**Instrucciones Tarea 3.**

Para esta tarea se utilizo el dataset llamado hotel\_booking.csv disponible en Kaggle en <https://www.kaggle.com/datasets/mojtaba142/hotel-booking>.

Se define como la variable objetivo ‘hotel’ la cual es una variable binaria que puede tomar dos valores: ‘Resort Hotel’ o ‘City Hotel’.

El objetivo de los modelos desarrollados es predecir a que tipo de hotel pertenece la reservación ingresada. La distribución de dicha variable es 66% para ‘City Hotel’ y 33% para ‘Resort Hotel’ la cual se considera aceptable.

El dataset contiene 119.390 observaciones y 36 columnas con información recabada entre 2015 y 2017. La información personal disponible en el dataset se construyó artificialmente por lo cual para este estudio se decide no utilizar. Así mismo se decide no utilizar las observaciones que tienen ausente el país de procedencia de la reservación (488) y todos los demás valores nulos se decide inputarlos con 0 ya que dada la naturaleza de las variables (‘children’, ‘agent’, ‘company’) el 0 representa la ausencia de uno de estos elementos lo cual se interpreta como adecuado. Esto deja el numero total de observaciones del dataset en

118.902.

Se presenta a continuación un resumen de las variables del dataset:

Tabla

Descripción generada automáticamente

Interfaz de usuario gráfica, Tabla

Descripción generada automáticamente

Distribución de la variable objetivo ‘Hotel’:

Tabla

Descripción generada automáticamente con confianza media

Se crea la matriz de correlación para las variables numéricas:

Interfaz de usuario gráfica, Aplicación, Teams

Descripción generada automáticamente

Se crean dos modelos de clasificación binaria (regresión logística y árbol de decisión) los cuales generan los siguientes resultados para el área bajo la curva ROC.

Imagen de la pantalla de un celular con texto e imagen

Descripción generada automáticamente con confianza media

Como puede observarse el área bajo la curva ROC (AUC) es más alta para el modelo de regresión logística que para el árbol de decisión lo cual implica que este modelo hace un mejor trabajo discriminando entre positivos verdaderos y falsos positivos para cada umbral de decisión. Por tanto el modelo de regresión logística es mejor.