Proyecto Final

Data Science

Dataset:

Encuesta de Satisfacción

de una aerolínea

Diego Sanchez Kramm

Septiembre 2023

CODERHOUSE

Índice

1. Descripción del caso de negocio

Es importante para toda aerolínea conocer a sus clientes, es decir, es indispensable obtener feedback sobre las opiniones y el nivel de satisfacción de los pasajeros que han viajado con dicha aerolínea. Esta información le puede otorgar una ventaja frente a sus competidores directos, ya que le brinda el conocimiento de lo realmente importante para el pasajero a la hora de seleccionar una aerolínea para volar.

Ante esta situación, el equipo de marketing de una gran aerolínea debe medir el nivel de satisfacción de los pasajeros de los distintos vuelos. Los pasajeros que son analizados pertenecen a una base de datos que posee clientes con diversas edades, que han realizado vuelos de diversas distancias (vuelos cortos y largos), han volado por diversos motivos y en diferente clase (Business, Económica, Económica Plus).

Se ha evaluado de forma particular diferentes aspectos o variables relacionadas con el vuelo (por ej. el servicio de entretenimiento a bordo, el servicio de wifi a bordo, etc.). Todo esto permite concluir el nivel de satisfacción ("satisfecho" o "neutro o insatisfecho") del pasajero referida a su experiencia volando con la aerolínea.

2. Objetivos del modelo

Teniendo en cuenta todas las variables del dataset relevado, se propone diseñar un sistema que permita detectar patrones en las elecciones de los diversos aspectos evaluados en el vuelo que permita identificar anticipadamente su nivel de satisfacción.

Para poder responder con el objetivo se plantean ciertas preguntas que sirven de guía:

Preguntas principales

¿Es posible clasificar a los pasajeros en "satisfechos" o "neutrales o insatisfechos" a través de un modelo? ¿Existen patrones particulares en los pasajeros que se encuentran "satisfechos" o "neutrales o insatisfechos"? ¿Qué particularidades tienen los pasajeros que tienen el nivel de satisfacción "satisfecho"? ¿Y cuáles el "neutral o insatisfecho"?

Preguntas secundarias

¿Qué edad tienen los pasajeros? ¿Qué distancia tienen los vuelos realizados por los pasajeros encuestados? ¿Qué clase utilizaron los pasajeros para viajar? ¿Qué nivel de satisfacción tienen los pasajeros? ¿Hay relaciones entre las variables del dataframe?

3. Data Acquisition

El dataset con el cual se trabajó es uno público que fue descargado del sitio Kaggle y que se puede acceder en el siguiente link.

El dataset contiene 24 variables con 103.904 registros, entre las que se encuentra información personal del pasajero, puntuación sobre diversas variables del vuelo, el nivel de satisfacción del pasajero, entre otros.

A continuación se lista una breve descripción de las variables que componen el dataset:

- **id**: Número de identificación
- **Gender**: Género
- **Customer Type**: Tipo de cliente (cliente fiel o desleal)
- **Age**: Edad (años)
- **Type of Travel**: Tipo de viaje (personal o negocios)
- **Class**: Clase (Business, Económico, Económico Plus)
- **Flight Distance**: Distancia del vuelo (Km)
- **Inflight wifi service**: Servicio del wifi a bordo (nivel de satisfacción del 1 al 5)
- **Departure/Arrival time convenient**: Hora de salida/llegada conveniente (nivel de satisfacción del 1 al 5)
- **Ease of Online booking**: Facilidad de reserva en línea (nivel de satisfacción del 1 al 5)
- **Gate location**: Ubicación de la puerta: nivel de satisfacción de la ubicación de la puerta (nivel de satisfacción del 1 al 5)
- **Food and drink**: Alimentos y bebidas (nivel de satisfacción del 1 al 5)
- **Online boarding**: Embarque en línea (nivel de satisfacción del 1 al 5)
- **Seat comfort**: Comodidad del asiento (nivel de satisfacción del 1 al 5)
- **Inflight entertainment**: Entretenimiento a bordo (nivel de satisfacción del 1 al 5)
- **On-board service**: Servicio general del vuelo (nivel de satisfacción del 1 al 5)
- **Leg room service**: Servicio referido al espacio para las piernas entre asientos (nivel de satisfacción del 1 al 5)
- **Baggage handling**: Manejo de equipaje (nivel de satisfacción del 1 al 5)
- **Checkin service**: Servicio de Check-in (nivel de satisfacción del 1 al 5)
- **Inflight service**: Servicio a bordo (nivel de satisfacción del 1 al 5)
- **Cleanliness**: Limpieza (nivel de satisfacción del 1 al 5)
- **Departure Delay in Minutes**: Retraso de salida en minutos (Minutos de retraso en la salida)
- **Arrival Delay in Minutes**: Retraso de llegada en minutos (Minutos de retraso en la llegada)
- **satisfaction**: Satisfacción (Nivel de satisfacción de la aerolínea: Satisfacción, neutral o insatisfacción)

Además se visualiza una tabla resumen de todas las variables:

```
#   Column                                Non-Null Count   Dtype
---  ------                                --------------   -----
 0   id                                    103904 non-null  int64
 1   Gender                                103904 non-null  object
 2   Customer Type                         103904 non-null  object
 3   Age                                   103904 non-null  int64
 4   Type of Travel                        103904 non-null  object
 5   Class                                 103904 non-null  object
 6   Flight Distance                       103904 non-null  int64
 7   Inflight wifi service                 103904 non-null  int64
 8   Departure/Arrival time convenient     103904 non-null  int64
 9   Ease of Online booking                103904 non-null  int64
 10  Gate location                         103904 non-null  int64
 11  Food and drink                        103904 non-null  int64
 12  Online boarding                       103904 non-null  int64
 13  Seat comfort                          103904 non-null  int64
 14  Inflight entertainment                103904 non-null  int64
 15  On-board service                      103904 non-null  int64
 16  Leg room service                      103904 non-null  int64
 17  Baggage handling                      103904 non-null  int64
 18  Checkin service                       103904 non-null  int64
 19  Inflight service                      103904 non-null  int64
 20  Cleanliness                           103904 non-null  int64
 21  Departure Delay in Minutes            103904 non-null  int64
 22  Arrival Delay in Minutes              103594 non-null  float64
 23  satisfaction                          103904 non-null  object
```

Se puede observar que la variable **Arrival Delay in Minutes** posee datos nulos.

4. Data Wrangling

4.1. Reemplazo de los valores nulos

Se determinó que existen 310 registros nulos en la variable **Arrival Delay in Minutes** (minutos de retraso en llegar). De estos 310 registros nulos, los mismo se dividen de la siguiente manera teniendo en cuenta la variable **Class** (clase en la que viajó el pasajero):

```
Class Business 132 Eco 152 Eco Plus 26
```

Luego se obtiene el promedio de retraso por clase:

```
Class Business 14.577272 Eco 15.672183 Eco Plus 16.088645
```

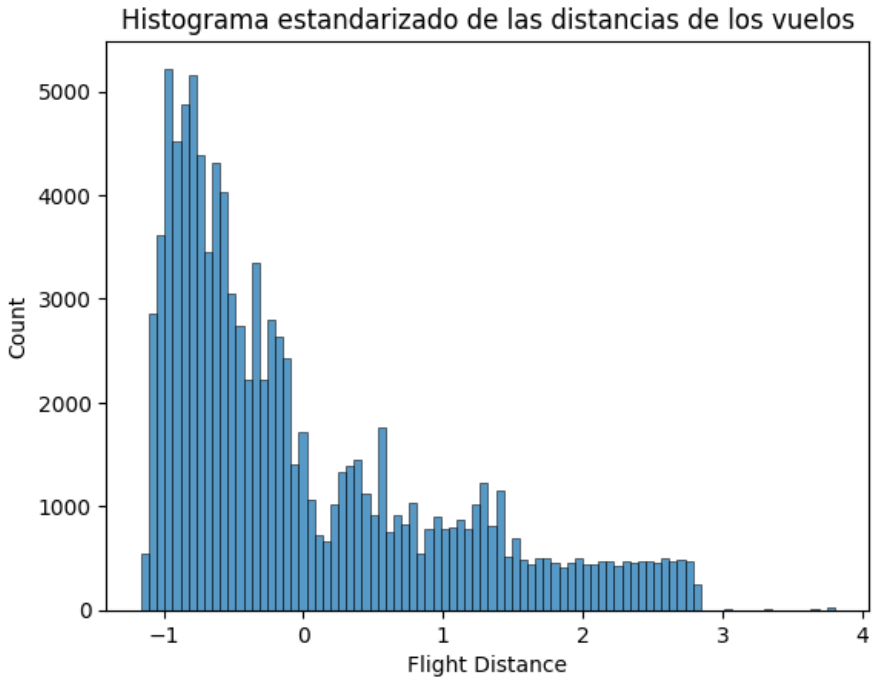
Finalmente, se reemplazan los valores nulos de cada uno de los 310 registros por el promedio por clase. De esta manera el dataset no tiene más valores nulos.

```
#   Column                                Non-Null Count   Dtype
---  ------                                --------------   -----
 0   id                                    103904 non-null  int64
 1   Gender                                103904 non-null  object
 2   Customer Type                         103904 non-null  object
```

```
3   Age                                103904 non-null  int64
4   Type of Travel                     103904 non-null  object
5   Class                              103904 non-null  object
6   Flight Distance                    103904 non-null  int64
7   Inflight wifi service              103904 non-null  int64
8   Departure/Arrival time convenient  103904 non-null  int64
9   Ease of Online booking             103904 non-null  int64
10  Gate location                      103904 non-null  int64
11  Food and drink                     103904 non-null  int64
12  Online boarding                    103904 non-null  int64
13  Seat comfort                       103904 non-null  int64
14  Inflight entertainment             103904 non-null  int64
15  On-board service                   103904 non-null  int64
16  Leg room service                   103904 non-null  int64
17  Baggage handling                   103904 non-null  int64
18  Checkin service                    103904 non-null  int64
19  Inflight service                   103904 non-null  int64
20  Cleanliness                        103904 non-null  int64
21  Departure Delay in Minutes         103904 non-null  int64
22  Arrival Delay in Minutes           103904 non-null  float64
23  satisfaction                       103904 non-null  object
```

4.2. Eliminación de valores outliers

Realizando un pre análisis de la variable **Flight Distance** (distancia de vuelo) se observan valores outliers en la misma. Para visualizar esta realidad, se realiza un histograma en el cual se puede apreciar la existencia de observaciones con más de 3 desviaciones estándar (outliers univariados):



Para obtener mayor especificación en la determinación de los valores outlier a continuación se detectan los mismos a través del criterio del IQR sobre la variable **Flight Distance**.

```
Int64Index([ 80, 173, 201, 215, 379, 388, 421, 446, 458, 473, ... 103357,
103448, 103512, 103534, 103553, 103565, 103648, 103727, 103865, 103889],
dtype='int64', length=2291)
```

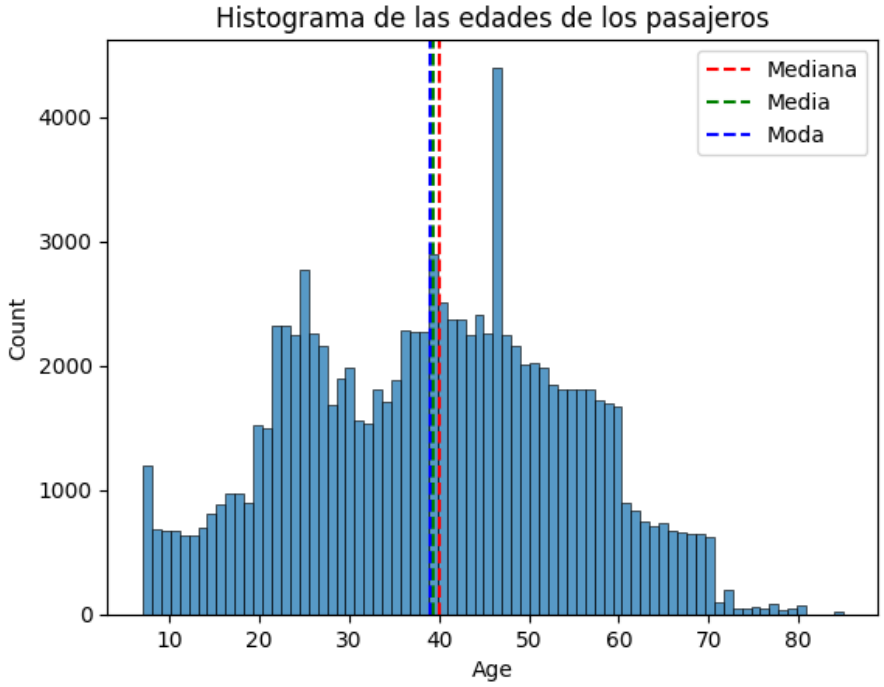
Ante esta situación, se decide eliminar los valores atípicos del dataset para continuar con el análisis exploratorio de datos (EDA) y la selección del algoritmo. Dando como resultado un dataset con 24 variables con 101.613 registros.

5. Exploratory Data Analysis (EDA)

Para guiar el Análisis Exploratorio de Datos (EDA) se utilizaron diversas preguntas secundarias.

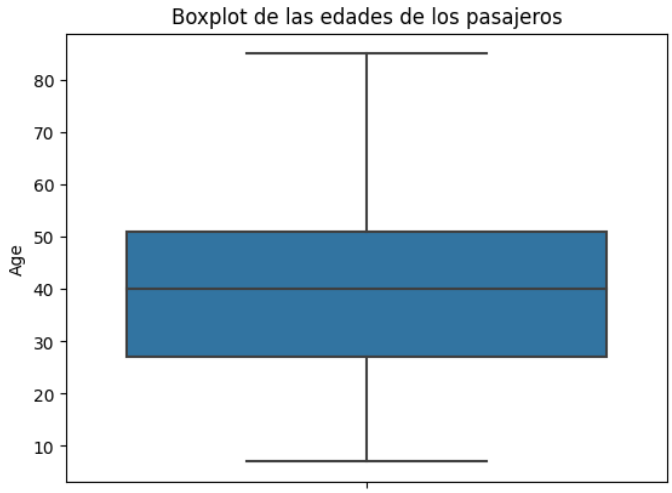
5.1. ¿Qué edad tienen los pasajeros?

Se realiza un histograma donde se puede visualizar la dispersión de las edades de los pasajeros que han volado en la aerolínea.



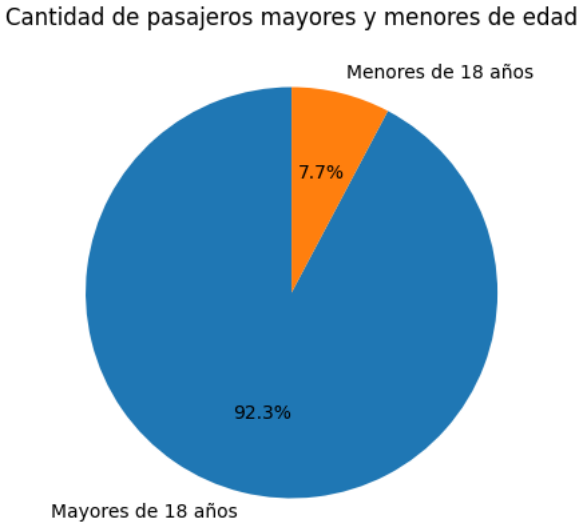
Además se grafican 3 medidas de tendencia central como la mediana (40 años), la media (39 años) y la moda (39 años).

También se grafica un boxplot para visualizar la distribución de los registros.



En este caso se puede decir que el 50% de los pasajeros que volaron tienen entre 28 y 51 años aproximadamente. Además no se visualizan valores atípicos que afecten el cálculo de resultados estadísticos.

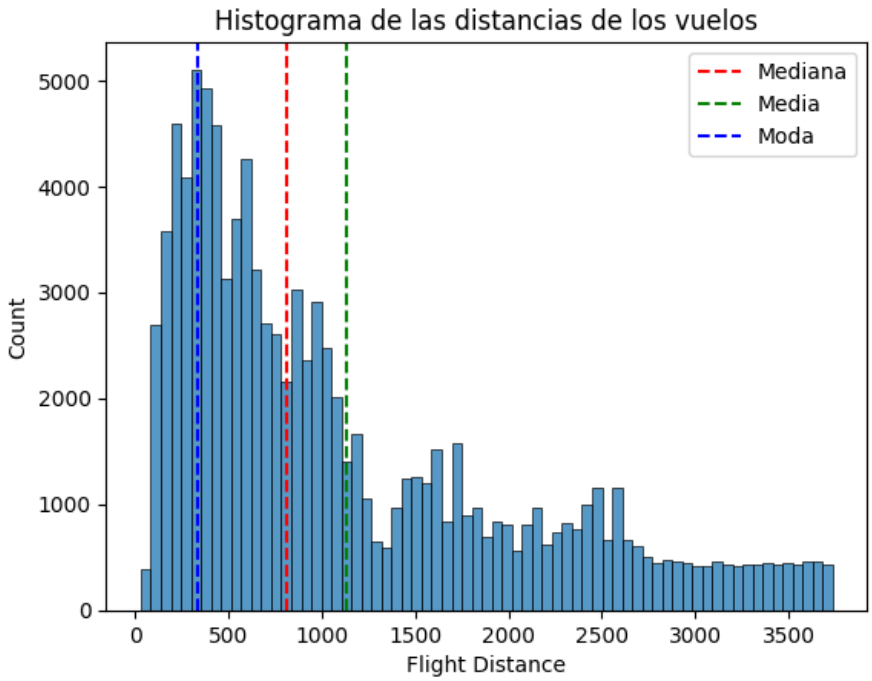
Buscando un poco más de especificaciones en torno a la variable edad, se dividió la misma entre mayores y menores. Los mayores se componen de aquellos pasajeros con 18 años o más. Los menores son aquellos con 17 años o menos.



Teniendo en cuenta esta división se observa que el 92.3% de los pasajeros es mayor de 18 años.

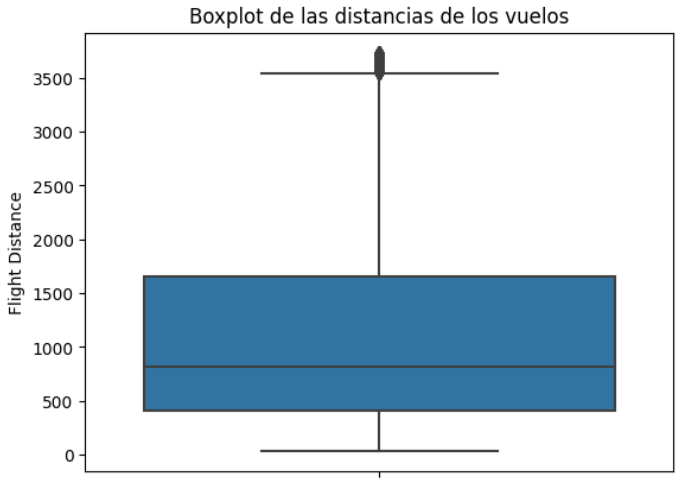
5.2. ¿Qué distancia tienen los vuelos realizados por los pasajeros encuestados?

Se realiza un histograma donde se puede visualizar la dispersión de las distancias de vuelo que han realizado los pasajeros.



Además se grafican 3 medidas de tendencia central como la mediana (814 km), la media (39 años) y la moda (39 años).

También se grafica un boxplot para visualizar la distribución de los registros.

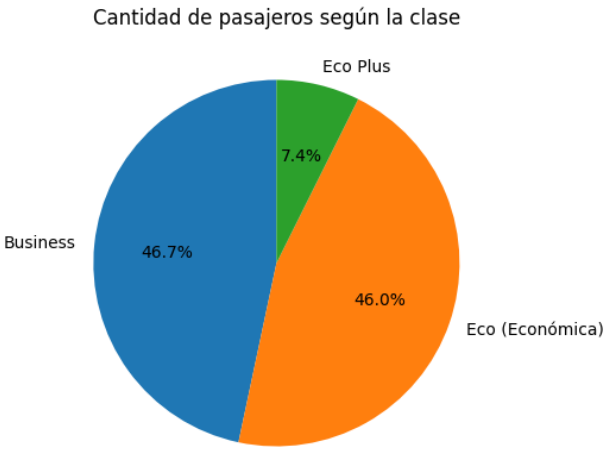


En este caso se puede decir que el 50% de los pasajeros encuestados volaron distancias entre 490 y 1700 km aproximadamente. En dicho gráfico se pueden visualizar valores atípicos.

Igualmente se obtiene un promedio de distancia de vuelo de los mayores y los menores. Para los mayores el promedio de las distancias de los vuelos es de 1148 km aproximadamente. A diferencia de los menores que este promedio es menor y es 899 km aproximadamente.

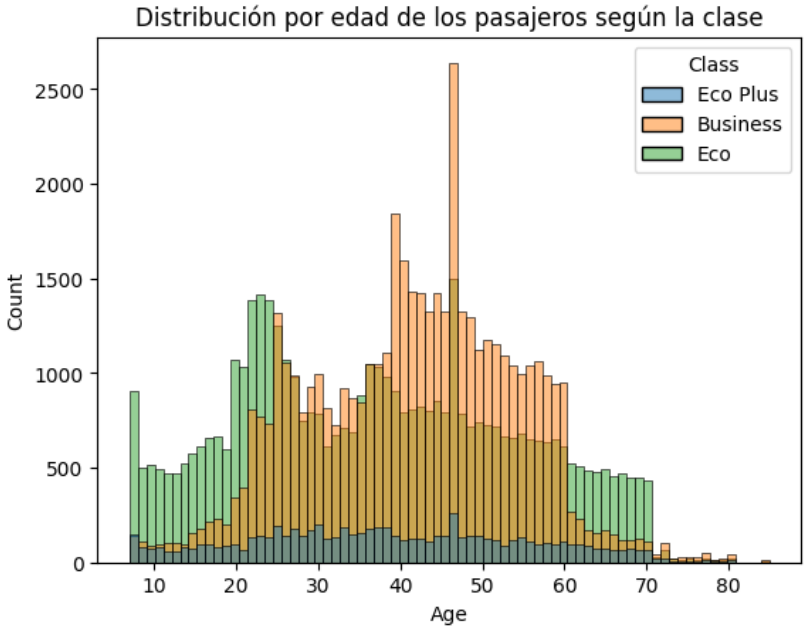
5.3. ¿Qué clase utilizaron los pasajeros para viajar?

Para iniciar con el análisis se realiza un gráfico de torta para graficar el porcentaje de cada clase de aquellos pasajeros encuestados.



En el mismo se visualiza que la clase en la que mayormente viajaron los pasajeros encuestados es la de Business con un 46.7%. Seguida por la clase Económica con un 46% y finalmente la clase Económica Plus con un 7.4%.

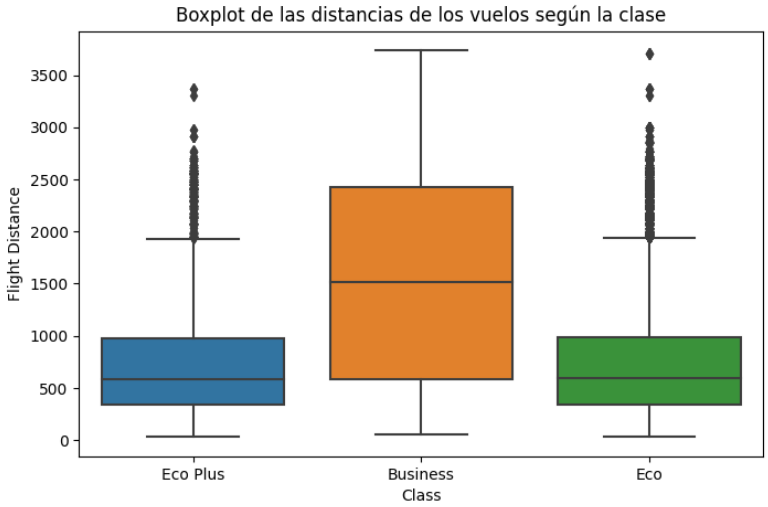
Luego para continuar con el análisis se realiza un histograma en el cual se observa la distribución por edad del pasajero según la clase en la cual viajó. El objetivo de este gráfico es obtener la distribución de la edad dependiendo de la clase.



Se observa en el gráfico que la clase Eco (Económica) es la más utilizada por todos los menores de 18 años. Igualmente, se puede visualizar que la clase Business tiene un crecimiento en cantidad de pasajeros menores.

Además se visualiza que la clase Business es la más elegida por los pasajeros del rango de edad que va de los 25 a los 60 años aproximadamente. Igualmente esta elección también es predominante en pasajeros de aproximadamente 70 años o más.

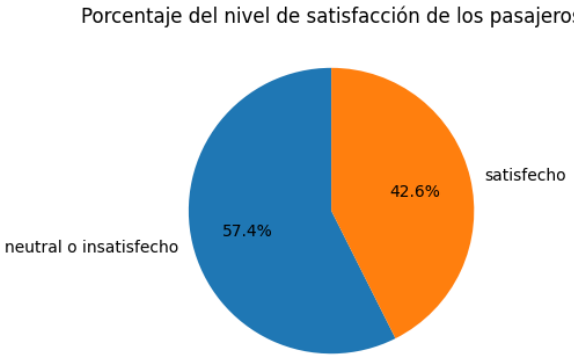
Para finalizar el análisis, se presenta un boxplot en donde se utiliza la variable de las distancias de los vuelos por clase.



Se visualiza que el 50% de los viajes realizados en clase Business varían entre 600 y 2500 km aproximadamente. En el mismo no observan valores atípicos. Para la clase Económica Plus y Económica, el 50% de las distancias de vuelo varían entre 400 y 1000 km aproximadamente; y se observan valores outliers. Es decir que para ambas clases el comportamiento de las observaciones es similar.

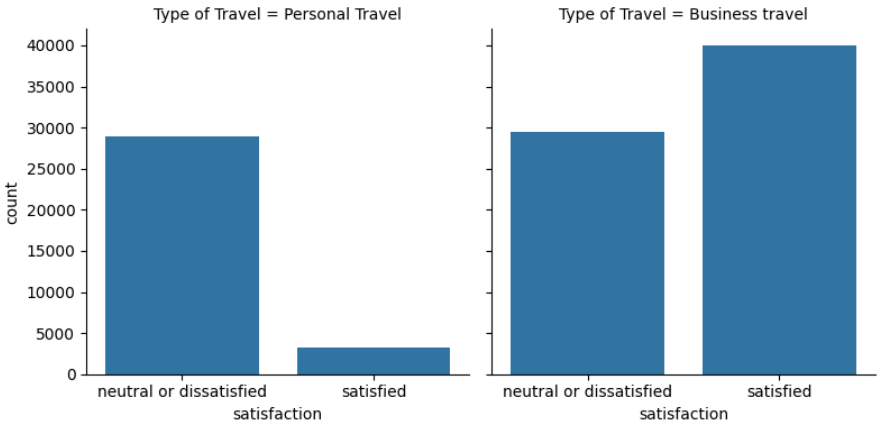
5.4. ¿Qué nivel de satisfacción tienen los pasajeros?

Para iniciar el análisis se realiza un gráfico de torta para obtener el porcentaje de pasajeros con los niveles de satisfacción "satisfecho" y "neutral o insatisfecho".



Se observa en el gráfico que el 42.6% de los pasajeros se encuentra "satisfecho" en su nivel de satisfacción.

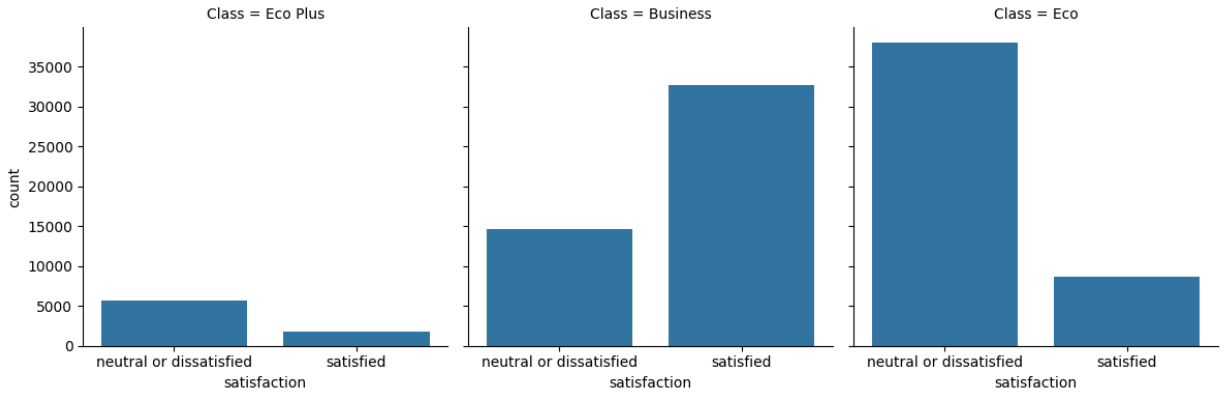
Luego para seguir con el análisis se determina qué tipo de viaje realiza cada pasajero y su nivel de satisfacción.



Se verifica en el gráfico que los pasajeros que viajaron por motivo personal, en su gran mayoría tienen el nivel de satisfacción "neutral o insatisfecho". En cambio los pasajeros que

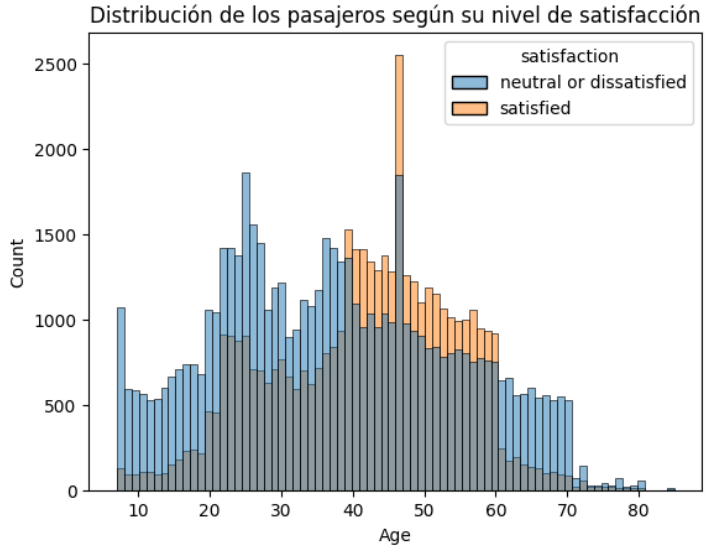
viajaron por motivos de trabajo, mayoritariamente se encuentran "satisfechos", pero el nivel "neutral o insatisfecho" también es alto.

Continuando con el análisis, se realiza una apertura del nivel de satisfacción según la clase en la que viajaron los pasajeros encuestados.



Se visualiza que los pasajeros con nivel de satisfacción "neutral o insatisfecho", la mayoría volaron en clase Económica. En cambio los pasajeros cuyo nivel de satisfacción es "satisfecho", en su mayoría viajaron en clase Business.

Para complementar el análisis anterior, se realiza un histograma para visualizar la distribución de la edad de los pasajeros encuestados según el nivel de satisfacción.



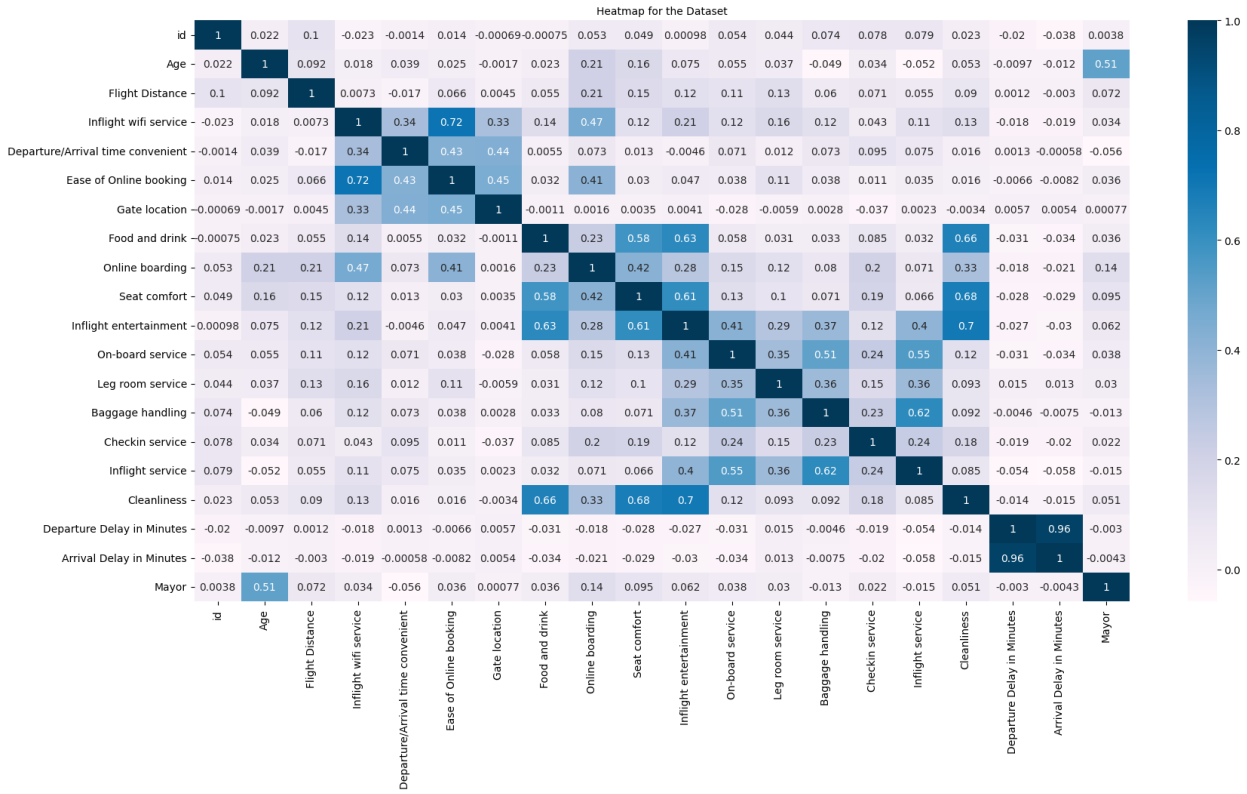
Se puede observar una gran cantidad de neutrales o insatisfechos con respecto a su nivel de satisfacción, pero en el rango de 39 a 60 años el nivel de satisfacción "satisfecho" es mucho mayor que el "neutral o insatisfecho".

5.5. ¿Hay relaciones entre las variables del dataframe?

Como el dataset posee variables numéricas y categóricas (ordinales y nominales). Para iniciar con el análisis de las relaciones entre estas variables se plantea un heatmap del dataset. Antes de analizar el gráfico, se establece que las únicas variables numéricas son:

- **Age**
- **Flight Distance**
- **Departure Delay in Minutes**
- **Arrival Delay in Minutes**

Teniendo en cuenta estas variables, el heatmap permite determinar el nivel de relación entre variables.



En el caso de la relación entre **Age** y **Flight Distance** el coeficiente es bajo, por lo que indica que no hay una fuerte relación entre ambas variables. Por el contrario sucede entre las variables **Departure Delay in Minutes** y **Arrival Delay in Minutes**, donde el coeficiente es cercano a 1, lo que indica una fuerte relación entre ambas.

En el caso de variables categóricas nominales, podemos realizar una tabla de contingencia para poder establecer la existencia de relación entre las siguientes variables binarias como:

- **Type of Travel**
- **satisfaction**

satisfaction	neutral or dissatisfied	satisfied
Type of Travel		
Business travel	29405	39995
Personal Travel	28940	3273

Analizando el coeficiente phi, se puede establecer que algun tipo de relación negativa existe entre ambas variables, no es una relación perfectamente negativa ya que su valor debería ser -1.

```
-0.44670611635665197
```

También podemos realizar una tabla de contingencia para establecer alguna relación entre las variables:

- **Gender**
- **satisfaction**

satisfaction	neutral or dissatisfied	satisfied
Gender		
Female	29917	21644
Male	28428	21624

Analizando el coeficiente phi, se puede establecer que casi NO existe relación entre ambas variables ya que el valor del coeficiente es cercano a 0. Es decir que es indistinto el género del pasajero teniendo en cuenta su nivel de satisfacción.

```
0.012391867291117433
```

5.6. Insights

- El promedio de los pasajeros tiene 39 años aproximadamente.
- El 50% de los pasajeros que volaron tienen entre 28 y 51 años aproximadamente.
- El 92.3% de los pasajeros es mayor de 18 años.
- En promedio los pasajeros encuestados viajaron 1128 km.
- El 50% de los pasajeros encuestados volaron distancias entre 450 y 1700 km aproximadamente.
- Para los mayores el promedio de las distancias de los vuelos es de 1148 km aproximadamente. A diferencia de los menores que este promedio es menor y es 899 km aproximadamente.
- La clase en la que mayormente viajaron los pasajeros encuestados es la de Business con un 46.7%. Seguida por la clase Económica con un 46% y finalmente la clase Económica Plus con un 7.4%.
- La clase Business es la más elegida por los pasajeros del rango de edad que va de los 25 a los 60 años aproximadamente.
- El 50% de los viajes realizados en clase Business varían entre 700 y 2500 km aproximadamente.
- El 50% de las distancias de vuelo varían entre 300 y 1000 km aproximadamente para la clase Económica Plus y Económica.
- El 42.6% de los pasajeros se encuentra "satisfecho" en su nivel de satisfacción.
- Los pasajeros que viajaron por motivo personal, en su gran mayoría tienen el nivel de satisfacción "neutral o insatisfecho".
- Los pasajeros que viajaron por motivos de trabajo, mayoritariamente se encuentran "satisfechos".
- En el rango de 39 a 60 años el nivel de satisfacción "satisfecho" es mucho mayor que el "neutral o insatisfecho".
- Los pasajeros con nivel de satisfacción "neutral o insatisfecho", la mayoría volaron en clase Económica.
- Los pasajeros cuyo nivel de satisfacción es "satisfecho", en su mayoría viajaron en clase Business.

6. Machine Learning model

Teniendo en cuenta que el dataset con el que venimos trabajando tiene un total de 25 variables, de las cuales 24 son independientes, del total se seleccionan aquellas determinantes para predecir el target.

Como el dataset ya posee la "etiqueta" en la variable **satisfaction** ("satisfecho" o "neutral o insatisfecho") quiero aplicar un problema de clasificación en el cual se utiliza el método de machine learning para predecir la clase más probable, ya que la variable es de tipo categórica. Se utilizarán la mayoría de las variables, pero las variables categóricas es necesario que se transformen en variables numéricas (dummies) para que los modelos funcionen (Árbol de decisión y Random Forest).

Antes de aplicar alguno de los algoritmos seleccionados, se transformaron las variables categóricas **Gender**, **Customer Type**, **Type of Travel** y **Class** en variables numéricas (dummies).

Además, como la fuente de datos de donde se extrajo el dataset posee una base de datos para train y una para test, el dataset de train se utiliza por completo para los modelados.

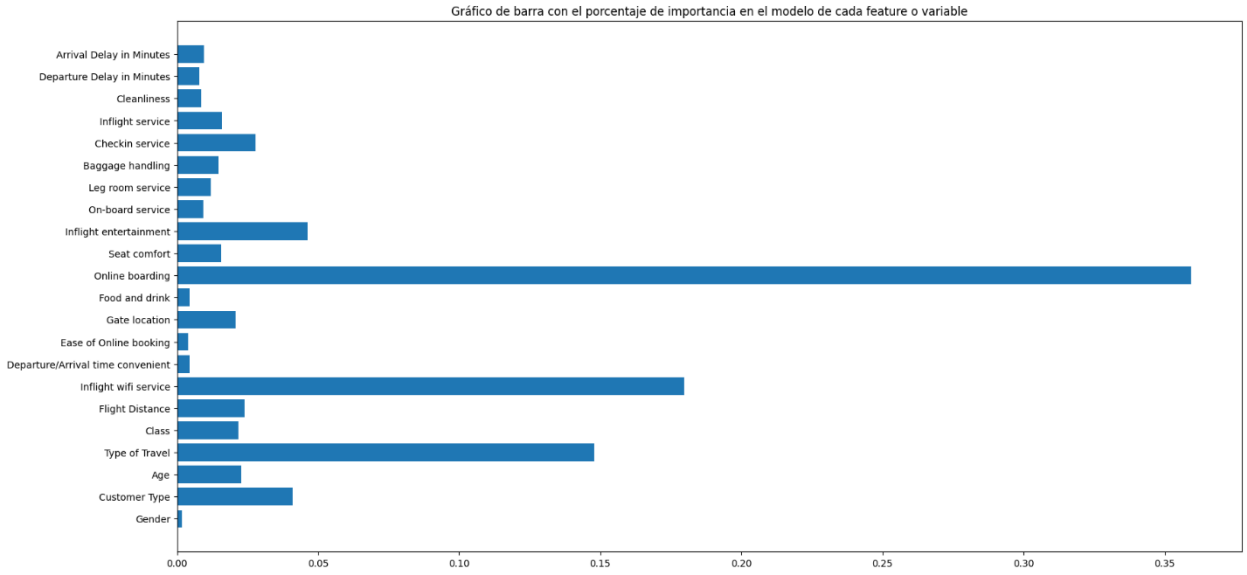
En el dataset de train, inicialmente se elimina la variable etiqueta a predecir. Luego, se eliminan de la misma aquellas variables innecesarias (**id** y **mayor**) y finalmente, se seleccionan las etiquetas.

En el dataset de test se realizan los ajustes necesarios (realizado en el dataset de train) para utilizarlo en el entrenamiento de los algoritmos.

6.1. Decision Tree

Se entrena el modelo utilizando los datos de train y se utiliza el método kfold para obtener el accuracy del árbol.

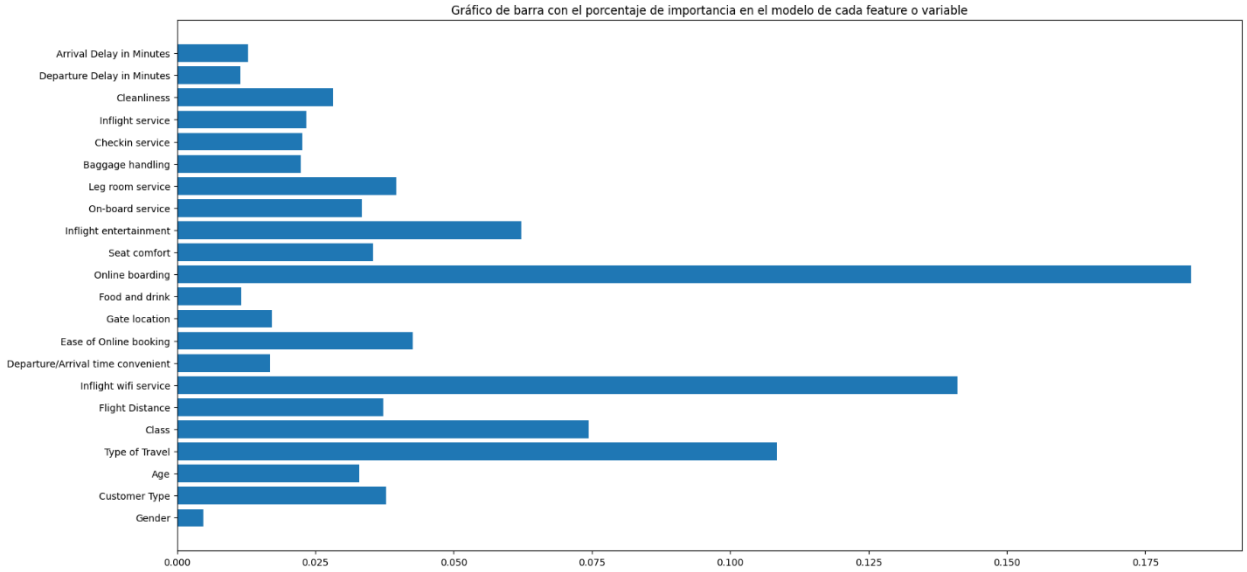
Luego, se determina a través de un gráfico de barra qué variables del modelo son las más importantes.



Teniendo en cuenta el gráfico, las features o variables más importantes del modelo son **Online boarding**, **Inflight wifi service** y **Type of Travel** ya que son las que tienen el porcentaje más alto.

6.2. Random Forest

Se entrena el modelo utilizando los datos de train y luego, se determina a través de un gráfico de barra qué variables del modelo son las más importantes.



Teniendo en cuenta el gráfico, las features o variables más importantes del modelo son **Online boarding**, **Inflight wifi service**, **Type of Travel**, **Class** y **Inflight entertainment** ya que son las que tienen el valor más alto.

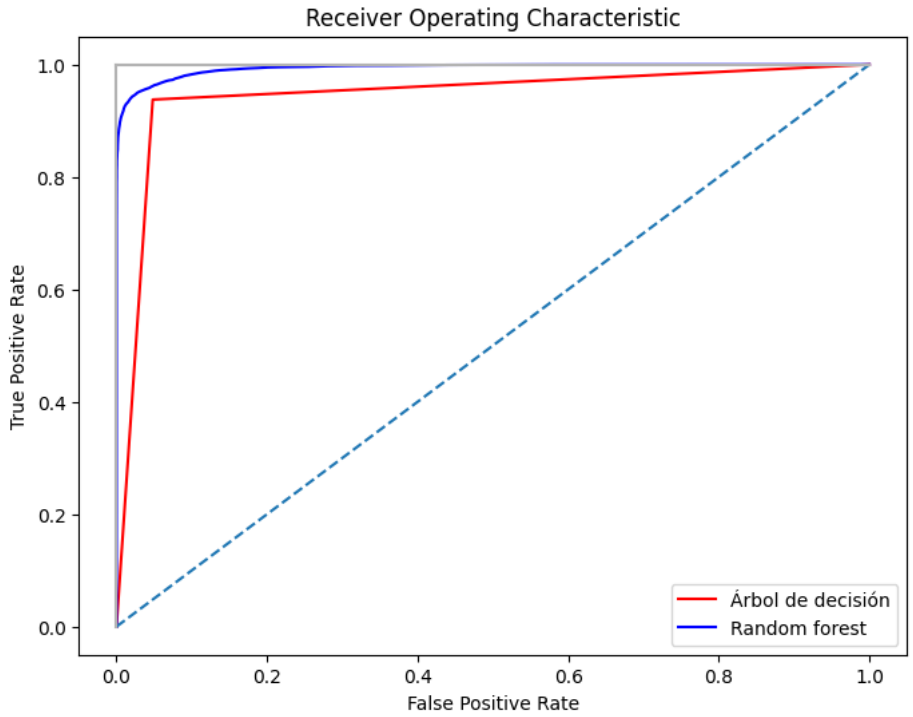
6.3. Métricas

Métrica	**Decision Tree**	**Random Forest**
Accuracy	`0.9456036341238065`	`0.96311980289498`
Matriz de confusión	`array([[13871, 702], [ 711, 10692]])`	`array([[14282, 291], [ 667, 10736]])`
Precisión	`0.9383886255924171`	`0.9736102294368368`
Recall	`0.9376479873717443`	`0.9415066210646321`
F1 score	`0.938018160284248`	`0.9572893446277306`

Analizando las métricas presentadas anteriormente en el cuadro, el modelo de Random Forest tiene mejores números en todas las métricas calculadas en comparación con el Árbol de decisión.

6.4. Curva ROC

Como la curva ROC calcula el área bajo la curva de cada uno de los modelos para identificar cuál es mejor que otro, para iniciar la construcción de la curva, se visualizan los valores de la variable y_test para verificar cómo son. Se observa que y_test estaba codificado como "satisfied" y "neutral or dissatisfied" y para calcular la curva ROC la función necesita recibir 0 y 1. Para continuar se convierte el valor "satisfied" en 1 y el valor "neutral or dissatisfied" en 0 y se crea una lista nueva llamada y_test_num. Luego, se calcula la curva ROC y se grafica la misma.



Finalmente para terminar con este análisis, se calcula la superficie de cada curva para determinar que modelo es superior al otro.

	Decision Tree	Random Forest
Superficie	0.9447383558625002	0.9938235966668552

En este caso, la curva del modelo Random Forest tiene un valor superior al Árbol, por lo cual significa que posee un mejor desempeño.