Checkpoint 2 - Grupo 24

## Introduccion

Comenzamos haciendo ingeniería de características en el dataset de train como por ej: juntar los países más relevantes, juntar la cantidad de personas total por habitación, entre otras cosas. Luego entrenamos el dataset , y armamos el primer árbol, luego realizamos un segundo árbol utilizando la técnica de split. Y para finalizar creamos un tercer modelo “optimizado” con Random search y K-fold Cross validation.

## Construcción del modelo

Detallar como mínimo los siguientes puntos del modelo que seleccionaron cómo su mejor predictor

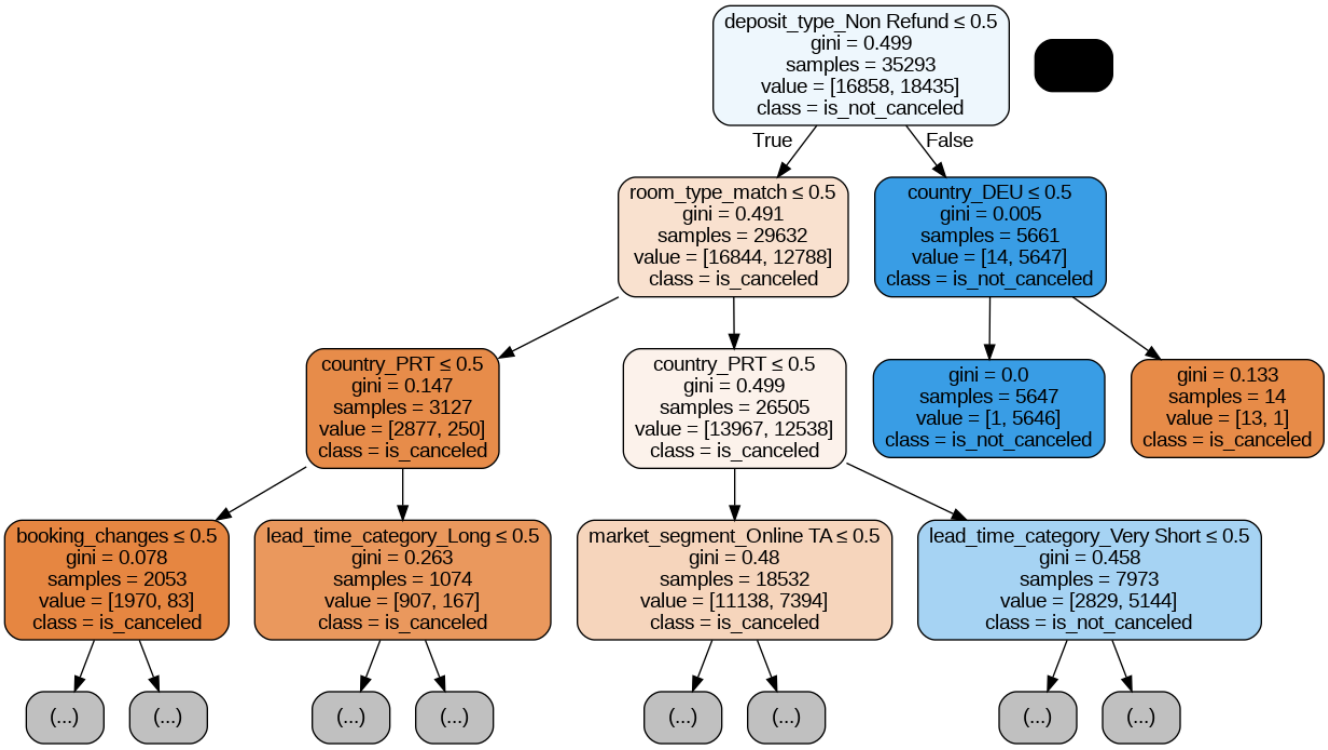
- Optimizacion de hiperparámetros:
  Como optimización de hiperparametros utilizamos random search y cross validation ,como bien mencionamos anteriormente , sin embargo no nos fue de mucha ayuda, ya que nos empeoró el f1 score.
  Los parámetros optimizados fueron los siguientes:
  ***min_samples_split, min_samples_leag, max_depth, criterion y ccp_alpha***

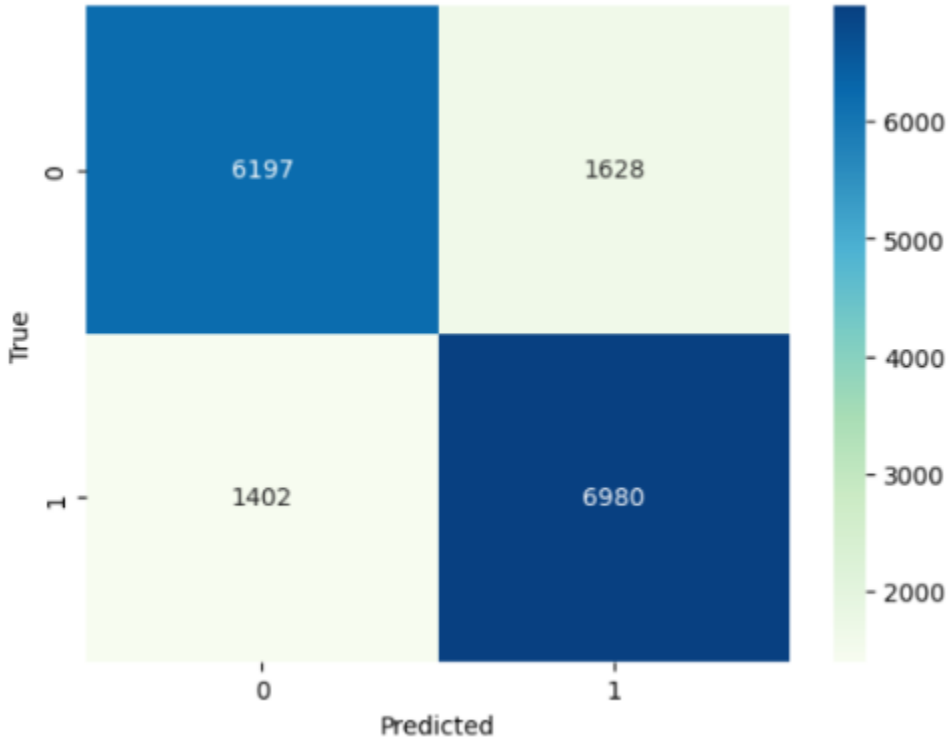
- Utilizamos K-fold Cross Validation y los fold utilizados fueron **15**.

- Nosotros consideramos adecuado optimizar la métrica de **f1 score** para buscar los mejores hiperparametros.

- Realmente no hubo una mejora significativa desde la métrica inicial a la final y hasta en casos nos ha empeorado, sin embargo esto solo nos paso en nuestro colab. En el kaggle mejoraba respecto de nuestro colab pero no superaba nuestra mayor puntuación(realizada sin optimización).

- Se utiliza un árbol de decisión para clasificar instancias. Comienza en el nodo raíz y se mueve hacia las hojas. La primera variable considerada es "Depost_type" debido a su alta importancia. Dependiendo de su valor, la instancia se dirige a la izquierda o derecha. Luego se evalúa "Room_type_match" y “country” respectivamente. En el último nodo, se decide la clase (“is_cancel”, “is_not_cancel”) basándose en"Depost_type". Cuanto más naranja sea el nodo, mayor es la confianza en la decisión.




## Matriz de Confusion

Esta es nuestra matriz de confusión de nuestra mejor predicción, en la cual logramos reducir un poco los falsos negativos.

Tabla con las las 3 mejores predicciones:

| Modelo | F1-Test | Precision Test | Recall Test | Accuracy | Kaggle |
|---|---|---|---|---|---|
| **modelo_1** | 0.819739 | 0.826044 | 0.8135289 | 0.814956 | 0.81967 |
| modelo_2 | 0.817819 | 0.828840 | 0.807086 | 0.81403 | 0.81851 |
| modelo_3 | 0.816732 | 0.826181 | 0.799986 | 0.81201 | 0.81703 |

## Tareas Realizadas

Indicar brevemente en qué tarea trabajo cada integrante del equipo, si trabajaron en las mismas tareas lo detallan en cada caso (como en el ejemplo el armado de reporte).

| Integrantes | Tarea |
|---|---|
| Lautaro Torraca, 108813<br>Marco Tosi, 107237<br>Gianluca Negrotti, 108184 | En este tp , trabajamos casi siempre juntos, a excepción del último día , que trabajamos en solitario intentando mejorar el f1 score, y no pudimos. |