De manera pre exploratoria, replicamos todos los cambios realizados anteriormente en nuestro *dataset* de *train* en el *dataset* de prueba. Además, los casos que antes preferíamos utilizar *true* o *false* los cambiamos a 0 y 1 para utilizar el árbol de decisión.

El primer paso realizado para la creación del modelo fue obtener columnas *dummies* para todas nuestras variables cualitativas. Sin embargo, esto nos dio como resultado un *dataframe* de 791 columnas que ralentiza enormemente el modelo. Para solucionar esto, primero imprimimos las características importantes de nuestro modelo para ver si podemos eliminar alguna columna que no nos sirva, y decidimos quitar las columnas *children* y *babies*. Luego de esto, vemos que la mayor cantidad de columnas pertenece a las *dummies* de *country*, *agent* y *company*, por lo que decidimos reemplazar las columnas por la probabilidad de que el valor tomado pertenezca a una reserva cancelada mediante la función *cambiar_columna_por_probabilidad_is_canceled*, y nos aseguramos de que esto no impacte negativamente a la precisión (en especial al puntaje F1) del modelo. De esta manera obtenemos un *dataframe* de tan solo 66 columnas.

Para buscar hiperparámetros creamos un diccionario con todos los parámetros que puede tomar una *RandomizedSearchCV* de scikit-learn y les dimos valores en un rango numérico o dentro de una lista de posibles valores en el caso de los parámetros que toman *strings*. Luego de pasarle una gran cantidad de iteraciones (entre 200 y 1000) al *RandomizedSearchCV* nos fijamos qué parámetros nos dieron un mejor *F1_score*, ya que nos permite comparar la *accuracy* y el *recall* de manera conjunta y equilibrada. A medida que encontramos mejores parámetros, ajustamos los valores para acercarnos a él y reducimos cada vez más la cantidad de iteraciones, hasta encontrar un puntaje que nos parezca adecuado.

En cuanto al árbol obtenido, decidimos mostrar una parte representativa con una profundidad de cinco nodos. En el mismo se puede ver como las variables más condicionantes para la toma de decisiones son *deposit_type, agent_prob_is_canceled,* si el *customer_type* es *Transient-Party*, *country_prob_is_canceled, lead_time, previous_cancellations,* entre otros (la lista entera se obtiene corriendo *caracteristicas_importantes(best_model, n)* al principio del trabajo, donde *n* es la cantidad de variables a listar).

En cuanto a la matriz de confusión, obtenemos que un ~18% de los valores predichos como positivos son falsos positivos, mientras que un ~16% de los valores predichos como negativos son falsos negativos. En total un 17% de los datos fueron predichos de manera incorrecta.