo que finalmente se realiza la selección de las variables.

Aparte de la investigación previa, también se tomó en cuenta un selector secuencial para que proporcionara las características que más beneficiarían a un modelo de regresión logística. Éstas ayudaron a la selección final de características en las que el equipo no se encontraba seguro sobre su elección.

Tras esto se eliminaron las características *SibSp*, *Parch* y *Fare*. Para finalmente realizar el guardado y la generación de la base de datos final que se utilizará en la siguiente etapa de este reto.

suficientes para qué?

No se nota.

Tecnológico de Monterrey

Data set actualizado

Variable	Definición	Key
Survived	supervivencia	No = 0.0, Yes = 1.0
Sex	sexo	Male = 0, Female = 1
Age	edad en años	
# relatives	número de parientes	
Embarked_C	Embarcado en Cherbourg	No = 0.0, Yes = 1.0
Embarked_Q	Embarcado en Queenstown	No = 0.0, Yes = 1.0
Embarked_S	Embarcado en Southampton.	No = 0.0, Yes = 1.0
Pclass_1	Primera clase	No = 0.0, Yes = 1.0
Pclass_2	Segunda Clase	No = 0.0, Yes = 1.0
Pclass_3	Tercera Clase	No = 0.0, Yes = 1.0

cuántos registros? Tabla de figura? Explicar

y cuántos por cada clase?