Fiuba
Facultad de Ingeniería
Universidad de Buenos Aires

Trabajo Práctico N°1
Reservas de Hotel

PRIMER CUATRIMESTRE DE 2023

Materia: Organización de Datos

Grupo N.º 5

Integrantes:

Padrón	**Apellido y Nombre**
104880	LOURENGO CARIDADE, LUCIA

Análisis Exploratorio y Preprocesamiento de Datos

Checkpoint 1:

A lo largo de este trabajo utilizaremos el dataset provisto por la cátedra, el archivo hotel_train.csv. El objetivo del trabajo será lograr predecir si una reserva será cancelada. Lo importamos mediante la librería pandas.
El dataset corresponde a un listado de las reservas de 2 hoteles: "City Hotel" y "Resort Hotel".
Analizamos cada una de las variables y las clasificamos en cualitativas, cuantitativas e irrelevantes.
Para las variables cuantitativas, buscamos su media, mediana y moda, y mediante histogramas analizamos sus distribuciones.
Para las variables cualitativas, estudiamos los valores que toman y cuán frecuentemente lo hacen mediante gráficos de barras y torta.
Por otro lado, analizamos la correlación existente entre las variables e hicimos un heatmap para visualizar la relación entre las variables y el target.
Por último, se detectaron los valores atípicos en los datos en forma univariada y multivariada. Para ello, estudiamos cómo se encuentran distribuidos los datos mediante gráficos boxplot y visualizamos los que se alejan de la distribución esperada.
Eliminamos las columnas company y agent. Company porque tenía aproximadamente el 95% de sus datos nulos y agent porque si bien solo tenia el 13% nulo, no consideramos que fuera una columna relevante para nuestro análisis.
Por otro lado, eliminamos las filas que tenían el valor outlier más alto de unas determinadas columnas, ya que al ser pocas, no creemos que estemos perdiendo muchos datos para el análisis.

Gráficos:

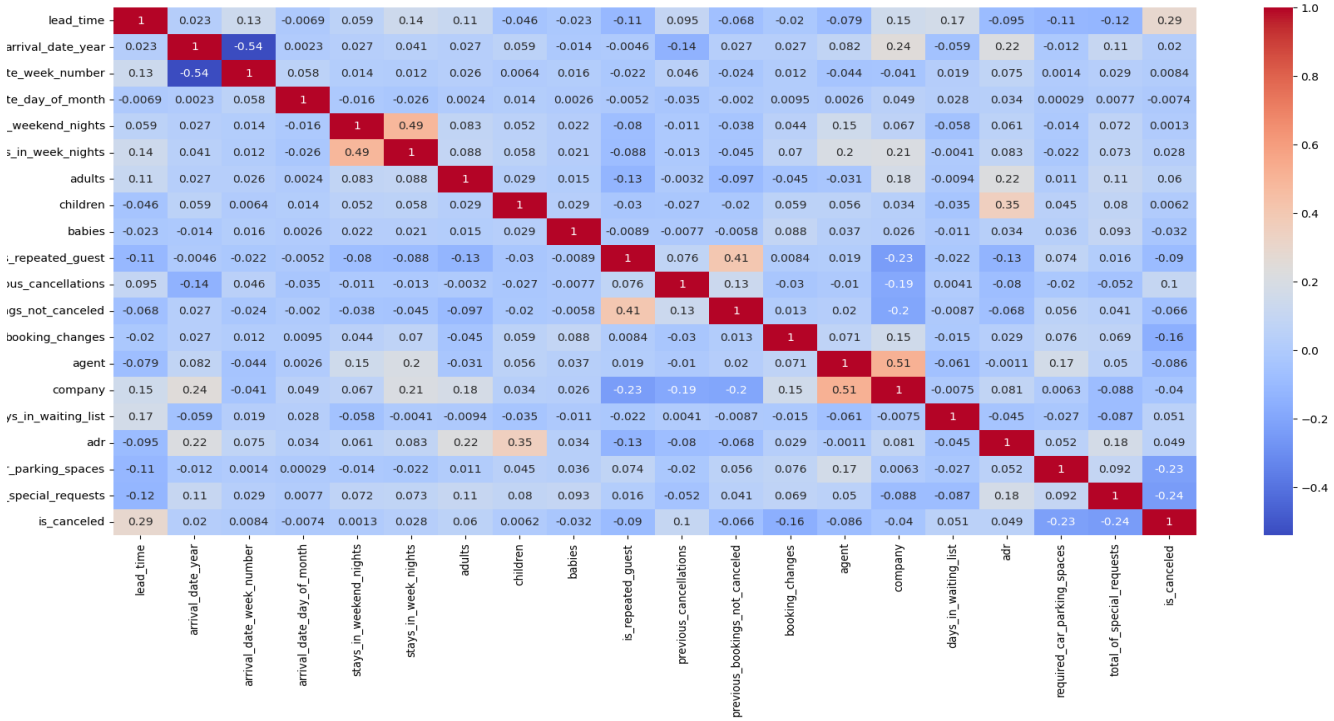

gráfico 1: de correlación entre las variables y el target

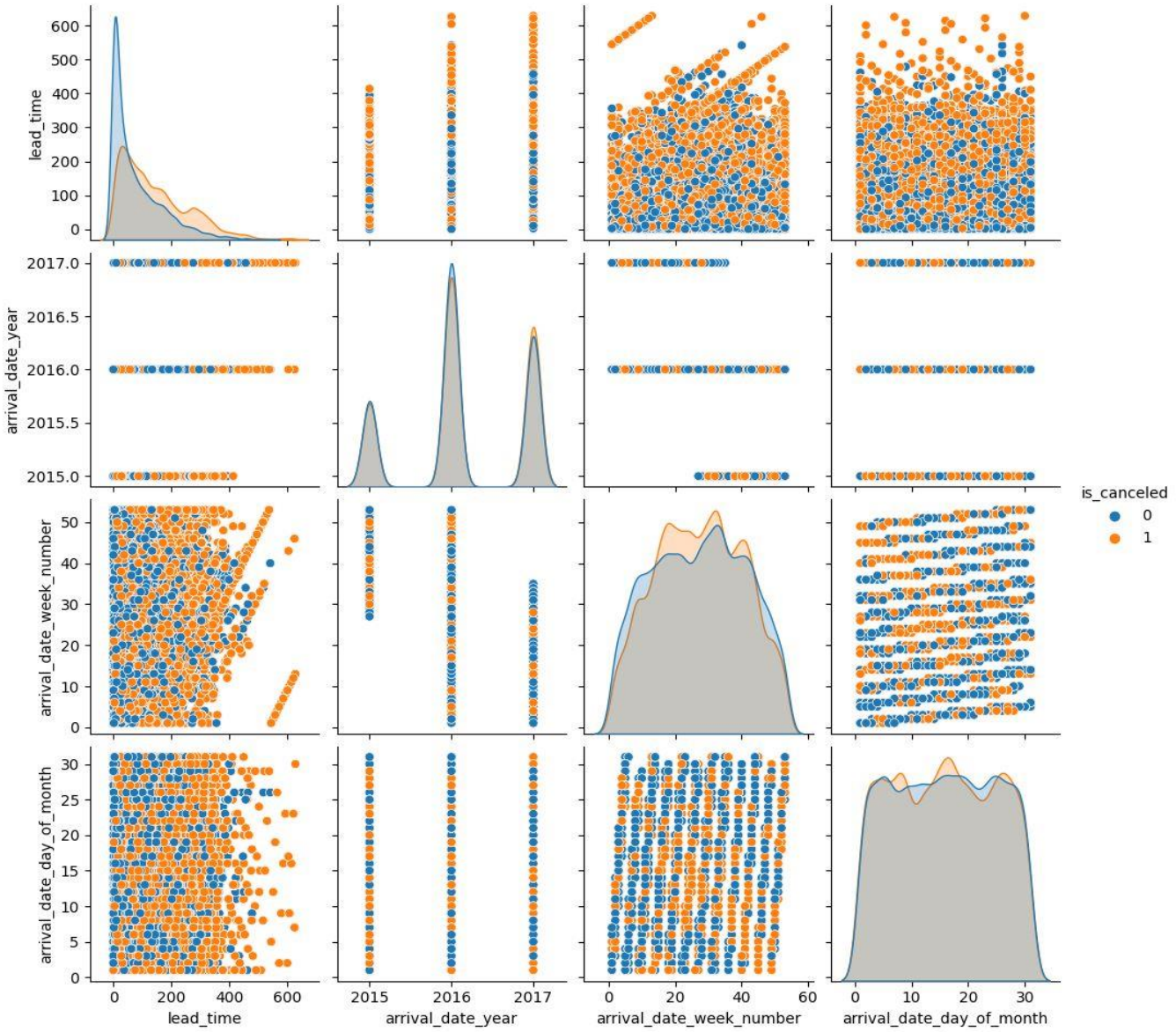

gráfico 2: Gráfico de a pares entre las variables lead_time y arrival year, week number y day_of_month

A partir del análisis encontramos que, entre las variables cuantitativas, las que mayor correlación mantienen con el target (is_canceled) son lead_time, total_of_special_requests y required_car_parking_spaces.
Por ello, decidimos mostrar los graficos de distribución de estas variables.

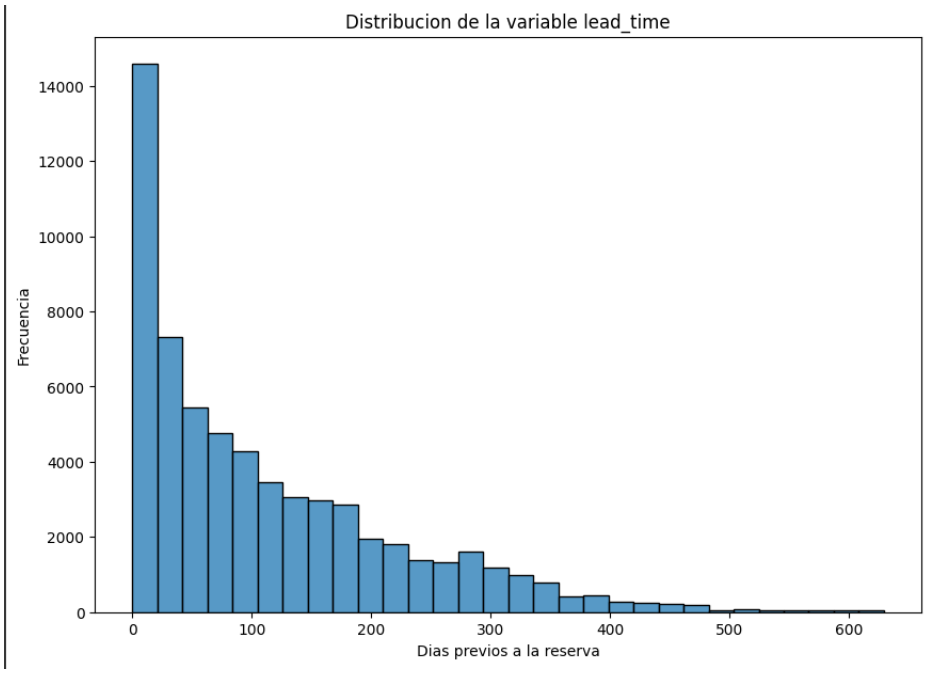

gráfico 3: Gráfico de distribución de la variable lead_time

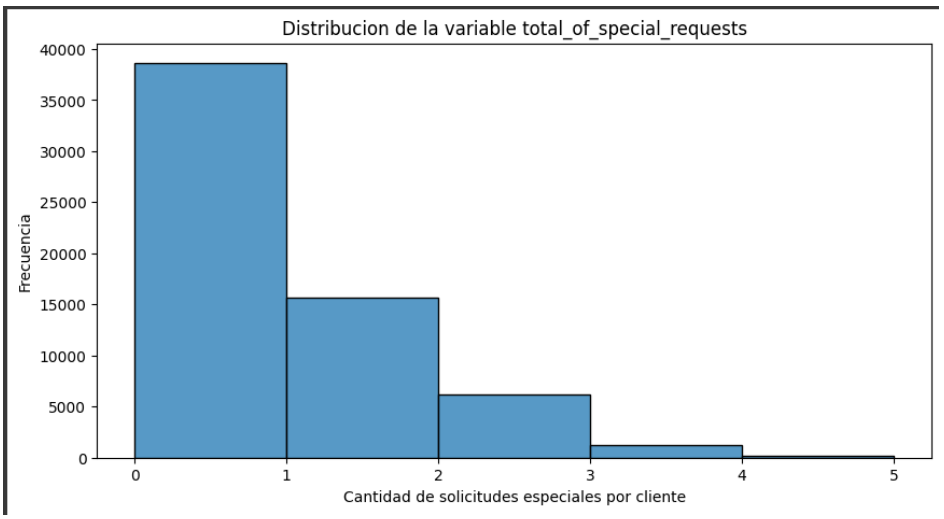

gráfico 4: Gráfico de distribución de la variable total_of_special_request

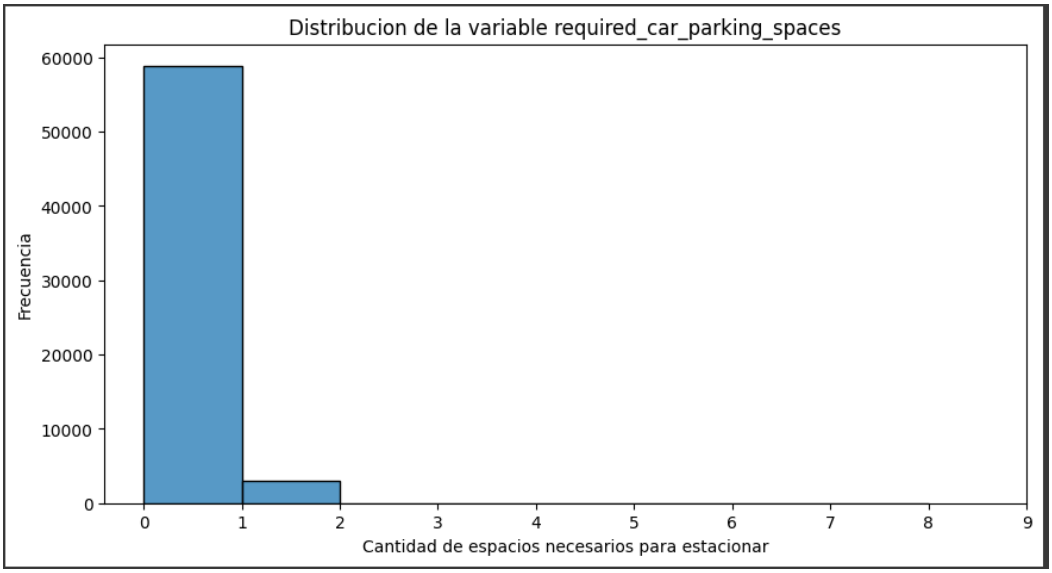

gráfico 5: Gráfico de distribución de la variable required_car_parking_spaces

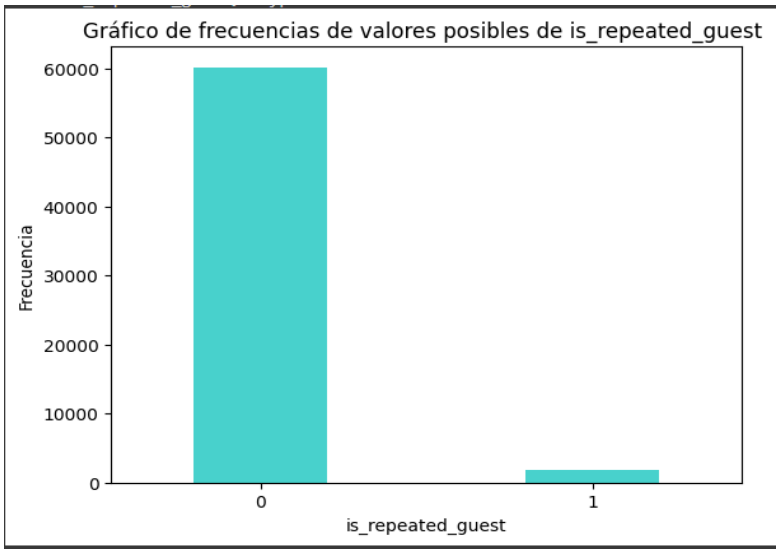

gráfico 6: Gráfico de frecuencia de la variable is_repeated_guest

Se observa que la mayoría de los huéspedes son nuevos en el hotel, lo que nos hace pensar que una posibilidad es que el servicio recibido no es muy bueno y por eso no quieren repetir su estadía en los establecimientos.

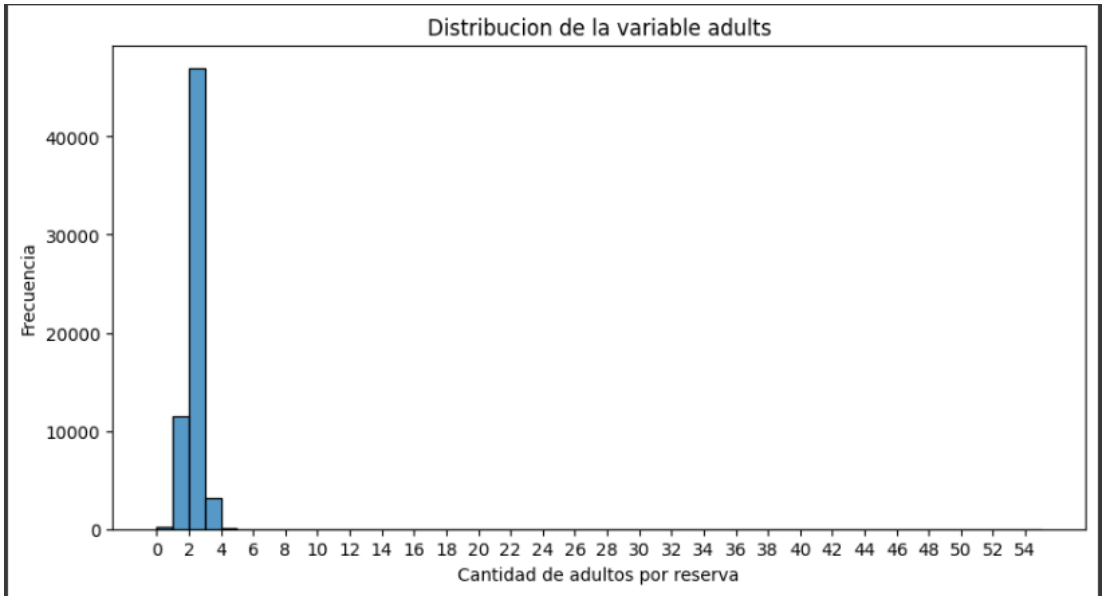

gráfico 7: Gráfico de frecuencia de la variable adults

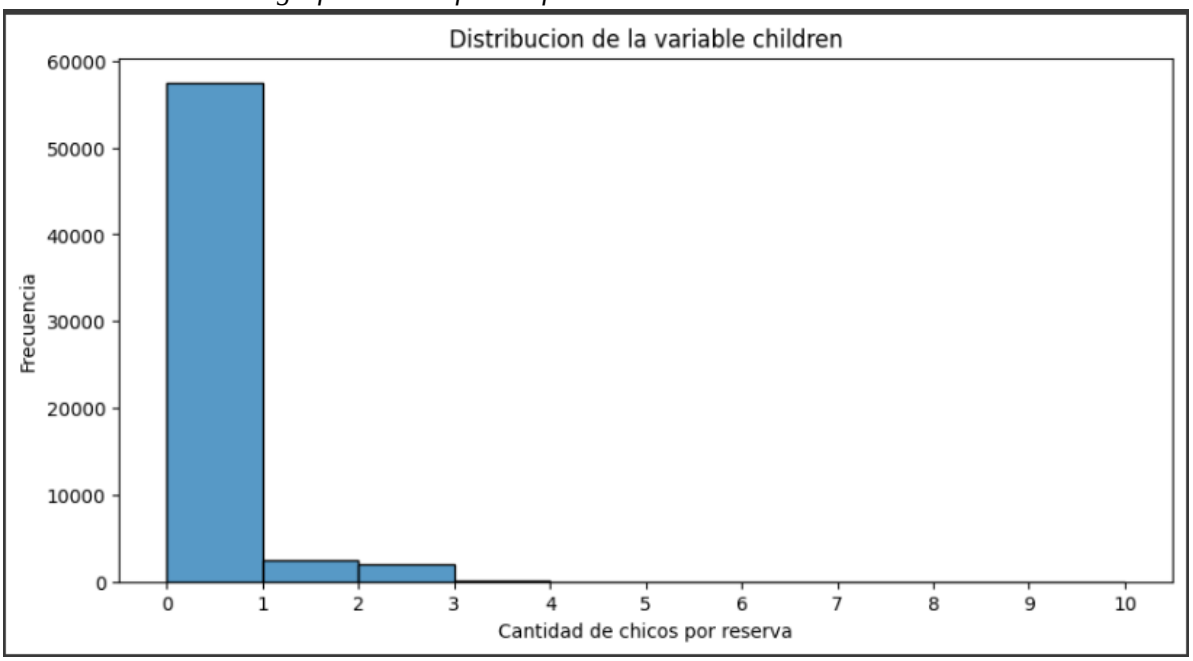

gráfico 8: Gráfico de frecuencia de la variable children

Analizando los graficos 7 y 8, llegamos a la conclusión que de que la mayoría de los huéspedes son adultos solos y que hay muy pocas reservas hechas por familias con 1 o 2 chicos.

Checkpoint 2:

En este checkpoint se realizó arboles de decisión con y sin poda. Se separo el dataFrame en test y train con proporción 70/30. Se realizo un análisis para encontrar la mejor métrica y poder armar los arboles con los mejores valores hiperparametros. Se hizo una predicción sobre el set de evaluación y luego se calcularon las métricas en el conjunto de evaluación. Comparamos la performance del árbol en el conjunto de train vs conjunto test. Repetimos el proceso con el árbol con poda. Se realizo una comparación entre ambos y se llego a la conclusión de que el árbol sin podar tiene mejor métrica que el árbol podado. Por último, se realizó una matriz de confusión (grafico 9).
Para finalizar se generaron el archivo submit.csv con las predicciones del archivo hotels_test.csv y un archivo Arbol_de_desicion_sin_poda.joblib con el modelo.

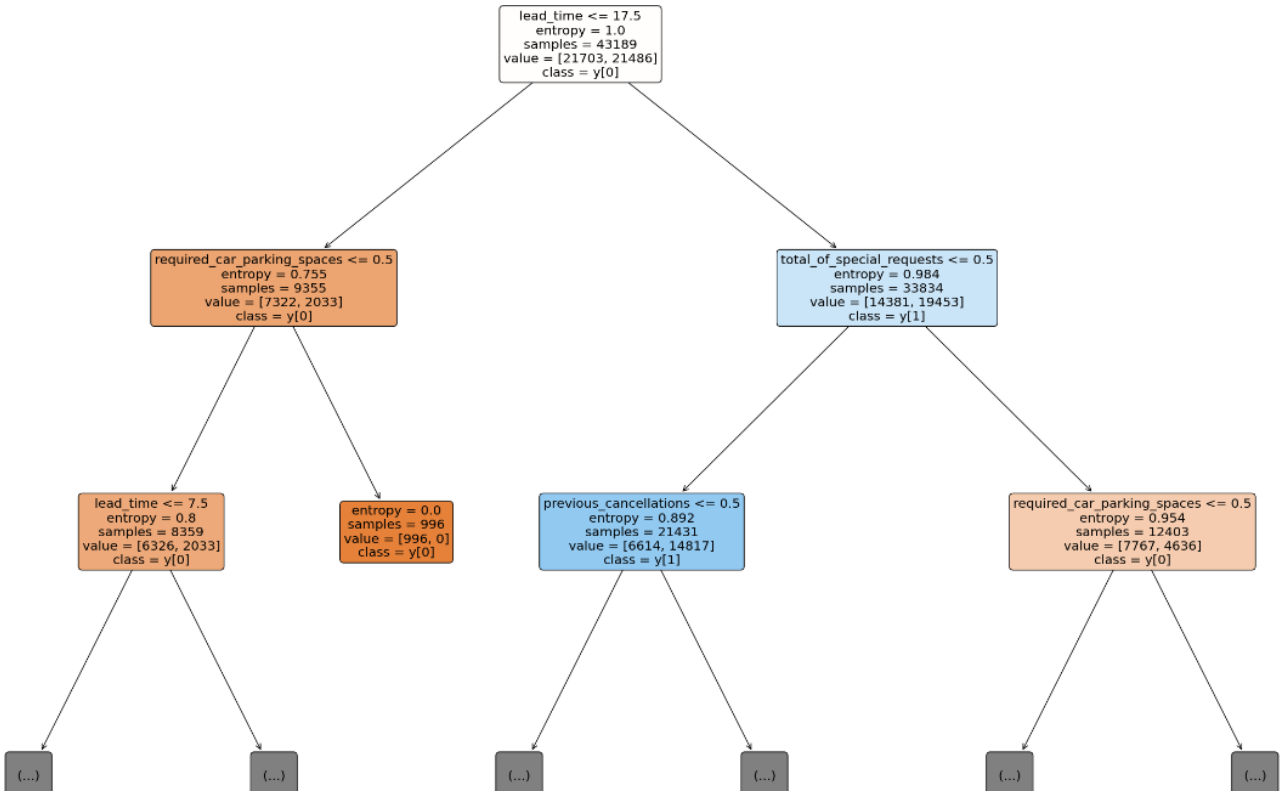

gráfico 9: Gráfico árbol sin poda versión reducida

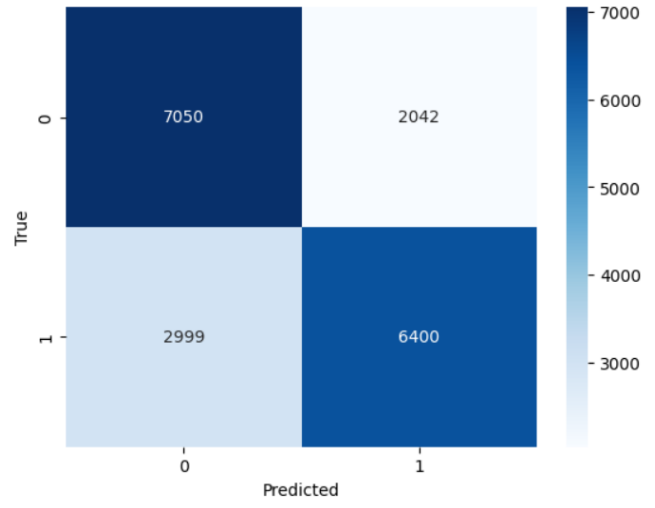

gráfico 9: Gráfico matriz de confusión