RECUPERACIÓN 2 EVALUACIÓN

SISTEMAS DE APRENDIZAJE AUTOMÁTICO

Andrei Alexandru Miu

Índice

Descripción de los datos

El dataset contiene un total de 401 registros (incluyendo cabecera) de clientes/pasajeros de aerolíneas, un total de 23 columnas, donde:

- **gender** – Género. Variable categórica.
- **customer_type** – Tipo de cliente. Variable categórica.
- **age** – Edad. Variable numérica.
- **type_of_travel** – Tipo de viaje. Variable categórica.
- **customer_class** – Clase de vuelo del cliente. Variable categórica.
- **flight_distance** – Distancia del vuelo. Variable numérica.
- **inflight_wifi_service** – Servicio de Wi-Fi a bordo. Variable numérica.
- **departure_arrival_time_convenient** – Comodidad del horario de salida/llegada. Variable numérica.
- **ease_of_online_booking** – Facilidad para reservar online. Variable numérica.
- **gate_location** – Ubicación de la puerta de embarque. Variable numérica.
- **food_and_drink** – Comida y bebida. Variable numérica.
- **online_boarding** – Embarque en línea. Variable numérica.
- **seat_comfort** – Comodidad del asiento. Variable numérica.
- **inflight_entertainment** – Entretenimiento a bordo. Variable numérica.
- **onboard_service** – Servicio a bordo. Variable numérica.
- **leg_room_service** – Espacio para las piernas. Variable numérica.
- **baggage_handling** – Manejo del equipaje. Variable numérica.
- **checkin_service** – Servicio de check-in. Variable numérica.
- **inflight_service** – Servicio durante el vuelo. Variable numérica.
- **cleanliness** – Limpieza. Variable numérica.
- **departure_delay_in_minutes** – Retraso en la salida (en minutos) . Variable numérica.
- **arrival_delay_in_minutes** – Retraso en la llegada (en minutos) . Variable numérica.
- **satisfaction** – Satisfacción. Variable categórica.

Datos inconsistentes y atípicos

Dato/s inconsistente/s:
En la fila 213, falta un valor en arrival_delay_in_minutes, ya que la fila está vacía.
Valor/es atípico/s:
En la fila 280, el valor 209 en departure_delay_in_minutes.

Codificación binaria

Se podrían analizar pares de variables como:

- Género vs Satisfacción
- Tipo de viaje y Satisfacción
- Retraso en salida (>15min) y Satisfacción
- Género y Edad

En el dataset se podría cambiar o interpretar de la siguiente manera:
En la columna de Género:

- Hombre: Valor 1
- Mujer: Valor 0

En la columna de Satisfacción:

- Satisfecho: Valor 1
- No satisfecho/Neutro: Valor 0

En la/s columna/s de Tipo de viaje:

- Business Travel: Valor 1
- Personal Travel: Valor 0

Estas son algunas de las columnas que podríamos cambiar o interpretar de otra manera, codificándolos así.

Matriz de confusión y cuartiles

Haremos una matriz de confusión de Genero y Edad, donde dividiremos la Edad en 2 grupos:

- Edad <= 40 -> Grupo joven
- Edad > 40 -> Grupo mayor

Así, podemos comparar la columna de Género con la Edad de forma binaria.

		Edad	
		Grupo Joven (Positivo)	Grupo Mayor (Negativo)
Género	Hombre	VP = 102	FN = 85
	Mujer	FP = 111	VN = 103

Interpretación de la tabla:

- **VP (Verdaderos Positivos):** Hombres jóvenes (≤ 40 años).
- **FN (Falsos Negativos):** Hombres mayores (> 40 años).
- **FP (Falsos Positivos):** Mujeres jóvenes (≤ 40 años).
- **VN (Verdaderos Negativos):** Mujeres mayores (> 40 años).

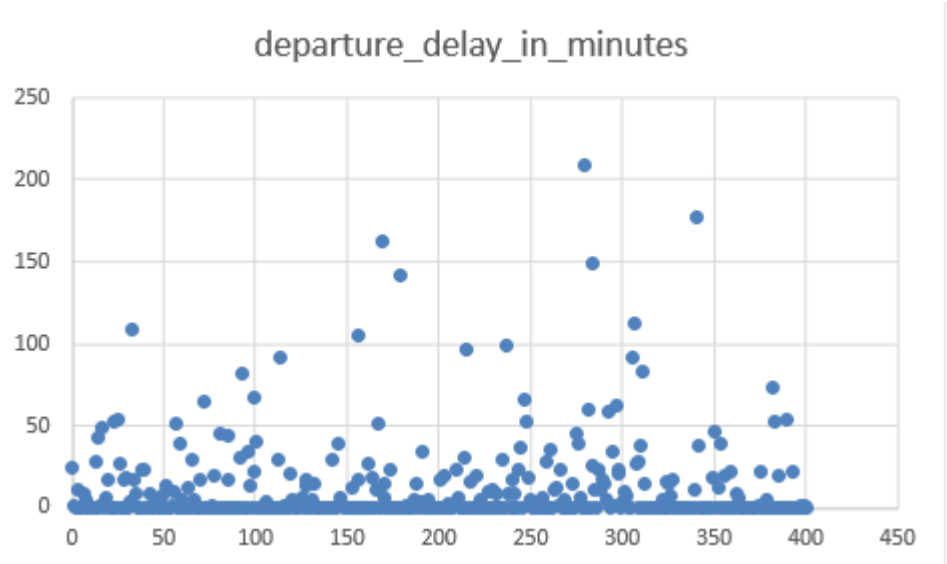
Cálculo de métricas

Exactitud (Accuracy):
(VP + VN) / (VP + VN + FN + FP) = (102 + 103) / (102 + 103 + 111 + 85) = 51.12%
Es el porcentaje de predicciones correctas. En este caso un 51.12% no es que sea un modelo muy fiable.
Precisión:
VP / (VP + FP) = 102 / (102 + 111) = 47.89%
De todas las veces que el modelo predijo la clase positiva, el 47.89% fueron correctas.
Sensibilidad (Recall):
VP / (VP + FN) = 102 / (102 + 85) = 54.55%
De todos los casos realmente positivos, el modelo detectó correctamente el 54.55%.
Especificidad:
VN / (VN + FP) = 103 / (103 + 111) = 48.13%
Indica qué tan bien el modelo detecta los negativos. En este caso un 48.13% es que hay muchos falsos positivos.
F1-Score:
2 x ((Precisión y Recall) / (Precisión+Recall)) = 51.06%
Esta métrica es útil cuando hay clases desbalanceadas o cuando queremos equilibrar precisión y recall.
Tasa de Error:
(FN + FP) / (VP + VN + FN + FP) = (111 + 85) / (102 + 103 + 111 + 85) = 48.88%
Porcentaje de predicciones incorrectas.
Prevalencia:
(VP + FN) / (VP + VN + FN + FP) = (102 + 85) / (102 + 103 + 111 + 85) = 46.63%
Proporción de casos positivos reales en el dataset.
Índice de Jaccard:
(VP + VN) / ((VP + VN + FN + FP) + (FN + FP)) = (102 + 103) / ((102 + 103 + 111 + 85) + (111 + 85)) = 34.34%
Es similar al F1-Score pero penaliza más los errores.
Tasa de Falsos Positivos (FPR):
FP / (VN + FP) = 75 / (75 + 75) = 51.87%
Indica el porcentaje de negativos mal clasificados como positivos. En este caso es un 51.87%, un porcentaje bastante alto.

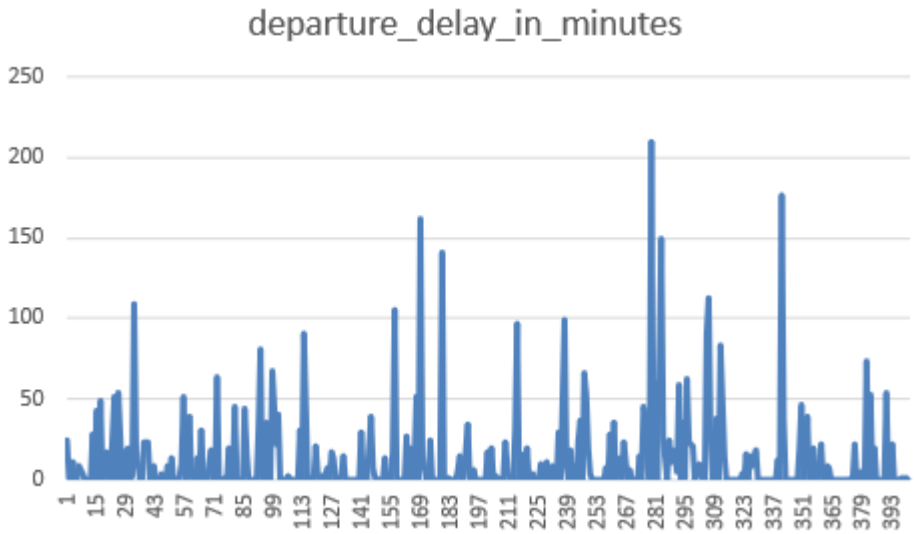
Tasa de Falsos Negativos (FNR):
FN / (VP + FN) = 85 / (102 + 85) = 45.45%
Indica el porcentaje de positivos mal clasificados como negativos, el cual también es alto.

Cálculo de cuartiles:
En base a la edad (age):
Valor mínimo: 8 años
Valor máximo: 72 años
Q1: 8 años
Q2: 39 años
Q3: 50 años
Q4: 72 años
IQR (Q3 - Q1): 50 – 8 = 42
El IQR es una medida estadística que sirve para saber qué tan dispersos están los datos en el centro de un conjunto.
Límite inferior: Q1 - 1.5 x IQR = 8 - 1.5 x 42 = -55
Límite superior: Q3 + 1.5 x IQR = 50 + 1.5 x 42 = 113

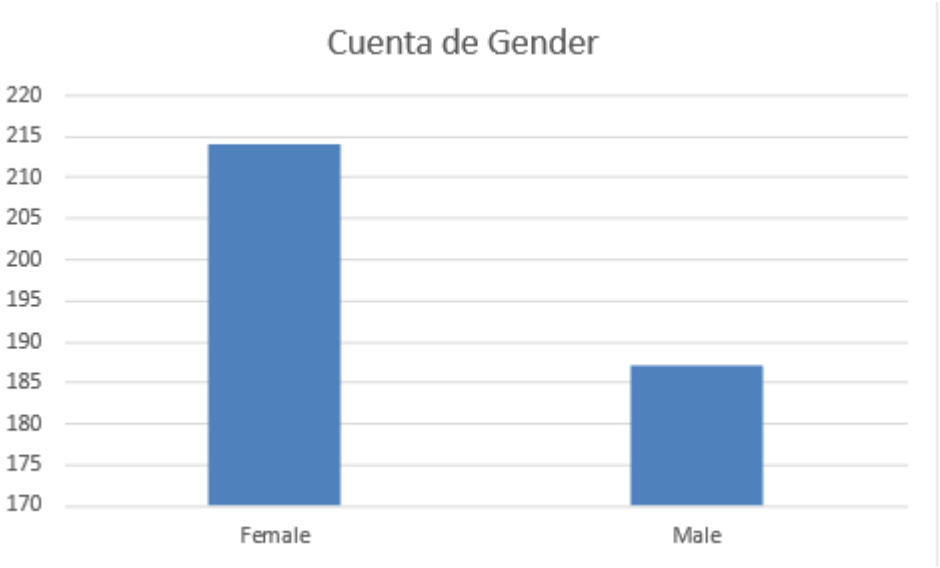
Gráficas



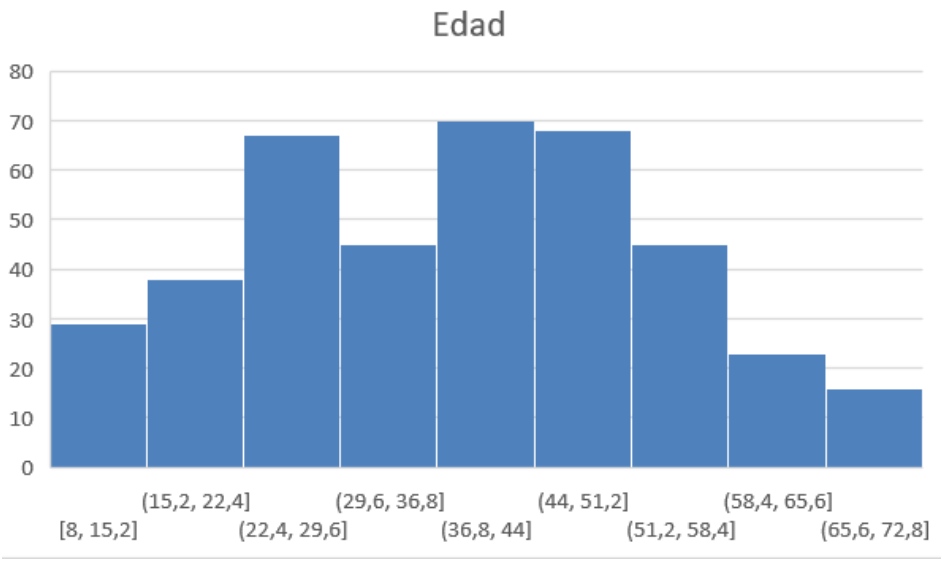
En la siguiente gráfica de dispersión muestra la variable departure_delay_in_minutes (retraso en la salida en minutos), donde se ve el valor atípico mencionado anteriormente (209)



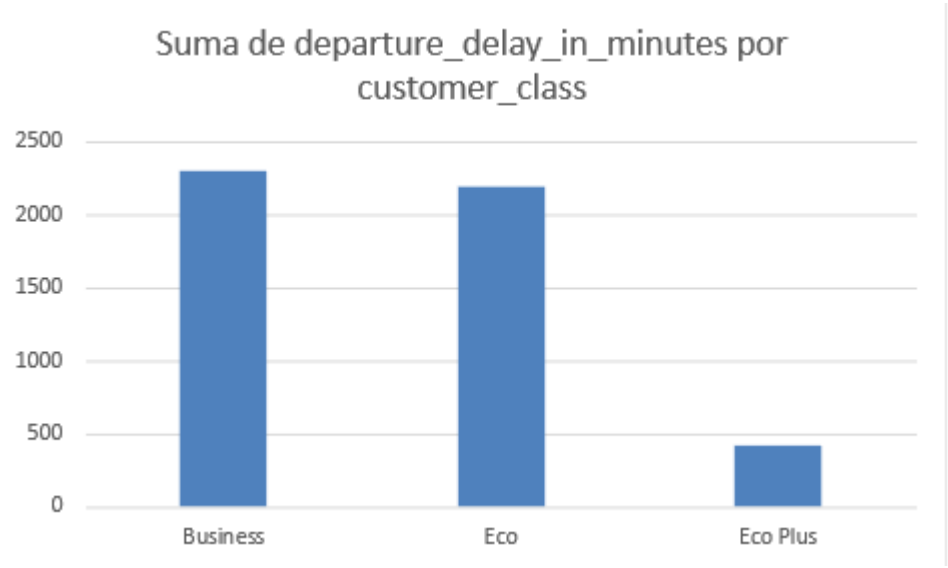
En este gráfico de líneas también se puede observar lo mismo.



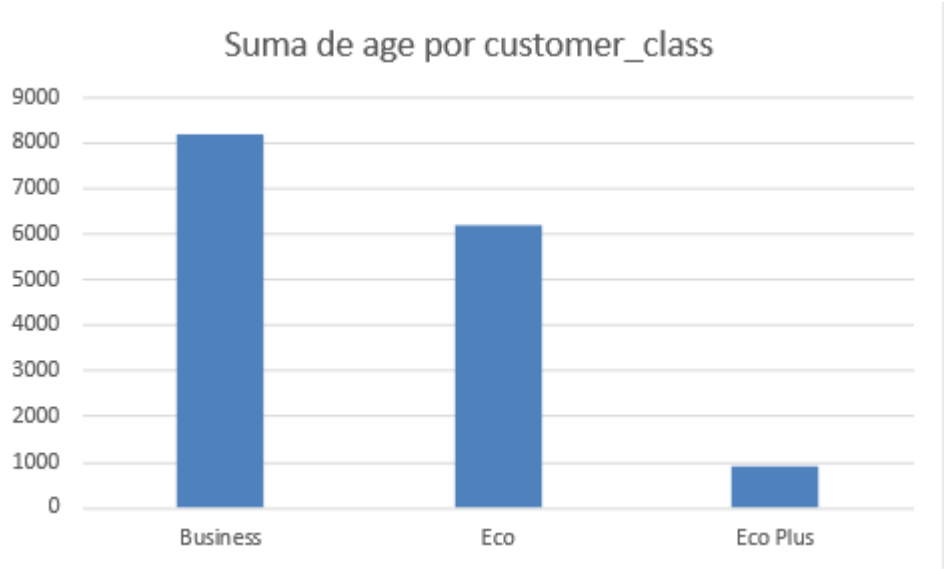
En este gráfico de barras se muestra la cantidad de hombres y mujeres que hay en el dataset.



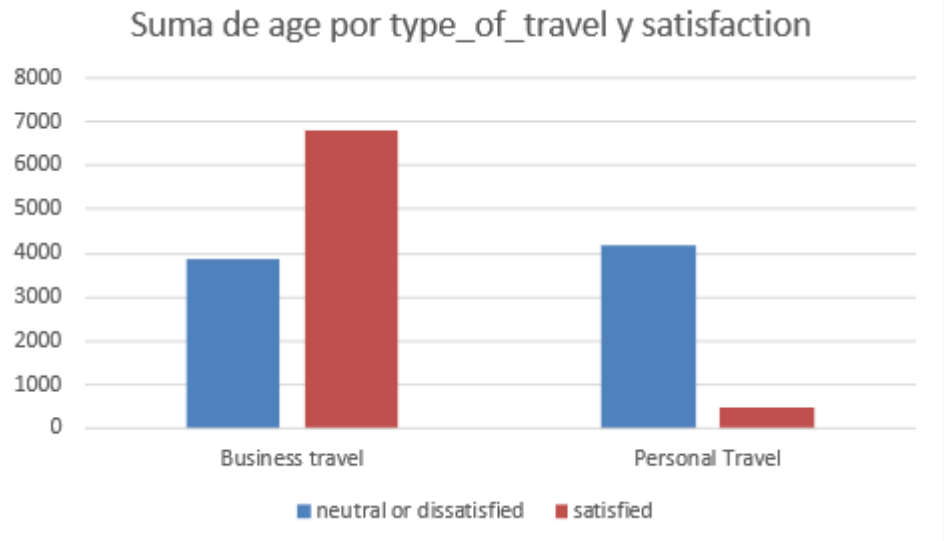
En este histograma, se muestra la distribución de la edad de los pasajeros agrupada en rangos.



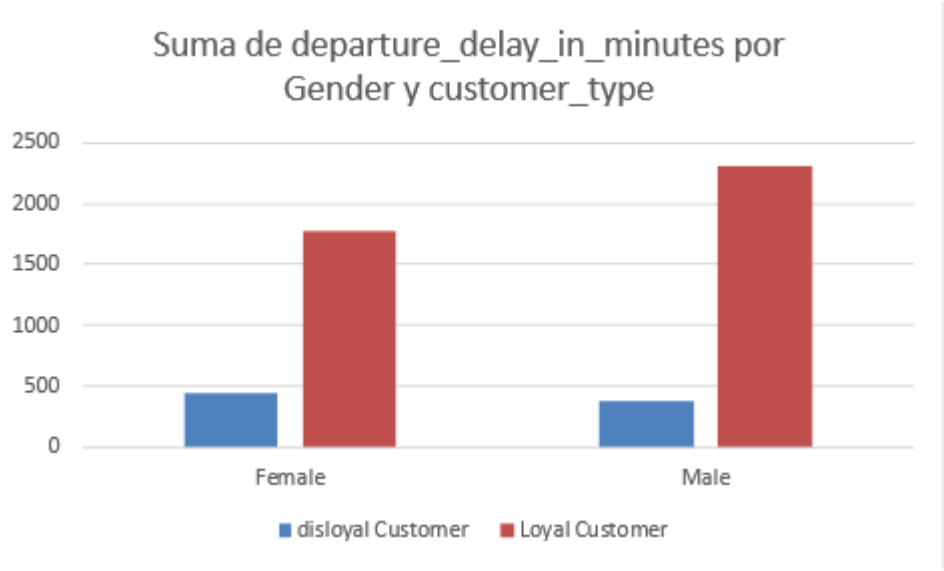
En la siguiente gráfica representa la suma total de los retrasos en la salida (departure_delay_in_minutes) agrupados por clase de viaje (customer_class).



En la siguiente grafica de barras representa la suma de edades (age) de los pasajeros, agrupada por la columna customer_class (clase de viaje).



En la siguiente gráfica de barras representa la suma de edades (age) agrupada por dos columnas: type_of_travel (tipo de viaje) y satisfaction (satisfacción).



En la siguiente gráfica de barras muestra la suma de los minutos de retraso en la salida de vuelos (departure_delay_in_minutes) agrupada por dos variables: género (gender) y tipo de cliente (customer_type).