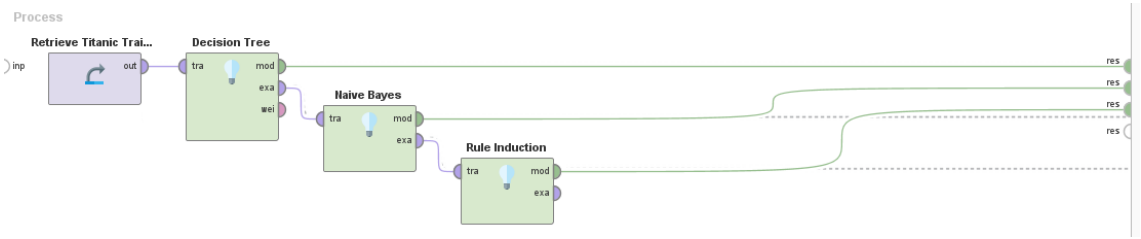
Modeling

El objetivo de este tutorial es el de aprender sobre el modelado predictivo, el cual nos permite, a través de técnicas de aprendizaje, encontrar patrones en grandes datasets de datos y con ellos realizar predicciones dada una situación nueva.



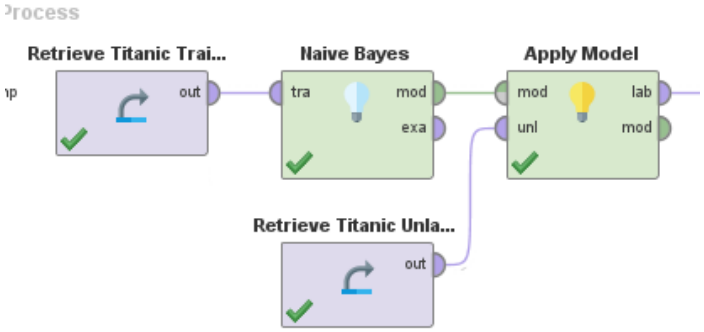
El nodo “Árbol de decisión”, como su nombre indica, crea un árbol donde realiza decisiones binarias basadas en la correlación con la salida (survival). Esta herramienta nos termina mostrando cómo hay patrones que son más significativos a la hora de sobrevivir siendo hombre o mujer, siendo primera clase o segunda, teniendo o no hijos, etc. Luego se utiliza el bloque “Naive bayes” para determinar, en general, la probabilidad a priori de sobrevivir. Finalmente, el bloque “Rule induction”, nos presenta un set de reglas provenientes de los pasos anteriores donde nos dice condicionalmente si , dadas ciertas condiciones, una persona sobrevive o no (con cierta certeza).

RuleModel

```
if Sex = Male and Passenger Fare ≤ 26.269 then No  (57 / 367)
if Sex = Female and Passenger Class = First then Yes  (97 / 4)
if Sex = Male and Passenger Fare > 31.137 then No  (33 / 90)
if Passenger Class = Second and Age ≤ 28.500 then Yes  (36 / 4)
if Passenger Fare ≤ 24.808 and Passenger Fare > 15.373 and Age > 29.441 then Yes  (18 / 3)
if Passenger Fare ≤ 14.281 then Yes  (68 / 40)
if Passenger Class = Third and Passenger Fare > 23.746 then No  (1 / 23)
if Passenger Class = Second and Passenger Fare > 30.375 then Yes  (4 / 0)
if No of Parents or Children on Board ≤ 0.500 and Age ≤ 30.441 and Passenger Fare ≤ 28.710 and Age > 28.500 then No  (1 / 8)
if Age ≤ 54 then Yes  (33 / 22)
if Age ≤ 71 then No  (0 / 6)
else Yes  (0 / 0)

correct: 750 out of 915 training examples.
```

Scoring



En el tutorial de Scoring, aprendemos como utilizar las características obtenidas de la data para poder predecir el resultado de un conjunto de datos sin clasificar. Utilizamos Naive bayes para las predicciones y aplicamos el aprendizaje del bloque sobre un conjunto sin etiquetas.

Row No.	prediction(S...	confidence(...	confidence(...	Age	Passenger ...	Sex	No of Sibling...	No of Parent...	Passenger F...
1	Yes	0.987	0.013	0.917	First	Male	1	2	151.550
2	Yes	0.714	0.286	53	First	Female	2	0	51.479
3	Yes	0.542	0.458	71	First	Male	0	0	49.504

Finalmente se llega a clasificar toda la data del conjunto unlabeled con cierto grado de certeza para cada una de las entradas.

Test Splits and Validation

Al igual que en el caso anterior, se demuestra el comportamiento de un modelo al intentar predecir como clasificar data desde el aprendizaje de otra. En este caso, la idea del tutorial es mostrar como podemos partir un dataset entero en dos partes donde utilizamos una de entrenamiento y otra de validación para nuestro modelo, de forma de poder probar el correcto funcionamiento del mismo antes de probarlo con data nueva.

Cross Validation

Cross Validation surge como solución a un problema que introduce la división y validación de data. ¿Qué pasa si nuestro dataset tiene grupos de datos muy distintos y siempre utilizamos la misma forma de dividir los datos? Esto nos puede llevar a crear sesgos y generar aprendizajes erróneos. Es por ello que la técnica de cross validation sirve para dividir el dataset en partes iguales y rotar varias veces las divisiones de los datos de forma de entrenar y validar el modelo de formas distintas en todas las iteraciones, reduciendo considerablemente el sesgo.

Visual model comparisson

En este tutorial, se aprende sobre como comparar nuestros modelos de entrenamiento con el bloque “ROC”. Este bloque nos permite graficar una curva de los verdaderos positivos contra los falsos positivos de forma de poder visualizar la performance de cada método de aprendizaje para el dataset utilizado.

Resultado: