1. Evaluación para perfil QA Analítica

Instrucciones:

Se dispone de un modelo analítico el cual predice a sobrevivientes al naufragio del Titanic. Se realizarán una serie de preguntas con respecto a la creación del proyecto; dicho proyecto contempla las siguientes etapas: Importación de datos, Análisis exploratorio de datos, Ingeniería de características, Modelamiento y Resultados.

Etapa 1: Importación de Datos

El conjunto de datos esta compuesto por 10 variables cuyo diccionario se muestra a continuación:

Data Dictionary

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	

Data Dictionary

cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

El proyecto importó un archivo “csv” con 1309 registros los cuales se los separó en 891 (68%) registros para datos de entrenamiento (train) y 418 (32%) registros para pruebas (test)

PREGUNTA 1. ¿Qué opinas de la distribución de registros entre los conjuntos de datos train y test, como QA harías alguna observación al desarrollador?

No hay una proporción única y fija que deba ser utilizada en todos los casos para la división de conjuntos de (train) y (test).

- La proporción puede variar según el tamaño del conjunto de datos, la complejidad del problema y otros factores. Sin embargo, hay algunas prácticas comunes que se utilizan en la división de datos para (train) y (test):
- Una división usada es 70% para (train) y 30% para prueba. Esto implica que el 70% de los datos se utiliza para (train) el modelo y el 30% restante se utiliza para evaluar su rendimiento.

Etapa 2: Análisis exploratorio de datos

PREGUNTA 2. Para realizar el análisis exploratorio se tomó el conjunto de datos de entrenamiento, a continuación, se muestran los primeros 5 registros de este.

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25	NaN	S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38	1	0	PC 17599	71.3	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.92	NaN	S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05	NaN	S

A este conjunto de datos se le realizó un análisis de datos nulos, el resultado del mismo es el siguiente:

Column	Nan_cnt	Nan_%
PassengerId	0	0
Survived	0	0
Pclass	0	0
Name	0	0
Sex	0	0
Age	177	0.2
SibSp	0	0
Parch	0	0
Ticket	0	0
Fare	0	0
Cabin	687	0.77
Embarked	2	0

¿Según la cantidad de valores nulos en las columnas age, cabin y embarked, crees que pueda existir un problema en la etapa de creación del modelo?, si es así, que sugerirías realizar al desarrollador para solucionar este inconveniente en cada columna.

- Los valores nulos pueden ser problemáticos durante la etapa de creación del modelo, ya que algunos algoritmos de modelado pueden requerir que todas las variables tengan valores válidos.
- Yo le sugeriría al desarrollador para la columna de age:

Calcular la media o mediana de la columna age para obtener un valor de referencia, posterior reemplazar los valores nulos en la columna age por la media o mediana calculada en el paso anterior. Esto permitiría retener los registros con valores nulos y utilizar la información disponible en otras columnas para el análisis y el modelado

- Para la columna cabin:
  Primero evaluará de la relevancia de la columna cabin:
  Se puede evaluar la relevancia de la columna mediante el uso del peso de la columna correspondiente a las columnas de entrada en una escala de cero a uno más relevante se considera, es decir mientras el valor del peso se acerque más a uno, se relaciona que la columna será más importante
  Después de determinar la importancia se puede eliminar la columna: Dado que la columna cabin tiene una alta proporción de valores nulos si decide mantener la columna cabin y los valores nulos son significativos, puede asignar un valor especial que indique que la información de la cabina está ausente para esos registros reemplazar los valores nulos en la columna Cabina con el valor especial elegido.
  Esto se puede hacer utilizando las funciones o métodos adecuados según la herramienta
  Por ejemplo, en Python con la librería pandas, puede usar la función `isnull()` para identificar los registros con valores nulos en la columna cabin y por ultimo reemplazarlos.
- Para la columna embarked:
  Si la cantidad de filas con valores faltantes en la columna embarked es pequeña en comparación con el tamaño total del conjunto de datos, puedes optar por eliminar esas filas.

  Sin embargo, esto solo es recomendable si la pérdida de datos no afecta significativamente el análisis o el modelo posterior para determinar la importancia de la columna es necesario considerar lo siguiente
  Examina la conexión entre la columna con valores faltantes y la variable objetiva (en este caso, "Sobrevivido").
  Si la columna "embarked" tiene una conexión baja con la variable objetiva, es posible que la falta de valores no afecte significativamente el rendimiento del modelo.

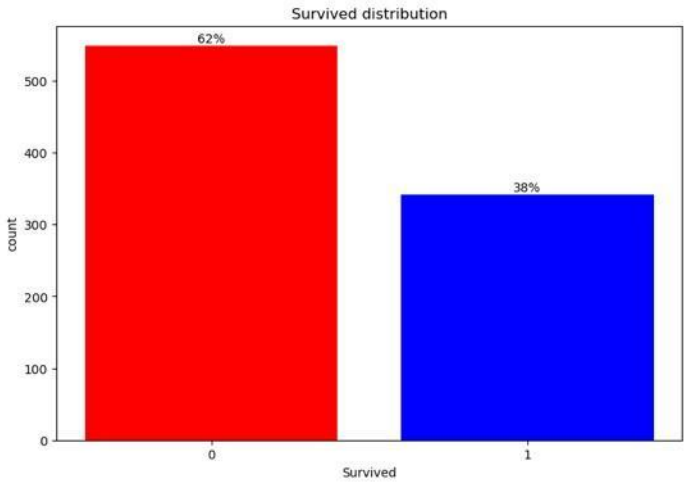
PREGUNTA 3. Adicionalmente se realizó un análisis de valores únicos por columna y este es el resultado:

Column	*number*	*percentage*
PassengerId	891	1
Name	891	1
Ticket	681	0.76
Fare	248	0.28
Age	88	0.1
SibSp	7	0.01
Parch	7	0.01
Pclass	3	0
Embarked	3	0
Survived	2	0
Sex	2	0

¿Crees que existe algún error u observación con respecto a valores únicos de cada variable que afecte posteriormente en la creación del modelo?, Eliminarías alguna variable, ¿Por qué?

- Según los resultados del análisis de valores únicos, no encuentro ningún error según las variables. Sin embargo, la decisión de eliminar una variable depende del contexto del problema y del objetivo del modelo.
- Para el modelo de estudio, se podrían eliminar algunas variables PassengerId: Esta variable representa simplemente un identificador único para cada pasajero y no proporciona información relevante para predecir la supervivencia. Por lo tanto, se podría considerar eliminarla, ya que no aporta valor predictivo las siguientes variables a primera vista, no parece tener una relación directa con la supervivencia y es poco probable que proporcione información útil para el modelo.

PREGUNTA 4. El modelo de predicción para este proyecto usó el algoritmo Gradient Boosting Regressor el cual utiliza arboles de decisión para predecir una variable, construyendo un conjunto de árboles de decisión en serie, donde cada árbol corrige los errores del árbol anterior. La variable objetivo que se desea predecir es *survived* la cual indica si el pasajero sobrevivió (0) o no sobrevivió (1), la distribución de dicha variable es la siguiente:



Observando la distribución de la variable objetivo (survived) y tomando en cuenta el algoritmo usado para construir el modelo, tú como QA ¿Deberías sugerir algún procesamiento adicional a dicha variable?, si es así, ¿cuál sugerirías?

Sí, considerando la distribución desequilibrada de la variable objetiva "Survived" (62% de sobrevivientes y 38% de no sobrevivientes) y el algoritmo utilizado (Gradient Boosting Regressor), se podría sugerir aplicar un método de Re muestreo para abordar el desequilibrio de clases. El desequilibrio de clases puede afectar el rendimiento del modelo, ya que puede haber un sesgo hacia la clase mayoritaria y dificultades para aprender patrones de la clase minoritaria. Al aplicar un método de Re muestreo, se busca equilibrar las clases de la variable objetiva, lo que puede mejorar la capacidad del modelo para generalizar correctamente los casos de ambas clases. Hay dos enfoques comunes para abordar el desequilibrio de clases: sub muestreo y sobre muestreo.

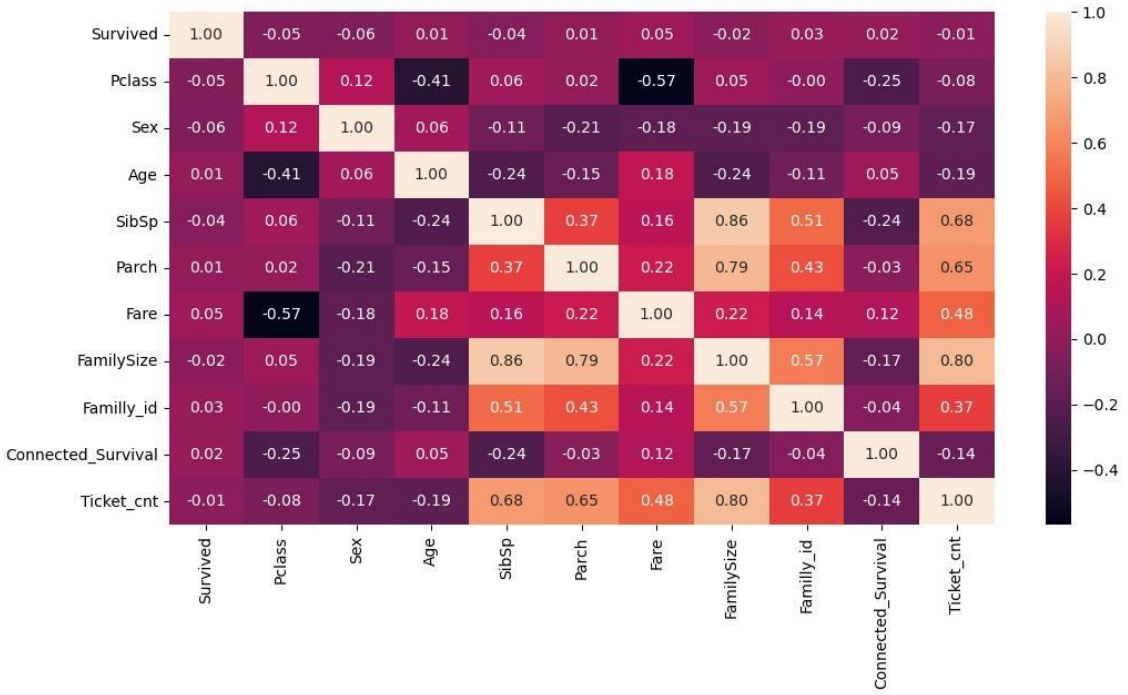
Sub muestreo consiste en reducir la cantidad de instancias de la clase mayoritaria para que se igualen con la clase minoritaria. Esto implica eliminar aleatoriamente ejemplos de la clase mayoritaria.

Etapa 3: Ingeniería de características

El desarrollador del modelo decidió crear varias variables sintéticas a partir del conjunto de datos original las cuales con:

- FamilySise: A partir de las variables *sibsp* que representan hermanos o parejas y *parch* que representan a padres o hijos se pudo determinar el tamaño de una familia
- FamilyId: identificada la familia completa con la variable anterior que indica la relación entre padres e hijos o hermanos y pareja, y su número de ticket se creó un identificador de familia
- Connected_survival: variable creada a partir de si el sobreviviente tenía familia o no, asignándole los valores (1 ; 0,5) correspondientemente.
- Ticket_cnt: cantidad de pasajeros con el mismo número de ticket

Con las variables originales y las sintéticas (creadas por el desarrollador) se realizó un análisis de correlaciones el cual se muestra a continuación:



PREGUNTA 5. A partir del grafico de calor correspondiente a correlaciones entre variables ¿Crees que existes variables correlacionadas?, ¿debería eliminarse alguna variable del conjunto de datos?

- Para determinar si existen variables correlacionadas y si alguna variable debería ser eliminada del conjunto de datos, es necesario realizar un análisis de correlación entre las variables.
  Esto permitirá identificar si hay una alta conexión entre algunas variables, lo que podría indicar una redundancia en la información que aporten al modelo.
  El análisis de conexiones puede realizarse utilizando técnicas como el cálculo del coeficiente de conexiones de Pearson o visualizar una matriz de conexiones. Si se encuentra una alta conexión entre dos variables, podría ser apropiada eliminar una de ellas para evitar problemas de multicolinealidad y reducir la complejidad del modelo. En el caso específico de las variables mencionadas, podemos analizar las conexiones entre ellas: 1. FamilySize y FamilyId: Estas variables están relacionadas, ya que FamilySize se calcula a partir de la combinación de sibsp y parch, mientras que FamilyId utiliza sibsp, parch y el número de ticket. Podría haber una conexión alta entre ellas, por lo que sería recomendable eliminar una de las dos para evitar redundancia en la información.
  Además de las variables mencionadas anteriormente, es importante realizar un análisis de conexión entre todas las variables del conjunto de datos para identificar posibles relaciones

fuertes o redundancias. Algunas otras combinaciones de variables que podrían evaluarse incluyen:

- Se pueden eliminar pclass y Fare: La clase del billete (Pclass) y el precio del billete (Fare) podrían estar correlacionados, ya que es probable que los billetes de clase más alta tengan precios más altos. Si se encuentra una conexión fuerte entre estas variables, podría ser redundante incluir ambas en el modelo y sería recomendable eliminar una de ellas

Etapa 4: Modelamiento y resultados

Luego de entrenar el modelo con el conjunto de train y probarlo sobre el conjunto de prueba se obtuvieron los siguientes resultados:

Train

	precision	recall	f1-score	support
0	0.86	0.87	0.87	134
1	0.80	0.79	0.80	89
accuracy			0.84	223
macro avg	0.83	0.83	0.83	223
weighted avg	0.84	0.84	0.84	223

Test

	precision	recall	f1-score	support
0	0.93	0.87	0.90	143
1	0.80	0.89	0.84	80
accuracy			0.88	223
macro avg	0.87	0.88	0.87	223
weighted avg	0.88	0.88	0.88	223

PREGUNTA 6. En base a los resultados anteriores, que interpretación puedes dar, ¿Aprobarías el modelo como válido?

- El modelo ha obtenido buenos resultados en términos de precisión, recuperación, F1-score y precisión en ambos conjuntos de datos (entrenamiento y prueba). Estas métricas se utilizan para evaluar el rendimiento de un modelo de clasificación. - Precisión: La precisión es la proporción de casos positivos correctamente identificados sobre el total de casos clasificados como positivos. En este caso, tanto en el conjunto de entrenamiento como en el conjunto de

prueba, la precisión para ambas clases (sobrevivientes y no sobrevivientes) es alta, lo que significa que el modelo logra predecir correctamente la etiqueta de supervivencia en la mayoría de los casos y considero al modelo como válido.