

TERCERA ENTREGA PROYECTO FINAL



Base de datos: Encuesta de satisfacción del usuario de una aerolínea





TABLA DE CONTENIDO



01

Conformación del equipo de trabajo.

02

Presentación de la empresa

03

Objetivo del proyecto



04

Descripción de la temática de datos (Data Acquisition)

05

Generación del primer Data Wrangling y EDA, univariado, bivariado y multivariado.

06

Algoritmos de clasificación

07

Algoritmos de optimización



CONFORMACIÓN DEL EQUIPO



**Victoria
Giay**

Ingeniera en alimentos.
Actualmente trabajando
en la industria
alimenticia.



**Xiomara Pillaca
Alarcon**

Salubrista público.
Actualmente trabajando
en el sector salud



**Julieta
Windischbauer**

Ingeniera Industrial.
Actualmente trabajando
en consultoría de RPA.

PRESENTACIÓN DE LA EMPRESA

La base de datos pertenece a una aerolínea.

La misma ofrece vuelos de corta y larga distancia, con tres tipos de clases de asientos: Eco, Eco plus y Business.

Actualmente está buscando mejorar su servicio, con lo cual realizó una encuesta a aquellos clientes que volaron recientemente con ellos, para intentar analizar cuáles son aquellas variables en las que destacan y cuáles son aquellas en las cuales podrían mejorar.

OBJETIVO DEL PROYECTO

Determinar cuales son las variables que mejor califican en los vuelos, a fin de poder utilizar las mismas en las próximas campañas de marketing.

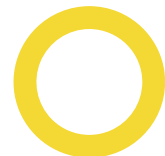
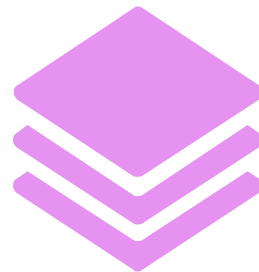
Obtener el perfil promedio de los clientes de una aerolínea, a través de las variables que aplican a la descripción del pasajero.

DESCRIPCIÓN DE LA TEMÁTICA DE LOS DATOS

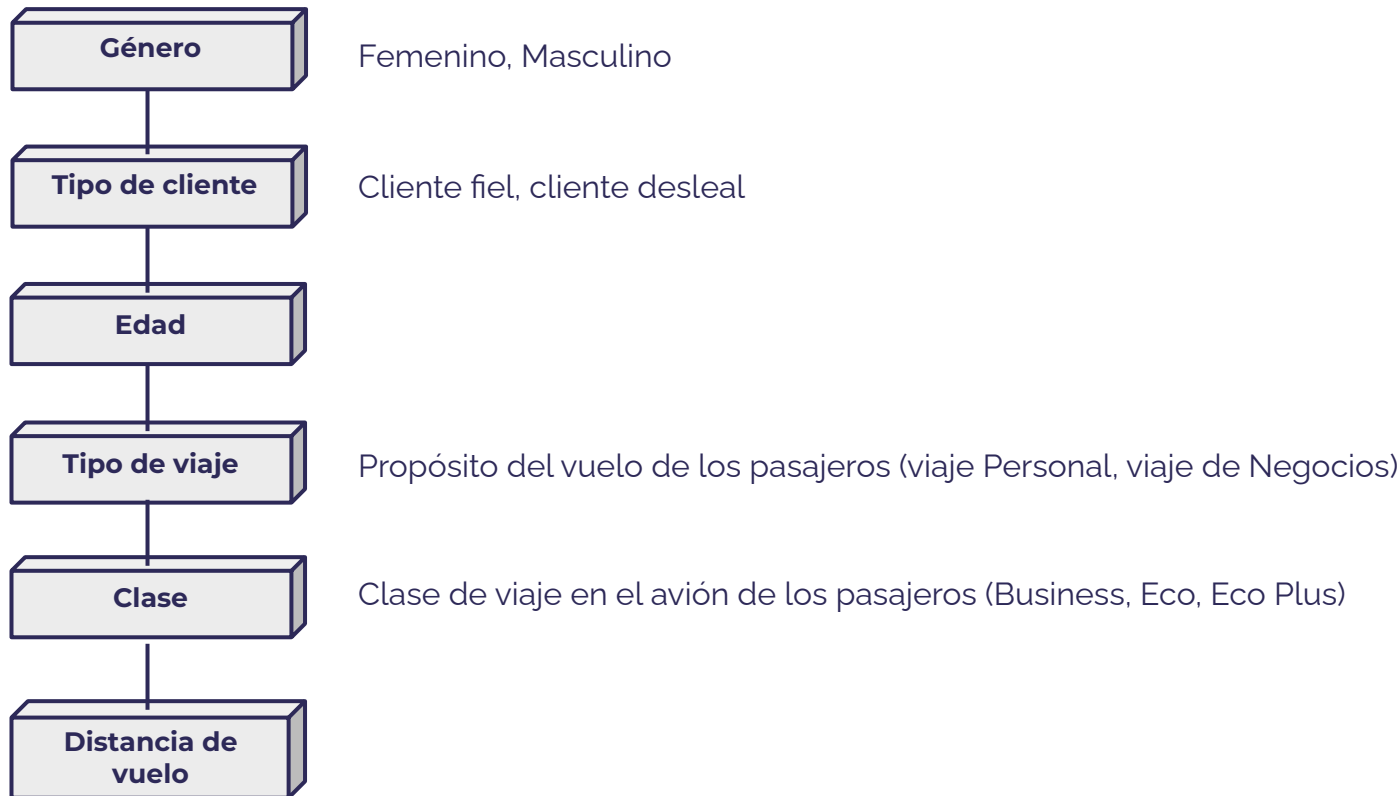
Data acquisition

La base de datos está conformada por una única tabla, en formato csv. Para su análisis se utilizó el programa google collaboratory.

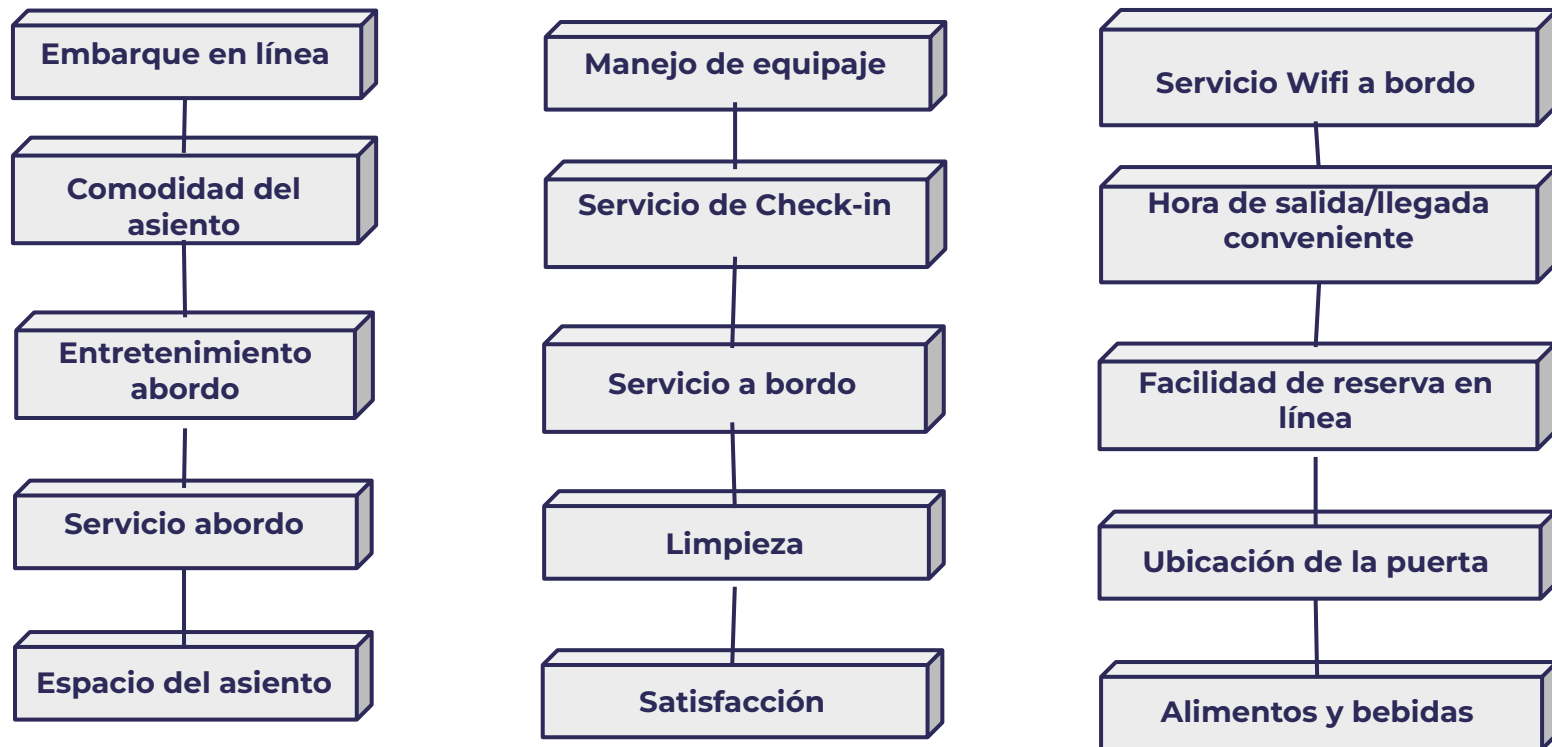
Nota: La escala de la encuesta va entre 1 (malo) y 5 (bueno), siendo 0 (cero) cuando no aplica al vuelo realizado.



Los 6 campos que clasifican al cliente son:



Los 15 campos que integran la encuesta son:



Data Wrangling y EDA: análisis de componentes principales



De la base de datos original se eliminaron las columnas que no se iban a considerar en el análisis y se cambió el nombre de un campo. No se tuvo que realizar ningún tratamiento de nulos pues todos los campos tenían la información completa (no se requirió hacer tratamiento de los mismos).

De la encuesta se observó que el único campo con outliers es check in service

El perfil del cliente lo analizamos a través de los campos: edad, género, tipo de cliente, tipo de viaje y clase. Hay dos categorías para clasificar la conformidad del vuelo: satisfecho y neutral o insatisfecho

Respecto a los vuelos, la mayoría son de corta distancia (menos de 1000 millas); la media es de 1200 millas.



Data Wrangling y EDA: análisis de componentes principales

PERFIL DEL CLIENTE

- Género: indistinto (viajan hombres y mujeres por igual).
- Edad promedio: 40 años (van entre los 20 y 50 años)
- De los clientes leales la mayoría viaja en business mientras que de los desleales la mayoría viaja en económica (Eco).
- Los clientes más jóvenes tienden a elegir clase Eco o Eco Plus.
- Los clientes desleales hacen, en su mayoría, viajes de negocios.
- Los clientes desleales suelen estar más insatisfechos o estar indiferentes frente al servicio.



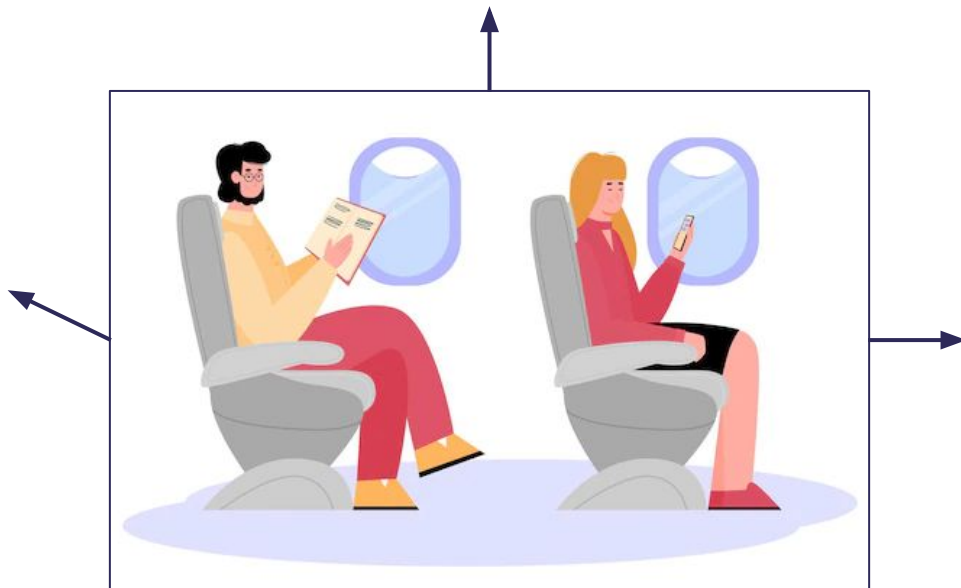
Data Wrangling y EDA: análisis de componentes principales



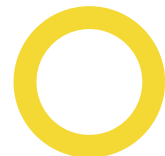
PERFIL DEL VUELO

En los vuelos de corta distancia se venden más asientos de Eco, y en los vuelos de larga distancia aumentan las personas que viajan en Business.

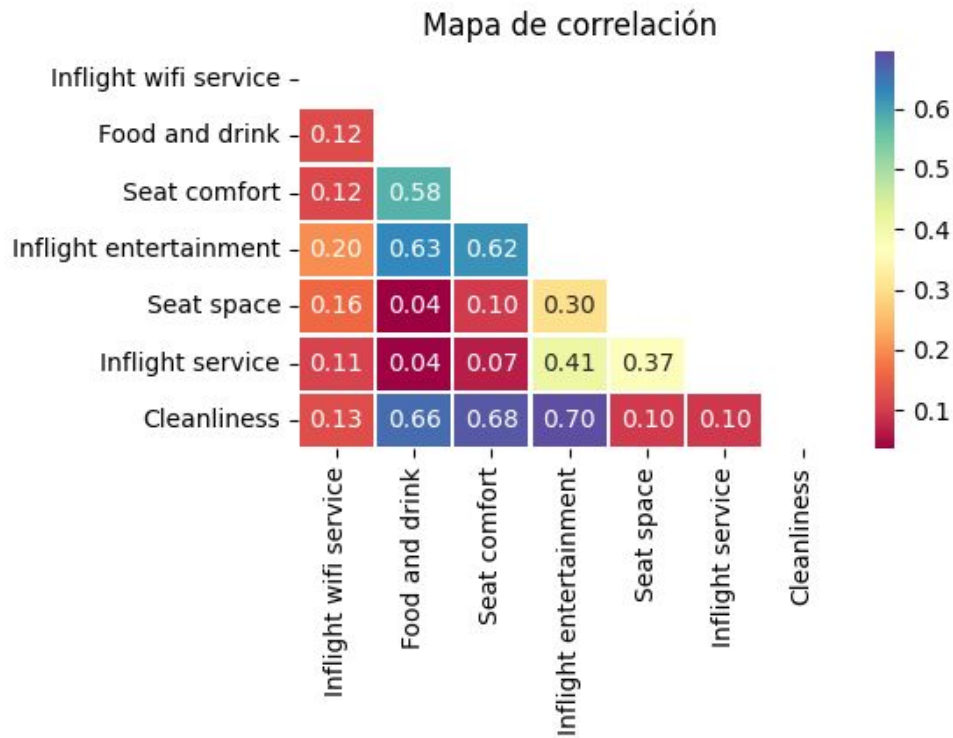
Las personas adultas no suelen elegir vuelos de larga distancia.



En los vuelos de larga distancia viajan principalmente clientes leales.



Data Wrangling y EDA: análisis de componentes principales



Analizando únicamente las variables propias del vuelo, se observa que la mayoría de las relaciones son débiles y directas, siendo la variable limpieza la que tiene correlación más fuerte con las variables alimentos y bebidas, comodidad del asiento y entretenimiento en el vuelo.

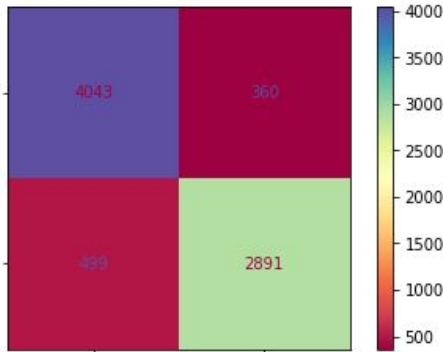
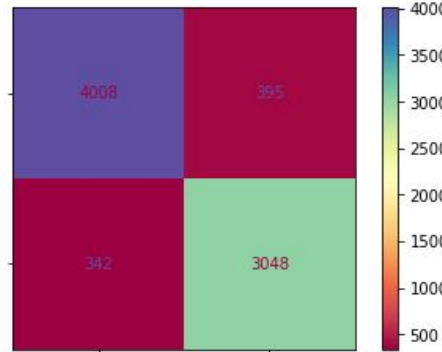
Algoritmo de clasificación

Analizamos 4 algoritmos de clasificación: KNN, árbol de decisión, random forest y support Vector Machine's.

Modelo	KNN	Arbol de decision	Random Forest	Support Vector Machine's
Accuracy	0.92	0.94	0.96	0.95
Precision	0.94	0.92	0.97	0.95
Recall	0.87	0.93	0.94	0.93
F1	0.90	0.93	0.95	0.94

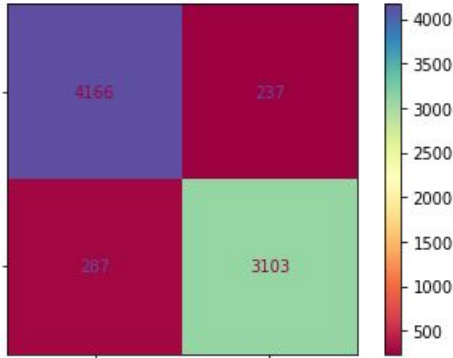
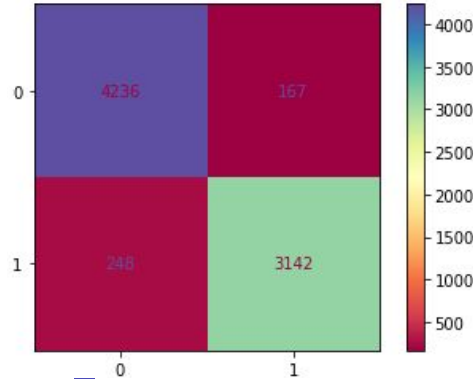
Algoritmo de clasificación

Analizamos 4 algoritmos de clasificación: KNN, árbol de decisión, random forest y support Vector Machine's.

Modelo	KNN	Arbol de decision								
Matriz de confusión	 <table><tr><td>4043</td><td>360</td></tr><tr><td>489</td><td>2891</td></tr></table>	4043	360	489	2891	 <table><tr><td>4008</td><td>395</td></tr><tr><td>342</td><td>3048</td></tr></table>	4008	395	342	3048
4043	360									
489	2891									
4008	395									
342	3048									

Algoritmo de clasificación

Analizamos 4 algoritmos de clasificación: KNN, árbol de decisión, random forest y support Vector Machine's.

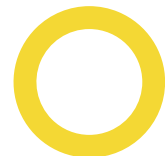
Modelo	Random Forest	Support Vector Machine's								
Matriz de confusión	 <p>Confusion matrix for Random Forest:</p> <table><tr><td>4166</td><td>237</td></tr><tr><td>267</td><td>3103</td></tr></table> <p>Color scale: 500 to 4000</p>	4166	237	267	3103	 <p>Confusion matrix for Support Vector Machine's:</p> <table><tr><td>4236</td><td>167</td></tr><tr><td>248</td><td>3142</td></tr></table> <p>Color scale: 500 to 4000</p>	4236	167	248	3142
4166	237									
267	3103									
4236	167									
248	3142									

Algoritmo de clasificación



Se obtuvo que el Random Forest es **el modelo que mejor se ajusta** a la base de datos que se está analizando.

El método **Random Forest** tiene mayor exactitud a la hora de predecir si una cliente va a estar satisfecho o no, según las respuestas que se obtengan de la encuesta.





Algoritmo de optimización

Para optimizar el modelo comparamos dos tipos de métodos: K-Fold Cross-Validation y Stratified K-Fold. Los resultados obtenidos se muestran a continuación.

Observando los valores obtenidos para Train y Test con el método de StratifiedKFold y con el método Kfold ambos son muy similares entre sí.

Modelo	K-Fold	Stratified K-Fold
Train score	1.0	1.0
Test score	0.933	0.931



**Muchas gracias
por la atención
prestada !**

