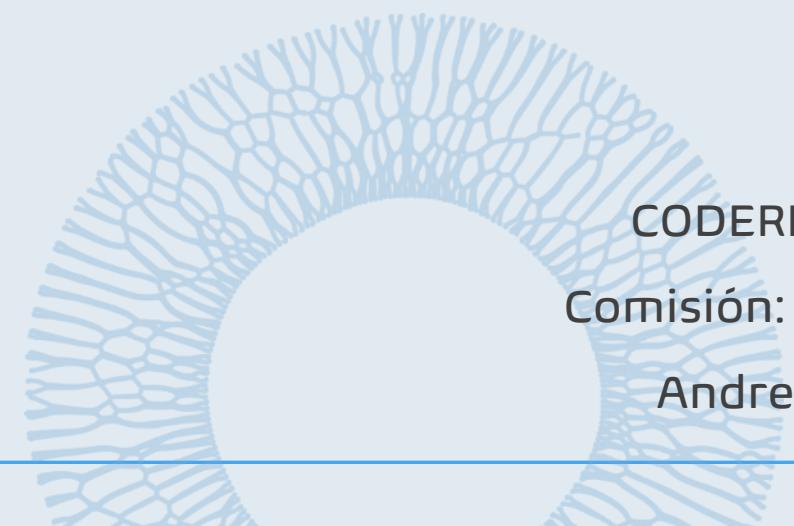


PROYECTO DATASCIENCE: CANCELACIONES EN RESERVAS HOTELERAS



CODERHOUSE
Comisión: 49150

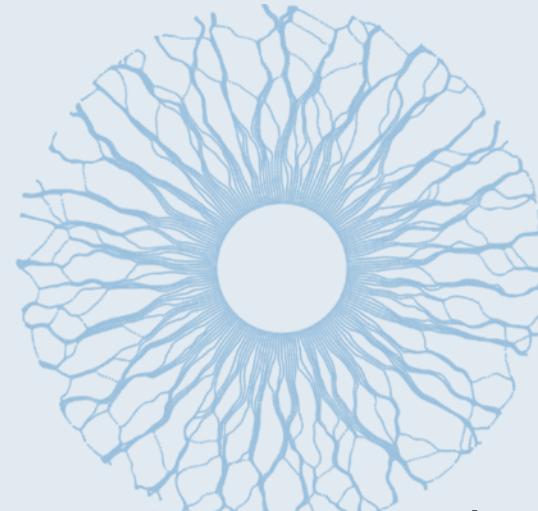
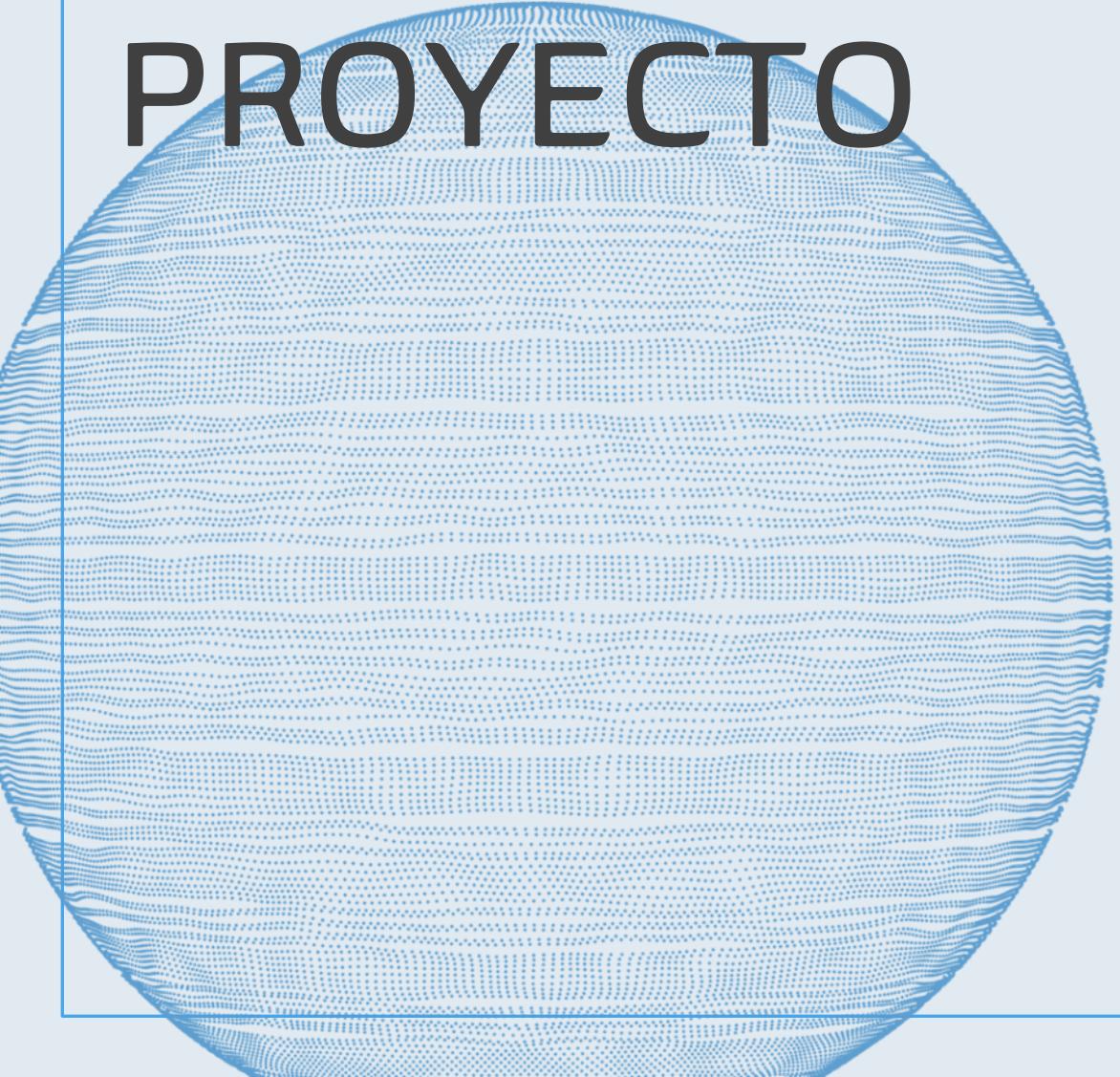
Andrea Brito

ÍNDICE

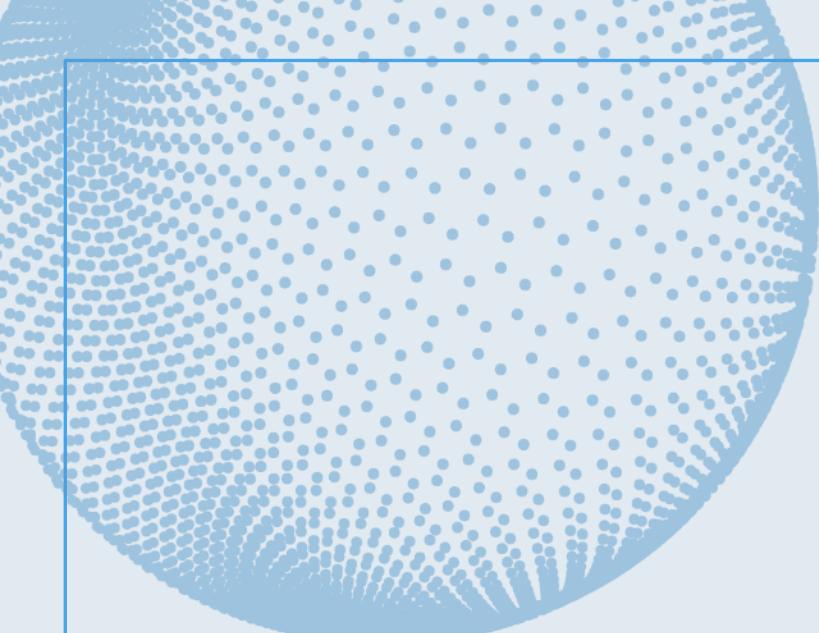
- A. OBJETIVO DEL PROYECTO
- B. PREGUNTAS Y RESPUESTAS
- C. HIPÓTESIS Y VERIFICACIÓN
- D. PREDICCIÓN DE CANCELACIONES
- E. CONCLUSIONES



OBJETIVO DEL PROYECTO

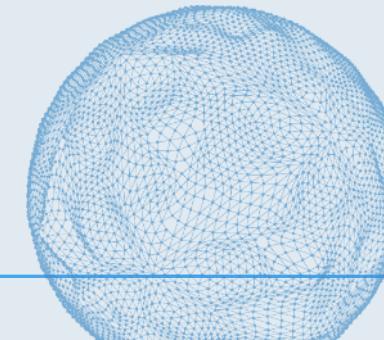
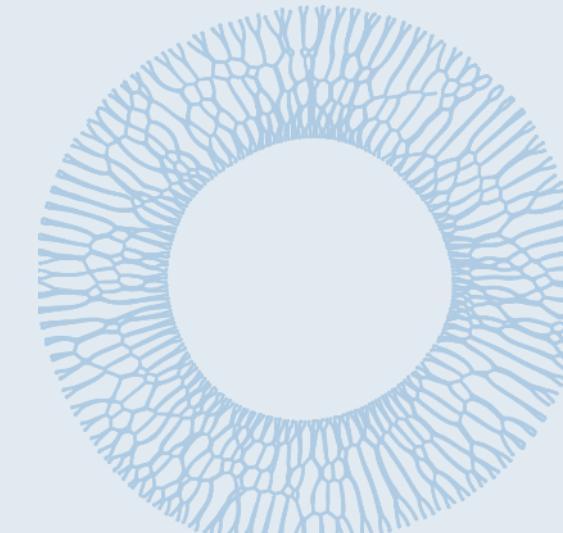


- Este proyecto tiene como objetivo poder predecir qué reservas hoteleras tienen mayor posibilidad de cancelarse, para que de esta forma las empresas puedan tomar las medidas que consideren correspondientes para minimizar el impacto de las cancelaciones en las ganancias.



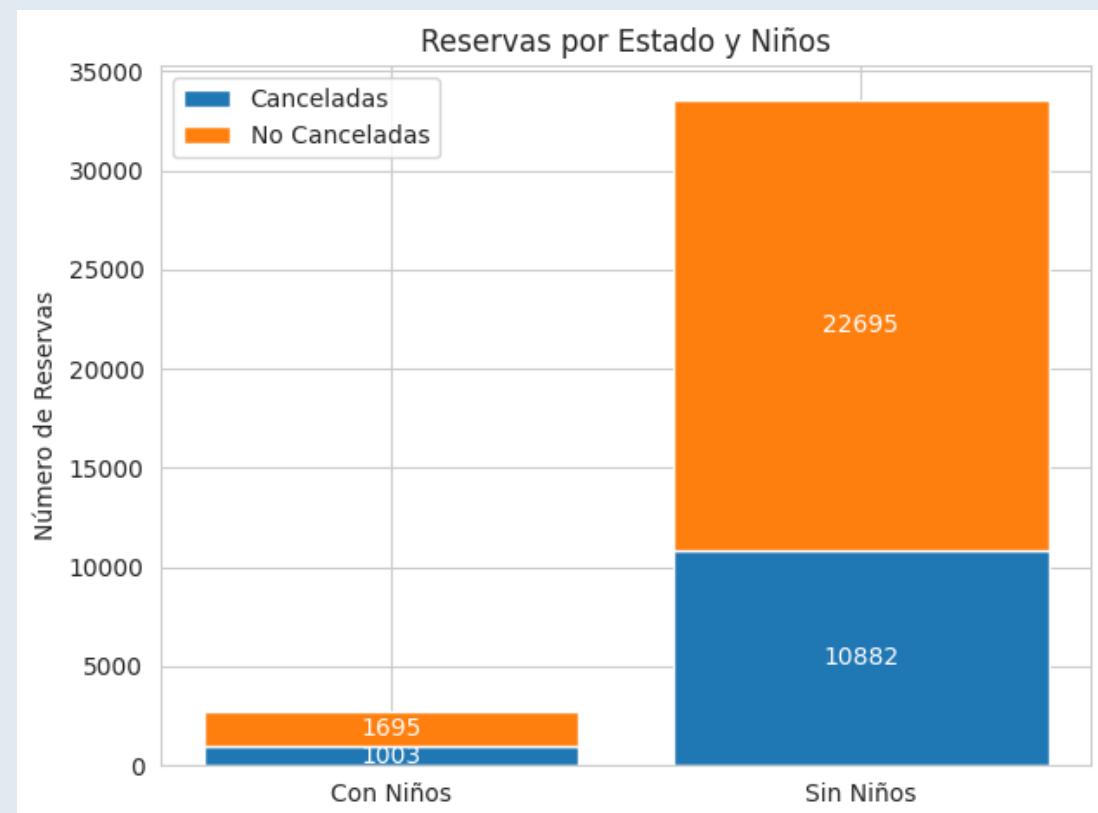
EDA (ANÁLISIS EXPLORATORIO DE DATOS)

PREGUNTAS Y RESPUESTAS



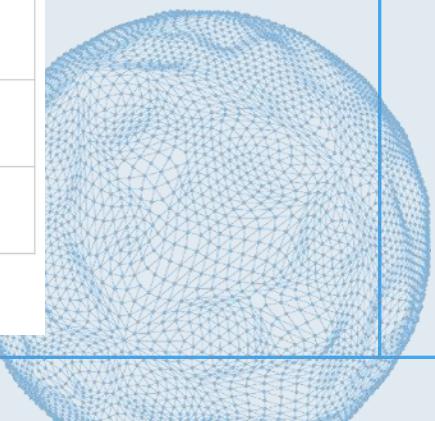
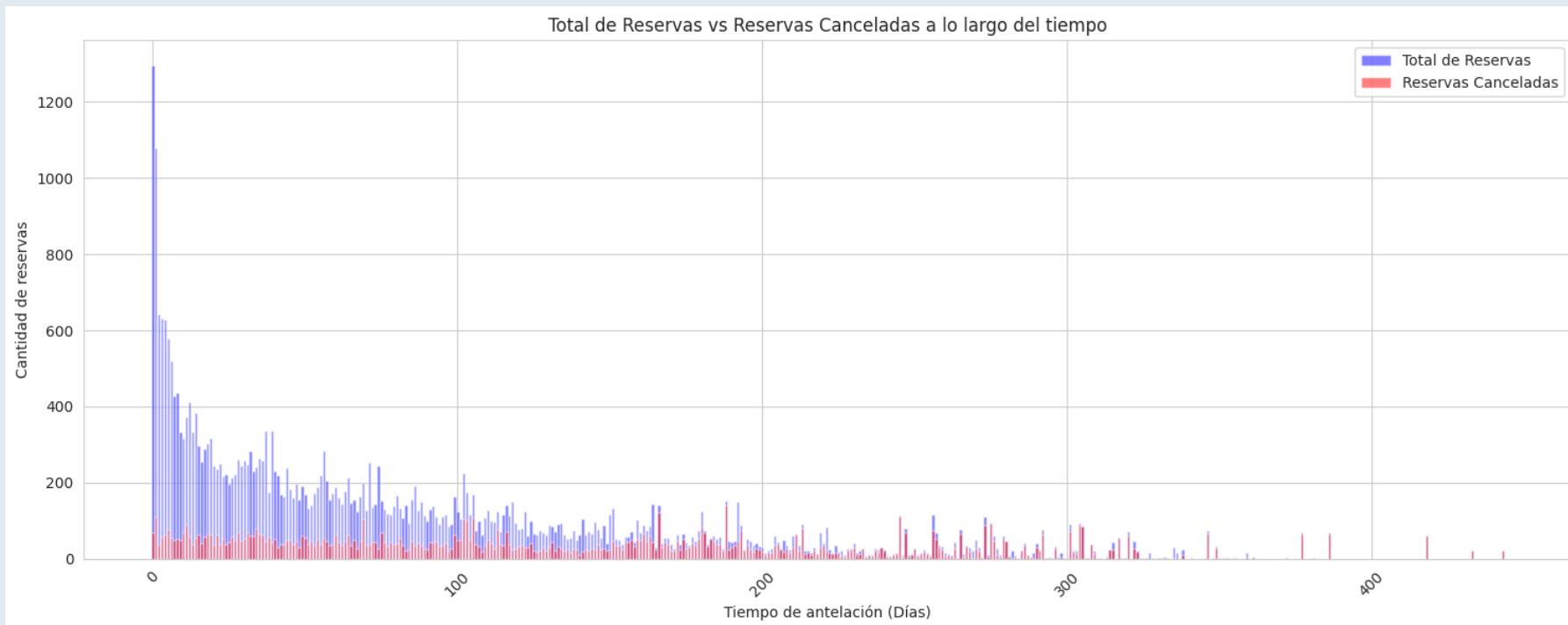
¿LOS CLIENTES QUE VIENEN CON NIÑOS TIENEN MAYOR O MENOR ÍNDICE DE CANCELACIÓN?

- Si bien las cancelaciones en las que hay niños son mayores, dado que representan una parte menor de las reservas no podemos decir que sea un indicador determinante a la hora de predecir las cancelaciones.



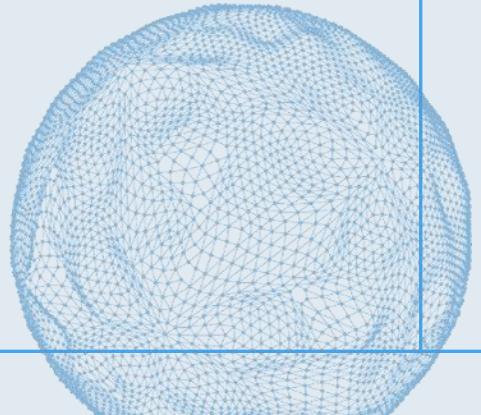
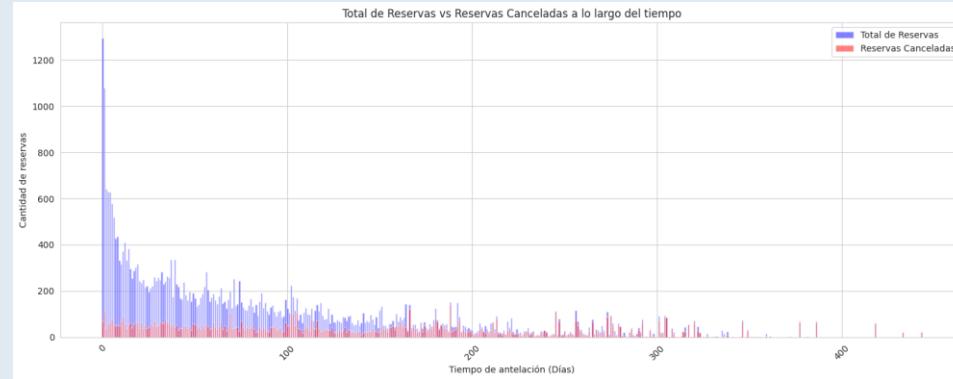
¿EN QUÉ INDICADOR ESTÁ EL MAYOR RIESGO DE CANCELACIÓN?

- Si bien no podemos definir un solo indicador, vemos claro que, a mayor tiempo de antelación con la que se produce la reserva mayor es el riesgo de cancelación.



¿QUÉ FACTORES INCIDEN DE FORMA DIRECTA EN LAS CANCELACIONES?

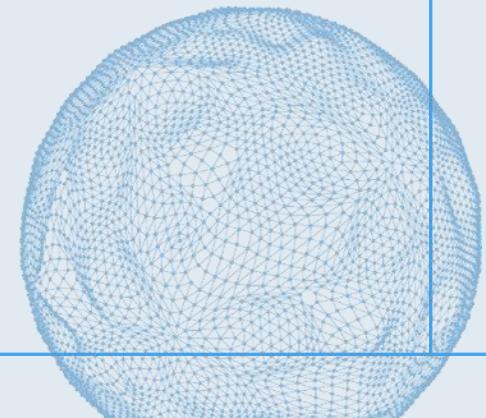
- Tiempo de antelación con el que se realiza la reserva
- Cancelaciones previas en clientes repetidos



HIPÓTESIS A:

- H_0 : EL TIEMPO DE ANTELACIÓN EN QUE SE REALIZA LA RESERVA INFUYE DE MANERA DIRECTA EN LA CANCELACIÓN DE LA MISMA
- H_1 : EL TIEMPO DE ANTELACIÓN EN QUE SE REALIZA LA RESERVA NO INFUYE DE MANERA DIRECTA EN LA CANCELACIÓN DE LA MISMA

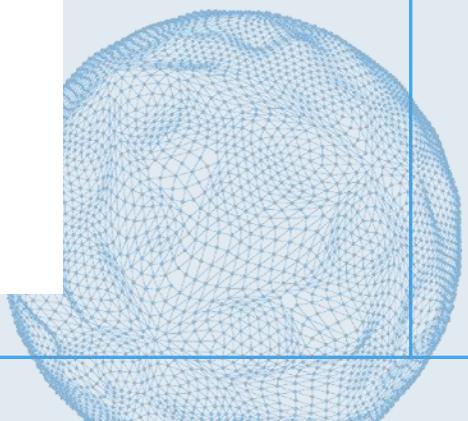
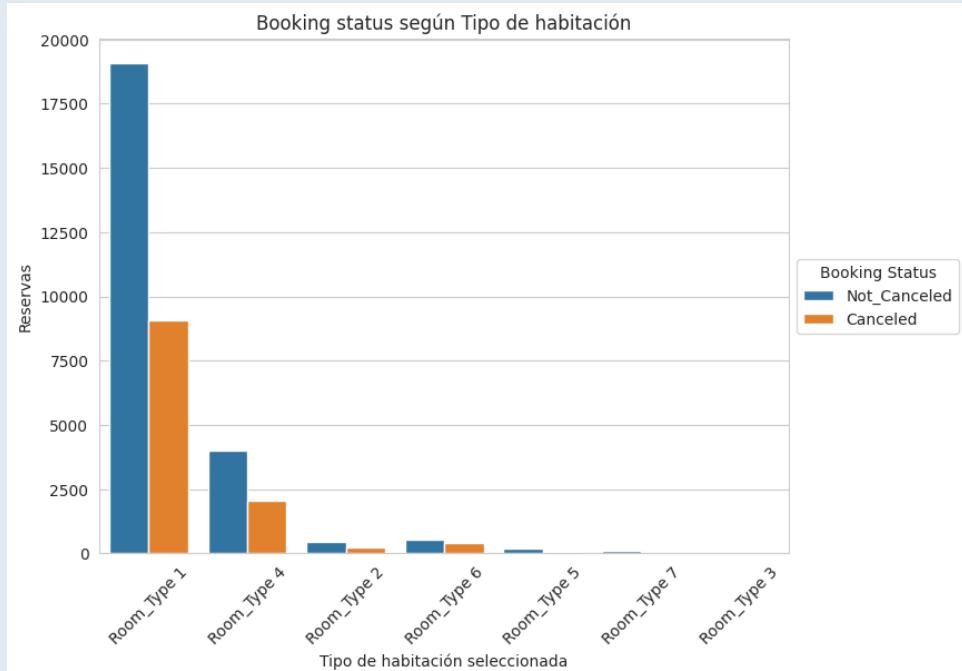
- Esta hipótesis se verifica en la respuesta a la segunda reserva planteada, en este caso H_0 se acerca más a la realidad.



HIPÓTESIS B:

- H_0 : SEGÚN EL TIPO DE HABITACIÓN SELECCIONADA, LAS RESERVAS SON CANCELADAS EN MAYOR O MENOR MEDIDA
- H_1 : SEGÚN EL TIPO DE HABITACIÓN SELECCIONADA, LAS RESERVAS NO SON CANCELADAS EN MAYOR O MENOR MEDIDA

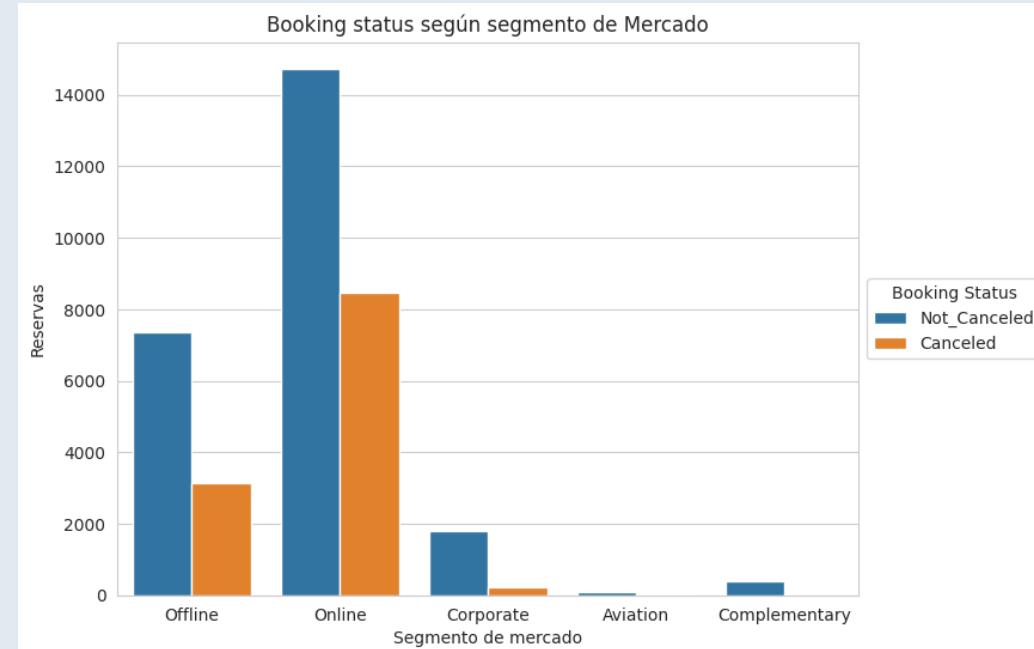
- En este caso, observando el gráfico vemos que H_1 se acerca más a los datos que tenemos.



HIPÓTESIS C:

- H_0 : EL SEGMENTO DE MERCADO INFLUYE DE MANERA DIRECTA EN LA CANCELACIÓN DE LAS RESERVAS
- H_1 : EL SEGMENTO DE MERCADO NO INFLUYE DE MANERA DIRECTA EN LA CANCELACIÓN DE LAS RESERVAS

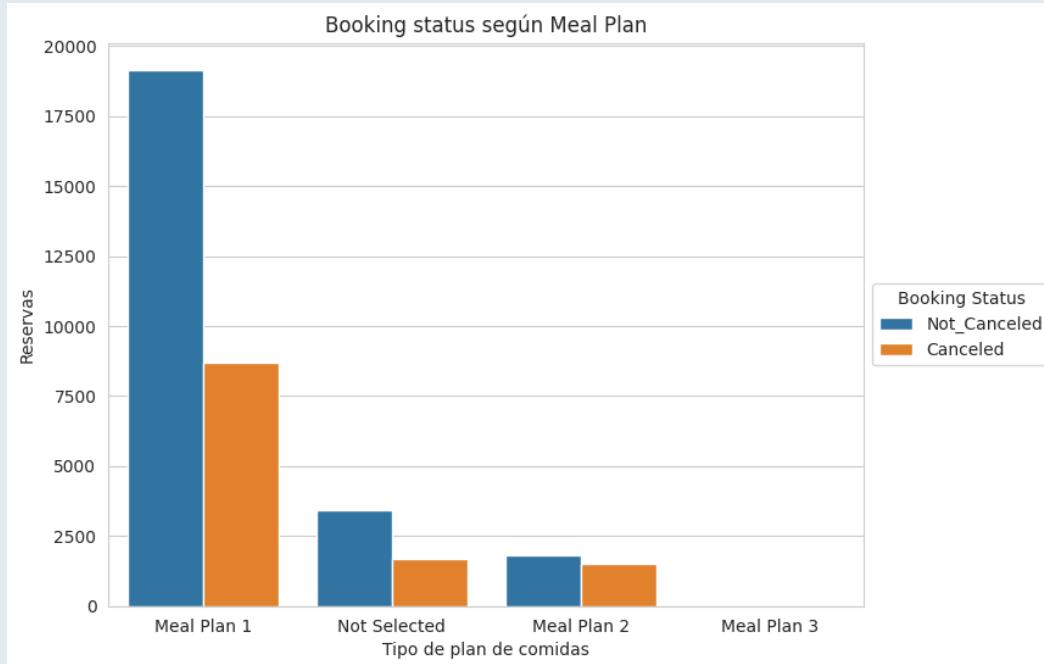
- Si bien vemos que las cancelaciones en las reservas corporativas son menores, no podemos decir que se cumpla ninguno de nuestros dos supuestos.



HIPÓTESIS D:

- H_0 : SEGÚN EL TIPO DE PLAN DE COMIDAS SELECCIONADO, LAS RESERVAS SON CANCELADAS EN MAYOR O MENOR MEDIDA
- H_1 : SEGÚN EL TIPO DE PLAN DE COMIDAS SELECCIONADO, LAS RESERVAS NO SON CANCELADAS EN MAYOR O MENOR MEDIDA

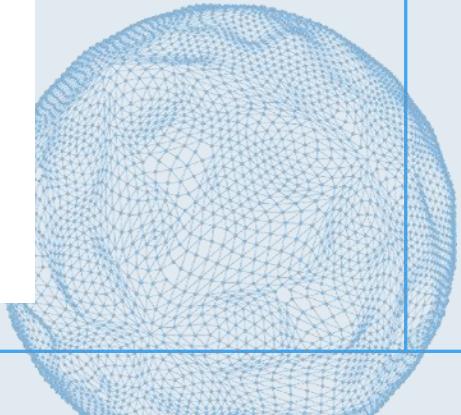
- Viendo el gráfico de la derecha, vemos que nuestro H_1 se acerca más a la realidad: el plan de comidas seleccionado no parece afectar directamente a la cancelación de las reservas.



HIPÓTESIS E:

- H_0 : SI HAY CANCELACIONES PREVIAS, LAS NUEVAS CANCELACIONES SON MÁS FACTIBLES
- H_1 : SI NO HAY CANCELACIONES PREVIAS, LAS NUEVAS CANCELACIONES SON MENOS FACTIBLES

- En este caso, y como ya vimos en nuestra tercera pregunta, nuestro H_0 pareciera ser correcto. Quienes ya han cancelado tienen más probabilidades de volver a hacerlo.

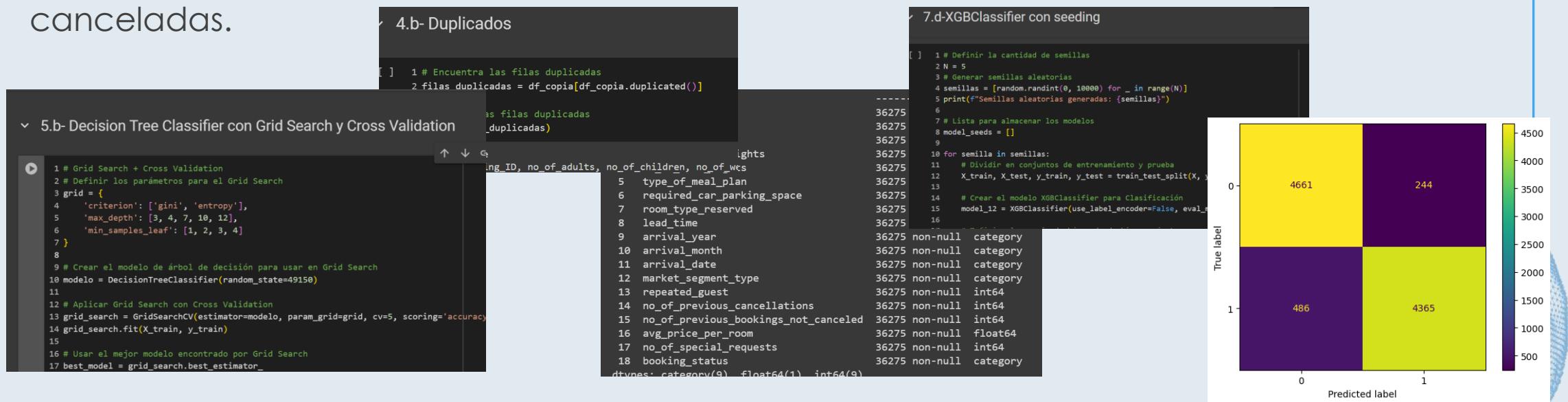


PREDICCIÓN DE CANCELACIONES

SELECCIÓN DE MODELOS DE MACHINE LEARNING

¿CÓMO SABER QUE RESERVAS SE CANCELARÁN?

Para ello vamos a intentar llevar a nuestros datos al mejor punto posible para que al aplicar modelos de Machine Learning seamos capaces de predecir con la mayor precisión posible cuáles reservas serán canceladas.



ANÁLISIS DE MÉTRICAS DE LOS MODELOS APLICADOS

Luego de aplicar varios métodos para mejorar nuestros datos y probar diferentes modelos de Machine Learning obtuvimos estas métricas finales:

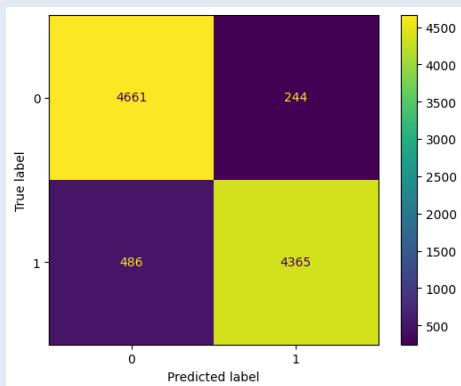
Model	Accuracy_Test	Accuracy_Train	Precision_Test	Precision_Train	Recall_Test	Recall_Train	F1_Test	F1_Train
LGBM	0.9026240262402624	0.9442138171381714	0.9027324632952692	0.9412304398797084	0.903469387755102	0.947460236018471	0.9031007751937984	0.9443350635403616
Logistic_Regression	0.7736777367773677	0.7733189831898319	0.7735577910080066	0.7723444438772906	0.7732403037143444	0.7752728390633806	0.7733990147783251	0.7738058709215506
D_Tree_Classifier	0.8995489954899549	0.9500820008200082	0.9190163239347042	0.9556075982305492	0.8787755929454693	0.9437220537595724	0.8984455958549222	0.949627637567232
Random Forest	0.9039565395653956	0.9444444444444444	0.9061728395061728	0.939735200121747	0.9015353121801432	0.949756472699308	0.9038481272447408	0.9447192615635676
XGBClassifier	0.9251742517425174	0.9883148831488316	0.9470600998047298	0.9894331879969224	0.8998144712430427	0.9872050770254364	0.922832980972516	0.988317876722857
Stacking	0.9186141861418614	0.9754766297662977	0.9175891758917589	0.9729936305732484	0.9194741166803615	0.9781272410613666	0.9185306792530268	0.9755536822745037

CONCLUSIONES

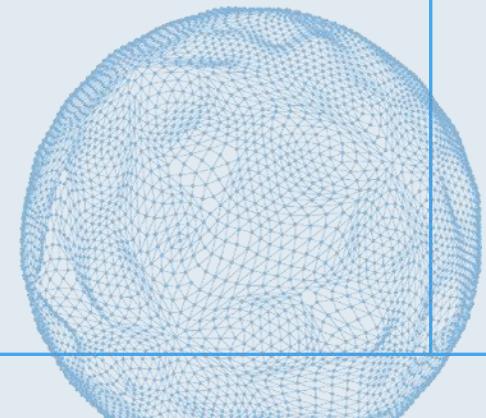
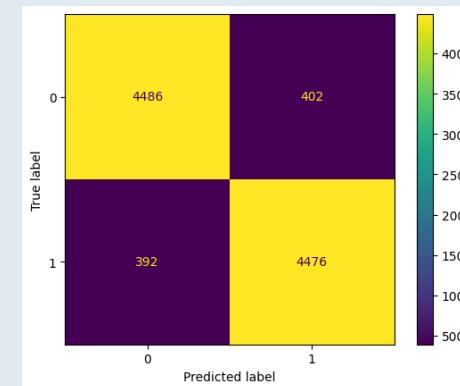
¿CUÁL ES EL MEJOR MODELO PARA PONER EN PRODUCCIÓN?

Por los resultados observados en cuanto a las métricas de Testeo, el mejor modelo de los que probamos para poner en producción sería el de XGB Classifier, seguido del Stacking de modelos, en ambos casos las predicciones fueron buenas y tanto en Accuracy, Precision, Recall y F1 (métricas de evaluación de modelos de M. L) los resultados fueron superiores a 0.9. Abajo las imágenes de CM de ambos modelos, como ya indicamos XGB obtuvo resultados superiores

Confusion Matrix XGB



Confusion Matrix
Stacking



GRACIAS

ANDREA BRITO

COMISIÓN 49150