Marcela Melgar

20200487

**Laboratorio 8 – Feature Engineering**

1. **Missing Data para las columnas**

* Según el dataset que contenía los datos de pasajeros del Titanic, se encontraban 183 individuos con 12 distintas variables de observación. Dentro de estas observaciones, existen 6 en las cuales existían valores que se encontraban faltantes tales como el sexo, la edad, la cantidad de parientes a bordo, la cantidad de hijos, la tarifa de pasajero y el puerto de embarcación.

“Sex” la variable que diferencia el sexo de cada uno de los pasajeros, contenía 51 valores faltantes. En términos porcentuales, los valores no registrados eran del 28% a comparación de todas las observaciones registradas. Es una variable categórica en la que los valores que no se encontraban registrados estaban identificados con ‘?’.

“Age”, la variable numérica que identificaba la edad de los pasajeros mostraba 25 observaciones con espacio vacío. Siendo este un 14% de la información dentro de esta variable.

“SibSp”, la variable numérica que mostraba la cantidad de parientes con las cuales contaba el pasajero. Esta variable contaba con únicamente 3 valores faltantes, siendo este un 1.63% de la información.

“Parch”, la variable numérica que muestra la cantidad de hijos que cada uno de los pasajeros contaba a bordo. Esta variable contaba con 12 valores faltantes, siento este un 6.55% de la data presentada.

“Fare”, la variable doble que representa la tarifa que pagó cada pasajero para poder abordar en el Titanic. Mostraba 8 observaciones faltantes, siendo este el 4.37% de información presentada para esta variable.

“Embarked”, la variable categórica que mostraba el puerto en la cual cada uno de los pasajeros abordó el barco. Contaba con 12 valores faltantes siendo este un 6.55% de información para la variable.

1. **Modelo utilizado por columna**

* Sex: Imputación sectorizada por moda, al ser la observación que más valores faltantes contenía, y únicamente existían 2 posibles opciones de variable categórica, se me hizo que asignar los valores faltantes según tendencias sería la mejor opción. También se le aplicó una estandarización.
* Age: imputación sectorizada por media, al ser el 25% de datos los que se encontraban faltantes dentro de esta y contaban con una relación entre otras variables, era una imputación adecuada para asignar valores a los datos perdidos.
* Parch: imputación sectorizada por moda, para mantener la coherencia en los datos y no tener valores atípicos dentro de la cantidad de hijos con las que cuentan los pasajeros.
* SibSp: imputación sectorizada por moda, al igual que la variable de cantidad de hijos, para imputar los valores faltantes, se hizo por medio de moda para seguir una tendencia y no contar con datos atípicos o que no consideraran a toda la persona.
* Fare: Regresión lineal, modelo de predicción lineal, utilicé este modelo para que en función de los demás valores dentro de la variable y poder predecir cuál sería el precio pagado por los pasajeros que contaban con este dato faltante.
* Embarked: imputación sectorizada por moda, al ser una variable categórica en donde solo se contaban con cierta cantidad de opciones por las que el pasajero pudo haber ingresado al barco, era más optimo imputar los datos según las tendencias que mostraban las demás observaciones. De esta manera no se encontrarían valores atípicos dentro de las variables.

1. **Reporte de filas que se encuentran completas**

* Dentro de la base de datos de pasajeros del Titanic brindada, se encontraban variables que no contaban con ningún valore faltante dentro de ellas. Las variables sin ningún valor vacío eran la identificación de cada pasajero, si sobrevivió, la clase, nombre, ticket y el número de cabina.

“PassengerId”, variable numérica descriptiva y categórica, mostraba el número único de identificación por pasajero.

“Survived”, variable numérica y categórica con opciones de valor 1 y 0. Describe si el pasajero fue de los que sobrevivió el incidente del hundimiento del barco (1) o si el pasajero es de los que murió por el hundimiento (0).

“Pclass”, variable numérica y categórica que muestra la clase en la que se encontraba hospedado el pasajero dentro del barco. 1 representa primera clase, 2 clase media y 3 clase económica.

“Name”, variable string categórica con el nombre de cada uno de los pasajeros.

“Ticket”, variable string en la que representa el código del ticket adquirido por cada uno de los pasajeros.

“Cabin”, variable string y numérica en la que se muestra la identificación de la cabina en donde se encontraba hospedado el pasajero.

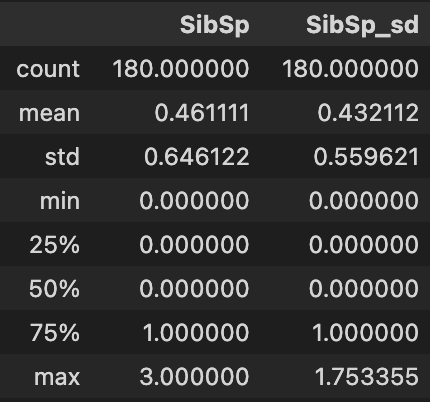
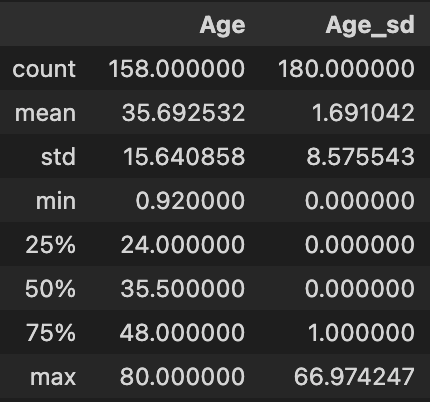
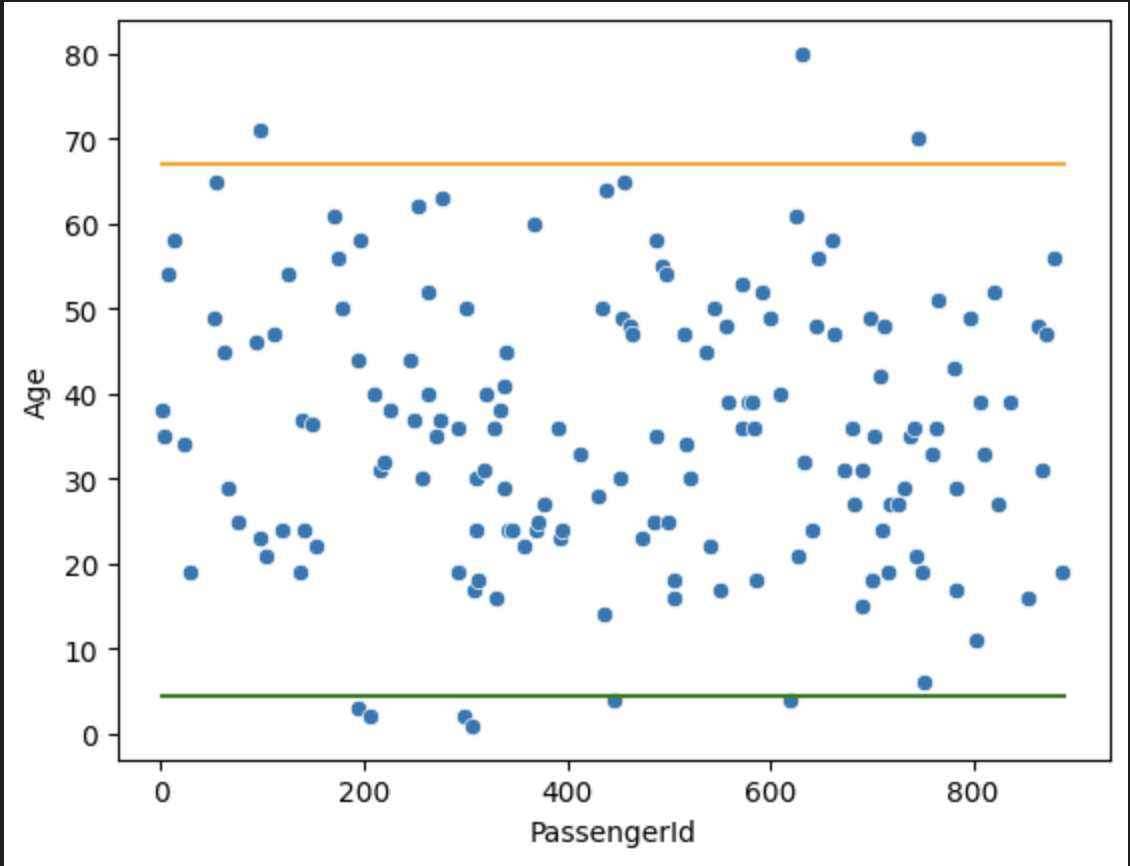
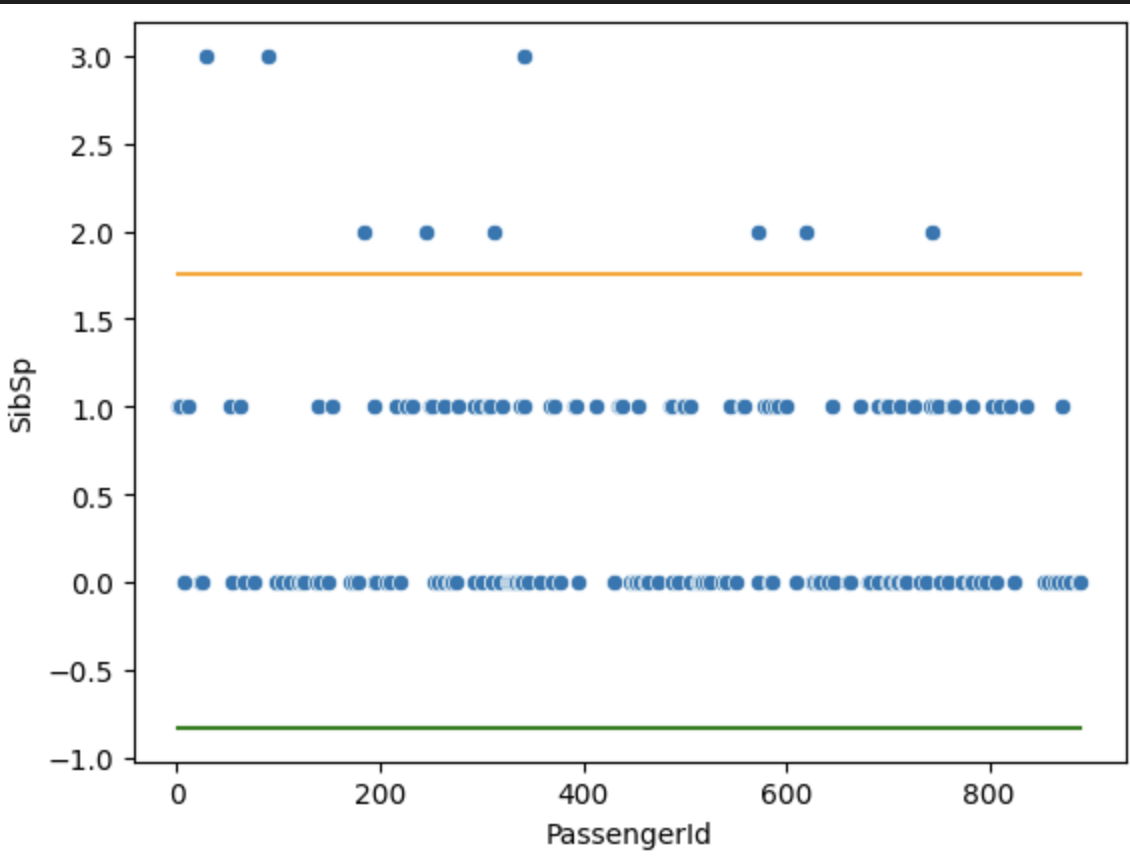
1. **¿Qué método se acerca más a la realidad?**

* Para la variable Sex, el mejor método para utilizar es el de imputación por moda porque es una de las imputaciones más exactas para las variables categóricas. Para la variable Age, entre los mejores métodos se encuentra la imputación sectorizada por media, ya que sería la integración de datos según el promedio de edades que se encuentran registradas dentro de las observaciones. Para la variable que muestra la cantidad de parientes y para la variable que presenta la cantidad de hijos de cada pasajero, la mejor opción sería el modelo de desviación estándar porque permite la visualización de la dispersión entre valores y poder definir el mejor valor para cada uno. Para fare, la más realista sería el modelo de regresión lineal, ya que compara los demás valores dentro de la variable y ayuda a encontrar una correlación entre ellas y brindar el valor adecuado predicho por el mismo. Por último, para encontrar el puerto en donde cada uno de los pasajeros embarcó el barco, la imputación por moda sería el modelo más confiable pues, al ser una variable categórica, es bastante adecuada y asignaría un valor según la tendencia mostrada por las demás observaciones dentro de la variable.

1. **Conclusiones**

* Para el análisis de datos dentro de una base de datos, es de suma importancia saber qué es lo que cada variable está mostrando para entender la información general. También, es importante encontrar cuáles son las variables que cuentan con información faltante o valores que se encuentren incompletos. Esto es parte de la preparación de la data para poder trabajar con ella. Es fundamental entender cómo se pueden trabajar las variables categóricas a diferencia de las variables numéricas para saber cuáles son los modelos que trabajarían mejor para imputar su información faltante. Al tener varias herramientas y modelos para asignar valores dentro de los que se encuentran faltantes, el entendimiento de las variables y el conocimiento de los modelos son los que permitirían encontrar la mejor manera para encontrar e incluso asignar los valores que se encuentran faltantes dentro del modelo y así, poder trabajar con el mismo.

1. Normalización de columnas

* 
* La desviación estándar en normalización sirve para asegurar que los datos estén dentro de un rango en específico y no sean demasiado altos o bajos. Como se puede vere en los datos encontrados dentro de la desviación en edad y cantidad de parientes, existe menos ruido dentro de los datos, la varianza logra ser menor.
* 
* La manera de lograr el escalar de datos dentro de 0 y 1 es por medio del min max scaler. Es útil en este caso ya que se encuentran datos con diferentes escalas y el min max scaler los transforma dentro de la misma escala. Los datos normalizados muestran información más clara, sin los datos faltantes y dentro de la misma escala.
*  
* 