Giordanelli, Ocampo, Olivera

TRABAJO PRÁCTICO FINAL BIG DATA

Airplane satisfaction

Encontramos un conjunto de datos de una encuesta de satisfacción de pasajeros. Donde las variables clave son:

- género
- edad
- tipo de viaje
- clase de vuelo
- servicios a bordo
- retrasos
- satisfacción general.

Definición del problema

¿Cuáles son los factores más influyentes en la satisfacción de los pasajeros? y ¿cómo podemos predecir si un pasajero estará satisfecho?

"Mejorar la satisfacción de sus pasajeros para aumentar la fidelidad y retención de clientes."

Análisis exploratorio de datos

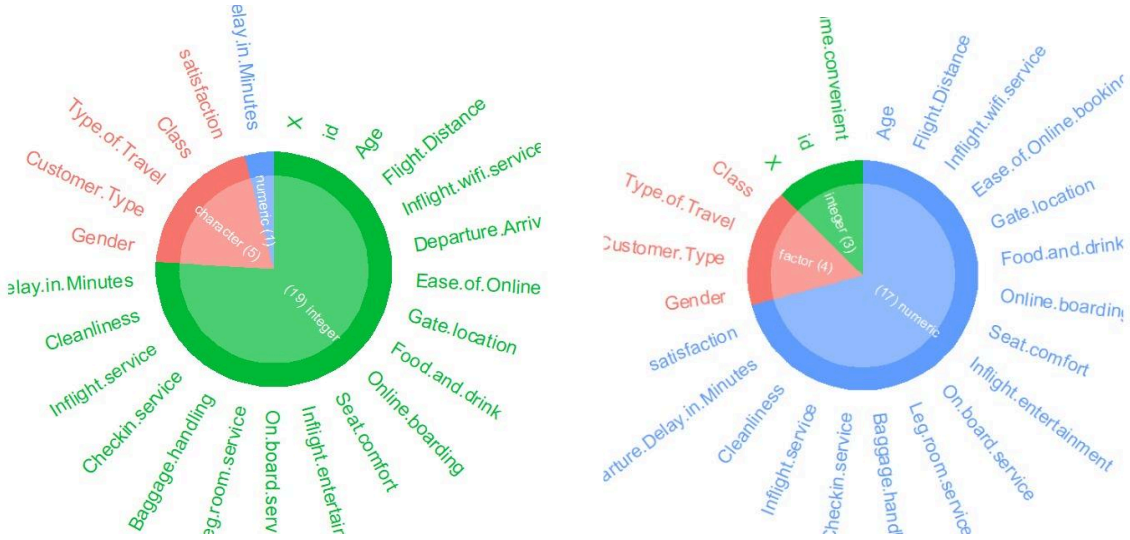


Fig 1. En la izquierda tenemos los datos como vinieron dados, siendo la mayoría integer. Por la derecha como nosotras decidimos categorizarlos.

Fig 2. Representa el nivel de satisfacción general de los pasajeros, siendo menor la cantidad de satisfechos con solo 43.4%.

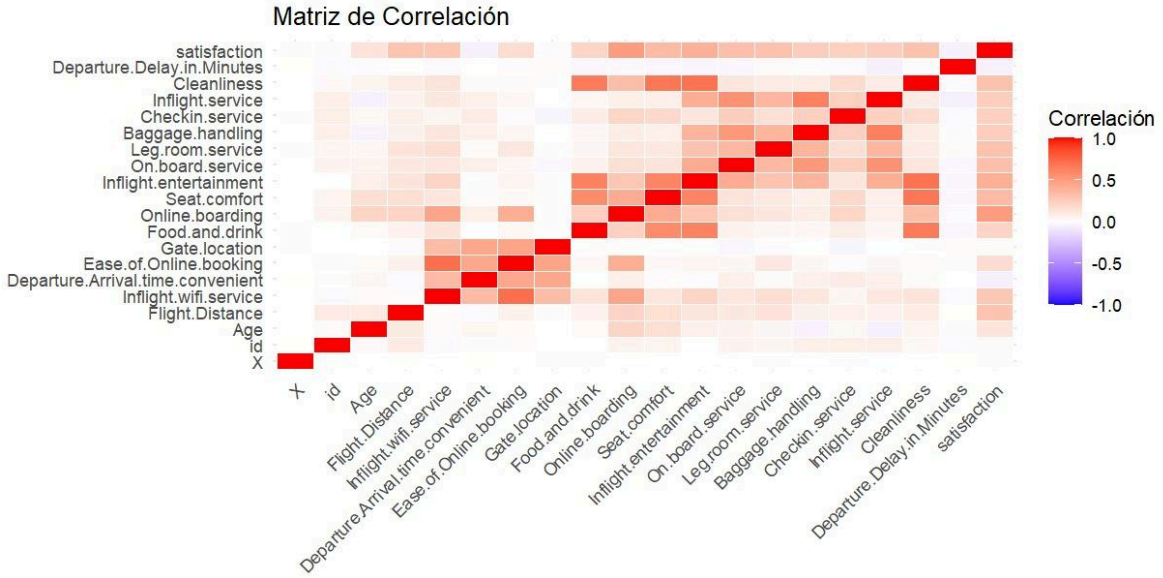


Fig 3. Tenemos una matriz de correlación, que representa la relación entre las distintas variables del dataset.

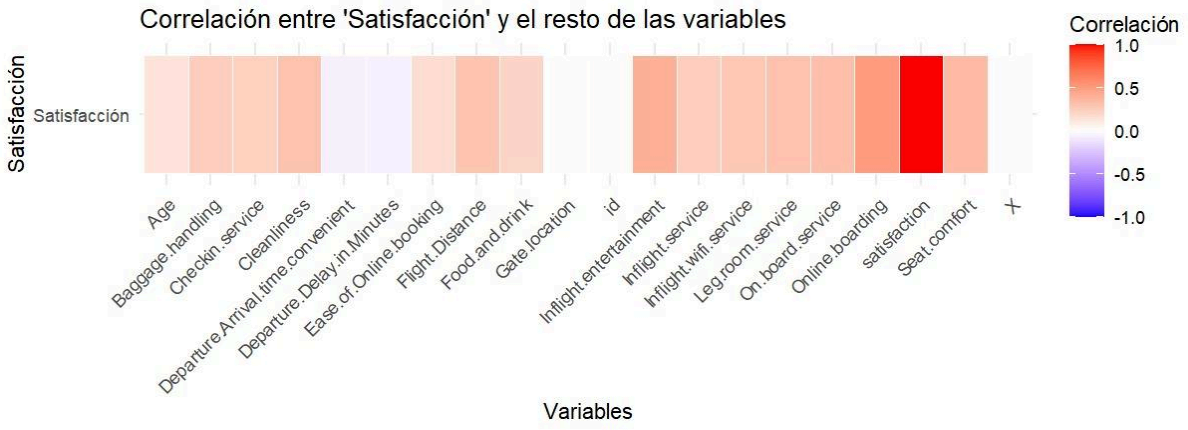


Fig 4. Tenemos la correlación entre satisfacción y el resto de las variables. Siendo las más relevantes: Online boarding, Inflight entertainment y Seat confort.

Hipótesis

- **Edad del cliente asociada con nivel de satisfacción:** Clientes mayores muestran una tendencia a reportar mayor satisfacción.
- **La clase de servicio influye en la satisfacción del cliente:** Clientes en Business son los más satisfechos. Esto sugiere que invertir en mejoras en las clases Economy podría incrementar la satisfacción.
- **La frecuencia de satisfacción es mayor en clientes insatisfechos:** Hay un predominio de clientes insatisfechos o neutrales, lo que indica que otros factores podrían estar generando esta tendencia y deberían investigarse para mejorar la satisfacción general.

Gráficos

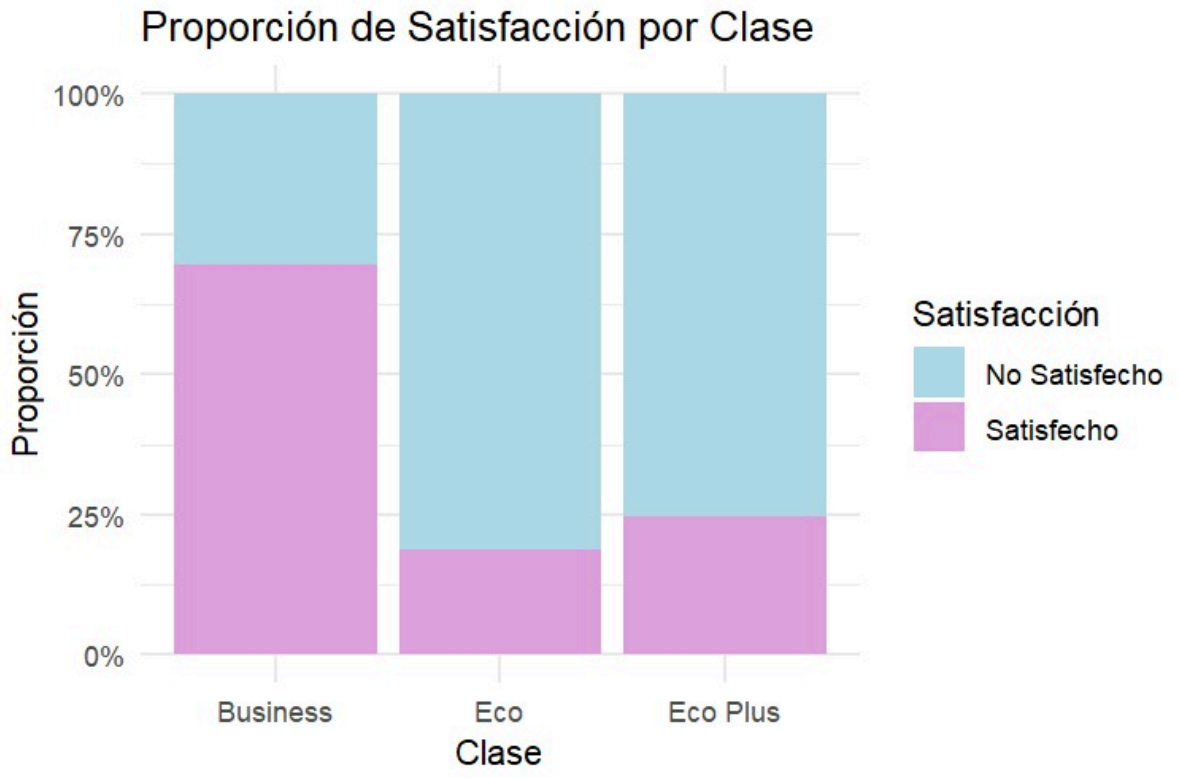


Fig 5. Con este histograma podemos ver como la clase de servicio influye con la satisfacción del cliente.

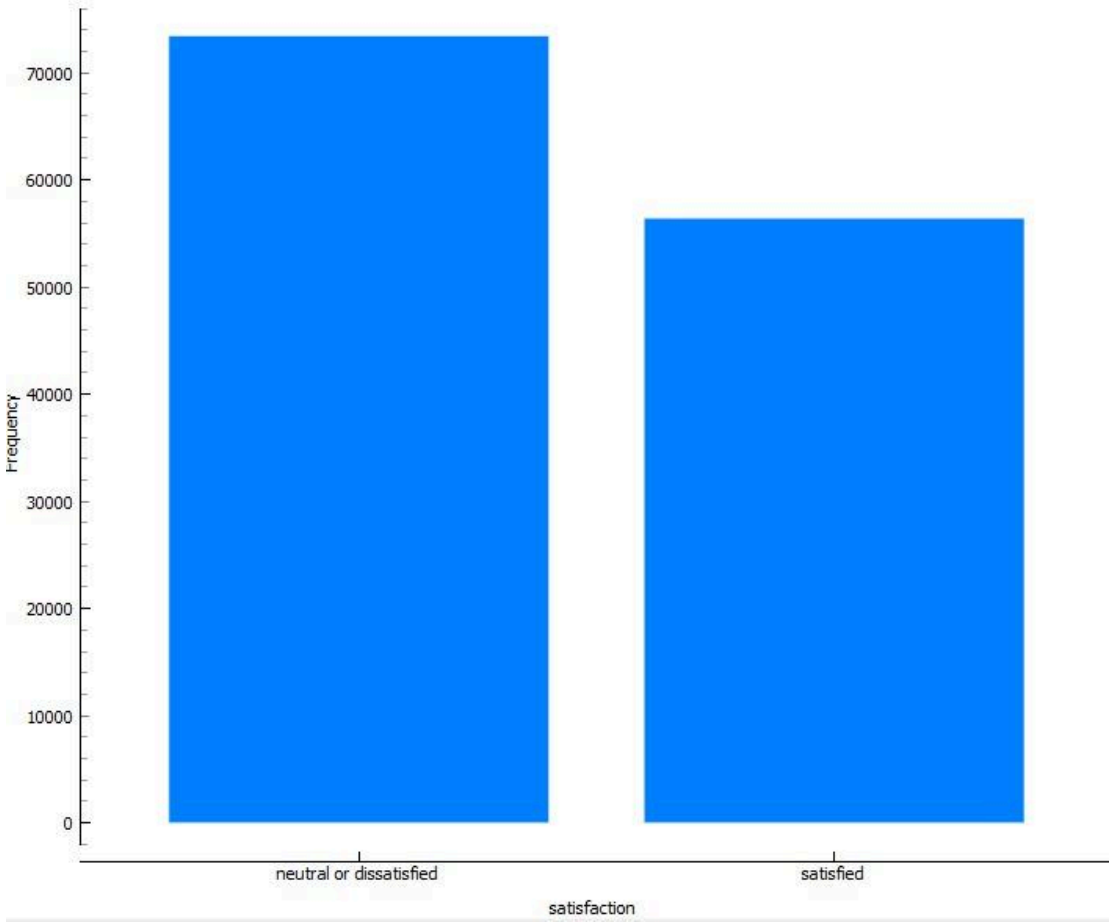


Fig 6: En este histograma podemos ver como hay más clientes insatisfechos.

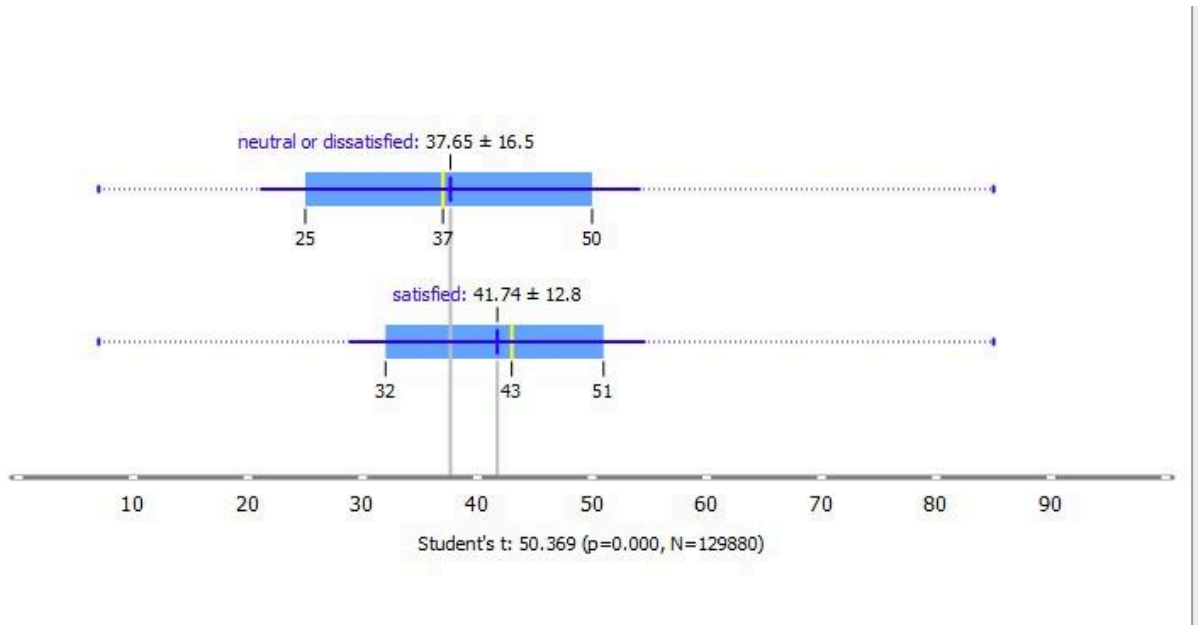


Fig 7. Con el boxplot podemos ver cómo los clientes satisfechos están entre los 32 y 51 años siendo la media de 43 años. Por el otro lado, los neutrales o no satisfechos están entre los 25 y 50 años con una media de 37 años.

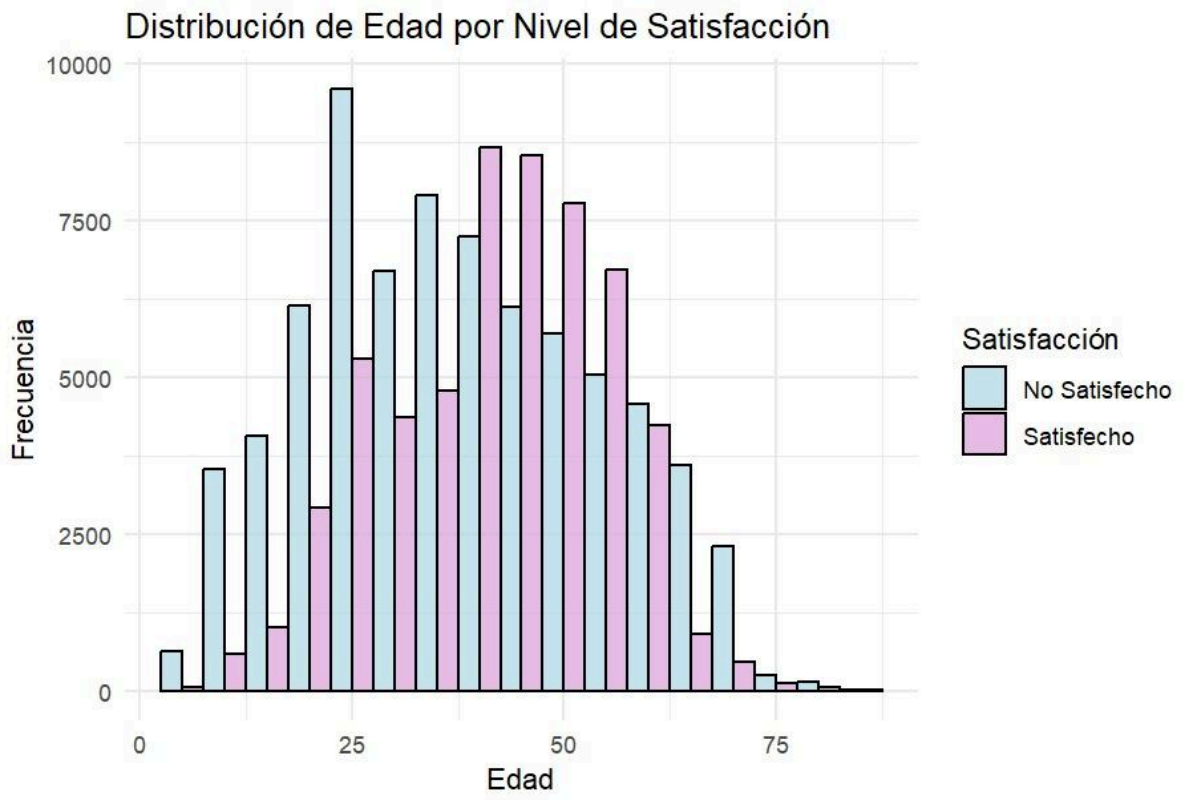


Fig 8. En este histograma podemos ver con mayor claridad que las edades mayores tienen mayor nivel de satisfacción, mientras los no satisfechos son los más jóvenes.

Preprocesamiento y Selección de Variables

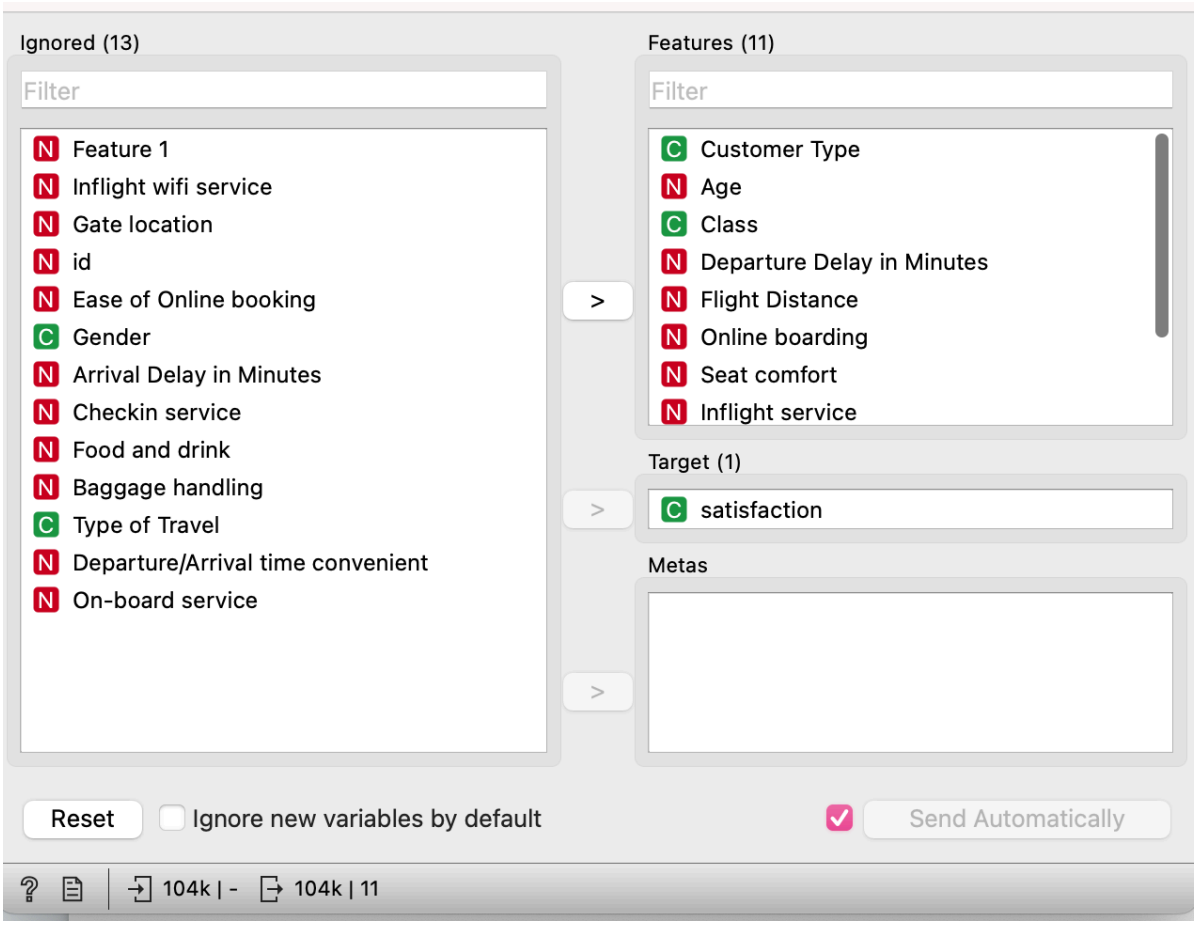
Features:

- Customer Type
- Age
- Class
- Departure delay in minutes
- Flight distance
- Online boarding
- Seat comfort
- Inflight service
- Satisfacción
- Cleanliness

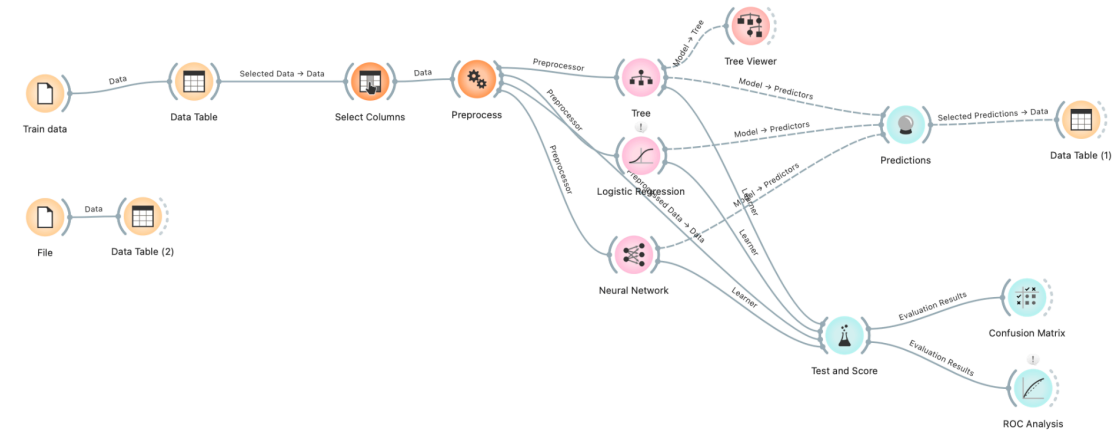
Ignored:

- Feature 1
- Inflight wifi service
- Gate location
- Id
- Ease of online booking
- Gender

- Arrival delay in minutes
- Leg room service
- Checkin service
- Food and drink
- Baggage handling
- Type of travel
- Departure/Arrival time convenient
- On-board service
- Inflight entertainment

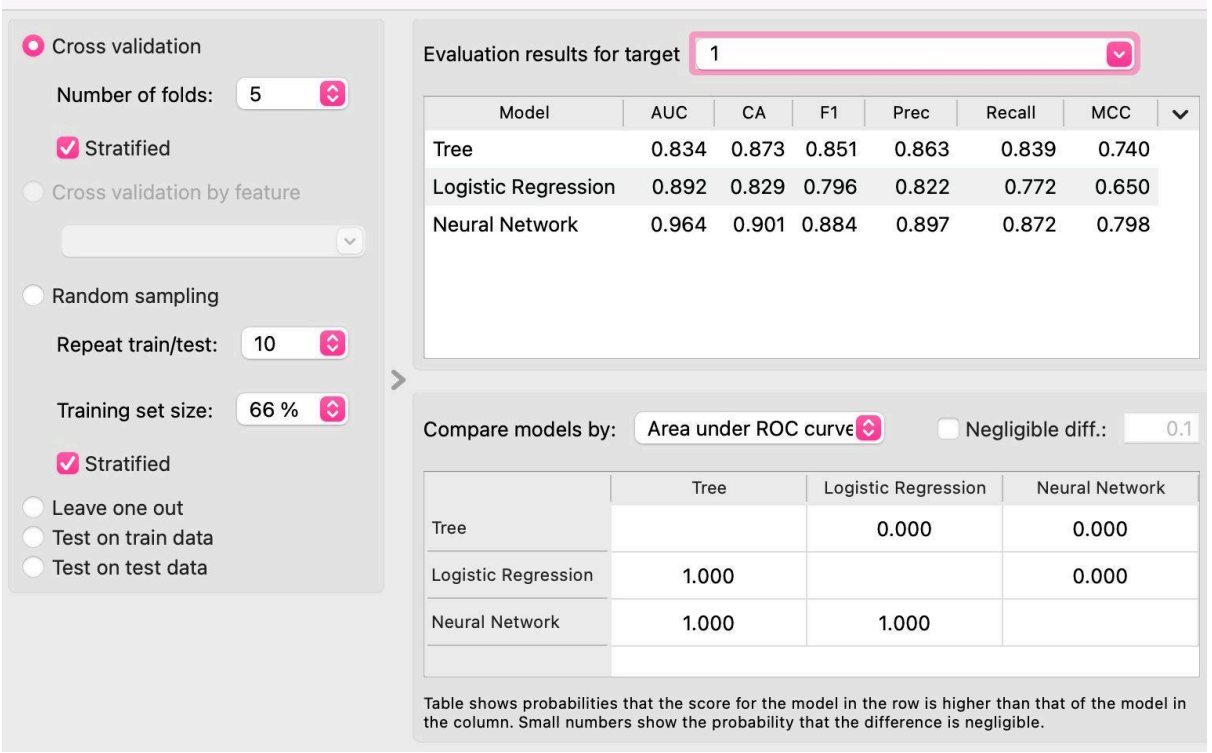


Modelado



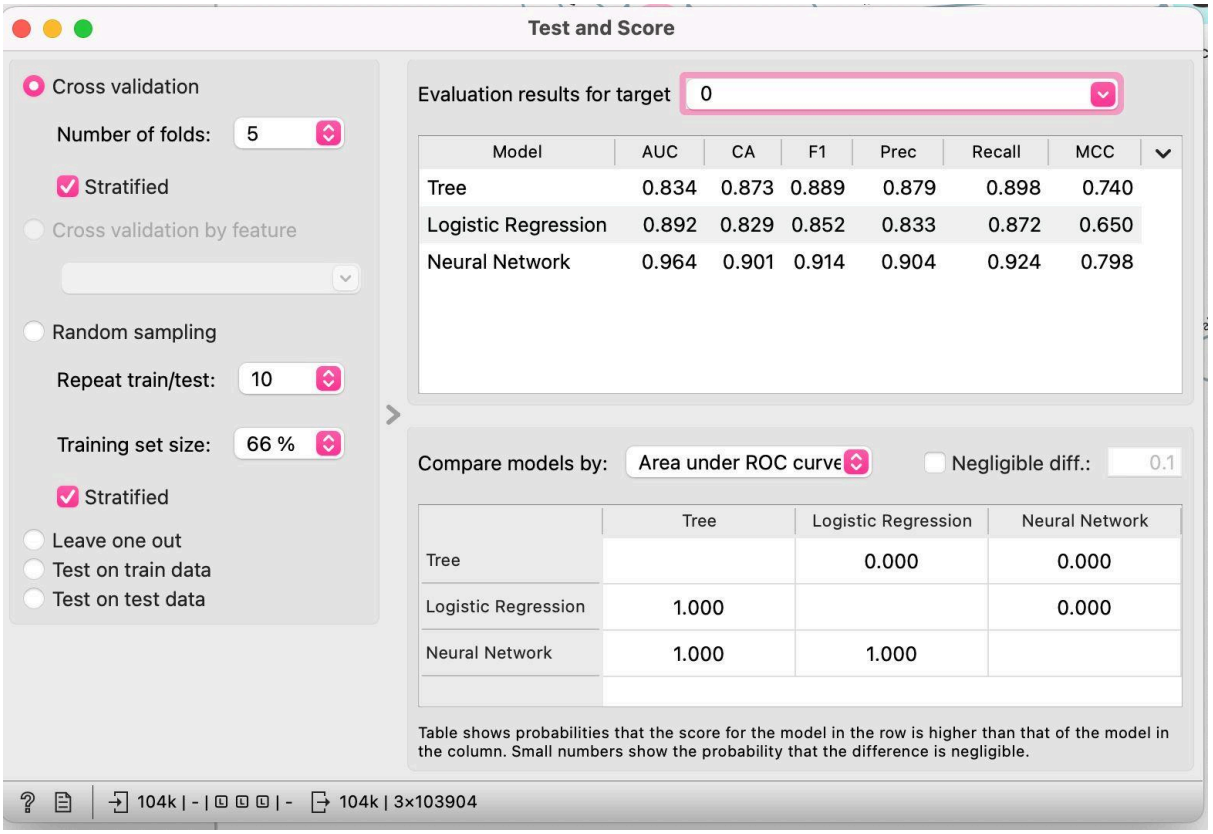
Test and score

Satisfecho

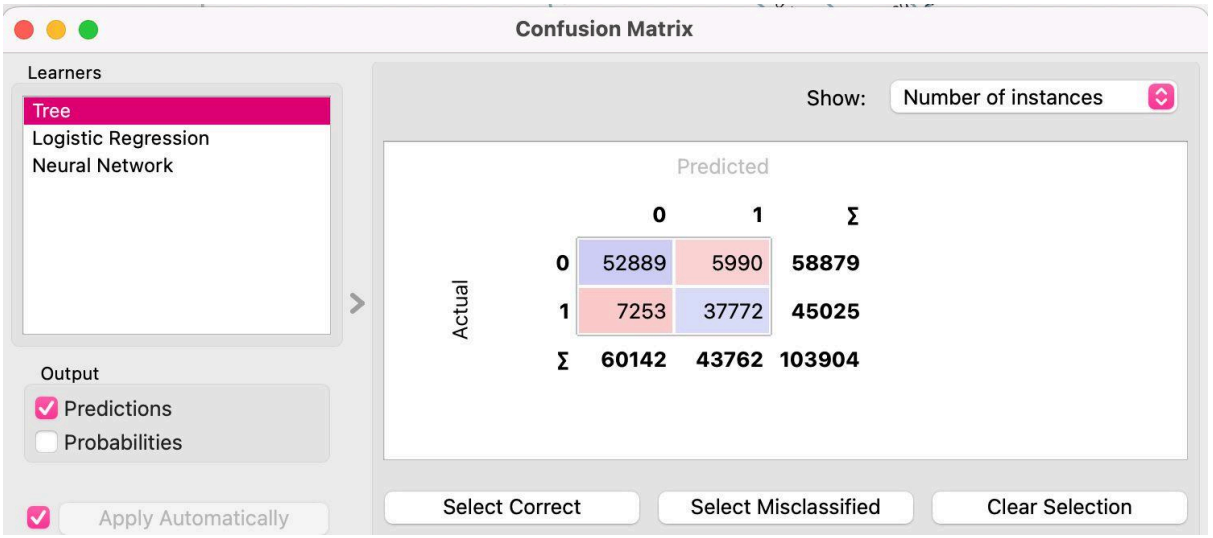


- Neural Network siempre es el mejor modelo al predecir e identificar clientes como satisfechos.

No satisfecho



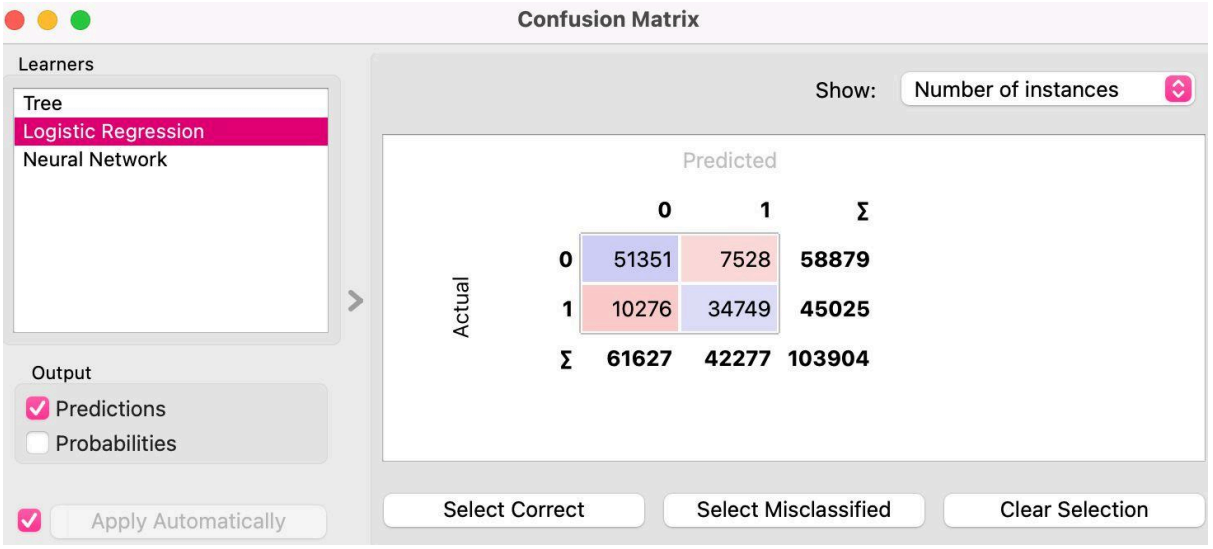
Confusion matrix



Puntos clave a considerar:

- Precisión General: El modelo alcanzó una precisión global del 87.2% nos indica un buen rendimiento general del modelo.
- Desbalance de Clases: Hay un mayor número de falsos negativos (7253) en comparación con los falsos positivos (5990).

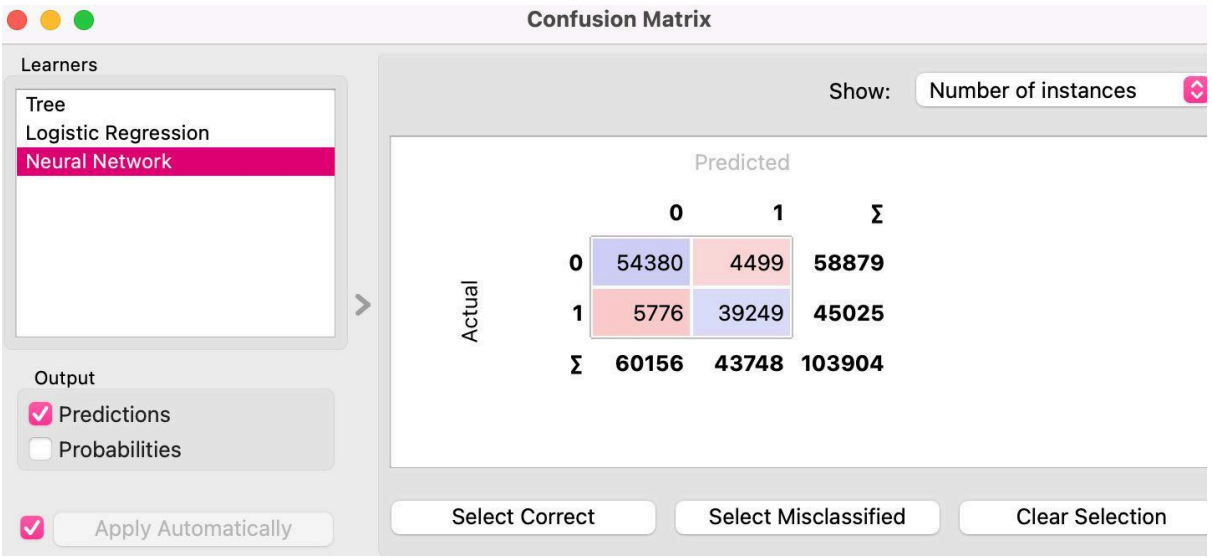
Logistic regression



El modelo tiene una precisión del 83% por lo que podemos decir que es aceptable para muchas aplicaciones.
El modelo de regresión logística está haciendo un buen trabajo en general. Nos dio un valor de 83% en precisión, lo cual es bastante bueno para varios proyectos.

- **Detecta bien a los pasajeros "normales" (clase 0):** El modelo es muy bueno para identificar a los pasajeros que no son tan exigentes (clase 0).
- **Se le escapan algunos pasajeros "exigentes" (clase 1):** Aunque detecta a muchos pasajeros de la clase 1 (los más satisfechos), se le escapan algunos. Esto significa que los está clasificando como si fueran de la clase 0 cuando en realidad son de la clase 1. Si queremos identificar a todos los pasajeros de la clase 1.

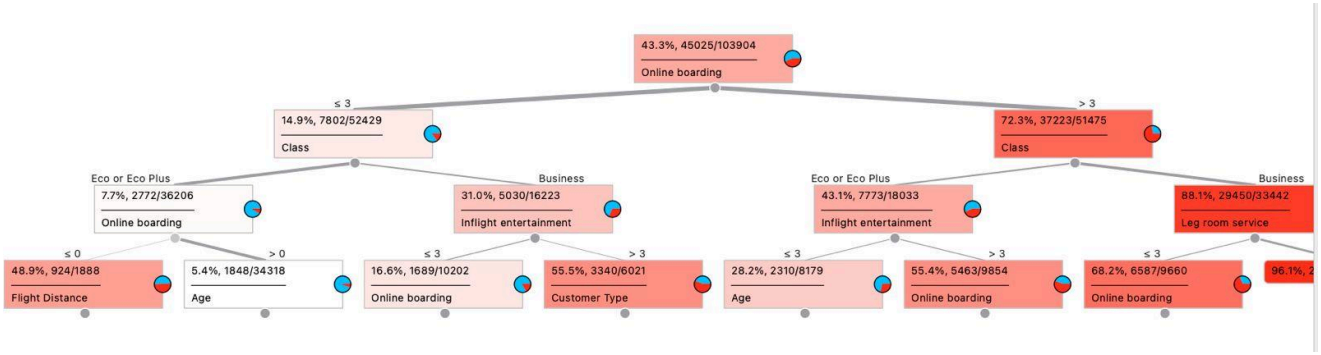
Neutral network



- La precisión del modelo (90.2%) sugiere un buen desempeño general. Además, cuenta con buena sensibilidad y especificidad: La red neuronal tiene un balance entre sensibilidad (87.2%) y especificidad (92.3%). Esto indica que el modelo reduce tanto los falsos negativos como los falsos positivos. La clase 1 cuenta con una precisión positiva de 89.7%..

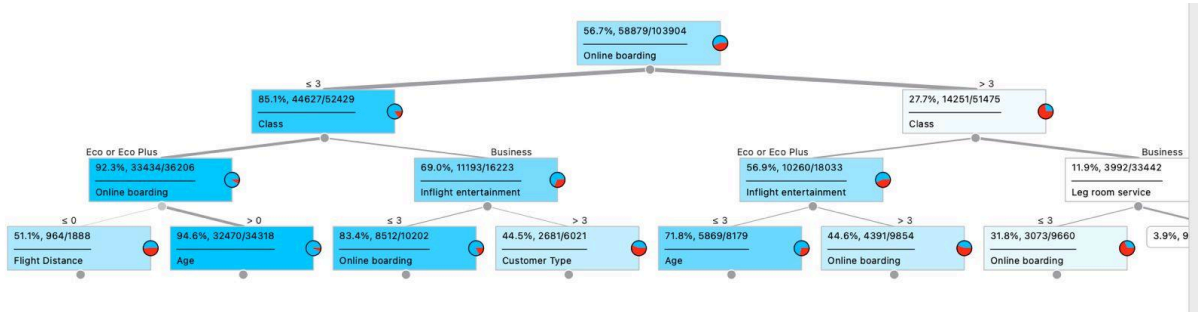
Tree viewer

Con satisfacción



- Clase del pasajero:
  - Clase Business: Más nivel de satisfacción que el resto de las clases.
  - Clase Eco o Eco Plus: Menor satisfacción en especial cuando se trata de inflight entertainment u online boarding.
  - Inflight entertainment es clave para la satisfacción del cliente, cuando el mismo es mayor a tres la satisfacción aumenta al 55%
- Online Boarding:
  - Es un punto clave en las clases eco y eco plus.
- Servicio de espacio para las piernas:
  - Factor clave en la clase business, cuando es mayor a 3 la satisfacción es muy alta a diferencia de si es menor.
  - Para los pasajeros de clase Business, el servicio de espacio para las piernas es un factor decisivo. Si es mayor a 3 la satisfacción es muy alta
- Tipo de cliente:
  - Los clientes frecuentes son de tener un nivel de satisfacción alto cuando los puntos anteriores son mayor a 3...

Sin satisfacción



- Los pasajeros de la clase economy plus por lo general están menos satisfechos, más que nada cuando el online boarding y el entertainment inflight son bajos. La satisfacción en general es alta cuando el online boarding es bajo con una% proporción de no satisfacción de
- En la clase Business, la no satisfacción es un poco menor, pero aumenta bastante si el servicio de espacio es malo. Ellos muestran una no satisfacción del 11.9% cuando sucede lo nombrado anteriormente.