UNIVERSIDAD AUSTRAL | CIENCIAS EMPRESARIALES

# ***Trabajo Práctico Big Data***

## Fecha de entrega: 8/11

## Análisis y Modelado de Datos con Orange Data Mining

Integrantes: Valentino Torlasco, Ian Carlomagno, Joaquin Curat.

1) **Selección del Dataset:** Se les proporcionarán tres datasets diferentes. Elijan uno que les interese y que crean que puede ser desafiante y relevante para desarrollar un modelo predictivo.

   RTA: Con el grupo elegimos el dataset de la aerolínea

2) **Definición del Problema:** Una vez seleccionado el dataset, deberán definir un problema específico de negocio o investigación que quieran resolver. Esta definición debe incluir una clara formulación de qué quieren predecir o clasificar y por qué es importante o relevante.

   RTA: El objetivo que nos proponemos es poder predecir la satisfacción de los pasajeros de la aerolínea, analizando distintas características como su experiencia de vuelo, tipo de viaje, clase en que viajan, duración de sus retrasos y el servicio recibido.

   La satisfacción al cliente es un indicador muy importante para las aerolíneas, porque influye directamente en la lealtad de los clientes y en su reputación, y a largo plazo puede afectar su rentabilidad. Por ejemplo, un pasajero satisfecho puede llegar a volver a usar la aerolínea y recomendarla a otros, mientras que uno insatisfecho puede elegir a un competidor y hasta dar publicidad negativa.

En base a esto la pregunta que nos hacemos es, ¿Cuáles son los factores que más afectan la satisfacción del cliente, y cómo podemos anticipar si un pasajero estará satisfecho o no en base a estos factores?
Nos parece que este tipo de predicción es importante ya que le permite a las aerolíneas poder identificar sus áreas de mejora en la experiencia de sus clientes. Al poder conocer cuales son los aspectos que más influyen en la satisfacción, la aerolínea puede tomar decisiones para poder optimizar sus servicios, y poder tener un impacto positivo en el pasajero.

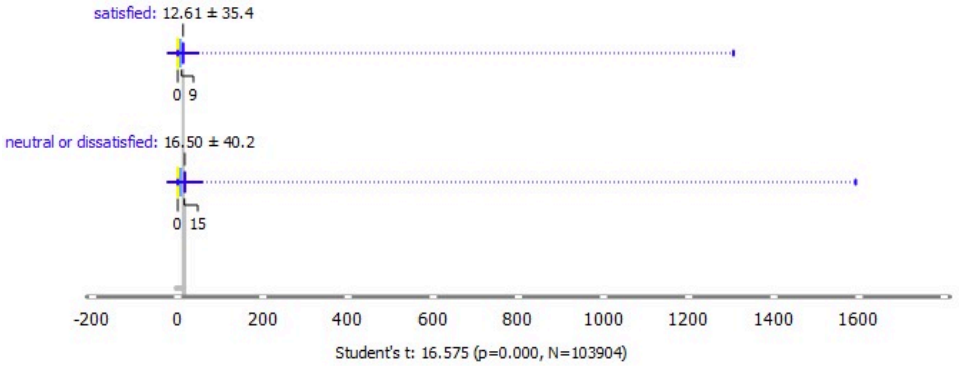
3) **Análisis Exploratorio de Datos:**
   A. Realicen un análisis exploratorio inicial para familiarizarse con los datos.
   **B. Planteen hipótesis sobre qué variables creen que son más relevantes para predecir el resultado de interés.**
   C. Utilicen gráficos y estadísticas para apoyar estas hipótesis y documenten sus hallazgos.

B) Después del análisis exploratorio de los datos pudimos plantear las siguientes hipótesis:

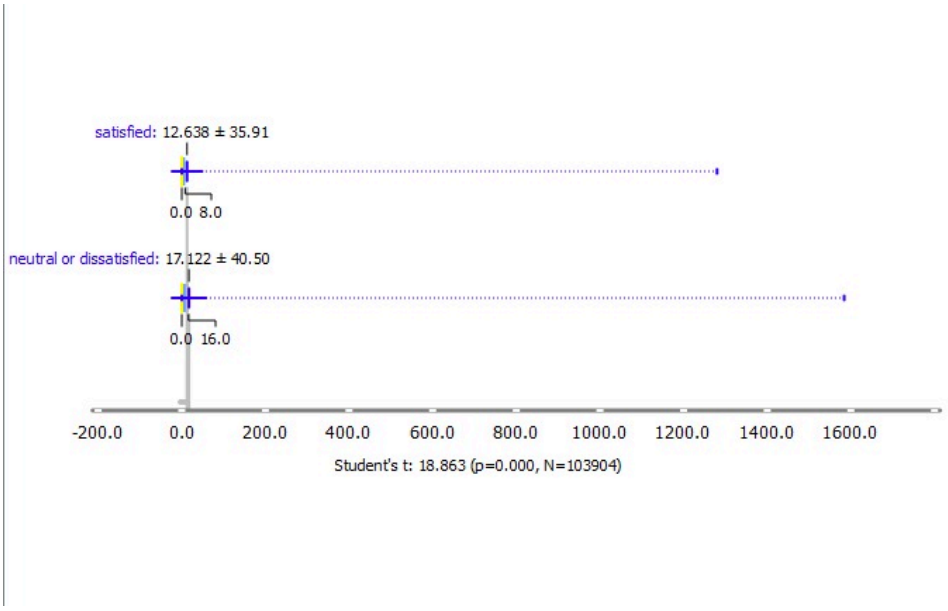
- **Hipótesis I:** Las variables relacionadas con los retrasos de despegue o aterrizaje afectan negativamente la satisfacción de los pasajeros.
- **Hipótesis II:** El tipo de viaje (type of travel), ya sea por negocio o viaje personal influye en la satisfacción, porque los pasajeros que viajan por negocios pueden tener mayores expectativas del vuelo.
- **Hipótesis III:** La clase en la que viajan los pasajeros está relacionada con la satisfacción, ya que los que están en mejores clases pueden tener un mejor servicio y esto lleva a tener una mejor experiencia.
- **Hipótesis IV:** El servicio de WIFI del vuelo y el la comodidad de los asientos también pueden impactar en la satisfacción de los pasajeros, ya que pueden mejorar la experiencia si estos son buenos o hacer que tengan una mala experiencia debido a que no van cómodos o el WIFI no anda bien.

C)

- **Hipótesis I:**




Los pasajeros insatisfechos tienen un retraso promedio de 16 minutos con 50 segundos que es ligeramente mayor que los satisfechos que su media es de 12.61 minutos, lo cual sugiere que mayores retrasos en el despegue pueden estar relacionados con menor satisfacción.



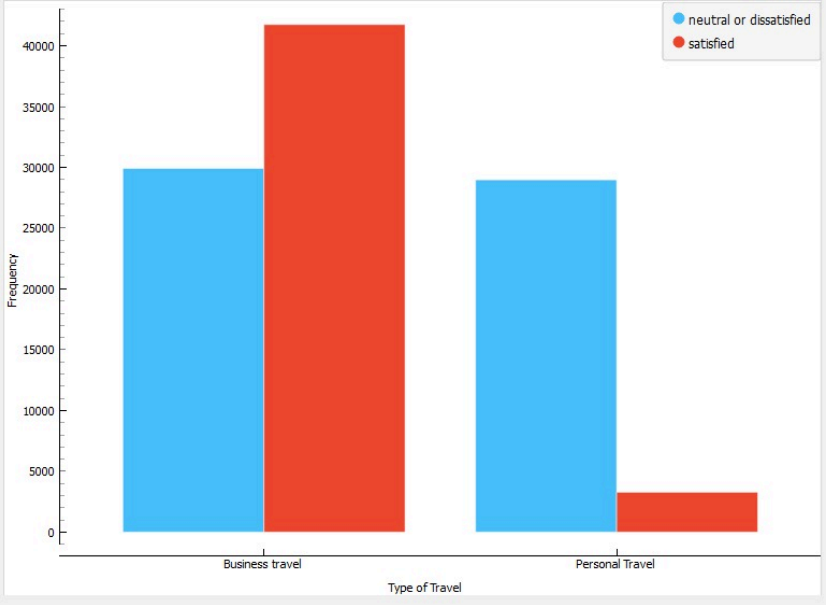

Acá también se observa que los pasajeros insatisfechos experimentan un retraso promedio mayor en los aterrizajes que los satisfechos.

Ambos gráficos muestran que los pasajeros insatisfechos tienen, en promedio, mayores retrasos en despegues y aterrizajes que los pasajeros satisfechos. La prueba estadística respalda que estas diferencias son significativas. Por lo tanto, estos resultados apoyan la

hipótesis de que los retrasos en despegues y aterrizajes afectan negativamente la satisfacción de los pasajeros. Demostrando que mientras más largos sean los retrasos, es más probable que los pasajeros se sientan insatisfechos.

También si analizamos las medias de cada boxplot, podemos ver que el promedio de los pasajeros que se les retrasa el vuelo y están insatisfechos es 3.89 minutos más que de los que están satisfechos y 4.49 minutos en el caso de los aterrizajes. Aunque estas diferencias no sean muy grandes son significativas para nuestro análisis, ya que sugiere que los retrasos en los vuelos están relacionados con los niveles de satisfacción. Entonces estos resultados apoyan la hipótesis de que los retrasos en los despegues o aterrizajes afectan negativamente la experiencia de los pasajeros y resaltan la importancia de reducir los tiempos de espera para mejorar las satisfacción de los clientes.

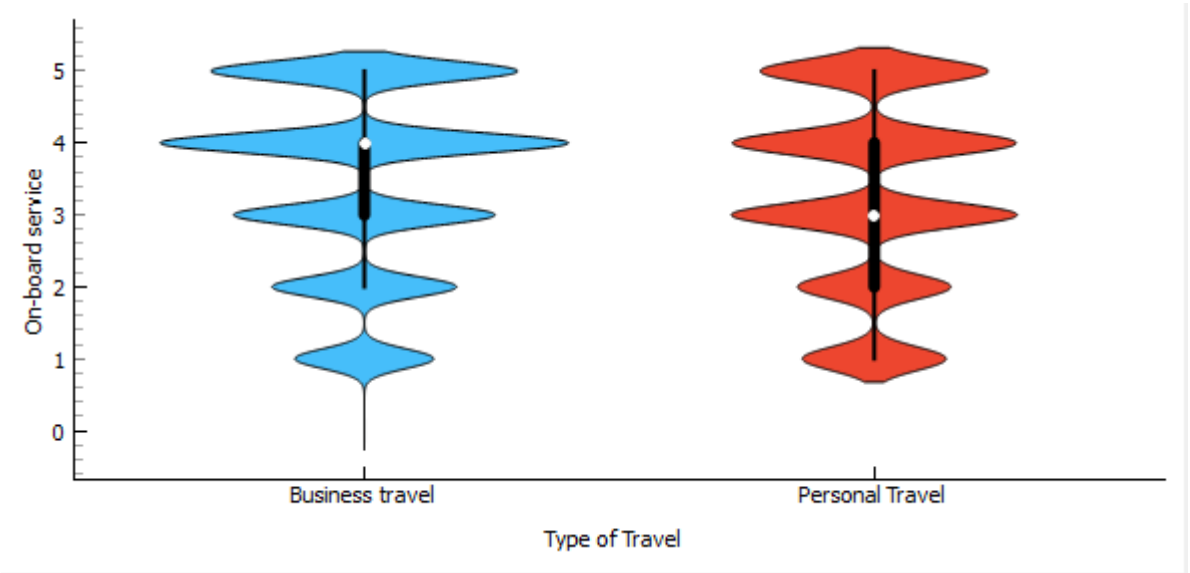
- **Hipótesis II:**



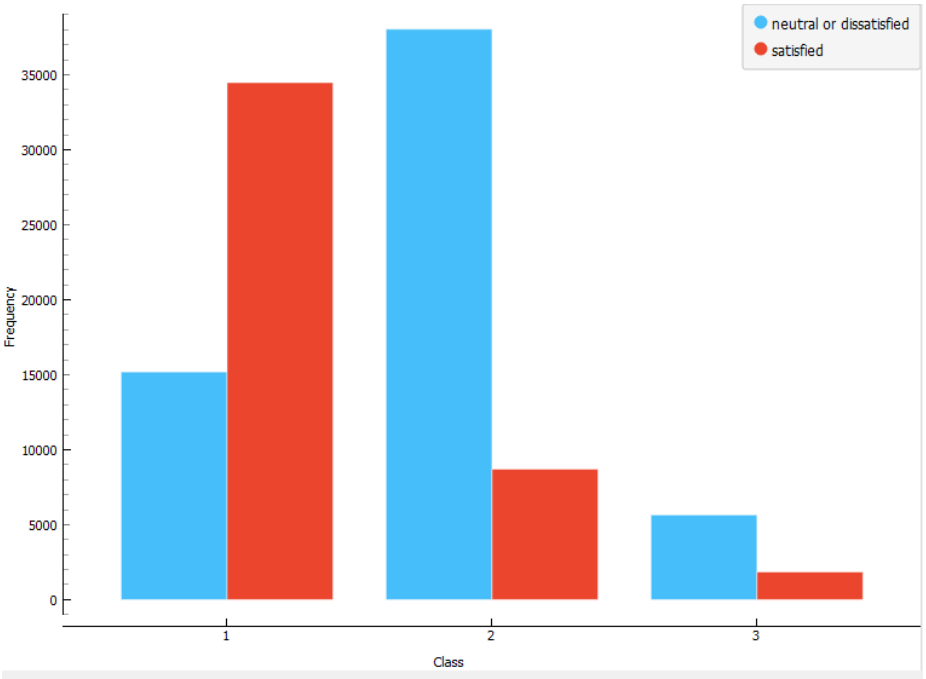

- **Viaje de negocios**:
  - La barra roja, que representa a los pasajeros satisfechos, es notablemente más alta que la azul, lo que indica que una mayor proporción de pasajeros en viajes de negocios están satisfechos en comparación con los que están neutrales o insatisfechos.
- **Viaje personal**:
  - En contraste, para los pasajeros en viaje personal, la barra azul (neutrales o insatisfechos) es significativamente más alta que la roja (satisfechos), lo que indica que hay una mayor proporción de insatisfacción en este grupo.

Este gráfico apoya la hipótesis de que el tipo de viaje influye en la satisfacción. Los pasajeros en viajes de negocios tienen una proporción más alta de satisfacción en comparación con los pasajeros en viajes personales, lo que podría deberse a que los viajeros de negocios suelen tener mayores expectativas y probablemente reciben un mejor servicio para satisfacer estas expectativas.

Para concluir, el gráfico sugiere que el tipo de viaje sí afecta la satisfacción de los pasajeros, siendo los viajeros de negocios, en general, más satisfechos que aquellos en viajes personales. Esto puede ser debido a que en los viajes por negocios, vemos que la mayoría de estos pasajeros tienen un mejor servicio a bordo del avión. Esto lo podemos ver con este gráfico de violín.



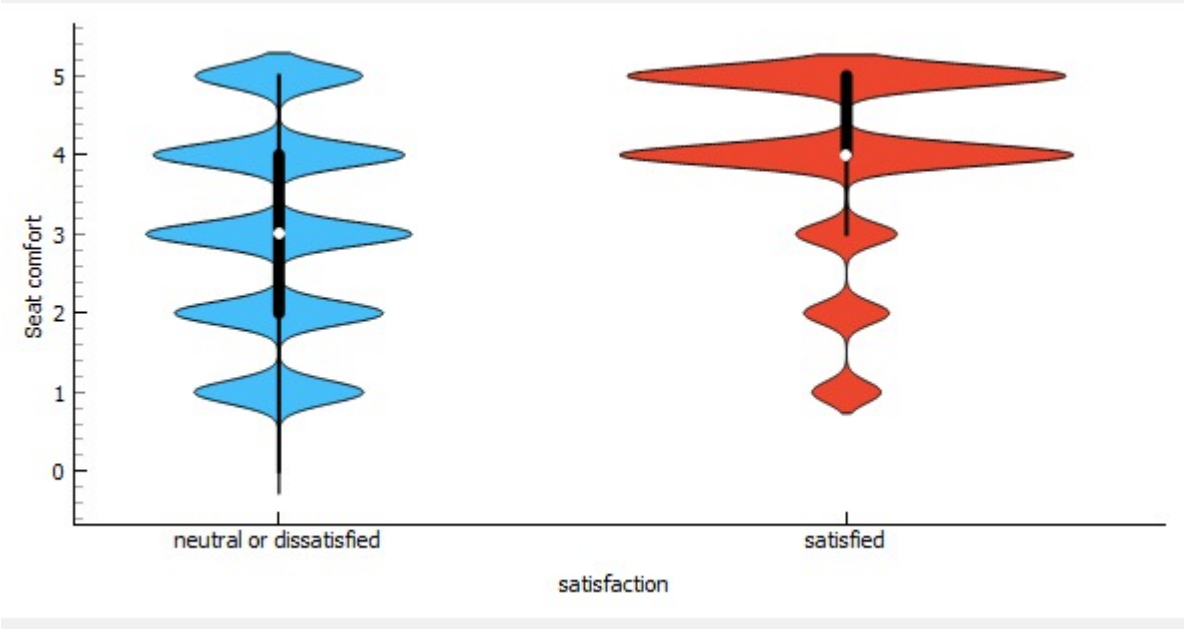

- **Hipótesis III:**



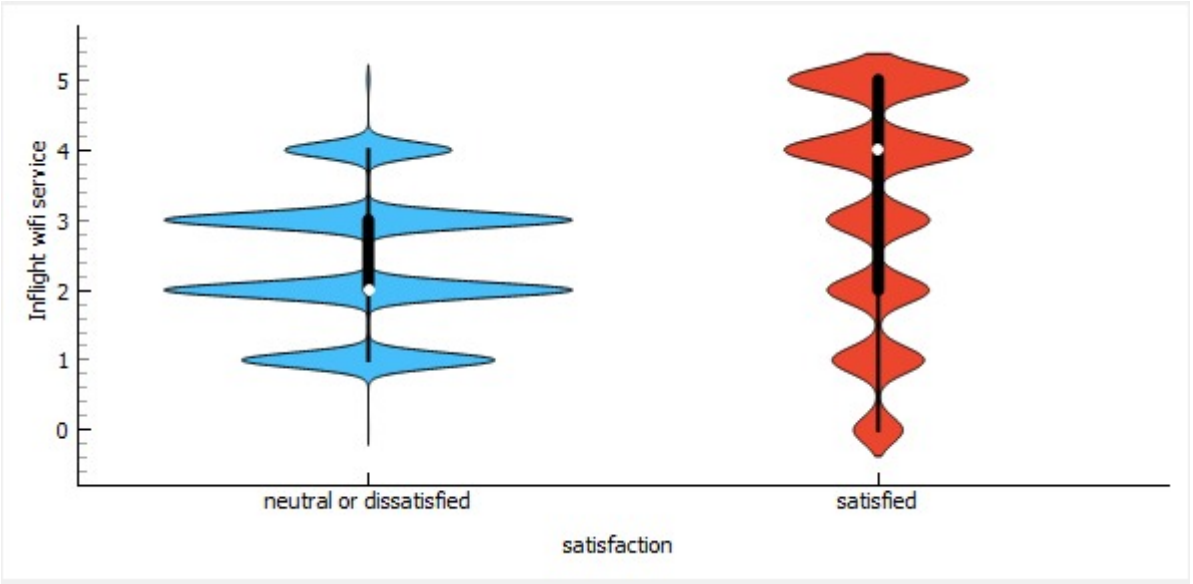

El gráfico ilustra la relación entre la clase en la que viajan los pasajeros y su nivel de satisfacción, con dos categorías: "satisfecho" (en rojo) y "neutral o insatisfecho" (en azul). En la Clase 1 (business), se observa que hay más pasajeros satisfechos que neutrales o insatisfechos, lo que sugiere que los pasajeros de esta clase, posiblemente debido a un mejor servicio, tienden a tener una experiencia más positiva. En la Clase 2(economy), en cambio, el número de pasajeros neutrales o insatisfechos es significativamente mayor que el de satisfechos, lo cual podría indicar que el nivel de servicio es inferior al de la Clase 1, afectando la percepción de la experiencia. Por último, en la Clase 3(economy plus), los pasajeros neutrales o insatisfechos también superan en cantidad a los satisfechos, y la satisfacción en esta clase parece ser la más baja, lo que podría estar asociado a un servicio más básico. En conjunto, los datos respaldan la hipótesis de que los pasajeros en clases superiores (como la Clase 1) muestran mayores niveles de satisfacción, probablemente debido a un mejor nivel de servicio.

- **Hipótesis IV:**

Los gráficos muestran la relación entre la satisfacción de los pasajeros y dos variables de la experiencia de vuelo: la comodidad de los asientos y el servicio de WiFi en el vuelo. Ambos gráficos son diagramas de violín que representan la distribución de la percepción de comodidad del asiento y de la calidad del WiFi en función de si los pasajeros están satisfechos o no.




En el primer gráfico, que analiza la comodidad de los asientos, se observa que los pasajeros satisfechos (en rojo) tienen una media de comodidad de asientos cercana a 4 en una escala de 0 a 5, con una concentración hacia los niveles superiores de comodidad. Por otro lado, los pasajeros neutrales o insatisfechos (en azul) muestran una distribución más uniforme, con una media que parece rondar los 3 puntos, y mayor dispersión en los niveles inferiores de comodidad. Esto sugiere que una mayor comodidad del asiento está asociada con una mayor satisfacción.



En el segundo gráfico, que evalúa el servicio de WiFi, se ve una tendencia similar. Los pasajeros satisfechos tienden a tener una media de satisfacción con el WiFi cercana a 4, con una concentración hacia niveles altos de puntuación, mientras que los pasajeros neutrales o insatisfechos tienen una media alrededor de 2 o 3, con mayor dispersión en los valores bajos. Esto indica que la calidad del WiFi también tiene un impacto significativo en la satisfacción de los pasajeros: quienes reportan mejor servicio de WiFi tienden a estar más satisfechos.

Para finalizar, estos datos apoyan la hipótesis de que tanto la comodidad de los asientos como la calidad del WiFi influyen en la satisfacción de los pasajeros. Los niveles altos de comodidad y de calidad de WiFi parecen estar asociados con una mejor experiencia general.

4) **Preprocesamiento y Selección de Variables:**

Las variables que se incluyen en el dataset son: Gender (género del pasajero), Customer Type (tipo de cliente, leal u ocasional), Age (edad), Type of Travel (tipo de viaje, ya sea personal o de negocios) y Class (clase de vuelo: Business, Eco, o Eco Plus). Además, se incluyen factores relacionados con el vuelo, como Flight Distance (distancia del vuelo en millas) y evaluaciones del servicio, tales como Inflight wifi service (servicio de wifi a bordo), Departure/Arrival time convenient (conveniencia del horario de salida y llegada), Ease of Online booking (facilidad para reservar online), Gate location (ubicación de la puerta de embarque), Food and drink (calidad de comida y bebida), Online boarding (experiencia de embarque en línea), Seat comfort (comodidad del asiento), Inflight entertainment (entretenimiento a bordo), On-board service (calidad del servicio a bordo), Leg room service (espacio para las piernas), Baggage handling (manejo del equipaje), Checkin service (servicio de check-in), Inflight service (servicio general a bordo), y Cleanliness (limpieza). También se miden Departure Delay in Minutes y Arrival Delay in Minutes (retrasos en salida y llegada, en minutos). Finalmente, la variable Satisfaction indica si el pasajero está "satisfied" o "neutral or dissatisfied", y es la variable objetivo en este análisis.

De estas variables, las columnas de **departure delay in minutes** y **arrival delay in minutes** contenían datos nulos entonces lo que hicimos fue mediante el widget Impute, hacer que estos datos nulos sean reemplazados por las medias de cada columna. Además con el widget Edit domain cambiamos los datos de género y clases. En género pusimos que femenino es 1 y masculino es 0, y para clases pusimos que business es 1, eco es 2 y eco plus es 3. Por último ignoramos las columnas de ID, Gate location, Ease of online booking y Departure arrival convenient. Este procedimiento se hizo tanto para el archivo de entrenamiento como para el de testeo.

6)**Evaluación de Modelos**

Para este análisis, definimos la F1 score como la métrica más importante, porque representa un equilibrio entre precisión y recall. En nuestro caso, es importante captar tanto a los pasajeros satisfechos como a los insatisfechos, evitando errores en la clasificación de cualquiera de las clases. Esto ayuda a que el modelo no solo sea preciso sino que también minimiza los falsos positivos y falsos negativos, lo que es fundamental para una predicción adecuada de la satisfacción del cliente.

Se evaluaron cuatro modelos: Árbol de Decisión, Red Neuronal, Regresión Logística y k-Nearest Neighbors (kNN), usando las métricas de precision, recall, accuracy y F1 score.

1. **Árbol de Decisión**: Este modelo logró una puntuación de 0.94 en todas las métricas (precision, recall, accuracy y F1 score), mostrando un desempeño balanceado. Su principal ventaja es la interpretabilidad, ya que permite observar la importancia de cada factor a través de la estructura del árbol. Esto lo convierte en una excelente opción si el objetivo es identificar claramente los factores que influyen en la satisfacción del cliente, manteniendo un buen rendimiento.
2. **Red Neuronal**: Con 0.957 en todas las métricas, la Red Neuronal presentó el mejor rendimiento en precisión y recall, destacándose en la F1 score. Este modelo sería ideal si el objetivo fuera maximizar la exactitud en la predicción, ya que predice con mayor acierto la satisfacción del cliente. Sin embargo, su desventaja radica en la falta de interpretabilidad, dificultando el entendimiento del peso de cada factor. Por lo tanto, aunque es el modelo más preciso, no permite desglosar fácilmente cómo cada variable contribuye a la satisfacción.
3. **Regresión Logística**: Obtuvo un rendimiento de 0.873 en todas las métricas. Si bien su precisión es inferior a los otros modelos, la regresión logística es bastante interpretable y permite entender el impacto de cada factor en la satisfacción. Esto la convierte en una buena opción si el propósito principal es comprender cómo cada variable contribuye a la satisfacción, aunque el modelo sacrifique algo de precisión en comparación con los otros modelos.
4. **k-Nearest Neighbors (kNN)**: Con resultados de 0.744 en precision y 0.745 en recall, accuracy y F1 score, este modelo tuvo el rendimiento más bajo en todas las métricas. Esto indica que kNN no es adecuado para este problema, probablemente debido a la complejidad de los datos y la relación entre las variables. Además, kNN carece de interpretabilidad, por lo que no resulta útil para identificar factores relevantes en la satisfacción del cliente.

**Conclusión:** El modelo de Árbol de Decisión representa la mejor opción si el objetivo es un equilibrio entre precisión e interpretabilidad, ofreciendo métricas consistentes y suficientes para identificar los factores clave en la satisfacción del cliente. Aunque la Red Neuronal presenta la mayor precisión y una F1 score ligeramente superior, su falta de interpretabilidad limita su utilidad si es necesario entender cómo cada variable afecta la satisfacción. En resumen, el Árbol de Decisión cumple mejor con los requisitos del problema, permitiendo una predicción precisa al tiempo que destaca los factores de mayor impacto en la experiencia del cliente.