Predicción de Satisfacción de Vuelos

Lic. Agustín D'Alessandro

Quién Soy

Soy Agustín D'Alessandro.
Egresado del Instituto Superior
del Profesorado "Dr. Joaquín V. González"
como Profesor de Matemática y egresado de
la Universidad CAECE como Licenciado en
Matemática, busco adentrarme en el mundo
de la Ciencia de Datos para llevar a la
práctica la formación matemática con la que
cuento. Esta búsqueda reflota mi formación
básica en programación adquirida durante
mi educación media, habiendo obtenido
el título de Bachiller en Informática.

Datos

```
RangeIndex: 103904 entries, 0 to 103903
Data columns (total 25 columns):
 #   Column                            Non-Null Count   Dtype
---  ------                            --------------   -----
 0   Unnamed: 0                        103904 non-null  int64
 1   id                                103904 non-null  int64
 2   Gender                            103904 non-null  object
 3   Customer Type                     103904 non-null  object
 4   Age                               103904 non-null  int64
 5   Type of Travel                    103904 non-null  object
 6   Class                             103904 non-null  object
 7   Flight Distance                   103904 non-null  int64
 8   Inflight wifi service             103904 non-null  int64
 9   Departure/Arrival time convenient 103904 non-null  int64
 10  Ease of Online booking            103904 non-null  int64
 11  Gate location                     103904 non-null  int64
 12  Food and drink                    103904 non-null  int64
 13  Online boarding                   103904 non-null  int64
 14  Seat comfort                      103904 non-null  int64
 15  Inflight entertainment            103904 non-null  int64
 16  On-board service                  103904 non-null  int64
 17  Leg room service                  103904 non-null  int64
 18  Baggage handling                  103904 non-null  int64
 19  Checkin service                   103904 non-null  int64
 20  Inflight service                  103904 non-null  int64
 21  Cleanliness                       103904 non-null  int64
 22  Departure Delay in Minutes        103904 non-null  int64
 23  Arrival Delay in Minutes          103594 non-null  float64
 24  satisfaction                      103904 non-null  object
dtypes: float64(1), int64(19), object(5)
```

Datos de entrenamiento: train.csv

```
RangeIndex: 25976 entries, 0 to 25975
Data columns (total 25 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   Unnamed: 0                        25976 non-null  int64
 1   id                                25976 non-null  int64
 2   Gender                            25976 non-null  object
 3   Customer Type                     25976 non-null  object
 4   Age                               25976 non-null  int64
 5   Type of Travel                    25976 non-null  object
 6   Class                             25976 non-null  object
 7   Flight Distance                   25976 non-null  int64
 8   Inflight wifi service             25976 non-null  int64
 9   Departure/Arrival time convenient 25976 non-null  int64
 10  Ease of Online booking            25976 non-null  int64
 11  Gate location                     25976 non-null  int64
 12  Food and drink                    25976 non-null  int64
 13  Online boarding                   25976 non-null  int64
 14  Seat comfort                      25976 non-null  int64
 15  Inflight entertainment            25976 non-null  int64
 16  On-board service                  25976 non-null  int64
 17  Leg room service                  25976 non-null  int64
 18  Baggage handling                  25976 non-null  int64
 19  Checkin service                   25976 non-null  int64
 20  Inflight service                  25976 non-null  int64
 21  Cleanliness                       25976 non-null  int64
 22  Departure Delay in Minutes        25976 non-null  int64
 23  Arrival Delay in Minutes          25893 non-null  float64
 24  satisfaction                      25976 non-null  object
dtypes: float64(1), int64(19), object(5)
```

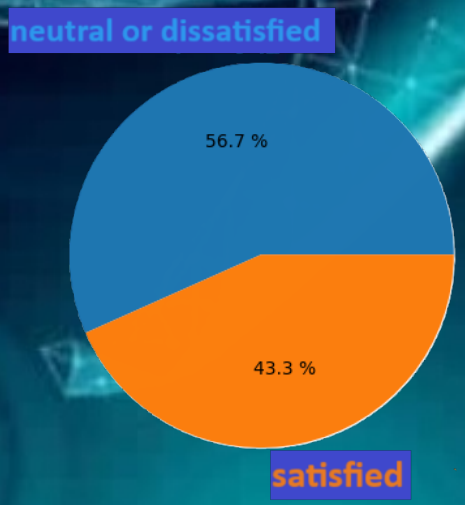
Datos de prueba: test.csv

Datos

Ambos datasets cuentan con las mismas veinticinco columnas, la mayoría (20) de tipos numéricos. La columna *Unnamed: 0* es una repetición del índice del dataframe, por lo que será descartada inmediatamente. La única columna en ambos datasets que tiene valores faltantes es *Arrival Delay in Minutes*, a la que le faltan 310 datos en el conjunto de entrenamiento y 83 en los datos de prueba. Para el primer modelo los datos faltantes serán completados, mientras que para el segundo modelo planteado no será necesario porque no se utilizará esa columna.

Datos

La variable objetivo en los dos modelos que se presentarán en el trabajo es la *satisfacción* de los usuarios. Esto permitirá identificar las variables más relevantes para el servicio y anticipar qué usuarios podrían mostrarse disconformes de antemano para ofrecerles un servicio personalizado que mejore su opinión.

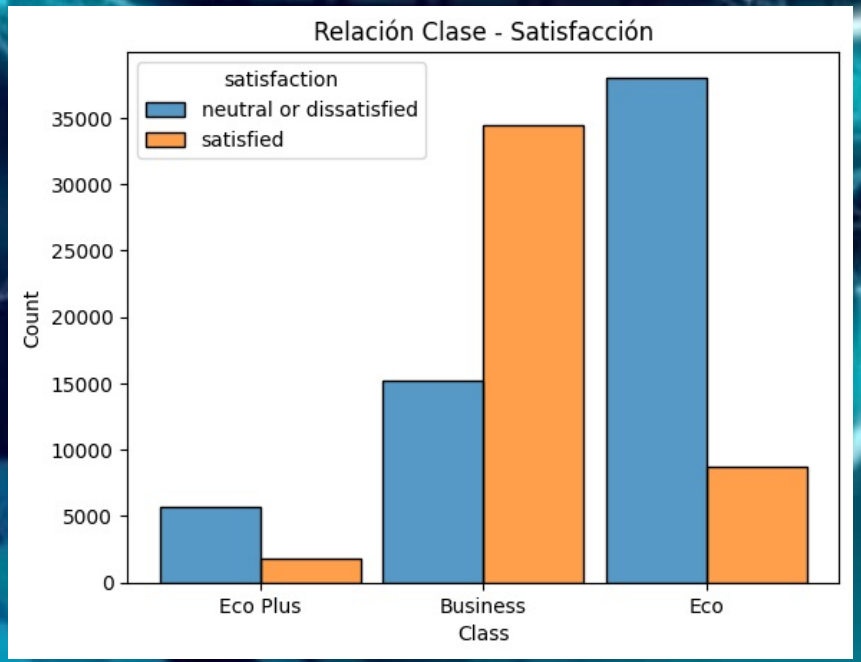


Porcentaje de Satisfacción

Hipótesis

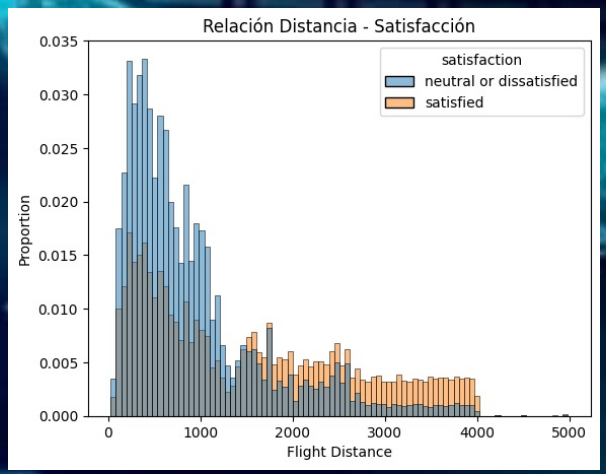
- ¿Es la **clase** un factor influyente en la satisfacción de los pasajeros?
- ¿La **distancia** de los viajes afecta la decisión de la **clase** para volar? ¿Afecta entonces a la satisfacción?
- ¿Qué **rangos etários** acceden a las mejores clases? ¿Termina la edad siendo relevante para la satisfacción de los pasajeros?
- ¿Las **demoras** logran que los pasajeros se muestren disconformes con el servicio?
- ¿Las **valoraciones promedio** son indicativos relevantes a la hora de medir la satisfacción de los usuarios?

EDA



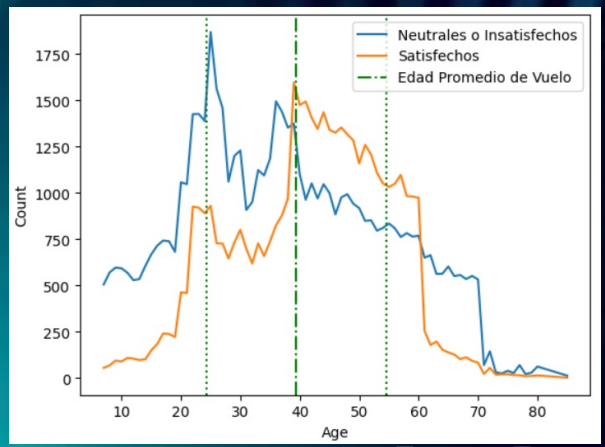
Hay una marcada relación entre las clases y la satisfacción de los pasajeros. La gran mayoría de los pasajeros satisfechos han volado en clase *Business*, que además es la única clase en la que hay más pasajeros satisfechos que insatisfechos.

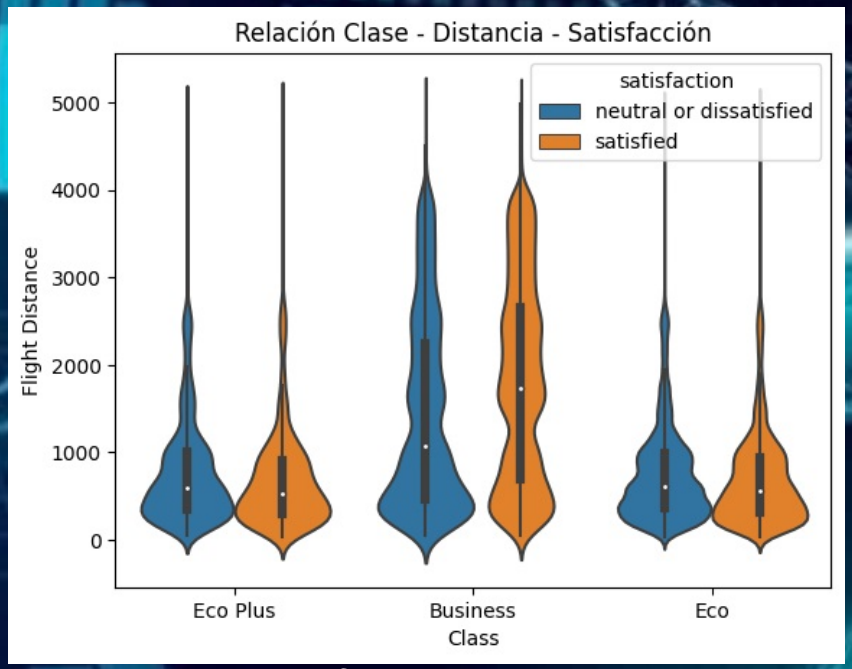
EDA

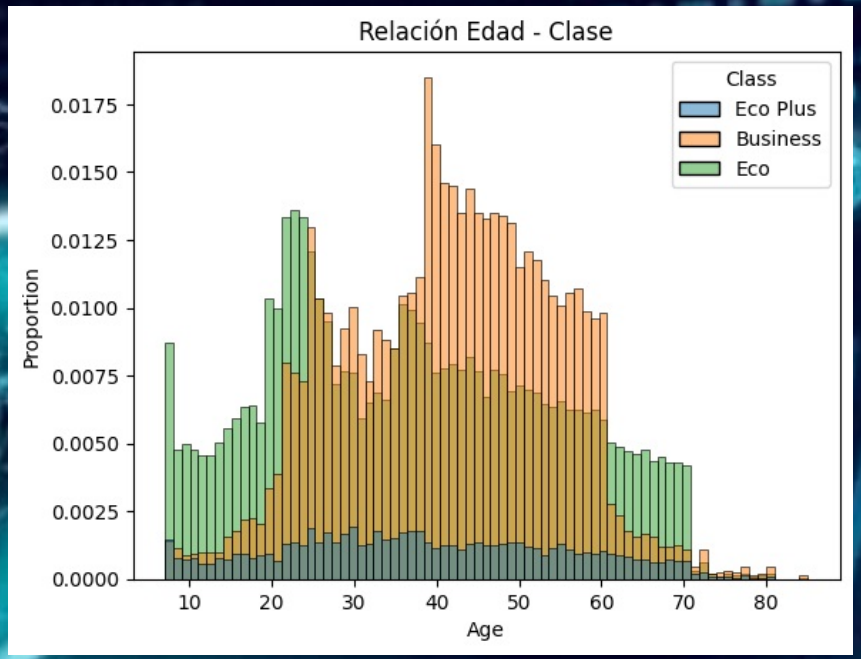


Vemos que a menor distancia hay mayor porcentaje de pasajeros insatisfechos mientras que a mayor distancia hay mayor porcentaje de usuarios satisfechos.

Un efecto similar pareciera apreciarse en la relación entre las edades y la satisfacción de los pasajeros.







Ambos gráficos muestran como las distancias y edades donde se tiene la mayor cantidad de pasajeros satisfechos coincide con las distancias y edades de pasajeros que vuelan mayormente en clase *Business*. Esto explica los aumentos en la satisfacción de los usuarios en estas franjas.

En el mapa de calor se pueden apreciar las correlaciones entre las variables del dataset. Pocas muestran un alto grado de correlación con la variable objetivo. Las más destacables son la *Class*, *Type of Travel*, *Online boarding*, *Inflight wifi service*.

Modelos

Algoritmos Empleados

Para el primer modelo se sometieron las variables a varios procesos de *Feature Selection*. Las variables no fueron escaladas ni normalizadas. Para este primer análisis se utilizó:

- Decision Tree

Para el segundo modelo se seleccionaron únicamente las variables identificables de manera previa al vuelo y se hizo un escalado y normalizacion de las mismas. Los modelos para este análisis fueron:

- Random Forest
- Support Vector Machine
- Logistic Regresion
- K-Nearest Neighbors

Modelos

En ambos modelos se toma en cuenta la *Especificidad* por la importancia que tiene para el problema el detectar aquellos usuarios que queden insatisfechos.
Las métricas obtenidas para el primer modelo, con los distintos métodos de *Feature Selection* son:

	Stepwise	Forward	Backward
Especificidad	0,94	0,94	0,88
Exactitud	0,88	0,88	0,89

Se puede ver que los modelos arrojan los mejores resultados con las variables obtenidas con los procesos de *Stepwise Selection* o *Forward Selection*.

Modelos

Con las variables seleccionadas se hizo el entrenamiento del modelo y se pasó a hacer las predicciones sobre el dataset *test.csv*. Los resultados obtenidos son:

	Modelo 1
Especificidad	0,9355
Exactitud	0,8749

Se puede apreciar que al pasar los nuevos datos por el modelo los resultados obtenidos tuvieron métricas altas, ambas por encima del 85%.

Modelos

Las métricas obtenidas con las variables previas al vuelo, luego del tuneo de hiperparámetros, son:

	Random Forest	Support Vector Machine	Logistic Regresion	K-Nearest Neighbors
Especificidad	0,9232	0,8455	0,8455	0,8844
Exactitud	0,8602	0,8286	0,8286	0,8711

El modelo con mejores resultados es el *Random Forest*, con una especificidad del $92,32\%$ y una exactitud del $86,02\%$.

Modelos

Con el modelo y los hiperparámetros escogidos se hizo el entrenamiento del modelo y se pasó a hacer las predicciones sobre el dataset *test.csv*. Los resultados obtenidos son:

	Modelo 2
Especificidad	0,9195
Exactitud	0,8572

Se puede apreciar que al pasar los nuevos datos por el modelo los resultados obtenidos tuvieron métricas altas, muy similares a las de los datos de entrenamiento.

Conclusiones

En base a lo analizado en estos datasets se puede concluir lo siguiente:

- Las variables con mayor influencia en la satisfacción parecen ser: *Gender*, *Customer Type*, *Age*, *Class*, *Flight Distance*, *Inflight wifi service*, *Gate location*, *Departure Delay in Minutes*, *Mean Satisfaction*.
- El modelo de *Random Forest* permite predecir con una alta especificidad la satisfacción de los usuarios en base a variables que pueden ser analizadas antes del vuelo, lo que le da la posibilidad a la aerolínea para compensar el servicio previo durante el vuelo.