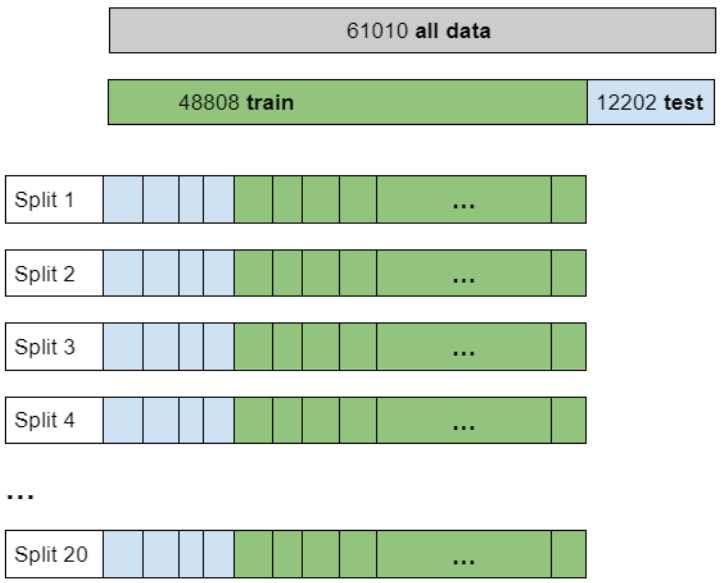
75.06 /95.58 Organización de Datos
Ing. Rodríguez - 1°C 2023

Informe TP1: Reservas de Hotel

Integrantes:
Agama Avila, Arely - 105829
Martinez, Selene Anahi - 100439
Meichtri, Melany - 102330

Checkpoint 2:

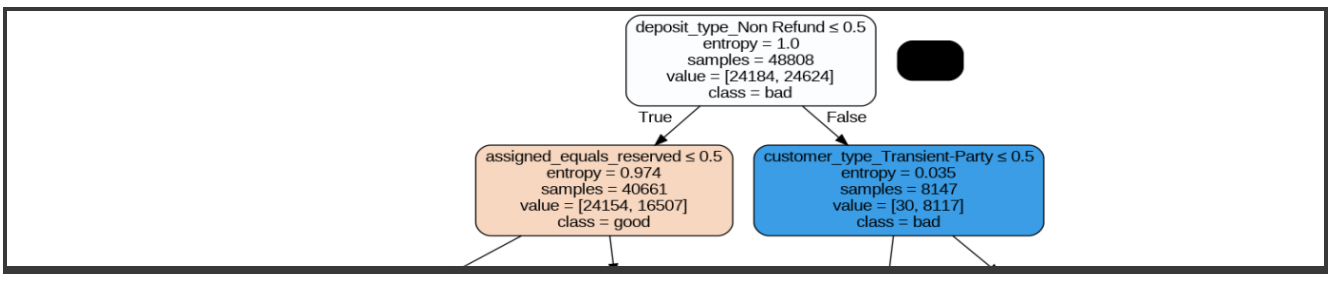
A. Primero decidimos dividir en 80/20 el dataset.Utilizamos el 80% para , es decir 48808 de 61010 filas de nuestro dataset para entrenar y el 20% (12202) para testear. Elegimos f1 score ya que es una métrica que tiene en cuenta tanto el recall como la precisión



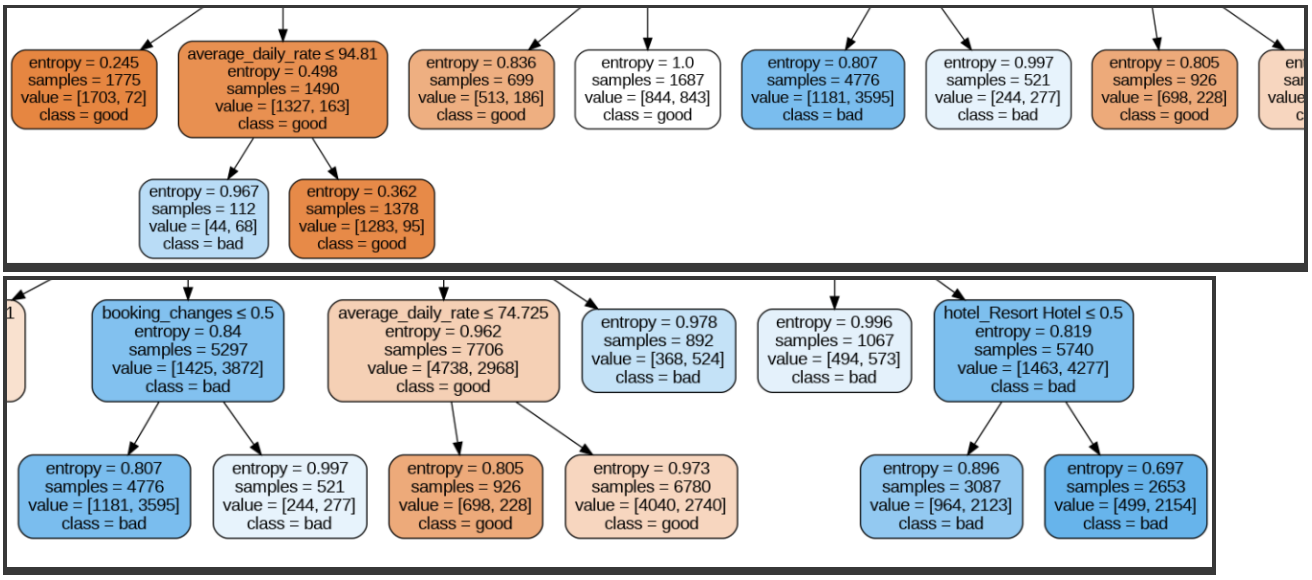
En un principio lo separamos en 20 folds y luego subimos a 45/40 .

B. Graficar el árbol de decisión con mejor performance

Raíz



Hojas



ÁRBOL 1

Elegimos las variables que estaban más relacionadas con el target en nuestro heatmap del CH1.

Transformamos las variables categóricas ‘meal’ y ‘hotel’ para poder utilizarlas en el árbol utilizando one-hot-encoding

Como eran 10 variables, decidimos que nuestro árbol (1, 8) de profundidad .

Al tener tantas observaciones, decidimos aumentar nuestra cantidad de folds,probamos con 20.

Del 1er árbol obtuvimos que las features más representativas para clasificar eran lead_time y previous_cancellations. Las reglas para este árbol fueron lead_time <= 17.5 y después clasificó por previous_cancellations >0.5.

Nuestro mejor parámetro obtenido solo tuvo profundidad de 2, por lo que, en los siguientes árboles, decidimos cambiar el mínimo para que sea un mejor clasificador.

ARBOL 2

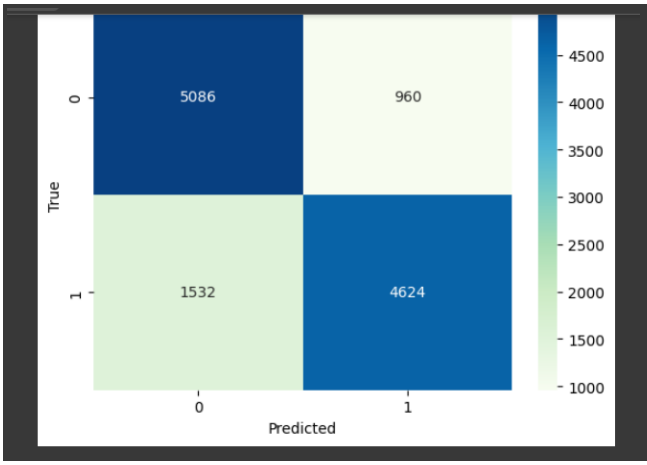
Llegamos a la conclusión que debíamos hacer clustering con algunas variables, como country dado que Portugal tiene mayor relación con el target. Hicimos clustering de assigned_room_type y reserved_room_type. Por último aplicamos one hot encoding en esas variables cualitativas, también en hotel, meal ,deposit_type, customer_type y country.

ÁRBOL 9

En el árbol 9 pudimos confirmar que las variables cualitativas mejoran mucho la predicción .

Los atributos más considerados en este árbol fueron:

```
hotel_Resort Hotel: 0.004064312174447036
customer_type_Transient-Party: 0.008417469018179847
booking_changes: 0.011756899399145369
total_of_special_requests: 0.030410535601930806
previous_cancellations: 0.033226819142557186
average_daily_rate: 0.03718664594185886
lead_time: 0.07234795917454122
required_car_parking_spaces: 0.08332980247837111
country_PRT: 0.08969252926235621
assigned_equals_reserved: 0.09841909682357101
customer_type_Transient: 0.10525721878643467
deposit_type_Non Refund: 0.4258907121966065
```




```
Accuracy: 0.7957711850516309
Recall: 0.751137102014295
Precision: 0.828080229226361
f1 score: 0.7877342419080067
```

Mejoramos el **recall**, es decir, que ahora nuestro último puede clasificar correctamente como positivos otros casos que antes tomaba como negativos.
Mantuvimos la **precisión:** se mantiene la cantidad de falsos positivos identificados
Mejoramos el **accuracy**