

## Entrega 2



LAURA CECILIA TOBÓN OSPINA

JUAN DAVID ARISMENDY PULGARÍN

Tutor: RAUL RAMOS POLLAN

Introducción Inteligencia Artificial 2023-1

Universidad de Antioquia

Medellín

2023

# 1. Planteamiento del problema

En la actualidad el comercio hotelero se ha venido manteniendo pese a las adversidades vividas por este sector en el período del COVID-19, implementando nuevas estrategia de servicio al cliente e innovando en tipos de negocios aleatorios relacionados como catering y planes cortos de fin de semana para locales[1]. Una de las estrategias que genera confianza ha sido permitirle al cliente cancelar una reserva con antelación a su visita sin tener que justificarlo y con un tiempo prudente de anticipación, lo cual dependerá del hotel segun sus politicas ya que pueden ser cancelaciones flexibles, moderadas o firmes, esta última puede incluir hasta reembolsos para el cliente [3]. Todos los metodos para conquistar y atraer huéspedes tienen un impacto económico en los hoteles, pero las cancelaciones y habitaciones que no son ocupadas puede acarrear costes que deben ser compensadas con las reservas que si son efectivas, razón por la cual tratar de implementar un modelo que ayude a la predicción de las posibles cancelaciones ayudaría a diseñar estrategias de amortización en estas circunstancias.

## 2. Dataset

El dataset a utilizar proviene de una competencia de kaggle en la cual se proporcionan datos que comparan información de varios booking comparada entre dos hoteles “ a city hotel and a resort hotel.” [1]

Un 34% de la información proveniente del resort y un 66% proveniente del hotel en la ciudad

hotel - (resort o ciudad)

is\_canceled (indica si la reserva fue cancelada o no)

lead\_time (días transcurridos entre la reserva y la fecha de llegada al hotel)

arrival\_date\_year

arrival\_date\_month

arrival\_date\_week\_number

arrival\_date\_day\_of\_month

stays\_in\_weekend\_nights

stays\_in\_week\_nights

adults

children

babies

meal

country

market\_segment

distribution\_channel  
is\_repeated\_guest  
previous\_cancellations  
previous\_bookings\_not\_canceled  
reserved\_room\_type  
assigned\_room\_type  
booking\_changes  
deposit\_type  
agent  
company  
days\_in\_waiting\_list  
customer\_type  
adr  
required\_car\_parking\_spaces  
total\_of\_special\_requests  
reservation\_status  
reservation\_status\_date

### 3. Métricas

La métrica que usaremos para evaluar el modelo principalmente será la medición del porcentaje de predicciones correctas o Accuracy

$$Accuracy (ACC) = \frac{TP + TN}{P + N}$$

En cuanto a la métrica de negocio sería de gran utilidad conocer la precisión del modelo para que la industria hotelera, o un hotel en particular pueda definir estrategias para apaciguar las pérdidas por cancelaciones, con nuevas ideas, ya sea como la mencionada anteriormente en la que un hotel puede ofrecer más habitaciones de las que realmente tiene confiando en la predicción de cancelaciones

### 4. Desempeño

La idea es poder **obtener la predicción de una cancelación de reserva**, sería de gran ayuda para las cadenas hoteleras saber si un cliente realmente va a llegar, esto podría ayudar al comercio para planear la gestión de recursos ya sea de personal o de alimentos.

Conociendo el promedio de cancelaciones la industria hotelera podría sobrevender habitaciones y así mitigar el impacto de las mismas y costear el valor del proyecto de predicción

Segun el dataset a usar, las cancelaciones están alrededor del 30%

## 5. Procesamiento de datos

Inicialmente pudimos realizar la carga del dataset y lo visualizamos en el codelab, pudimos crear un nuevo API token para descargar el dataset.

Según las respuestas la data quedó correctamente subida, pero nos dimos cuenta de algunos datos que no eran útiles y muchos nulos que había que procesar.

```
#checking for null values
dict_={}
for feature in data.columns:
    dict_[feature]=data[feature].isnull().sum()
pd.DataFrame(dict_,index=['null_values']).transpose()
```

Tomamos decisiones como eliminar los valores de “company”, ya que habían más de 100.000 que tenían resultados nulos después de la revisión.

Posterior a ello pudimos crear una copia de la tabla de datos para poder trabajar en ella

### Valores faltantes

Se evidencia que también tenemos valores faltantes, una cantidad no tan importante como lo sucedido en “company”, pero si para poner especial atención en ellos.

```
data_set.isnull().sum().sort_values(ascending=False)[:8]
```

agent	16340
country	488
Total_guests	4
children	4
reserved_room_type	0
assigned_room_type	0
booking_changes	0
deposit_type	0

dtype: int64

**agent:** en este caso se identifican y se cargan con ceros los que aparecen nulos, una decisión temporal, mientras analizamos las mejores opciones para nuestro dataset

**country:** Para estos valores faltantes, se procede a reemplazarlos por su moda

**Children:** Para estos valores faltantes, se procede a reemplazarlos por su número de promedio redondeado.

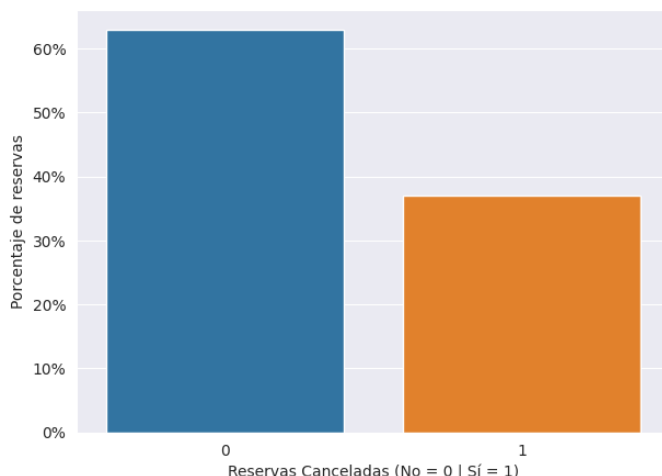
Se realizan verificaciones para comprobar que no se tengan 0 invitados

```
#double check of there are not 0 guesses in the dataset
((data_set.adults + data_set.babies + data_set.children)==0).sum()
```

0

Ahora, algunos de los tipos de datos parecen ser inadecuados por lo que tuvimos que realizar cambios a flotante y a entero en las columnas “children”, “company”, y “Agent”

Con muchos de esos datos útiles pudimos realizar graficaciones para visualizar de mejor manera la información representativa, como una comparativa de reservas y cancelaciones



También se pudo observar que la proporción de cancelaciones entre el hotel de ciudad es más alta que en el resort, que la proporción de reservas canceladas en espera parece ser relativamente estable a lo largo del mes y que las reservas con el tipo de comida seleccionado, es decir completa, tienen más probabilidades de ser canceladas y cosas como que los clientes frecuentes tienden a mantener las reservas, más que los clientes nuevos, entre otras.

**link colab:**

<https://colab.research.google.com/drive/1Vt2Fw4kWQpVfjMmocl5ombuw9W-djcfB?usp=sharing#scrollTo=JxPH3hSiyViM>

**link dataset:** <https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand>

## 6. Bibliografía

[1].| Kaggle.

<https://www.kaggle.com/code/santhosh77/hotel-booking-eda-and-cancellation-prediction/notebook>

[2].

[https://www.hosteltur.com/comunidad/005020\\_como-el-covid-19-ha-afectado-al-sector-hoteleroy-perspectiva-para-el-nuevo-ano.html](https://www.hosteltur.com/comunidad/005020_como-el-covid-19-ha-afectado-al-sector-hoteleroy-perspectiva-para-el-nuevo-ano.html)

[3] <https://www.airbnb.com.co/help/article/475/>