Checkpoint 2 : Árboles de Decisión

En este trabajo sobre árboles de decisión, primeramente se hizo un preprocesamiento de los dataset, luego un primer modelo de un árbol de decisión con su respectivo análisis de métricas y luego un modelo más avanzado con una optimización por CrossValidation de 5 kfolds.

En el preprocesamiento de datos, contamos con ya un dataset sin valores faltantes y una limpieza de outliers severos, sin embargo había que hacer ciertos filtros tanto para el dataframe de train como de test.

El primero de ellos, fue la eliminación de columnas irrelevantes:

- **id**: Se consideraba irrelevante desde el análisis exploratorio
- **reservation_status_date**: Esta variable no se podía utilizar ya que contaba con información del target
- **assigned_room_type** y **reserved_room_type**: Ambas variables las consideramos irrelevantes debido a la creación de missasigned_room en el análisis exploratorio.

Luego, se realizó un *Label Encoding* de la variable **country** ya que el modelo necesita de valores numéricos y country era una variable categórica en texto.Para una mejor evaluación de la única variable continua, se normaliza por z-score a la variable **adr**. Finalmente, como última modificación del dataset, se realizó *One Hot Encoding* para transformar las variables cualitativas restantes y puedan ser procesadas por los modelos. Cabe destacar que desde el dataset de train, separamos los datos en train y test locales para tomar nuestras propias médicas, ya que no contábamos con los valores reales del target para el dataset test.

Para optimizar el árbol, buscamos los mejores hiperparámetros para el modelo mediante k-fold CrossValidation para obtener la mejor performance. Decidimos buscar hiperparámetros que optimizaran accuracy ya que nos pareció la métrica más adecuada. Realizamos 20 iteraciones para buscar los mejores hiperparámetros, y dividimos el dataset en 5 folds para el Cross Validation. Elegimos usar 5 ya que nos pareció que es una cantidad razonable de subconjuntos.