
Materia: Introducción a Ciencia de Datos para Economía y Negocios
Profesor: Franco Mastelli
Alumno: Renzo Falciglia
Fecha: 24/10/2023
Mail institucional: 89FA34985709@campus.economicas.uba.ar
Mail personal: renzo.falciglia@gmail.com

Trabajo Práctico Final

ÍNDICE

Punto 1 - Análisis exploratorio

Inciso A

Los datos utilizados en este trabajo, que surgen de la página Kaggle, se refieren a la demanda de reservas de hoteles. Específicamente se trata de un dataset de un total de 32 columnas, con 119.390 observaciones. Cabe destacar que la información se presenta a nivel turista (sin datos personales). Hay tanto columnas en formato numérico, como las que tienen el formato "int64", como columnas con categorías.

Algunas de las columnas más importantes a los efectos de evaluar los modelos de este trabajo son:

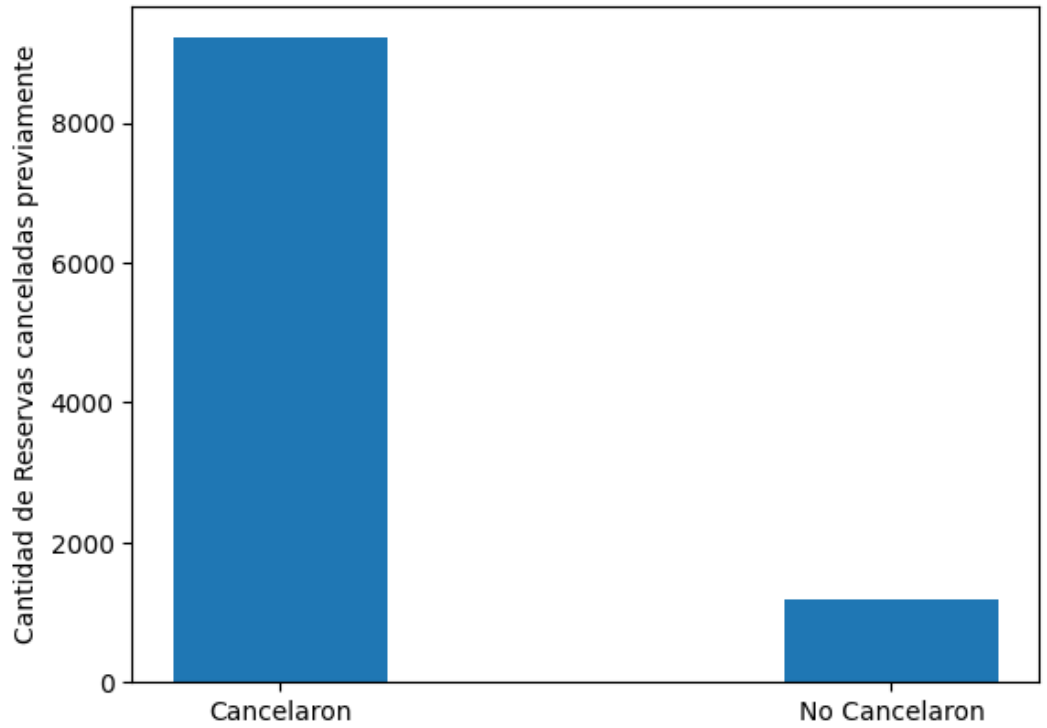
- **is_canceled:** hace referencia a si la reserva fue cancelada o no. Esta columna es importante ya que permite estimar los modelos de los incisos siguientes.
- **lead_time:** hace referencia al tiempo que transcurre entre que se realiza la reserva y se realiza el viaje. Es una variable importante en el sector turístico.
- **arrival_date:** son un total de 4 columnas donde se presenta el año, mes, semana y día de arribo de los turistas.
- **stays:** con 2 columnas se presenta la información del tiempo de estadía, ya sea en noches del fin de semana o en noches durante la semana.
- **market_segment:** hace referencia al segmento de mercado de los turistas que realizan las reservas. Puede ser por ejemplo corporativo, directo, en grupos, etc.
- **distributtion_channel:** hace referencia al canal de distribución por el cuál se realizó la reserva. Puede ser corporativo, directo, etc.
- **previous:** a partir de 2 columnas se presenta la información sobre cancelaciones previas y sobre no cancelaciones previas de cada cliente.
- **deposite_type:** hace referencia al tipo de depósito que realizaron los turistas, por ejemplo depósito reembolsable, no reembolsable, etc.
- **days_in_waiting_list:** hace referencia al tiempo de espera que tuvo el turista para contratar su reserva.
- **customer_type:** hace referencia al tipo de cliente.

A su vez, también se presentan algunas columnas que hacen referencia a las características de los turistas, por ejemplo si tienen hijos o bebés, si es un turista que repite la reserva, entre otras.

Inciso B

En cuanto a la correlación de las variables numéricas que tiene el dataset, puede diferenciarse los grupos de turistas de acuerdo a si canceló o no la última reserva y observar la cantidad de cancelaciones previas. El gráfico a continuación muestra estos datos.

Gráfico 1 - Cantidad de cancelaciones previas, por tipo de cliente (canceló o no)
En cantidad de cancelaciones previas

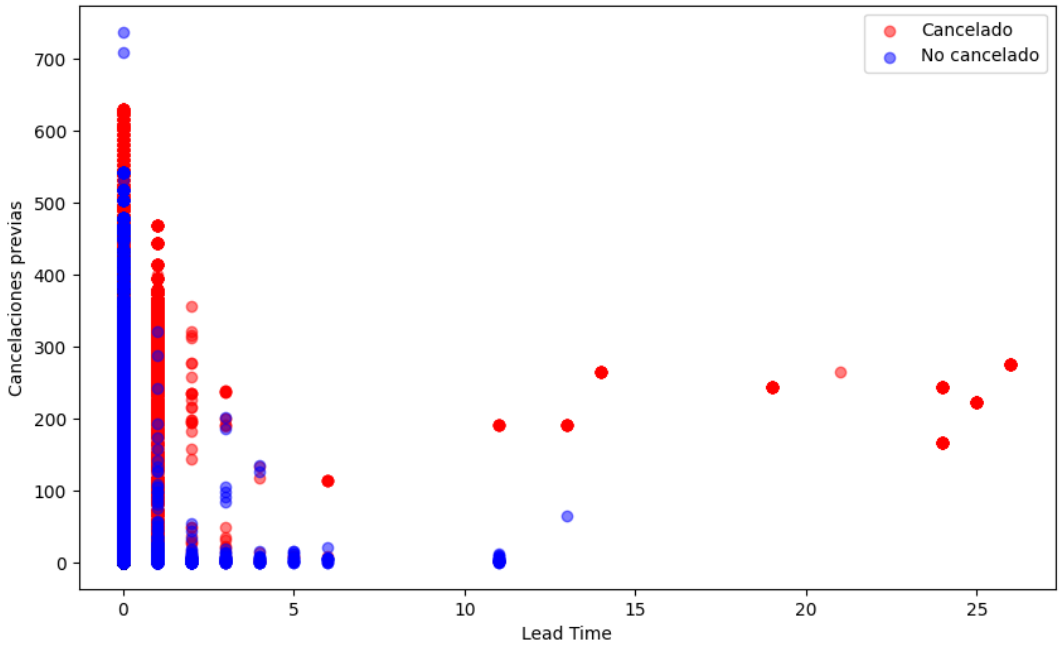


Fuente: elaboración propia en base a "hotel booking demand dataset"

Cómo se observa, **el grupo de turistas que cancelaron su última reserva tienen una mayor cantidad de cancelaciones previas que los que no la cancelaron**. Esto podría indicar que es más probable que cancele una reserva alguien que ya se caracteriza por haber hecho cancelaciones previas.

También puede analizarse la correlación entre el tiempo entre la reserva (lead time) y el viaje, y la cantidad de cancelaciones previas, para ambos grupos (cancelaron la última reserva y no cancelaron la última reserva). El gráfico a continuación muestra estos resultados.

Gráfico 2 - Correlación entre lead time y cancelaciones previas

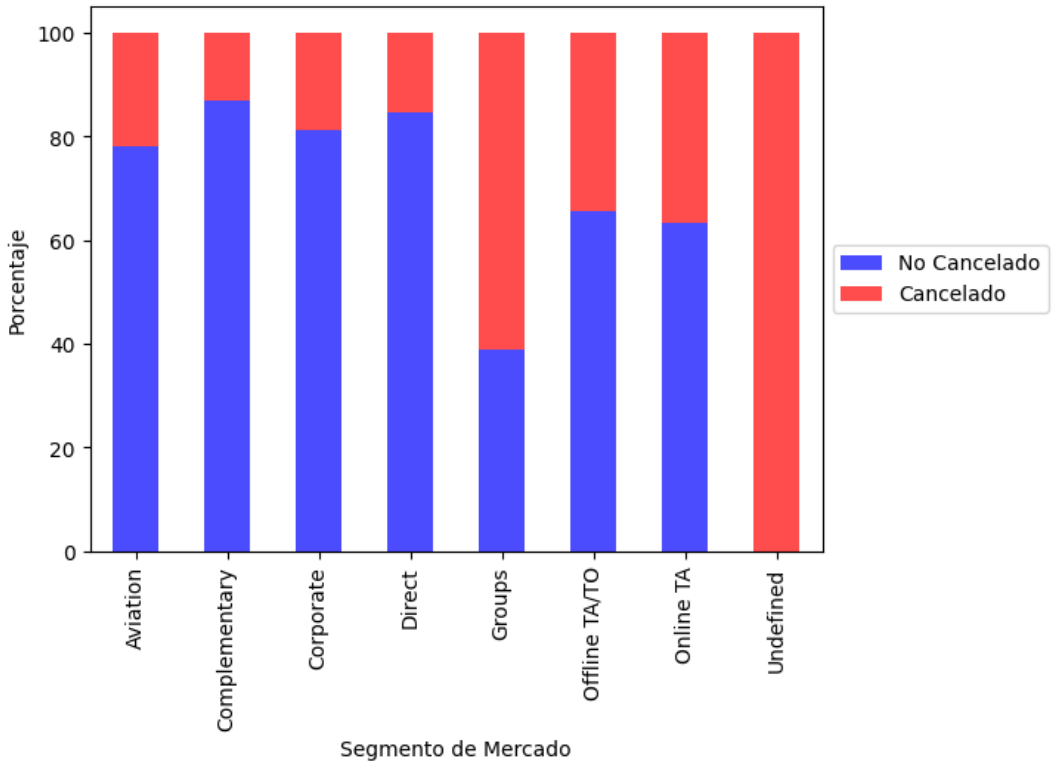


Fuente: elaboración propia en base a "hotel booking demand dataset"

En este gráfico se muestran distintos puntos de acuerdo al tipo de turistas (cancelaron su última reserva en rojo, no cancelaron su última reserva en azul). Lo que se observa es que **en general los turistas con un mayor lead time, es decir, que sacaron la reserva con mayor anticipación, son más proclives a cancelar las reservas** (puntos rojos a la derecha). En tanto, los que mayores cancelaciones registran son también los que cancelaron su última reserva (se puede aplicar el mismo razonamiento que en el gráfico 1), y son también los que registran un mayor plazo entre la reserva y el viaje (mayor lead time).

Por último, se puede observar como es la proporción de cancelaciones por segmento de mercado, para analizar si hay un segmento específico que registra una mayor proporción de cancelaciones con respecto al resto de los segmentos. En el gráfico a continuación se muestra esta relación.

Gráfico 3 - Proporción de cancelaciones por segmento de mercado
En proporción sobre el total



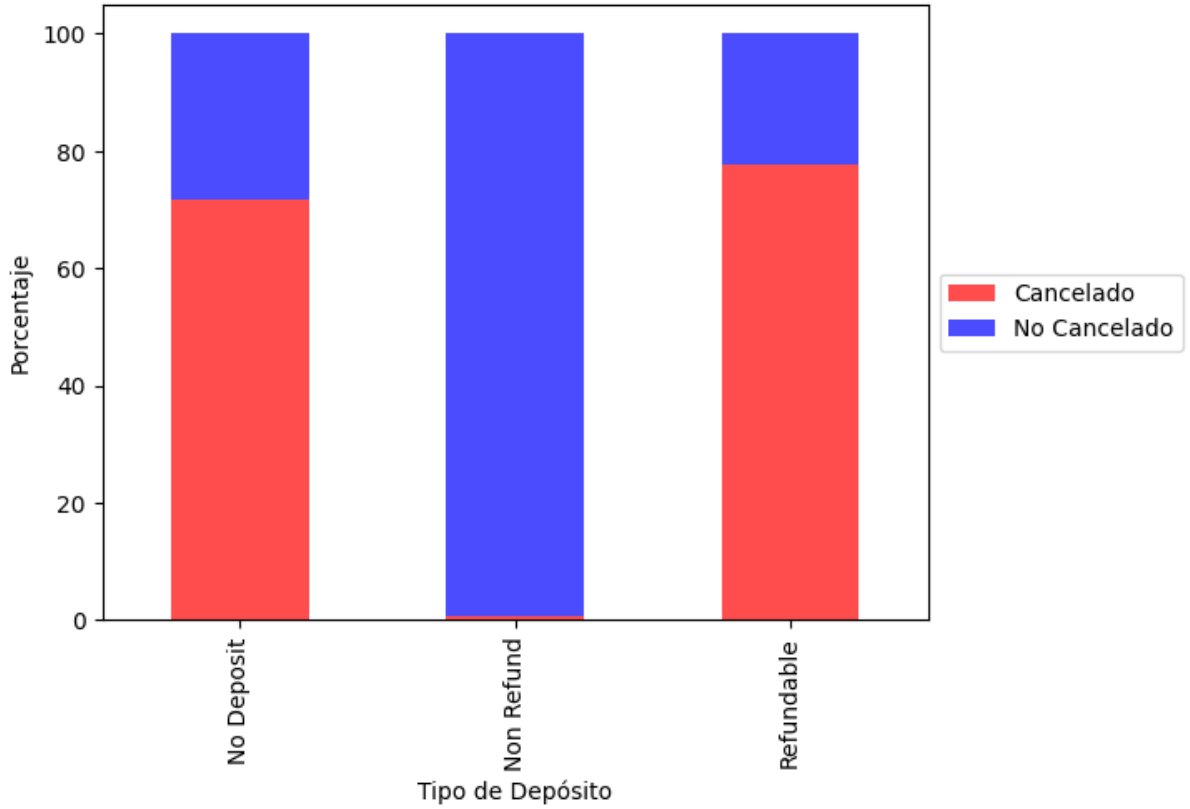
Fuente: elaboración propia en base a "hotel booking demand dataset"

Lo que se observa es que **el segmento que tiene una clara proporción mayor de cancelaciones, con más del 50%, es el segmento de "Grupos"**. Esto podría explicarse porque al viajar en grupo, es más probable que distintos miembros del grupo tengan algún problema que les impida viajar, provocando la cancelación completa del viaje.

Inciso C

Se presenta a continuación el mismo gráfico de proporciones que en el inciso anterior, pero esta vez desagregando por tipo de depósito.

Gráfico 4 - Proporción de cancelaciones por tipo de depósito
En proporción sobre el total



Fuente: elaboración propia en base a "hotel booking demand dataset"

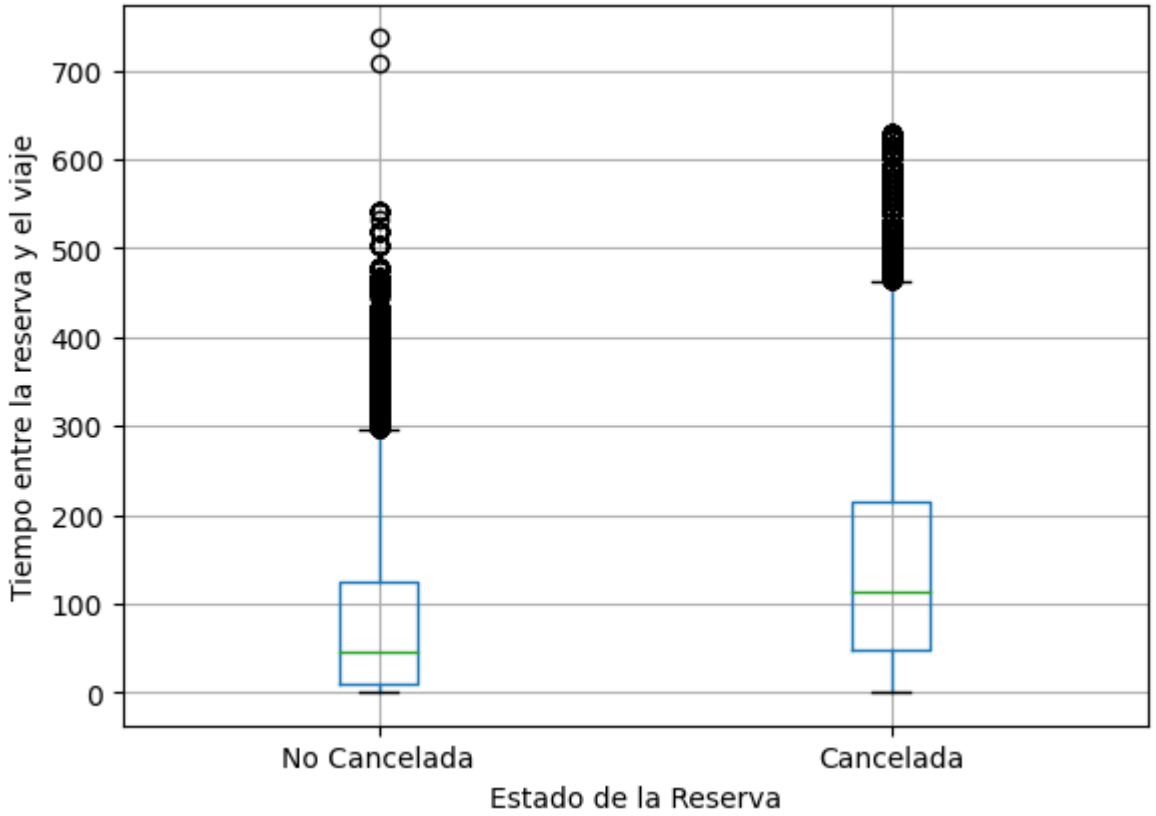
Lo que se observa gráficamente es un resultado esperable, ya que **la mayor proporción de cancelaciones se da cuando no hay depósito realizado o cuando el depósito es reembolsable**, mientras que una proporción mínima de cancelaciones se da en los casos de depósitos no reembolsables. Esto es lógico ya que los turistas al cancelar un depósito no reembolsable pierden dinero, por lo que tratarán de evitarlo. Cuando la cancelación no implica ningún costo para los turistas, están más predispuestos a realizarla.

También se efectuó un test de *chi-cuadrado* para evaluar la relación entre estas dos variables, lo que arrojó un p valor de 0, lo que implica que la relación entre el tipo de depósito y las cancelaciones es estadísticamente significativa.

Inciso D

A continuación se presenta un boxplot de la relación entre el tiempo entre la reserva y el viaje (lead time) y si la reserva fue cancelada o no. La ventaja de este tipo de gráficos es que aportan una buena cantidad de información sobre las variables, al mostrar el rango intercuartílico, la mediana y los valores atípicos.

Gráfico 5 - Boxplot de la relación entre lead time y cancelaciones



Fuente: elaboración propia en base a "hotel booking demand dataset"

Cómo se observa en el gráfico, **en el caso de las reservas no canceladas (caja de la izquierda), tanto el rango intercuartílico como la mediana y los bigotes de la caja presentan un lead time menor que en los casos de las reservas canceladas.** También en el caso de las reservas no canceladas, se presentan mayor cantidad de outliers. En tanto, las reservas que fueron canceladas presentan un rango intercuartílico y una mediana superior al otro grupo. Esta diferencia entre los dos grupos sugiere que las reservas canceladas son en general las que se realizan con un lead time mayor.

La explicación de estos resultados puede deberse a la mayor probabilidad de que les ocurran imprevistos a los turistas que les imposibiliten viajar. Cuando mayor es el tiempo entre la reserva y el viaje, mayor probabilidad de imprevistos y mayor probabilidad de tener que cancelar la reserva.

Punto 2 - Partición de datos

Inciso A

Se realiza una partición estratificada del dataset, entre un conjunto de entrenamiento (80%) y uno de test (20%), usando como semilla el número de documento de los turistas. Los resultados que arroja dicho test son los siguientes:

- **Forma de X_train (95.512, 31):** partición de entrenamiento de las características o variables explicativas. Representa el 80% del total y contiene 31 columnas.
- **Forma de X_test (23.878, 31):** partición de prueba de las características o variables explicativas. Representa el 20% del total y contiene también 31 columnas.
- **Forma de y_train (95.512):** partición de entrenamiento de la variable dependiente, representa el 80% del total.
- **Forma de y_test: (23.878):** partición de prueba de la variable dependiente, representa el 20% del total.

Inciso B

Se crearon también dos listas que contienen los atributos que se utilizarán para explicar el comportamiento de la variable dependiente. Estas dos listas tienen los siguientes elementos:

- **Atributos categóricos:** “segmento de mercado”, “canal de distribución”, “si es cliente frecuente”, “tipo de depósito”, “tipo de cliente”.
- **Atributos numéricos:** “tiempo entre reserva y viaje (lead time)”, “estadía en semanas”, “cantidad de niños”, “cantidad de cancelaciones previas”, “cantidad de reservas previas no canceladas”, “días en la lista de espera”.

Estos dos listados de atributos son los que se utilizan para predecir, en los modelos de los puntos siguientes, si un turista ha cancelado o no su reserva. Se seleccionaron específicamente estos atributos por considerarse de mayor interés. Por ejemplo, es esperable que el tipo de depósito o el tipo de cliente afecten la decisión de cancelar o no una reserva, así como el lead time, la estadía y las cancelaciones previas. Del análisis exploratorio del punto 1 surgen algunas de las ideas que se toman en este apartado para la construcción de las listas.

Punto 3 - Modelos

Inciso A

Se realizó un árbol de clasificación usando únicamente los atributos categóricos descritos en el punto anterior. Debido a la naturaleza categórica de los datos, fue necesario aplicarle una transformación *one-hot-encoding* para poder realizar el procesamiento.

Una vez confeccionado los distintos grupos (entrenamiento y control), para los atributos categóricos y la variable que se desea predecir (si canceló o no la reserva), se aplicaron las transformaciones y se entrenó el modelo.

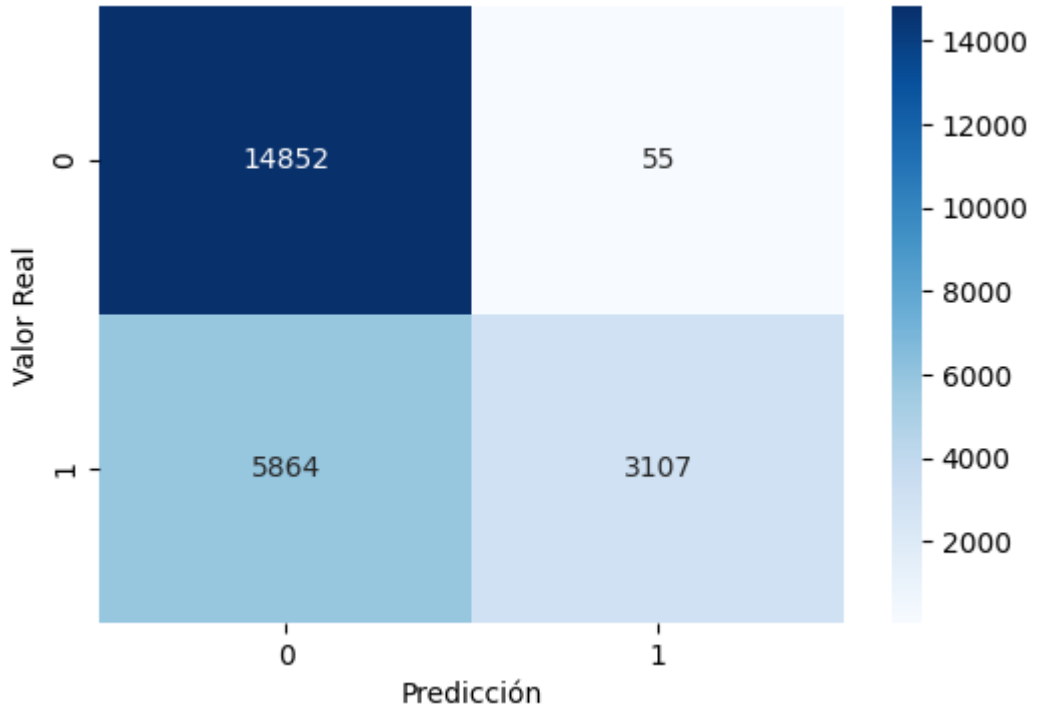
Inciso B

Una vez entrenado el modelo, se calcularon una serie de medidas de precisión para analizar los resultados. Las medidas de precisión arrojan lo siguiente:

- **Precisión del modelo (Accuracy):** 0.75
- **Precisión:** 0.98
- **Recall:** 0.35
- **F1-Score:** 0.51

A su vez, la matriz de confusión arroja los siguientes resultados:

Gráfico 6 - Matriz de confusión



Fuente: elaboración propia en base a "hotel booking demand dataset"

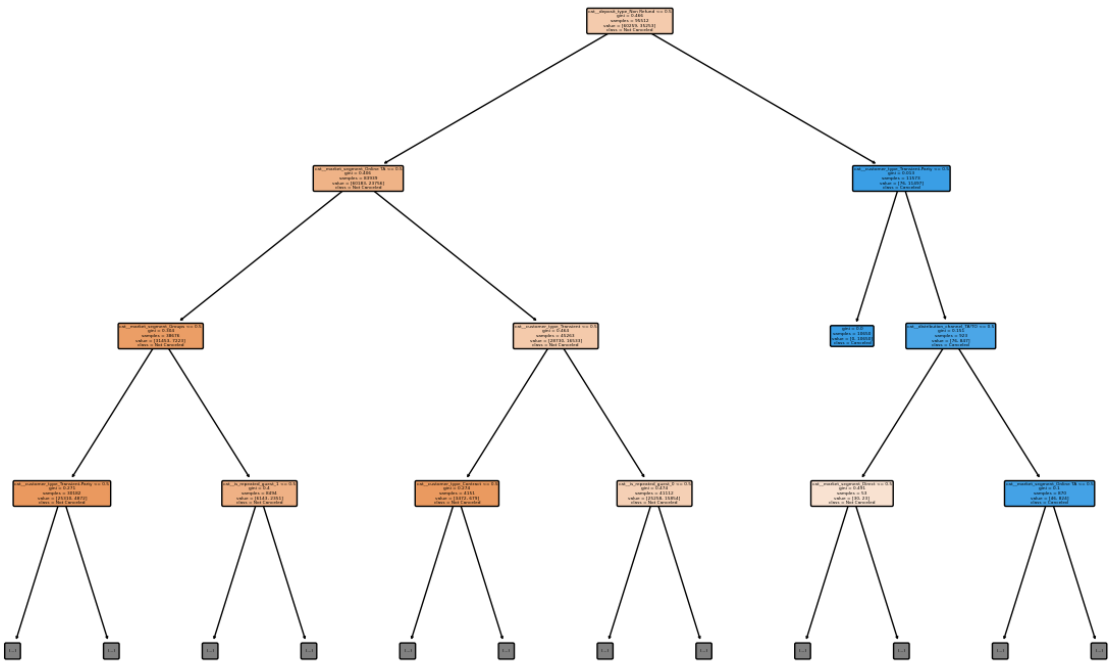
Para evaluar los resultados del modelo pueden utilizarse las distintas métricas encontradas. Sin embargo, el F1 score arroja información acerca de la precisión (proporción de predicciones positivas correctas (TP) con respecto a todas las predicciones positivas) y el recall (proporción de verdaderos positivos (TP) con respecto a todos los casos positivos reales). **En este caso, al arrojar un valor de 51%, puede decirse que el modelo tiene un adecuado equilibrio entre precisión y recall**. A su vez, si se evalúa la precisión total del modelo o Accuracy, **se observa que el modelo sirve para predecir correctamente en el 75% de los casos.**

Inciso C

En este inciso se realiza una búsqueda de hiper-parámetros mediante *cross-validation* (5 folds) y *grid-search*, para el árbol de clasificación elaborado en los incisos anteriores. Para esto se utiliza el set de entrenamiento.

La estructura del árbol que surge como óptimo luego de este ajuste de hiper-parámetros es la siguiente:

Esquema 1 - Árbol de decisión ajustado por hiper-parámetros (atributos categóricos)



Fuente: elaboración propia en base a "hotel booking demand dataset"

A su vez, si se calcula la Precisión promedio en *5-fold cross-validation*, el resultado es de 75,3%.

Inciso D

Se vuelve a realizar un árbol de clasificación, pero esta vez utilizando tanto los atributos categóricos como los atributos numéricos. A los datos categóricos se le aplicó la misma transformación de *one-hot-encoding* para poder realizar el procesamiento.

En el caso de los atributos numéricos, se aplicó la transformación *Standard-Scaler*, que es una técnica común para preparar los datos numéricos en modelos de aprendizaje automático. Lo que hace esta técnica es estandarizar las características numéricas, de tal manera que los datos tengan una media de 0 y una desviación estándar de 1. Este procedimiento ayuda a que los datos tengan una distribución normal y facilita el procesamiento.

Una vez realizadas las transformaciones, se siguen los pasos del inciso A y B, y se calculan las siguientes métricas de precisión:

- **Precisión del modelo (Accuracy):** 0.78
- **Precisión:** 0.79
- **Recall:** 0.57
- **F1-Score:** 0.66

Cómo se observa en estos resultados, la precisión del modelo (accuracy) se eleva levemente hasta un 78%, en tanto la precisión baja al 79% y la Recall y el F1 Score se incrementan hasta 57% y 66% respectivamente. Estos resultados sugieren que este modelo, que tiene en cuenta los atributos categóricos y los numéricos, aún sin optimizar por hiper-parámetros, arroja una mejor performance que el modelo del inciso 1. En el inciso siguiente se ajusta el modelo y se lo compara con el inciso C.

Inciso E

En este inciso se realiza entonces la optimización del modelo que contempla los parámetros categóricos y numéricos, a través de una búsqueda de hiper-parámetros mediante *cross-validation* (5 folds) y *grid-search*.

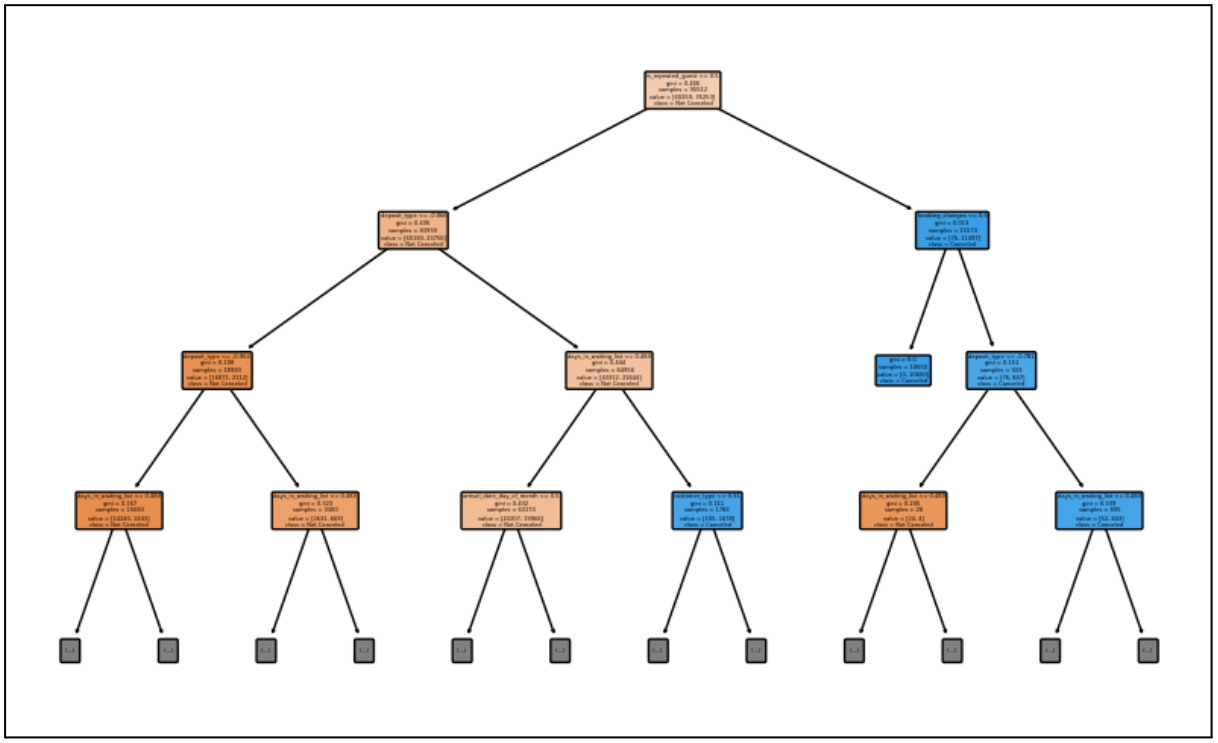
En este caso, la precisión promedio en 5-fold *cross-validation* es de un 77,5%, lo cuál mejora en dos puntos porcentuales la performance del modelo anterior. En el siguiente inciso se muestra la estructura de este modelo optimizado.

Punto 4 - Importancia de features y testing

Inciso A

Se presenta la estructura del mejor árbol de decisión, optimizado por la búsqueda de hiper-parámetros, tal como se detallo en el inciso anterior:

Esquema 2 - Árbol de decisión ajustado por hiper-parámetros (atributos categóricos y numéricos)



Fuente: elaboración propia en base a "hotel booking demand dataset"

Inciso B

La importancia de las features del árbol se calcula para determinar en qué medida cada característica tiene importancia a la hora del rendimiento del árbol. Cuanto mayor sea la importancia de una característica, más influyente es en las decisiones del árbol. A continuación se presenta una tabla con las características más importantes, junto a su importancia.

Tabla 1 - Importancia de las features del árbol

#	Feature	Importancia
1	market_segment	0.103641
2	lead_time	0.000527
3	is_repeated_guest	0.000342
4	children	0.000293
5	previous_cancellations	0.000027

Fuente: elaboración propia en base a "hotel booking demand dataset"

De esta manera, **las características más importantes para las decisiones del árbol son el segmento de mercado (categórica), el lead time (numérica), si es un turista repetido (categórica), la cantidad de hijos (numérica) y las cancelaciones previas (numéricas)**. Sin embargo, dentro de estas características la que destaca es el segmento de mercado, con una importancia de 0,10, muy superior al resto.

Inciso C

Finalmente, se entrena el árbol de decisión con todo el conjunto de datos de entrenamiento y se evalúa su performance. Cabe destacar que el árbol que se está evaluando es el árbol que contiene tanto los atributos categóricos como los atributos numéricos y que está optimizado con la búsqueda de hiper-parámetros. Los resultados encontrados son los siguientes:

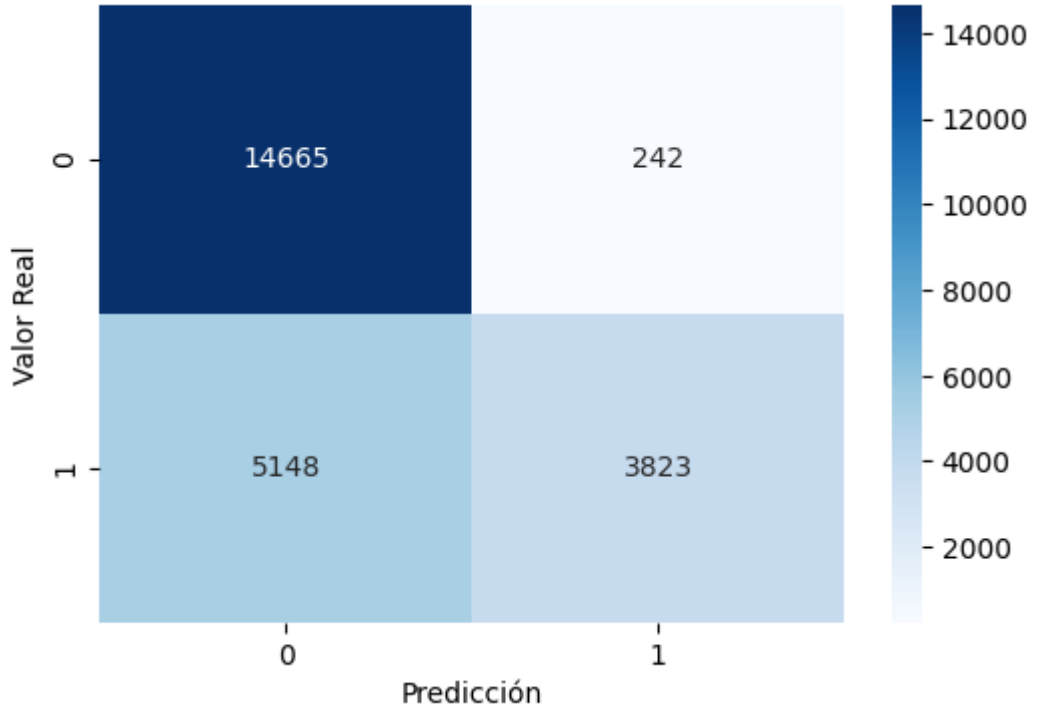
- Precisión (Accuracy): 0.77
- Recall: 0.43
- Precisión: 0.94
- F1-Score: 0.59

Estos resultados indican que el modelo tiene una precisión (Accuracy) del 0,77, lo que implica que **es capaz de clasificar correctamente cada observación en el 77% de los caso**s. A su vez, la capacidad que tiene el modelo de interpretar casos positivos reales (capacidad de estimar una cancelación) es de un 43% dado el resultado del recall. **Este resultado podría indicar que el modelo tiene cierta dificultad para clasificar correctamente algunas cancelaciones.**

También puede decirse que de las observaciones que el modelo clasifica como positivas (cancelaciones), en el 94% de los casos son efectivamente cancelaciones reales, de acuerdo

a la métrica de precisión. **Esto indica que si el modelo clasifica una observación como una cancelación, en el 94% de los casos resulta ser efectivamente una cancelación.** Por último, el F1-Score arroja un cierto equilibrio entre la precisión y el recall. También se presenta a continuación la matriz de confusión.

Gráfico 7 - Matriz de confusión



Fuente: elaboración propia en base a "hotel booking demand dataset"

En el caso del cuadrante de la parte superior izquierda, se presentan los resultados de los "verdaderos negativos", lo que implica que el modelo fue capaz de predecir en 14.665 casos que el cliente no cancelaría y efectivamente no canceló. El cuadrante de abajo a la derecha implica los "verdaderos positivos", e implica que el modelo fue capaz de predecir correctamente la cancelación de 3.823 casos.

Los otros dos cuadrantes recogen los datos de cuando el modelo no predice correctamente. El cuadrante de arriba a la derecha arroja los "falsos positivos", que implican que el modelo estimó una cancelación en 242 casos de clientes que no cancelaron. En tanto el cuadrante de abajo a la izquierda detecta los "falsos negativos", que implican que el modelo estimó para 5.148 que los clientes no cancelarían una reserva cuando en realidad lo hicieron. **Estos resultados sugieren que el modelo debería mejorar en su capacidad de reconocer las cancelaciones, para que se reduzcan los casos de falsos negativos.**

Punto 5 - Ensambles

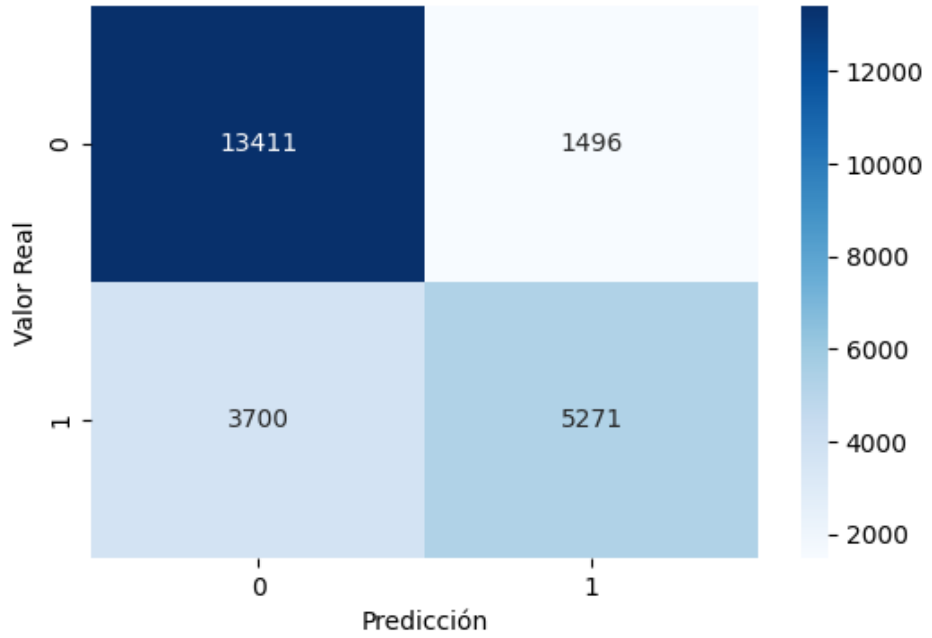
Inciso A

En este último inciso se calcula un modelo de Random Forest (sin buscar hiper-parámetros), sobre los datos de entrenamiento y se evalúa su performance.

- Precisión (Accuracy): 0.78
- Recall (Random Forest): 0.59
- Precisión: 0.78
- F1-Score: 0.67

El modelo de Random Forest arroja entonces una Accuracy de 78%, lo que sugiere que en general es un modelo eficiente para predecir los resultados. Esto implica que la precisión de este modelo es apenas superior al modelo entrenado en el punto 4, que tenía una Accuracy del 77%. En cuanto al recall, en este modelo de Random Forest se llega a un 59%, presentando un resultado superior al del modelo del inciso 4, que tenía un recall de 43%. **Esto implica que el modelo de Random Forest es mejor para identificar casos reales de cancelaciones.** En cuanto a la precisión, en el modelo de Random Forest se alcanza un 78%, lo que es relativamente bajo si se lo compara con el modelo del punto 4, que arrojaba una precisión del 94%. **Esto implica que el modelo de Random Forest tiene mayores problemas al predecir ciertas cancelaciones que luego no se cancelan**. Por último, el F1-Score arroja un resultado equilibrado de 59% entre recall y precisión. Se presenta a continuación la matriz de confusión.

Gráfico 8 - Matriz de confusión



Fuente: elaboración propia en base a "hotel booking demand dataset"

En el caso del cuadrante de la parte superior izquierda, se presentan los resultados de los "verdaderos negativos", lo que implica que el modelo fue capaz de predecir en 13.411 casos que el cliente no cancelaría y efectivamente no canceló. Este resultado es inferior al anterior modelo, que tenía un resultado de 14.665 en dicho cuadrante.

El cuadrante de abajo a la derecha se refiere a los "verdaderos positivos", e implica que el modelo de Random Forest fue capaz de predecir correctamente la cancelación de 5.271 reservas, lo que es superior a los 3.823 casos predecidos correctamente por el anterior modelo.

Los otros dos cuadrantes recogen los datos de cuando el modelo no predice correctamente. El cuadrante de arriba a la derecha arroja los "falsos positivos", que implican que el modelo estimó una cancelación en 1.496 casos de clientes que no cancelaron (el modelo anterior arroja un resultado más bajo y de sólo 242 casos). En tanto el cuadrante de abajo a la izquierda detecta los "falsos negativos", que implican que el modelo estimó para 3.700 casos que los clientes no cancelarían una reserva cuando en realidad lo hicieron, mientras en el modelo anterior se trataba de 5.148 casos.

En síntesis, **el modelo de Random Forest es menos eficiente para detectar casos de no cancelaciones (cuadrante de arriba a la izquierda), a expensas de estimar una mayor cantidad de efectivas cancelaciones (cuadrante de abajo a la derecha)**. **También es más**

eficiente a la hora de detectar falsos negativos, ya que el número de casos que detecta de no cancelaciones que luego se cancelaron es menor al número de casos del modelo anterior (cuadrante de abajo a la izquierda). Esto se da a expensas de estimar una mayor cantidad de falsos positivos, es decir, de casos en que predice cancelaciones cuando luego no las hay.

Este tipo de modelos, si bien arrojan una buena performance en términos de predicciones, podrían mejorarse a partir de la incorporación de nuevos atributos y de nueva información que no contiene el dataset utilizado.