

## 1. Dataset seleccionado:

Airline Passenger Satisfaction

### Fuente:

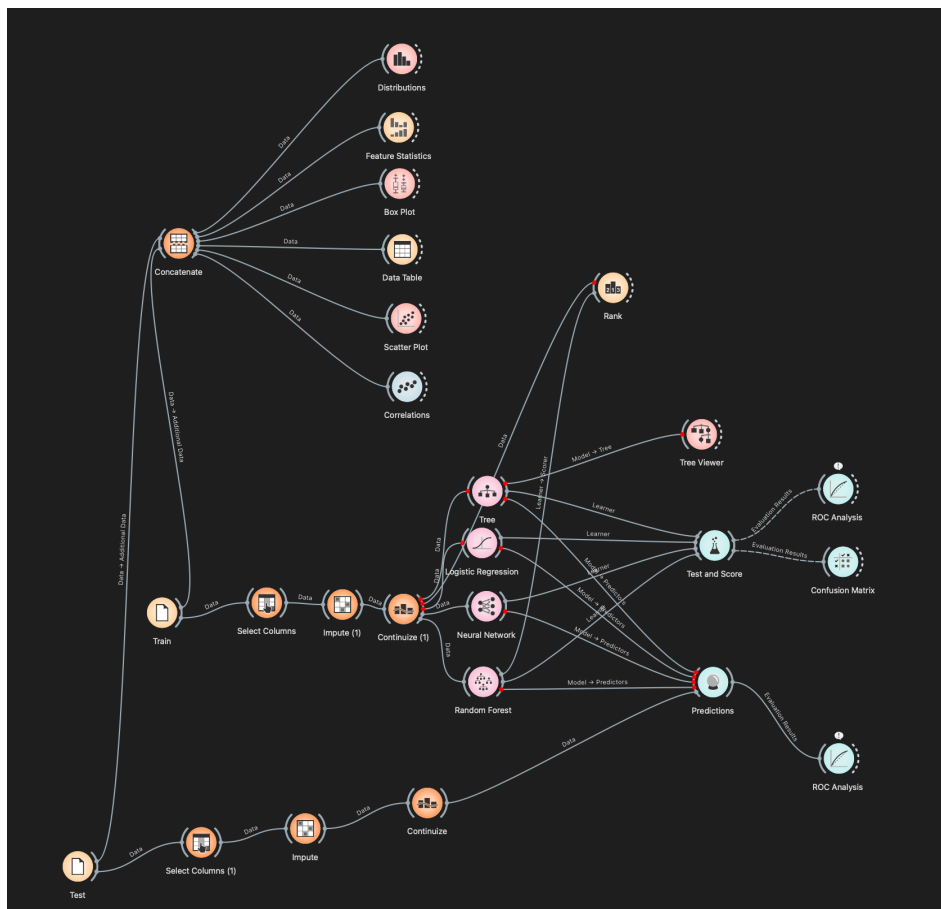
[Kaggle – Airline Passenger Satisfaction Dataset](#)

### Descripción:

Este dataset contiene información detallada sobre la experiencia de pasajeros de una aerolínea, incluyendo aspectos como el tipo de viaje, clase de vuelo, género, edad, distancia recorrida, nivel de satisfacción con distintos servicios a bordo (como wifi, entretenimiento, comida, embarque, entre otros) y una etiqueta final que indica si el pasajero quedó satisfecho o no con el vuelo.

### Objetivo del uso del dataset:

Se utilizará para entrenar modelos de machine learning que permitan predecir si un pasajero estará satisfecho o no, en base a los distintos factores disponibles. Esta predicción es clave para que la aerolínea pueda anticiparse a la insatisfacción y tomar medidas como enviar descuentos o mejorar la atención a tiempo para tener cada vez más clientes satisfechos y lograr fidelizarlos.



## 2. Problema a resolver:

En el contexto actual de alta competencia en el sector aeronáutico, las aerolíneas necesitan anticiparse a los factores que impactan negativamente en la experiencia del pasajero. Este proyecto busca resolver el siguiente problema:

**¿Es posible predecir si un pasajero estará satisfecho o no con su experiencia en el vuelo, a partir de variables como el tipo de cliente, clase del vuelo, servicios ofrecidos y características del viaje?**

### Importancia del problema:

Poder anticipar la insatisfacción de los pasajeros permite implementar acciones preventivas (como enviar descuentos, encuestas personalizadas o asistencia especial) que **mejoren la fidelización, reduzcan quejas y potencien la experiencia del cliente**. Esto representa una ventaja competitiva clave para cualquier aerolínea.

La satisfacción del cliente impacta directamente en la **fidelización y por ende en la rentabilidad**, ¿por qué? cliente fidelizado = mayor **LTV**

Nos permite identificar **puntos débiles** del servicio (check-in, entretenimiento, limpieza, etc.).

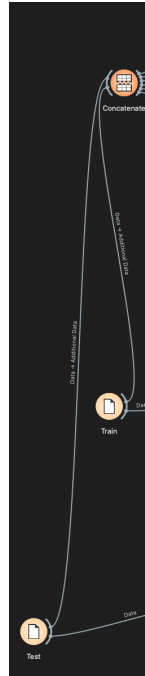
### Variable objetivo (target):

**satisfaction** → indica si el pasajero quedó “**satisfied**” o “**neutral or dissatisfied**”.

### Tipo de problema:










Clasificación binaria (clasificar a cada pasajero como satisfecho o insatisfecho).

- Primero para poder llevar a cabo el análisis exploratorio de datos (EDA), realizamos una concatenación de los datasets, porque ya nos dieron los datos separados un 80% en train, y un 20% en test, y queremos ver una primera “big picture” de cómo están el 100% de los datos.



Una vez concatenados pasamos a la parte descriptiva, donde utilizamos los módulos Feature Statistics y Box Plot de Orange, que nos permitieron visualizar la distribución de variables y comparar grupos según la variable objetivo (satisfaction). Con el objetivo de obtener una representación más completa del comportamiento general del dataset, concatenamos las bases de entrenamiento y testeo, trabajando con el conjunto completo de observaciones. Esta decisión se aplicó exclusivamente para esta etapa exploratoria, antes del entrenamiento de modelos, evitando cualquier riesgo de data leakage.

Name	Distribution	Mean	Mode	Median	Dispersion	Min.	Max.	Missing
N Age		39.43	39	40	0.38	7	85	0 (0 %)
N Flight Distance		1190.32	337	844	0.84	31	4983	0 (0 %)
N Inflight wifi service		2.73	2	3	0.49	0	5	0 (0 %)
N Departure/Arrival time convenient		3.06	4	3	0.50	0	5	0 (0 %)
N Ease of Online booking		2.76	3	3	0.51	0	5	0 (0 %)
N Gate location		2.98	3	3	0.43	0	5	0 (0 %)
N Food and drink		3.20	4	3	0.41	0	5	0 (0 %)
N Online boarding		3.25	4	3	0.42	0	5	0 (0 %)
N Seat comfort		3.44	4	4	0.38	0	5	0 (0 %)

	Name	Distribution	Mean	Mode	Median	Dispersion	Min.	Max.	Missing
N	Age		39.43	39	40	0.38	7	85	0 (0 %)
N	Flight Distance		1190.32	337	844	0.84	31	4983	0 (0 %)
N	Inflight wifi service		2.73	2	3	0.49	0	5	0 (0 %)
N	Departure/Arrival time convenient		3.06	4	3	0.50	0	5	0 (0 %)
N	Ease of Online booking		2.76	3	3	0.51	0	5	0 (0 %)
N	Gate location		2.98	3	3	0.43	0	5	0 (0 %)
N	Food and drink		3.20	4	3	0.41	0	5	0 (0 %)
N	Online boarding		3.25	4	3	0.42	0	5	0 (0 %)
N	Seat comfort		3.44	4	4	0.38	0	5	0 (0 %)

## Hipotesis

**La clase del vuelo influye en la satisfacción del cliente**



**Hipótesis:** Los pasajeros de clase Economy tienden a estar más insatisfechos que los de Business o Eco Plus.

- En **Feature Statistics**, la variable **Class** tiene su moda en **Business (90.2%)**, pero existe una gran proporción de pasajeros en clase Economy dentro del grupo insatisfecho.
- En el **Box Plot**, se evidencia que los pasajeros **satisfechos son principalmente de clase Business**, mientras que **los insatisfechos pertenecen en su mayoría a Economy**.
- También hay diferencia significativa ( $\chi^2 = 32906.17$ ,  $p < 0.001$ ).

**Interpretación:** El nivel de clase contratada **refleja la experiencia percibida**: los pasajeros con mejor servicio (Business/Eco Plus) tienden a valorar mejor el vuelo.

**El tipo de viaje afecta la percepción**



**Hipótesis:** Los pasajeros que hacen viajes personales están más insatisfechos que los de negocios.

- En el módulo **Feature Statistics**, la moda de la variable **Type of Travel** es **Business travel**, con un 61.9%. Sin embargo, observamos que **la proporción de pasajeros insatisfechos es mucho mayor en viajes personales**.
- Esto se confirma en el **Box Plot**, donde la mayoría de los **satisfechos viajan por negocios**, mientras que **los insatisfechos predominan en viajes personales (Personal Travel)**.
- La diferencia es estadísticamente significativa ( $\chi^2 = 26282.52$ ,  $p < 0.001$ ).

Los pasajeros por placer podrían tener **expectativas más altas**, o bien **menor tolerancia a fallas del servicio**, lo cual afecta su satisfacción.

### Servicios abordo influyen mucho en la satisfacción

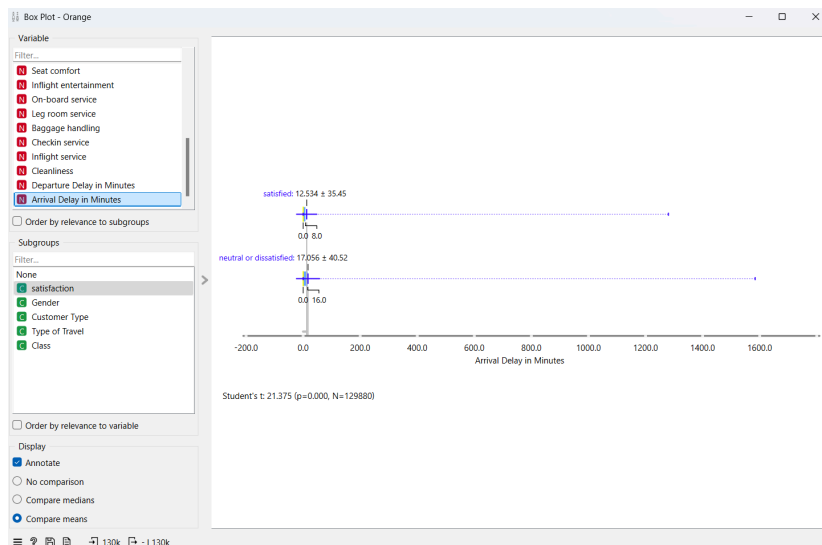
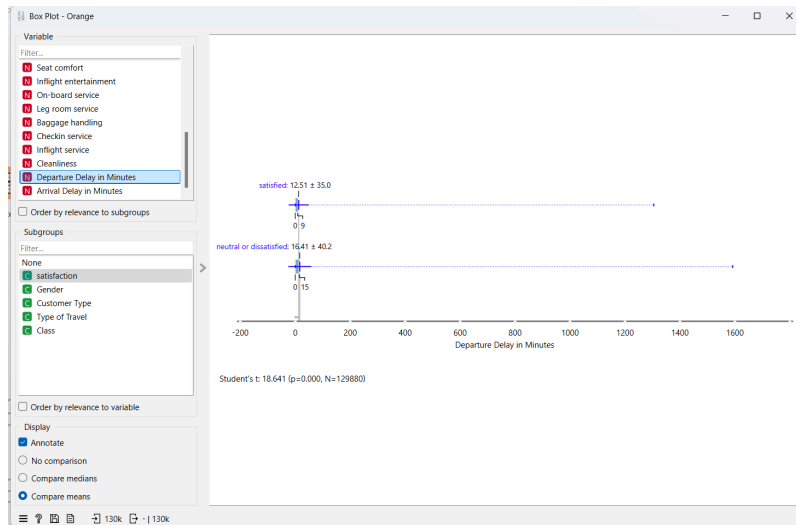
**Hipótesis:** La calidad del wifi, entretenimiento, comida y embarque están fuertemente correlacionadas con la satisfacción.

- En **Feature Statistics**, las variables **Inflight wifi service**, **Online boarding**, **Inflight entertainment**, **Food and drink** y **Seat comfort** presentan **medias entre 2.7 y 3.4** y **dispersión moderada (~0.4-0.5)**, lo cual sugiere **diversidad de experiencias**.

- Estas variables fueron luego validadas como relevantes por el modelo (Rank), confirmando que son **fuertes determinantes de satisfacción**.

**Interpretación:** La **variabilidad en la calidad de los servicios percibidos** influye directamente en la satisfacción, siendo un área crítica de mejora para la aerolínea.

## Las demoras impactan en la satisfacción



**Hipótesis:** Pasajeros con demoras altas están más insatisfechos.

- En **Feature Statistics**, las variables **Departure Delay in Minutes** y **Arrival Delay in Minutes** muestran **medias bajas (~15 min)** pero **máximos superiores a 1500 min**, lo que

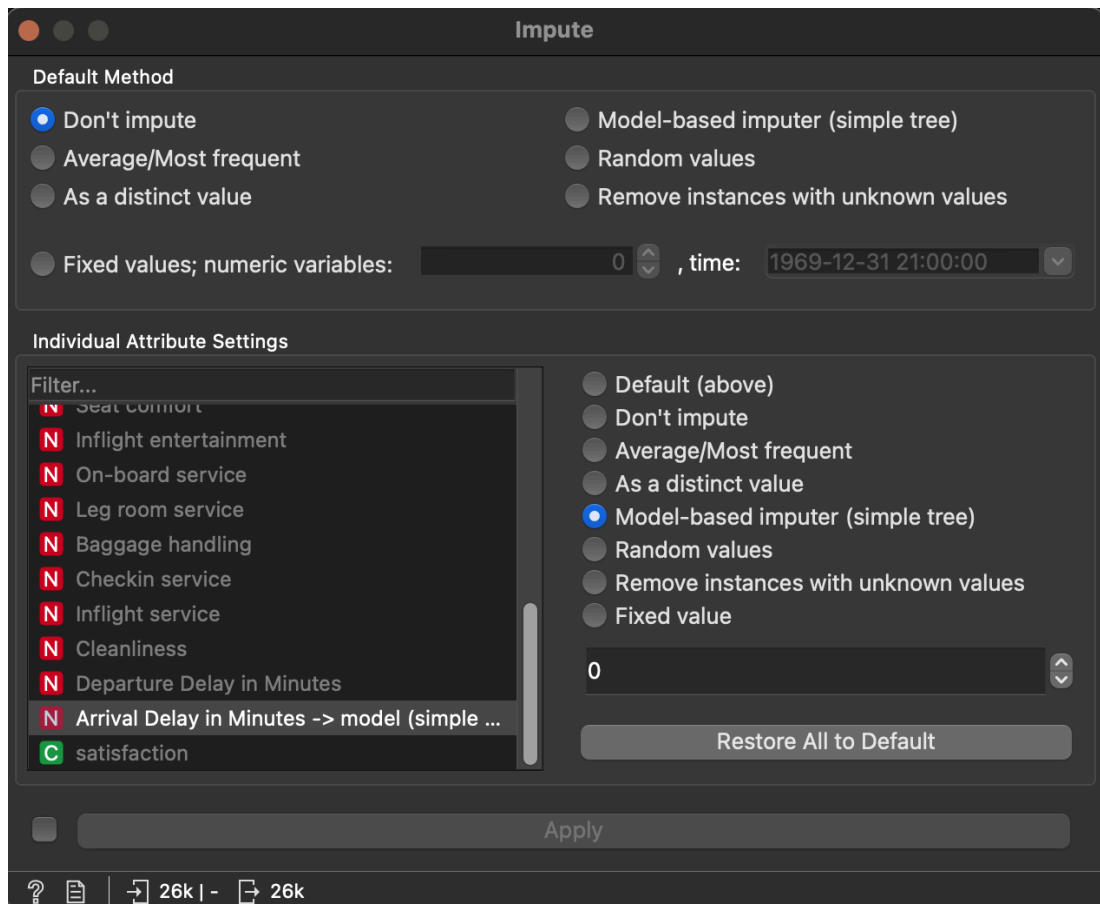
indica presencia de **outliers severos**.

- En los **Box Plots**, los pasajeros **insatisfechos tienen mayor promedio de demora**, tanto en despegue (16.4 min vs 12.5 min) como en llegada (17.0 min vs 12.5 min).
- En ambos casos, la diferencia es estadísticamente significativa ( $t = 18.641$  y  $21.375$ ,  $p < 0.001$ ).

**Interpretación:** Aunque el promedio general es bajo, los **casos extremos podrían tener un impacto desproporcionado** en la percepción del servicio y explicar insatisfacción.

#### 4. Preprocesamiento y Selección de Variables

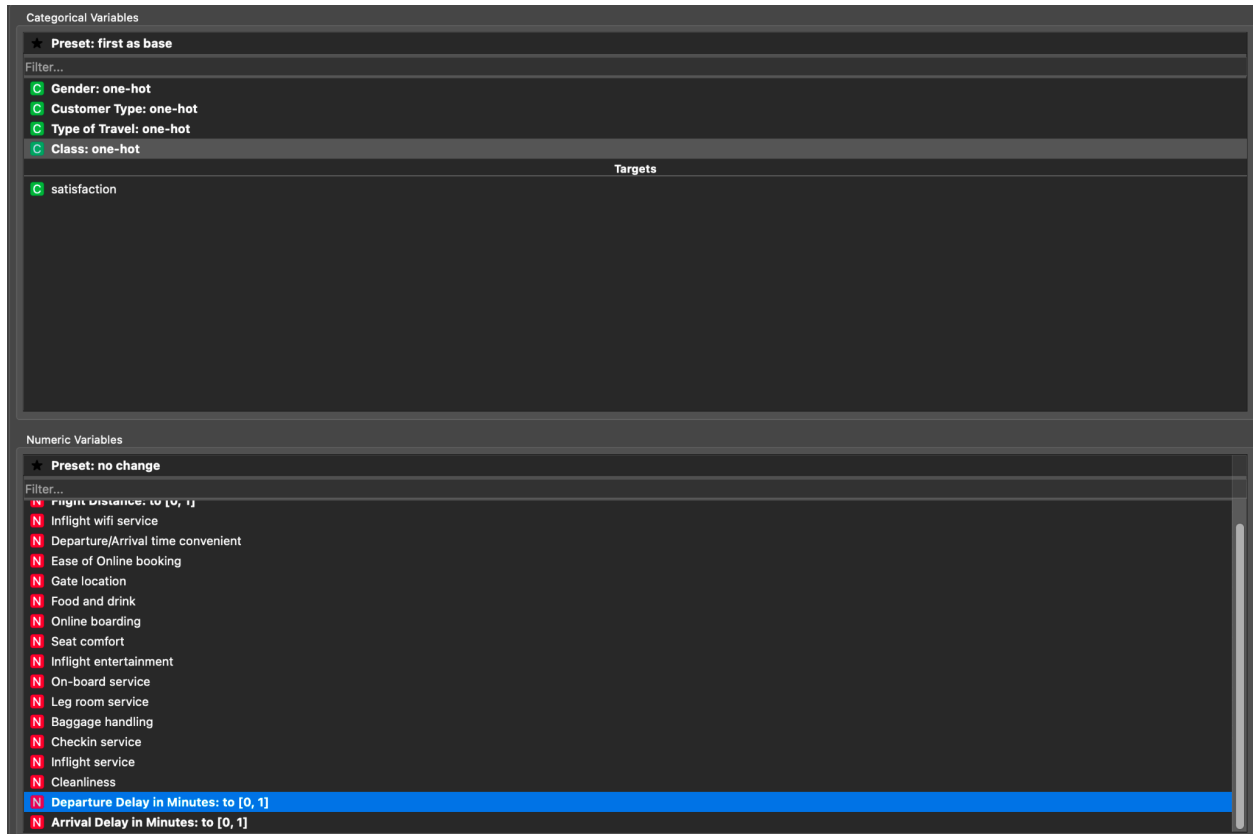
Nos dimos cuenta que había variables que tenían datos nulos y que debíamos hacer algo con ellos, era solo en la variable arrival delay in minutes, elegimos realizar **imputación** con el modelo tree, (no nos confiamos de imputar con la media porque su distribución no era normal ya que su media estaba sesgada por los outliers, había datos de un arrival delay de 1500 minutos..)



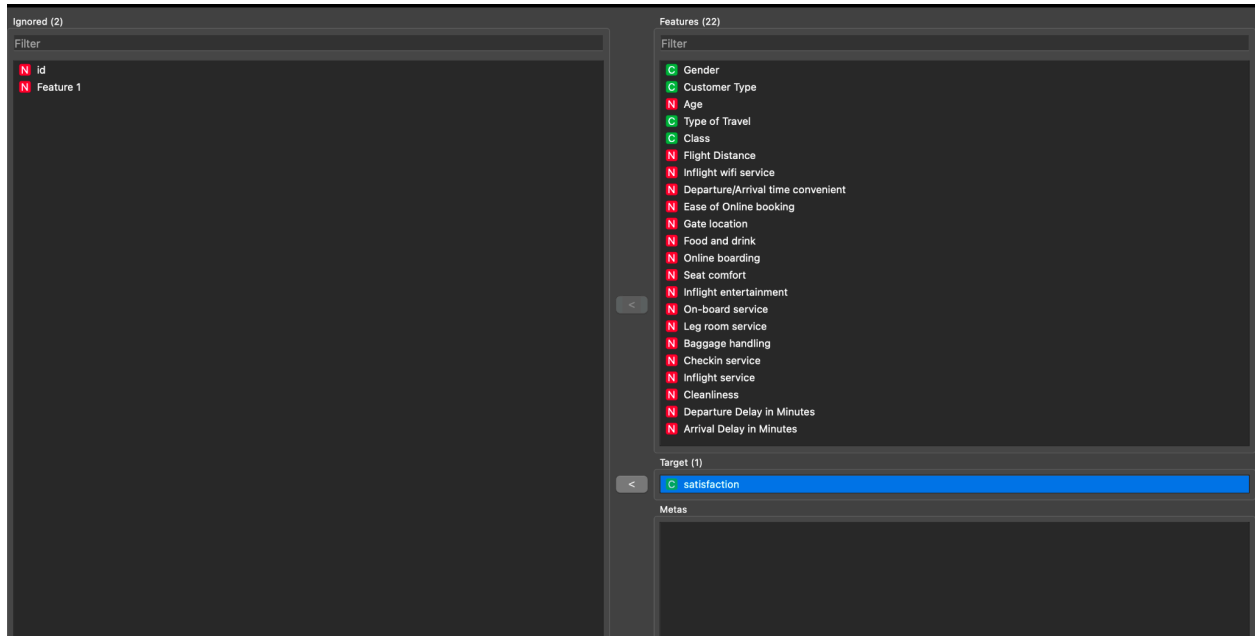
Después transformamos las categóricas a one hot encoding (como binarias pero que puedes tener + de 2 categorías)



Y también **normalizamos** Departure Delay in minutes, Arrival Delay in minutes, y Flight Distance porque tenían outliers y normalizadas mejoran la performance del modelo



Luego planteamos que nuestra variable target que estamos buscando predecir es **satisfaction**, ignoramos id, y feature 1 que no nos aportan valor para el modelo ya que son irrelevantes, y dejamos las otras 22 variables como features.



## 1. Eliminación de variables irrelevantes:

- Se descartó la columna **id y feature 1** ya que **no aporta valor predictivo** al tratarse de un identificador único sin relación con la satisfacción del pasajero.

## 2. Revisión de tipos de variables:

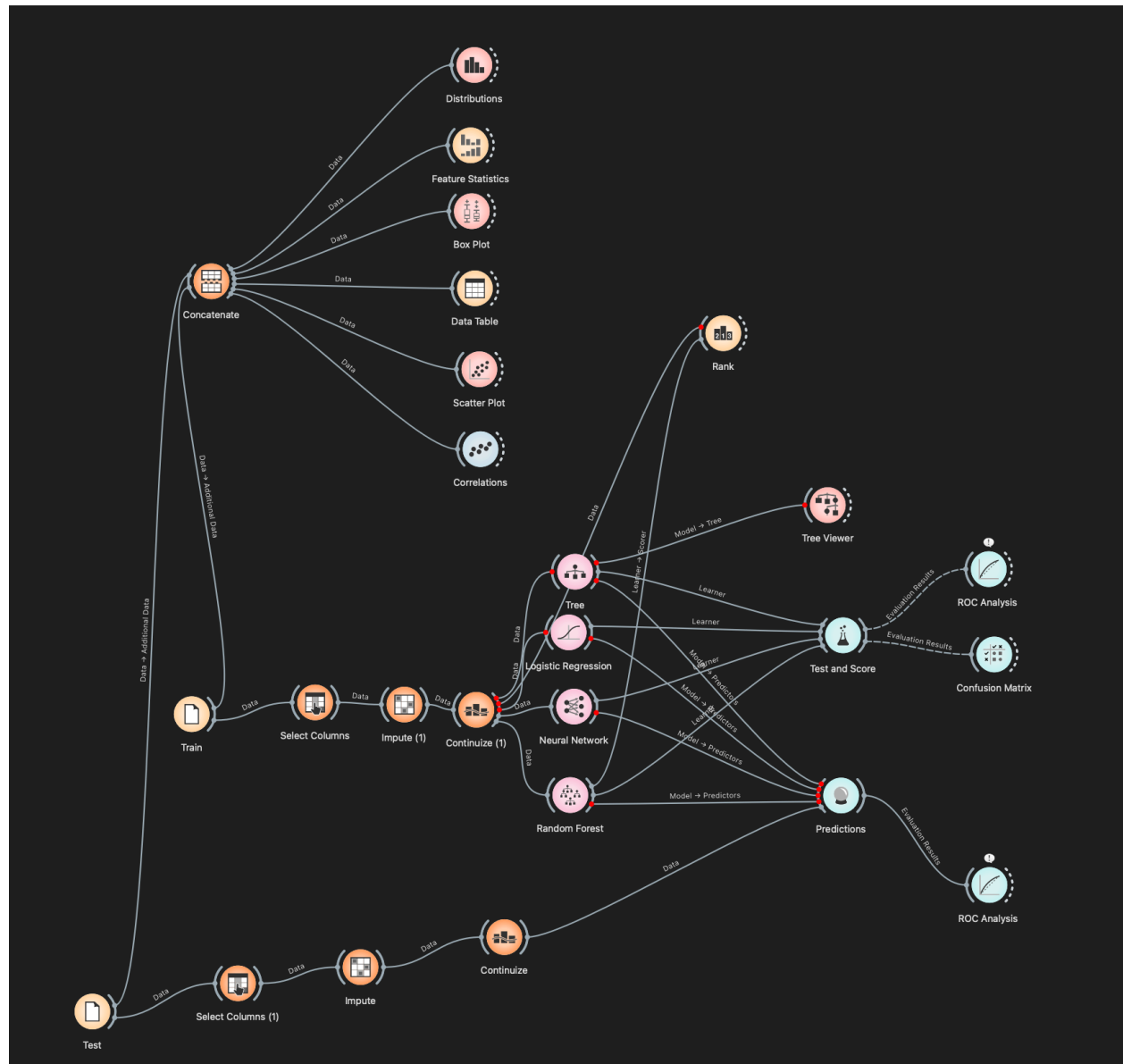
- Se transformó las variables categóricas (**Gender, Customer Type, Type of Travel, Class**, etc.) a un formato **one hot encoding**
- Las variables numéricas (**Age, Flight Distance, Inflight wifi service**, etc.) fueron mantenidas como continuas, ya que representan escalas o valores medibles útiles para los modelos
- Departure Delay in minutes, Arrival Delay in minutes, y Flight Distance fueron **normalizadas** para mejorar la performance del modelo

Cabe aclarar que esto supuso una diferencia significativa en la performance del modelo. Aquí es la performance sin realizar las transformaciones de variables e imputación

<input checked="" type="checkbox"/> Show performance scores		Target class: (Average over classes)					
Model	AUC	CA	F1	Prec	Recall	MCC	
Tree	0.844	0.819	0.819	0.841	0.819	0.659	
Logistic Regression	0.864	0.803	0.802	0.802	0.803	0.598	
Neural Network	0.881	0.779	0.779	0.792	0.779	0.569	
Random Forest	0.880	0.734	0.730	0.788	0.734	0.526	

Cuando realizamos las transformaciones pasamos a una mejora sustancial, pasando de un 81,2% de accuracy del mejor modelo (tree en este caso) a un 95,9% de accuracy de Random Forest

Model	AUC	CA	F1	Prec	Recall	MCC
Tree	0.931	0.944	0.944	0.944	0.944	0.886
Logistic Regression	0.927	0.875	0.875	0.875	0.875	0.745
Neural Network	0.993	0.956	0.956	0.956	0.956	0.909
Random Forest	0.992	0.959	0.959	0.959	0.959	0.917



## Evaluación de Modelos: ¿Cuál es el Mejor y Por Qué?

Model	AUC	CA	F1	Prec	Recall	MCC
Tree	0.931	0.944	0.944	0.944	0.944	0.886
Logistic Regression	0.927	0.875	0.875	0.875	0.875	0.745
Neural Network	0.993	0.956	0.956	0.956	0.956	0.909
Random Forest	0.992	0.959	0.959	0.959	0.959	0.917

Se entrenaron cuatro modelos de aprendizaje automático utilizando Orange Data Mining: **Random Forest**, **Neural Network**, **Tree** y **Logistic Regression**

Al someter nuestros cuatro modelos (**Random Forest**, **Neural Network**, **Tree** y **Logistic Regression**) a la prueba final con el conjunto de datos de test, obtuvimos una tabla comparativa de su rendimiento. El objetivo aquí es simple: determinar qué modelo es el más eficaz para predecir la satisfacción de un pasajero.

Basado en todas las métricas importantes, podemos afirmar con total seguridad que el modelo **Random Forest** es el claro ganador.

### 1. AUC (Área Bajo la Curva)

Esta es la métrica más importante y confiable para comparar modelos de clasificación. Mide la capacidad general del modelo para distinguir correctamente entre las dos clases (satisfecho vs. insatisfecho).

- Random Forest: 0.992
- Tree: 0.931
- Logistic Regression: 0.927
- **Neural Network: 0.993**

Un puntaje de 0.992 es extraordinariamente alto si bien el modelo Neural network es levemente superior con un 0.993 (el máximo perfecto es 1.0). Esto significa que el **Random Forest** es casi perfecto para diferenciar a un pasajero que estará satisfecho de uno que no lo estará. Su capacidad de discriminación es muy superior a la de los otros dos modelos.

## 2. CA (Classification Accuracy)

Esta es la métrica más intuitiva. Nos dice, del total de pasajeros en la prueba, ¿qué porcentaje fue clasificado correctamente?

- **Random Forest: 95,9%**
- Tree: 94.4%
- Logistic Regression: 87.5%
- Neural Network: 95,6%

**Random Forest** acertó en su predicción para el 95,9% de los pasajeros del conjunto de prueba. Es decir, de cada 1000 pasajeros nuevos, el modelo se equivocaría en menos de 40. Es un nivel de precisión altísimo y, de nuevo, el mejor de los cuatro.

## 3. F1 Score

Esta métrica es clave porque busca un balance entre dos tipos de errores: no predecir a un cliente como satisfecho cuando en realidad no lo está (Precisión) y no olvidarse de identificar a los que sí están satisfechos (Recall). Acá empataron random forest y neural network

- **Random Forest: 95,9%**
- Tree: 94.4%
- Logistic Regression: 87.5%
- **Neural Network: 95,6%**

Su alto F1-Score nos indica que el modelo es muy equilibrado. No solo es preciso en general, sino que es confiable tanto para identificar a los clientes satisfechos como a los insatisfechos, sin sesgarse hacia un lado.

## Conclusión de la Evaluación

El análisis comparativo no deja lugar a dudas. Aunque el Neural Network ofrece un rendimiento muy respetable y podría ser útil por su interpretabilidad, el **Random Forest** lo supera en algunos frentes. Es más robusto, más preciso y más confiable. La Regresión Logística, si bien es un modelo válido, no es competitiva para la complejidad de este problema en particular. Por lo tanto, para la fase final del proyecto, seleccionamos el **Random Forest** como nuestro modelo campeón.

Al aplicar el modelo sobre el conjunto de test, obtuvimos métricas altamente satisfactorias: una precisión del **95,9%**, un recall del **95,9%**, un AUC de **0.992** y un F1-score de **95,9%**. Esto indica que el modelo no solo acierta en la mayoría de las predicciones, sino que también mantiene un balance adecuado entre falsos positivos y falsos negativos.

		Predicted		$\Sigma$
		neutral or dissatisfied	satisfied	
Actual	neutral or dissatisfied	57363	1516	58879
	satisfied	2744	42281	45025
$\Sigma$		60107	43797	103904

La matriz de confusión revela que el modelo clasificó correctamente a 57.363 pasajeros como "neutral o insatisfechos" y a 42.281 como "satisfechos". Sin embargo, se identificaron 1.516 **falsos negativos** (insatisfechos clasificados incorrectamente como satisfechos) y 2744 **falsos positivos** (insatisfechos clasificados incorrectamente como satisfechos)(satisfechos mal clasificados como insatisfechos). Esta distribución de errores es favorable para nuestro objetivo, ya que minimizar los falsos positivos (es decir, no detectar a tiempo a un cliente insatisfecho) es clave para tomar acciones correctivas y evitar pérdida de fidelización.

Además, en el widget de predicciones se observó que más del 90% de las predicciones tienen un error menor a 0.1, lo cual demuestra que el modelo no solo acierta, sino que también lo hace con alto nivel de confianza. Esto refuerza su aplicabilidad práctica para asistir en decisiones de negocio orientadas a la experiencia del cliente.

### Interpretación de las predicciones – variables relevantes

Para interpretar el funcionamiento del modelo y entender qué factores influyen más en la predicción de la satisfacción, se utilizó el widget **Rank** de Orange. Este permite ordenar las variables según su importancia en la clasificación, aplicando dos criterios principales: **Gain Ratio** y **Gini Decrease**. El Gain Ratio mide cuánta información útil aporta una variable para hacer predicciones, pero además ajusta ese valor para evitar que se vean favorecidas las variables que tienen muchas categorías diferentes. Así ayuda a elegir mejor las variables realmente importantes. El Gini Decrease, en cambio, muestra cuánto ayuda una variable a dividir mejor los datos dentro de un árbol de decisión (como los usados en el modelo Random Forest). Cuanto más baja la impureza (mezcla de clases) gracias a una variable, más importante se considera.

Los resultados indican que las variables más influyentes para predecir la satisfacción de los pasajeros son, en primer lugar, **Online Boarding**, seguida por **Class = Business**, y luego **Type of Travel = Personal Travel**, **Type of Travel = Business Travel**, **Class = Eco**.

La variable que más influye es Online Onboarding, lo cual es consistente con las hipótesis planteadas en el análisis exploratorio. Además, quienes viajan en clase Economy muestran menores niveles de satisfacción en comparación con Business o Eco Plus, lo que refuerza la relevancia de la variable **Class**.

Por último, la calidad del proceso de embarque y los servicios a bordo como el WiFi y el entretenimiento tienen un impacto significativo en la percepción del pasajero.

Esta interpretación no solo refuerza la validez del modelo, sino que también aporta información estratégica para la aerolínea, al indicar qué aspectos priorizar para mejorar la experiencia del cliente.

Scoring Methods		#	Info. gain	Gain ratio	Gini	Rand...rest
<input checked="" type="checkbox"/> Information Gain						
<input checked="" type="checkbox"/> Information Gain Ratio						
<input checked="" type="checkbox"/> Gini Decrease						
<input type="checkbox"/> ANOVA						
<input type="checkbox"/> X <sup>2</sup>						
<input type="checkbox"/> ReliefF						
<input type="checkbox"/> FCBF						
1	N Online boarding		0.288	0.146	0.180	0.151
2	N Class=Business		0.192	0.192	0.125	0.065
3	N Type of Travel=Personal Travel		0.164	0.183	0.099	0.042
4	N Type of Travel=Business travel		0.164	0.183	0.099	0.058
5	N Class=Eco		0.155	0.156	0.100	0.092
6	N Inflight wifi service		0.138	0.069	0.091	0.159
7	N Inflight entertainment		0.133	0.067	0.087	0.034

## Conclusiones

A lo largo del proyecto se logró implementar de forma exitosa un proceso integral de análisis, exploración y modelado predictivo utilizando Orange Data Mining. Partimos de una problemática concreta del sector aeronáutico: anticipar la insatisfacción de los pasajeros con el objetivo de mejorar su experiencia y reducir potenciales pérdidas de clientes.

Durante el análisis exploratorio se detectaron patrones relevantes que permitieron formular hipótesis informadas, como el mayor grado de insatisfacción entre pasajeros de clase Economy y aquellos que viajan por motivos personales. Estas hipótesis fueron validadas mediante estadísticas descriptivas, gráficos y pruebas estadísticas, y se confirmaron posteriormente por el modelo como las variables más influyentes en la predicción.

Identificamos 2 variables que podrían ser útiles para el modelo: “timequeue” que mide el tiempo total que los clientes realizan filas (check-in, migraciones, scanners); y “oldcomplaints” que reúne valores de quejas anteriores hechas por los clientes. A nuestro parecer estas dos variables podrían ser de importancia a la hora de correr el modelo y poder capturar mejores resultados

Luego del preprocesamiento y la selección de variables, se entrenaron y evaluaron cuatro modelos distintos: árbol de decisión, regresión logística y neural network, Random Forest. Este último demostró ser el más efectivo, alcanzando un **recall del 95,9%**, un **AUC de 0,992**, y valores equilibrados de **F1-score, precisión** (todas iguales o superiores a 0,959). Esto lo convierte en una herramienta sólida para identificar correctamente a los pasajeros insatisfechos.

En cuanto a las métricas, se priorizó especialmente el **recall**, ya que en este contexto los **falsos negativos** (pasajeros insatisfechos que no son detectados) representan un mayor impacto para la aerolínea ya que se corre el peligro de perder un cliente. Detectarlos a tiempo permite aplicar acciones preventivas como descuentos, encuestas o mejoras específicas.



Finalmente, el modelo no solo ofreció resultados cuantitativos sobresalientes, sino también **coherencia con las hipótesis iniciales** y una lectura clara de qué atributos influyen en la experiencia del cliente, reforzando así la confianza en los resultados y su aplicabilidad en contextos reales.

### Lecciones Aprendidas

- **Explorar antes de modelar mejora los resultados:** El análisis exploratorio previo fue clave para detectar patrones y formular hipótesis sólidas. Esto permitió elegir variables relevantes y entender mejor el contexto, lo cual fortaleció la interpretación y aumentó la efectividad del modelo.
- **El preprocesamiento influye directamente en la calidad del modelo:** La eliminación de columnas irrelevantes (como ID y Feature 1), la revisión de tipos de datos y la unificación del dataset (concatenando train y test para el EDA) facilitaron un entrenamiento más robusto y representativo.
- **No todo es precisión: contexto + métricas:** La elección del modelo no se basó únicamente en su exactitud, sino también en su adecuación al problema. En este caso, se priorizó el **recall** para minimizar falsos negativos y no dejar pasar pasajeros insatisfechos, lo cual tiene mayor impacto en el negocio.
- **Orange combina simplicidad visual con profundidad analítica:** El entorno permitió crear flujos de trabajo comprensibles, reproducibles y potentes, facilitando la experimentación con distintos modelos, métricas y visualizaciones sin perder trazabilidad.
- **La visualización es más que presentación: es comprensión:** Herramientas como Box Plot, Feature Statistics, Rank y la matriz de confusión fueron fundamentales para validar hipótesis, interpretar resultados y comunicar hallazgos con claridad y evidencia.