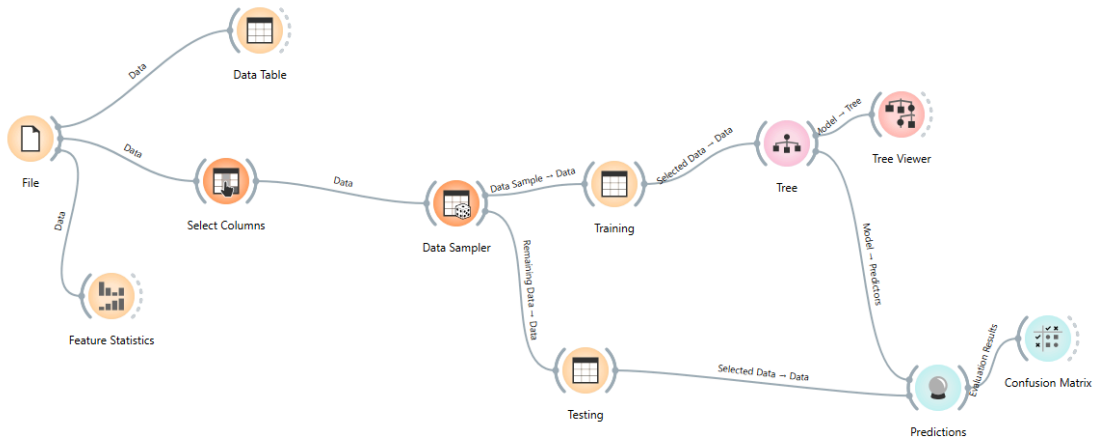
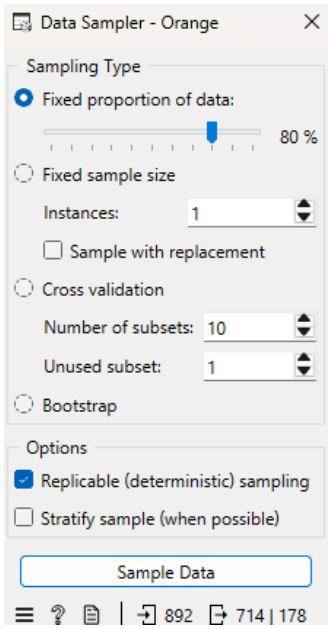
Orange - Titanic



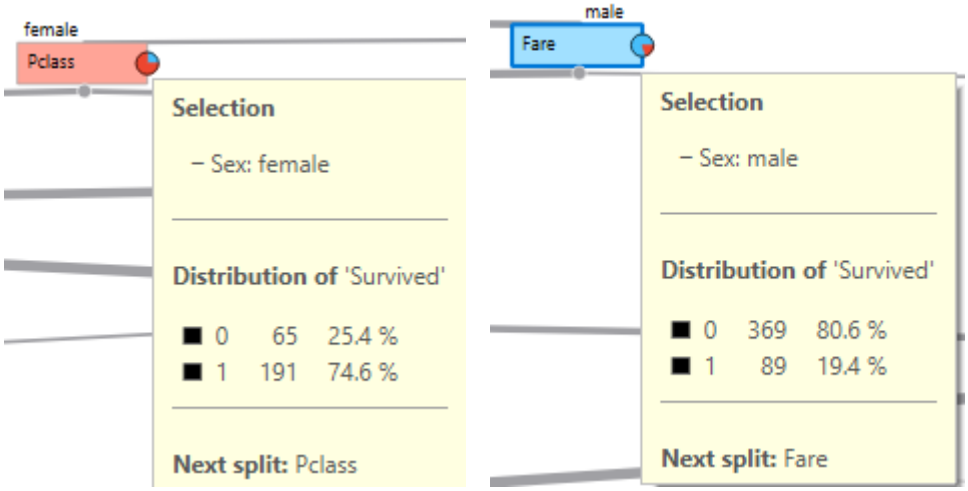
Para analizar el archivo, realizamos diferentes conexiones a través de Orange. En primer lugar, removimos las variables que no queríamos analizar con el módulo Select Columns, y definimos la variable Survived como el target de nuestro modelo. Luego, dividimos la información en dos porciones, una más grande, con el 80% de los datos, para entrenar al modelo, y otra más pequeña para luego testearlo, con el módulo Data Sampler.



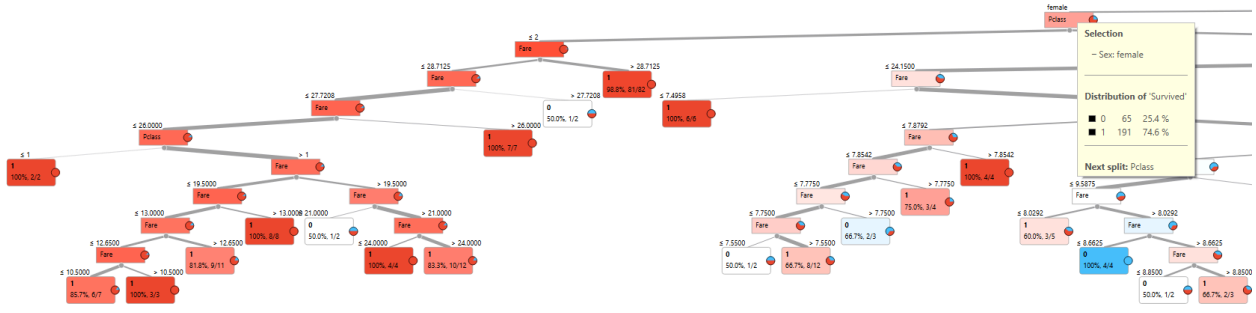
Del módulo “Training”, quisimos visualizar la información a través de un árbol, para entender las conexiones entre las diferentes variables.

En primer lugar, vemos como el árbol primero divide por sexo, lo cual indica que esta variable fue la más importante para predecir la supervivencia. Si sex = female, el camino lleva hacia la parte izquierda del árbol. y si si sex = male, hacia la derecha.

Esto ya nos dice que el modelo detectó una gran diferencia en supervivencia entre hombres y mujeres (lo que podría indicar una prioridad por la protección de mujeres y niños). En términos numéricos, el árbol muestra cómo el 74% de las mujeres sobrevivieron, mientras que tan solo el 19% de los hombres sobrevivieron.

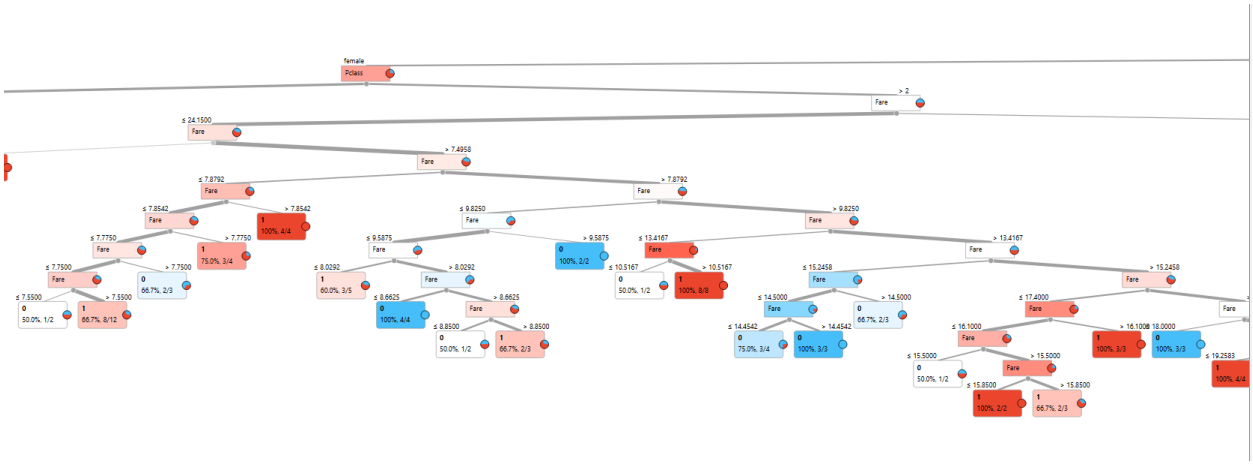


Para este informe, nos enfocamos principalmente en las pasajeras femeninas, y qué proporción de ellas sobrevivió, en relación al Pclass (primera, segunda o tercera clase) y al Fare (cantidad de dinero abonada para viajar). Obtuvimos la siguiente rama:



En el lado de las mujeres, el árbol sigue dividiendo según Pclass (clase del boleto: 1, 2 o 3) y Fare (tarifa abonada).

Si la clase de la pasajera era de las más altas, como 1 o 2, casi todos los nodos que representan a mujeres de primera o segunda clase son rojos oscuros, con valores cercanos a 100% de 1 (sobrevivieron). Esto significa que la gran mayoría de las mujeres de estas clases sobrevivieron según los datos.



Ahora bien, si la pasajera tenía tercera clase, se observan nodos azules o más claros, indicando mayor proporción de 0 (murieron), en relación también al Fare que abonaron estas pasajeras. Por ejemplo, si el Fare ≤ 7.75 o Fare ≤ 8.6625, el porcentaje de supervivencia baja drásticamente (valores como 50%, 60% o incluso 0%). Esto muestra que entre las mujeres de tercera clase, las que pagaron tarifas más bajas (probablemente ubicaciones más económicas del barco) tuvieron menos probabilidad de sobrevivir.

Luego, utilizamos el modelo entrenado para analizar con qué precisión puede predecir la supervivencia de un pasajero, en base a la información que utilizamos para entrenarlo. A continuación, mostramos un fragmento de lo predicho por el modelo, su grado de error y lo que ocurrió en la realidad:

	Tree	error	Survived	Pclass	Sex	Fare
1	0.90 : 0.10 → 0	0.900	1	3	male	7.2292
2	1.00 : 0.00 → 0	1.000	1	1	male	76.7292
3	1.00 : 0.00 → 0	1.000	1	2	male	14.5000
4	0.01 : 0.99 → 1	0.012	1	1	female	39.6000
5	0.75 : 0.25 → 0	0.250	0	3	male	7.7750
6	0.01 : 0.99 → 1	0.012	1	1	female	55.9000
7	0.00 : 1.00 → 1	0.000	1	3	female	15.8500
8	0.87 : 0.13 → 0	0.133	0	2	male	10.5000
9	0.75 : 0.25 → 0	0.250	0	1	male	110.8833
10	1.00 : 0.00 → 0	0.000	0	3	male	69.5500
11	0.90 : 0.10 → 0	0.100	0	3	male	7.2292
12	0.67 : 0.33 → 0	0.333	0	1	male	52.0000
13	0.01 : 0.99 → 1	0.012	1	1	female	71.0000
14	1.00 : 0.00 → 0	0.000	0	1	male	33.5000
15	0.00 : 1.00 → 1	1.000	0	3	female	6.7500
16	0.75 : 0.25 → 0	0.250	0	3	male	7.7750
17	0.75 : 0.25 → 0	0.250	0	3	male	7.7958
18	1.00 : 0.00 → 0	1.000	1	1	male	29.7000
19	0.81 : 0.19 → 0	0.194	0	3	male	8.0500
20	0.81 : 0.19 → 0	0.194	0	3	male	8.0500
21	0.75 : 0.25 → 0	0.250	0	3	female	14.4542
22	0.81 : 0.19 → 0	0.194	0	3	male	8.0500
23	1.00 : 0.00 → 0	1.000	1	1	male	57.0000
24	0.50 : 0.50 → 0	0.500	1	1	male	26.5500
25	0.89 : 0.11 → 0	0.889	1	3	male	7.2250
26	0.01 : 0.99 → 1	0.012	1	1	female	57.0000
27	0.33 : 0.67 → 1	0.667	0	3	male	15.8500
28	0.58 : 0.42 → 0	0.417	0	3	male	7.9250
29	0.75 : 0.25 → 0	0.250	0	2	male	26.0000
30	1.00 : 0.00 → 0	0.000	0	2	male	11.5000
31	1.00 : 0.00 → 0	0.000	0	2	male	27.0000
32	0.81 : 0.19 → 0	0.194	0	3	male	8.0500
33	1.00 : 0.00 → 0	0.000	0	3	female	27.9000
34	0.90 : 0.10 → 0	0.100	0	3	male	7.2292
35	0.50 : 0.50 → 0	0.500	1	3	female	15.5000

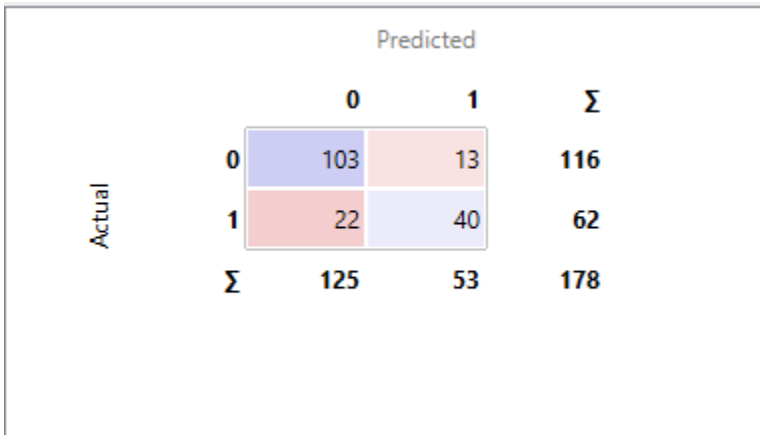
En base a lo obtenido, sacamos algunas conclusiones sobre el modelo:

- El modelo predijo con una muy alta precisión la supervivencia de mujeres de primera clase o con tarifas altas, con errores muy bajos (0,01 aproximadamente) o nulos. Esto refuerza lo visto en el árbol: las mujeres de clases altas casi siempre sobrevivieron.
- Comete errores en hombres de tercera clase, con un error alto de 1, en algunos casos, donde el modelo los clasifica a veces como sobrevivientes cuando en realidad no lo hicieron, o viceversa.
- En varios casos, el modelo duda, mostrando valores intermedios de error, indicando que las condiciones de los pasajeros no eran claras, y no pudo predecir con firmeza su supervivencia.

El modelo capta correctamente la tendencia principal: el sexo y la clase fueron los factores más determinantes en la supervivencia. Sin embargo, el modelo duda cuando analiza la variable Fare.

Model	AUC	CA	F1	Prec	Recall	MCC
Tree	0.808	0.803	0.799	0.800	0.803	0.555

También, al final del módulo Predictions, vemos algunas métricas que indican cuán preciso fue el modelo. En este caso, tiene un Accuracy del 80%, lo cual tiene sentido, ya que sorteamos el 80% de la información para entrenar al modelo. Es un modelo fuerte y preciso, pero dado que se trata de “Vida o Muerte”, probablemente se busque un modelo que tenga una precisión mayor al 90% como mínimo, ya que, posiblemente ningún futuro pasajero quiera abordar a un barco mientras le dicen que según el modelo predictivo, están 80% seguros que va a sobrevivir.



Por último, obtuvimos la matriz de confusión, que muestra que en los casos donde hubo muertes, el modelo predijo 103 de ellas correctamente (“Verdaderos negativos”), y 22 de ellas incorrectamente (“Falsos negativos”), y en cuanto a aquellos que sobrevivieron, el modelo predijo 40 de ellos correctamente (“Verdaderos positivos”) y 13 de ellos incorrectamente (“Falsos positivos”). En su totalidad, el modelo tuvo 143 aciertos de 178 casos, y 35 errores.

En conclusión, el modelo supo clasificar correctamente la mayoría de los pasajeros que no sobrevivieron (103/116). A su vez, identificó de manera correcta a 40 de los 62 sobrevivientes. Los errores se concentran en sobrevivientes clasificados como no sobrevivientes (22 casos), lo que sugiere que podríamos mejorar el modelo utilizando variables como Age o si tenía familia a bordo. El modelo tiene alto rendimiento general pero muestra leve tendencia a orientarse a la no supervivencia de los pasajeros, lo cual tiene sentido siendo que la mayoría de las personas a bordo no sobrevivieron (550 de un total de 892 pasajeros)