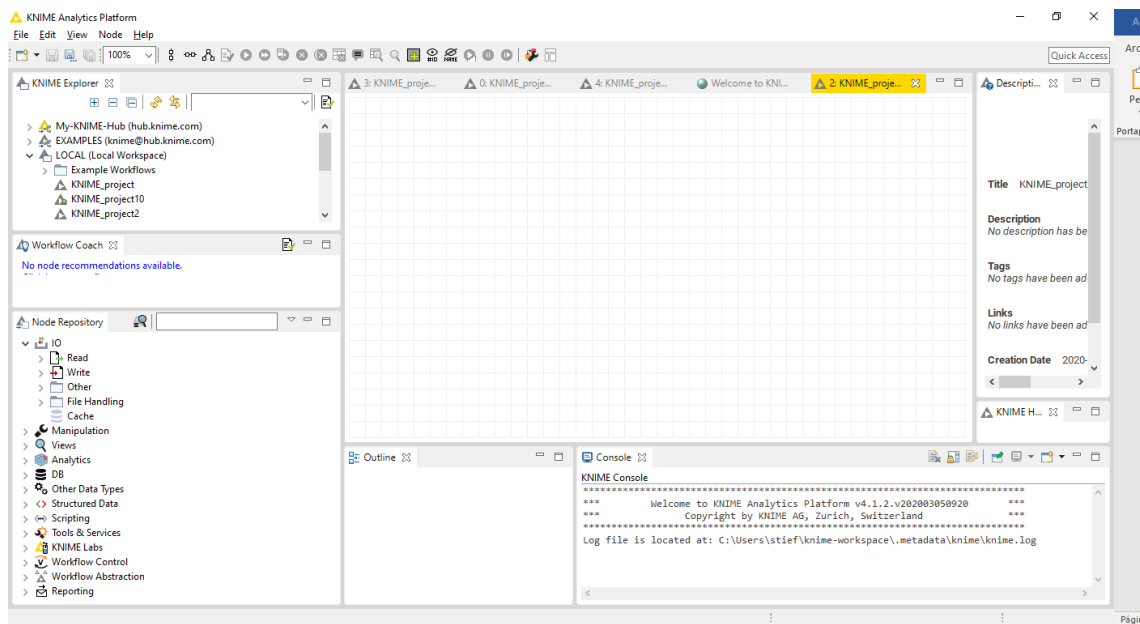


Tarea del 28 de abril: árboles de decisión con Knime.

En esta ocasión vamos a realizar un árbol de decisión con la herramienta Knime y el archivo de datos empresas.txt y por último con el archivo weather.arff.

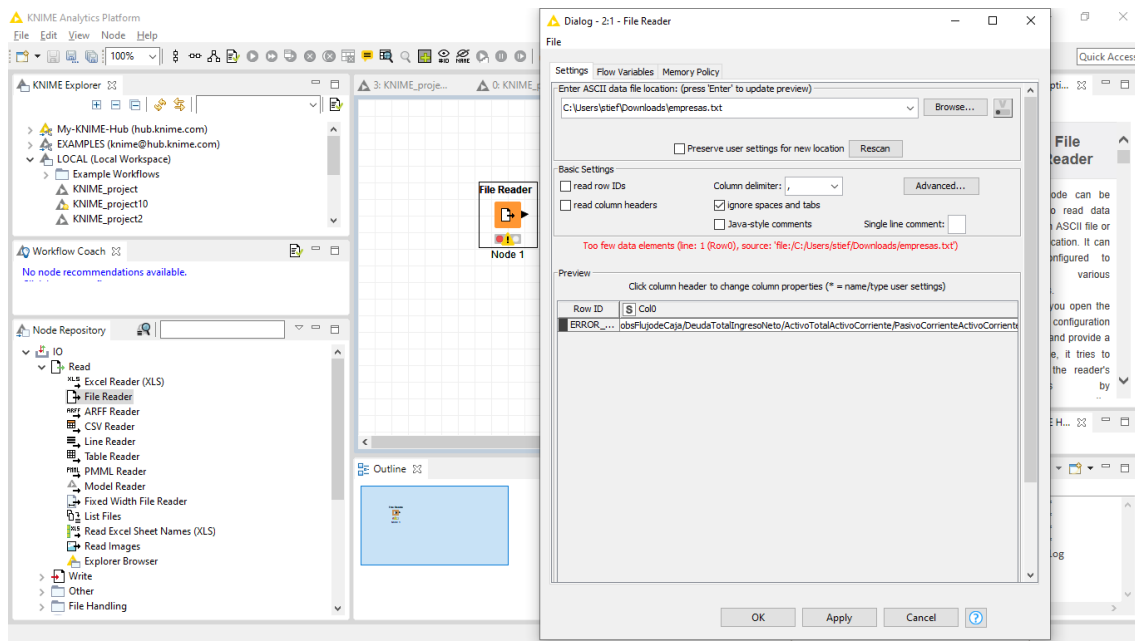
Comenzamos abriendo el programa Knime y creando un espacio de trabajo tal y como hemos hecho en las tareas anteriores.



Como queremos leer un archivo .txt, buscaremos en “Node Repository” un nodo capaz de leer archivos en este formato, es decir, un nodo “Filer Reader”. Lo seleccionamos y arrastramos al espacio de trabajo.

Para configurarlo, con el nodo seleccionado, pulsamos F6.

Navegamos por las carpetas hasta encontrar el archivo empresas.txt, el cual seleccionamos y abrimos. Observaremos que los datos no aparecen correctamente, y solo tenemos una columna de atributos.



Esto se debe a que tenemos como delimitador de columna “,”, cuando debería ser “<tab>”. Tras cambiar esto, observamos que nos aparecen correctamente las columnas pero los números no están bien.

Dialog - 2:1 - File Reader

File

Settings Flow Variables Memory Policy

Enter ASCII data file location: (press 'Enter' to update preview)

C:\Users\stief\Downloads\empresas.txt Browse...

☐ Preserve user settings for new location Rescan

Basic Settings

☐ read row IDs Column delimiter: <tab> Advanced...

☒ read column headers ☒ ignore spaces and tabs

☐ Java-style comments Single line comment:

Preview

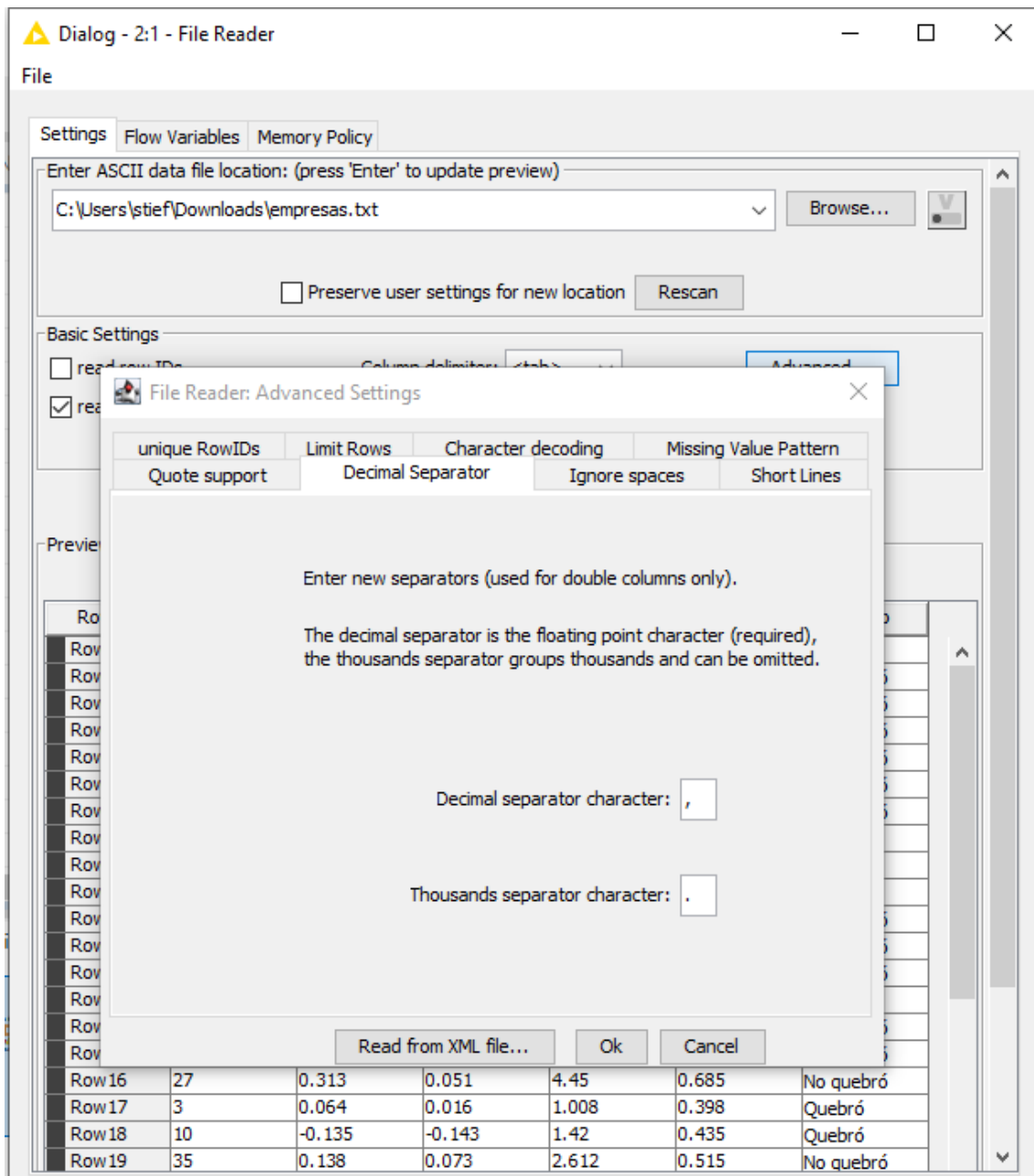
Click column header to change column properties (* = name/type user settings)

| Row ID | I obs | S Flujode... | S Ingreso... | S ActivoC... | S ActivoC... | S Grupo |
|--------|-------|--------------|--------------|--------------|--------------|-----------|
| Row0 | 8 | -,0653 | -,0566 | 1,3737 | ,4032 | Quebró |
| Row1 | 29 | -,0173 | ,0233 | 2,0538 | ,3484 | No quebró |
| Row2 | 25 | ,1933 | ,0473 | 2,2506 | ,3309 | No quebró |
| Row3 | 40 | -,333 | -,0854 | 3,0124 | ,473 | No quebró |
| Row4 | 42 | ,5603 | ,1112 | 4,2918 | ,4443 | No quebró |
| Row5 | 24 | ,3776 | ,1075 | 3,2651 | ,3548 | No quebró |
| Row6 | 41 | ,4785 | ,091 | 1,2444 | ,1847 | No quebró |
| Row7 | 13 | ,0109 | ,0011 | 2,1495 | ,6969 | Quebró |
| Row8 | 11 | -,2298 | -,2961 | ,331 | ,1824 | Quebró |
| Row9 | 20 | ,1227 | ,1055 | 1,1434 | ,1655 | Quebró |
| Row10 | 38 | ,2907 | ,0597 | 1,8381 | ,3786 | No quebró |
| Row11 | 34 | ,1398 | -,0312 | ,4611 | ,2643 | No quebró |
| Row12 | 23 | ,0769 | ,0195 | 2,0069 | ,5304 | No quebró |
| Row13 | 19 | ,0115 | -,0032 | 1,2602 | ,6038 | Quebró |
| Row14 | 22 | ,5135 | ,1001 | 2,4871 | ,5368 | No quebró |
| Row15 | 45 | ,1661 | ,0351 | 2,4527 | ,137 | No quebró |
| Row16 | 27 | ,3132 | ,0511 | 4,45 | ,6852 | No quebró |
| Row17 | 3 | ,0643 | ,0156 | 1,0077 | ,3978 | Quebró |
| Row18 | 10 | -,1353 | -,1433 | 1,4196 | ,4347 | Quebró |
| Row19 | 35 | ,1379 | ,0728 | 2,6123 | ,5151 | No quebró |

OK Apply Cancel ?

Esto se debe a que la variable es de tipo string, es decir es una cadena de caracteres. Pero nosotros lo que tenemos en esa variable son números, números con decimales, lo que causa que la coma que separa la parte entera de la decimal de causa conflicto.

Esto lo resolvemos yéndonos a la configuración avanzada, y en la pestaña "Decimal Separator" vamos a cambiar "Decimal separator character" por una coma, y "Thousands separator character" por un punto.



Tras pulsar "Ok" veremos como no sólo aparecen correctamente los números, si no que el tipo de variable que es ha cambiado de tipo string a tipo decimal.

Dialog - 2:1 - File Reader

File

Settings Flow Variables Memory Policy

Enter ASCII data file location: (press 'Enter' to update preview)

C:\Users\stief\Downloads\empresas.txt Browse...

☐ Preserve user settings for new location Rescan

Basic Settings

☐ read row IDs Column delimiter: <tab> Advanced...

☒ read column headers ☒ ignore spaces and tabs

☐ Java-style comments Single line comment:

Preview

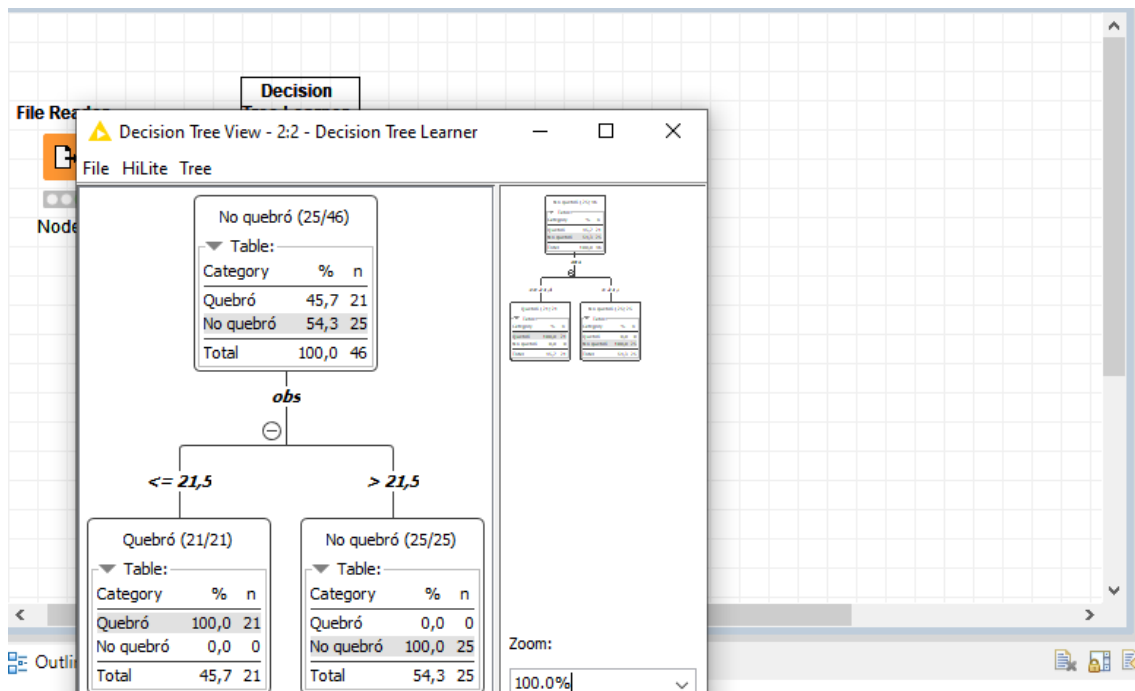
Click column header to change column properties (* = name/type user settings)

| Row ID | I obs | D Flujode... | D Ingreso... | D ActivoC... | D ActivoC... | S Grupo |
|--------|-------|--------------|--------------|--------------|--------------|-----------|
| Row0 | 8 | -0.065 | -0.057 | 1.374 | 0.403 | Quebró |
| Row1 | 29 | -0.017 | 0.023 | 2.054 | 0.348 | No quebró |
| Row2 | 25 | 0.193 | 0.047 | 2.251 | 0.331 | No quebró |
| Row3 | 40 | -0.333 | -0.085 | 3.012 | 0.473 | No quebró |
| Row4 | 42 | 0.56 | 0.111 | 4.292 | 0.444 | No quebró |
| Row5 | 24 | 0.378 | 0.107 | 3.265 | 0.355 | No quebró |
| Row6 | 41 | 0.478 | 0.091 | 1.244 | 0.185 | No quebró |
| Row7 | 13 | 0.011 | 0.001 | 2.15 | 0.697 | Quebró |
| Row8 | 11 | -0.23 | -0.296 | 0.331 | 0.182 | Quebró |
| Row9 | 20 | 0.123 | 0.105 | 1.143 | 0.166 | Quebró |
| Row10 | 38 | 0.291 | 0.06 | 1.838 | 0.379 | No quebró |
| Row11 | 34 | 0.14 | -0.031 | 0.461 | 0.264 | No quebró |
| Row12 | 23 | 0.077 | 0.019 | 2.007 | 0.53 | No quebró |
| Row13 | 19 | 0.011 | -0.003 | 1.26 | 0.604 | Quebró |
| Row14 | 22 | 0.513 | 0.1 | 2.487 | 0.537 | No quebró |
| Row15 | 45 | 0.166 | 0.035 | 2.453 | 0.137 | No quebró |
| Row16 | 27 | 0.313 | 0.051 | 4.45 | 0.685 | No quebró |
| Row17 | 3 | 0.064 | 0.016 | 1.008 | 0.398 | Quebró |
| Row18 | 10 | -0.135 | -0.143 | 1.42 | 0.435 | Quebró |
| Row19 | 35 | 0.138 | 0.073 | 2.612 | 0.515 | No quebró |

OK Apply Cancel ?

Tras esto pulsamos "Ok" y F7 para ejecutar el nodo.

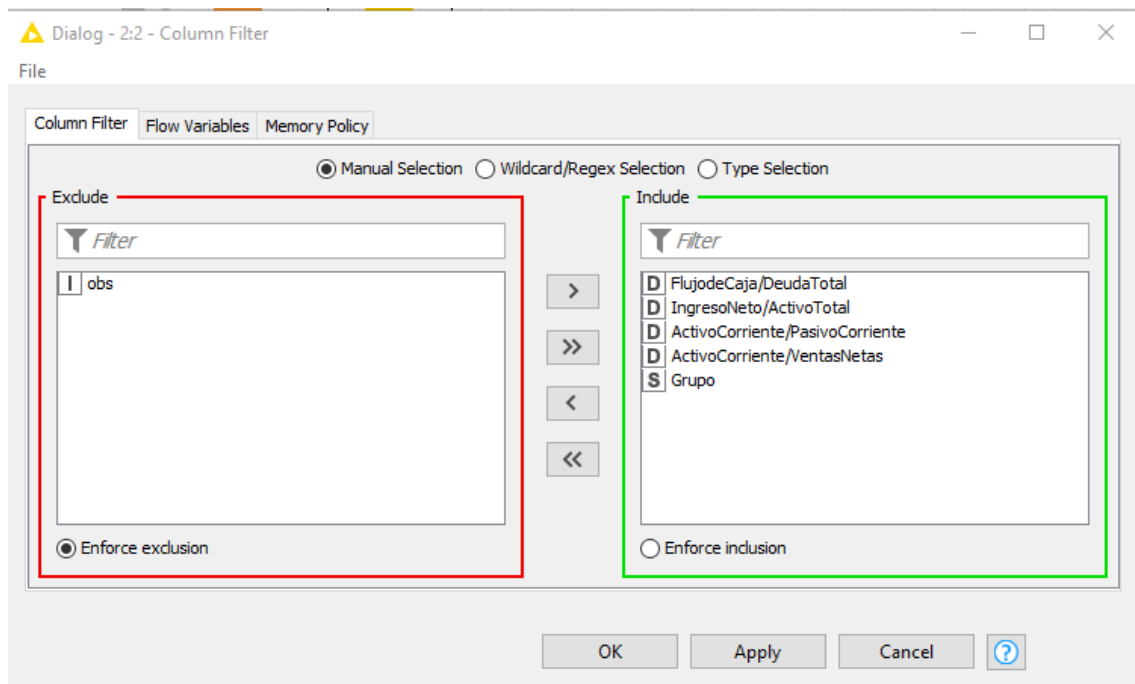
Si creásemos un nodo "Tree Decision Learner" y lo conectásemos con el nodo "File Reader", configurándolo para que la clase sea el atributo grupo, que es dónde se encuentran las etiquetas, observaríamos al ejecutarlo lo siguiente:



Observamos que todas quedaron bien clasificadas. Esto se debe a que el atributo “obs” está incluido como una de las columnas, cuando realmente este no es un atributo que tendremos realmente entre los datos, sino que simplemente ordena los datos. En este caso los ordenó poniendo primero las empresas que quebraron y después las que no quebraron. Por ello debemos quitar esa columna de los datos que usaremos para el árbol de decisiones.

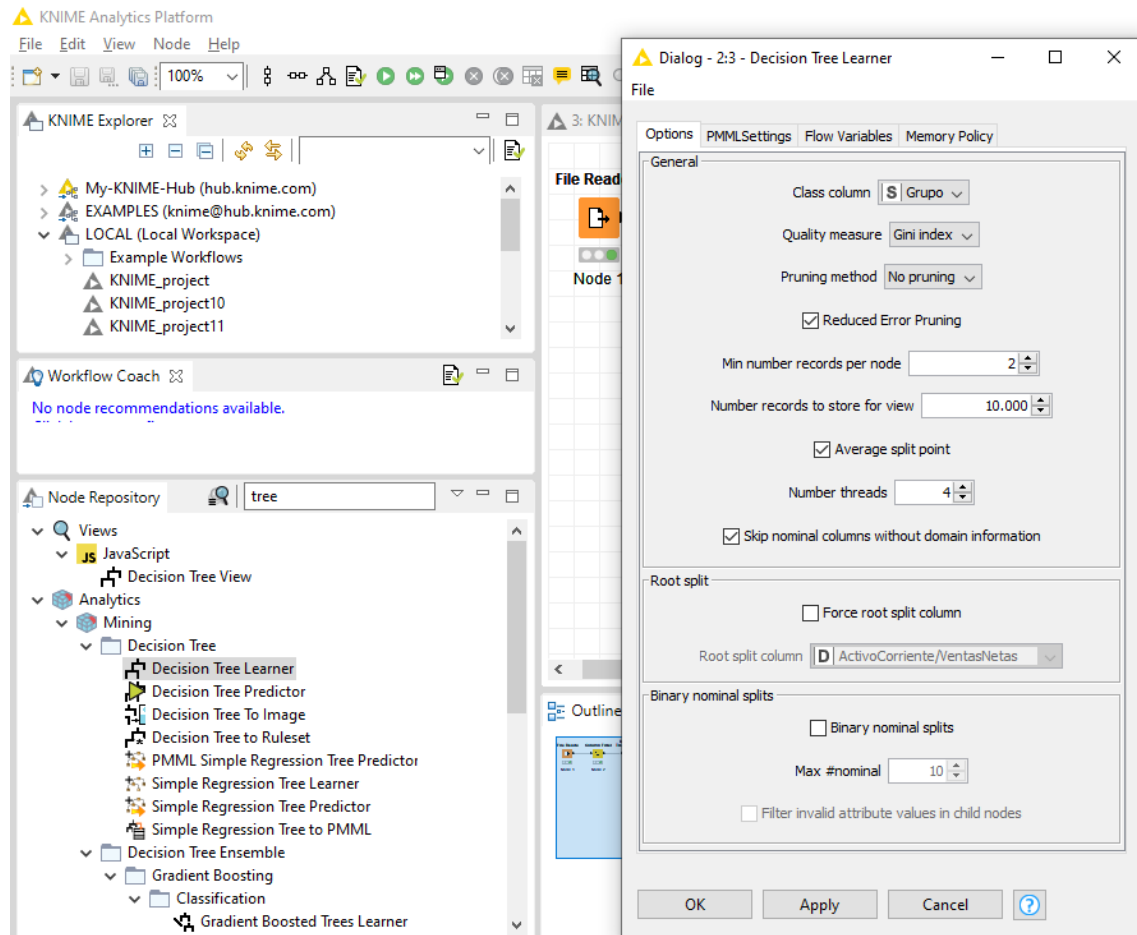
Para eliminar dicha columna, usaremos un nodo llamado “Column Filter”. Buscamos el nodo en “Node Repository” y lo arrastramos hasta nuestro espacio de trabajo. Una vez hecho esto, pulsamos F6 para configurarlo.

En esta pantalla elegiremos qué atributo queremos eliminar. En nuestro caso es el atributo “obs”.

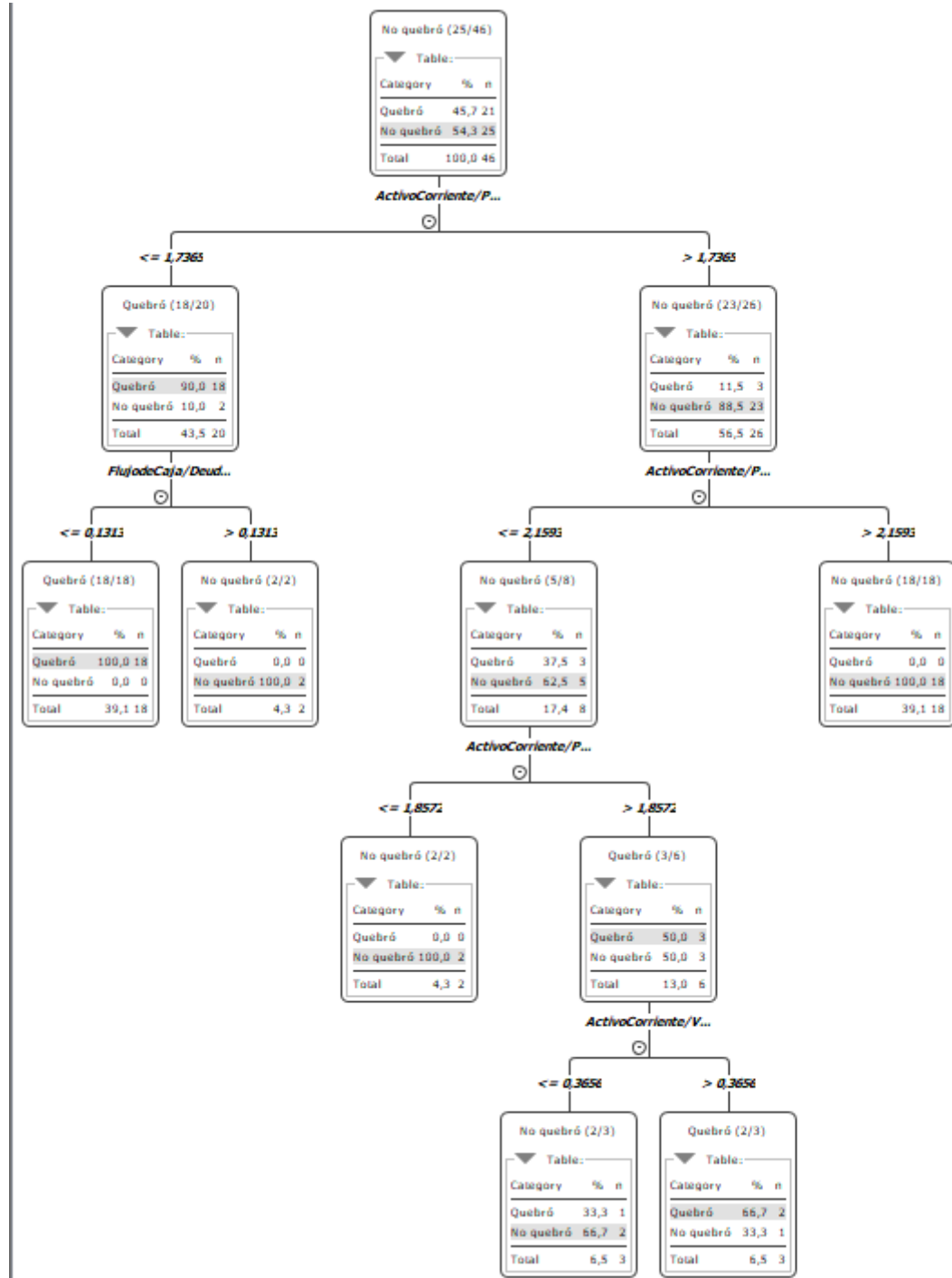


Tras quitar la columna “obs”, pulsamos “Apply” y “Ok”. Si ponemos el cursor encima de la salida del nodo “File Reader” podremos observar que comunica que hay 46 filas y 6 columnas, mientras que si lo colocamos a la salida del nodo “Column Filter” observamos que nos dice que hay 46 filas y 5 columnas.

Es el momento de buscar el nodo “Tree Decision Learner” y conectarlo a la salida de “Column Filter”. Para configurar el nodo pulsamos F6 y nos aseguramos de que en “Class column” está el atributo quebró, ya que es la variable que queremos explicada.



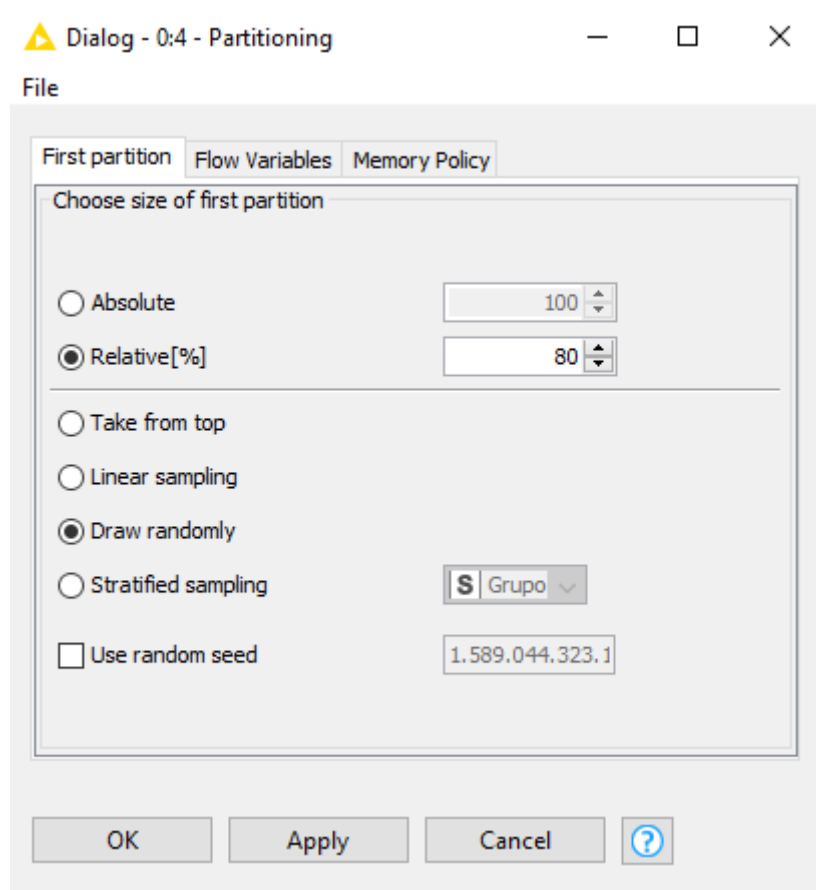
Ahora pulsamos SHIFT + F10 y de esta manera ejecutamos y abrimos la vista del árbol:



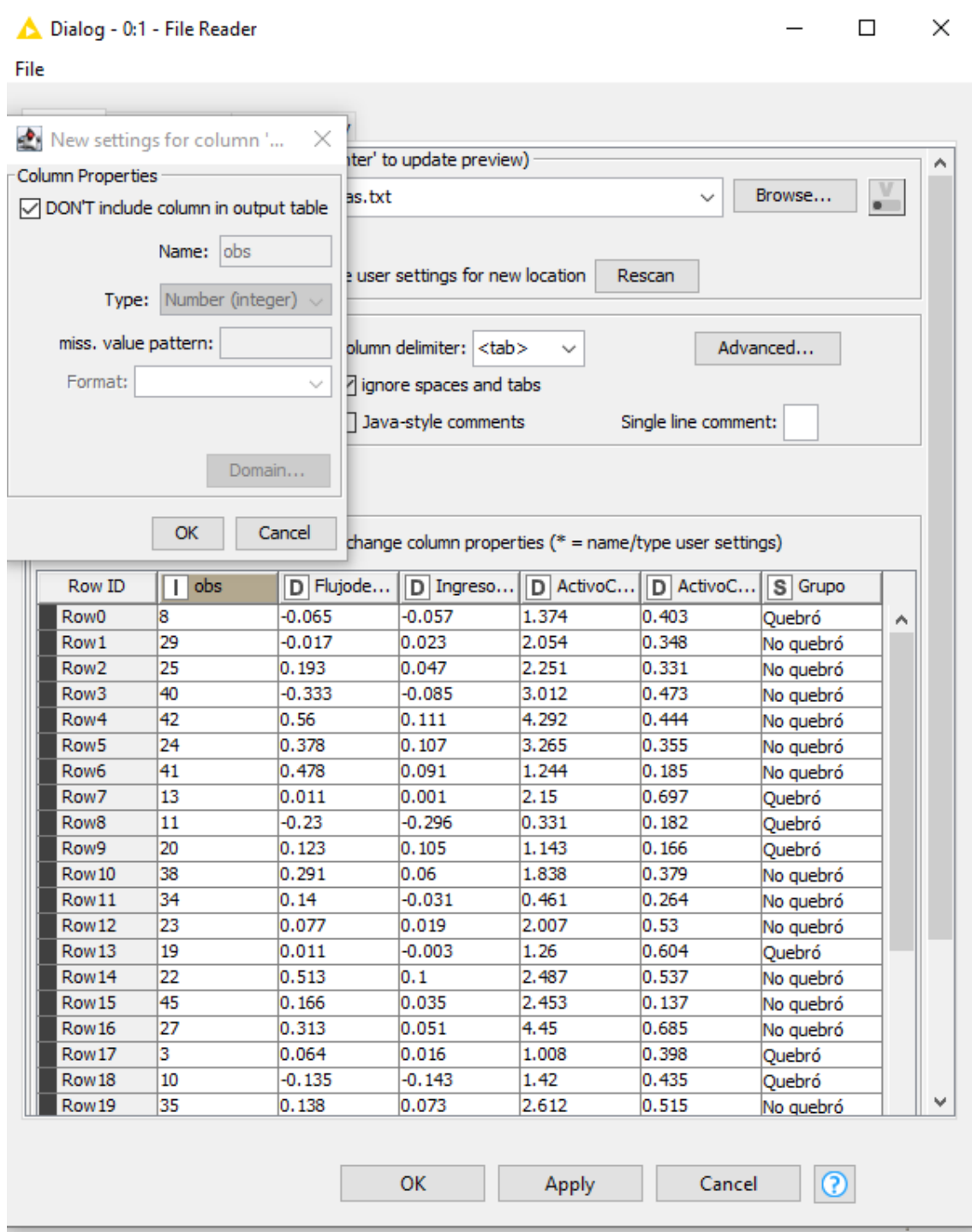
Como podemos observar, el árbol se ha desarrollado correctamente en esta ocasión. Podemos observar que el atributo que más discrimina es ActivoCorriente/PasivoCorriente.

Sin embargo no hemos testado el modelo. Por tanto vamos a eliminar la conexión del nodo "Column Filter" con "Decision Tree Learner" y vamos a meter entre ellos un nodo de partición de datos, dividiéndolos en el 80% de datos que usaremos para crear el modelo y un 20% que usaremos para el testeo del modelo.

Buscamos en "Node repository" el nodo "Partitioning" y lo configuramos para dividir los datos en el 80% y el 20%.

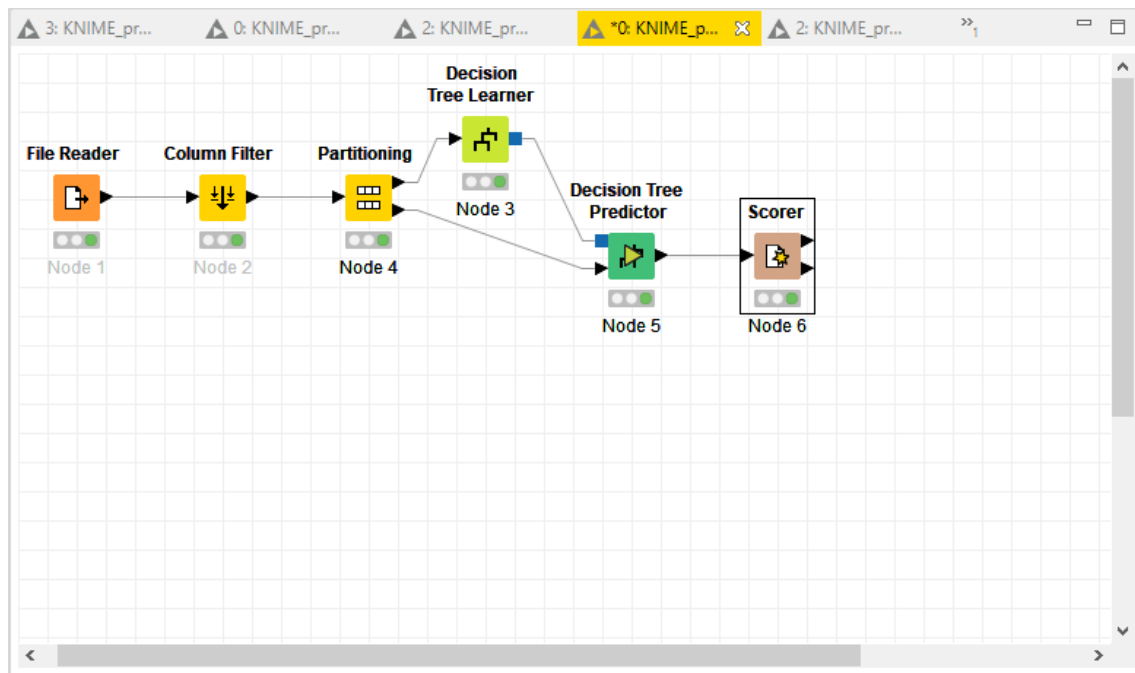


Realmente, para haber obtenido este arbol, podríamos haberlo hecho sin el uso del nodo Column Filter, ya que en el propio nodo File Reader podemos observar en el menu de configuración, que si pulsamos en la columna que queremos eliminar o filtrar, se despliega un submenu:



Si hacemos click en la opción “DON’T include column in output table”, conseguiremos que no aparezca esa columna en la salida.

La salida superior del nodo la conectaremos con el nodo “Decision Tree Learner”, y la salida inferior la conectaremos a la entrada inferior del nodo “Decision Tree Predictor”. La entrada superior de dicho nodo la conectaremos a la salida del “Decision Tree Learner”, ya que queremos testear el modelo creado a partir del 80% de los datos con el 20% de los datos restantes.



El nodo Scorer se encarga de crear la matriz de confusión. Seleccionamos el nodo y pulsamos F6 para configurarlo. En el menu de configuración solo nos tenemos que asegurar de que tenemos seleccionadas la variable objetiva en la pestaña “First Column”, y la predicción de la variable objetiva en “Second Column”. En este caso nos quedaría de la siguiente manera:

Dialog - 0:6 - Scorer

File

Scorer | Flow Variables | Memory Policy

First Column
[S] Grupo

Second Column
[S] Prediction (Grupo)


Sorting of values in tables
Sorting strategy: Insertion order [] Reverse order

Provide scores as flow variables
[] Use name prefix

Missing values
In case of missing values... [X] Ignore [] Fail

OK Apply Cancel ?

Pulsamos "OK" y ejecutamos el nodo. Haciendo click derecho sobre el, seleccionamos la opción "Confusion Matrix" y observamos el resultado:

 Confusion matrix - 0:6 - Scorer

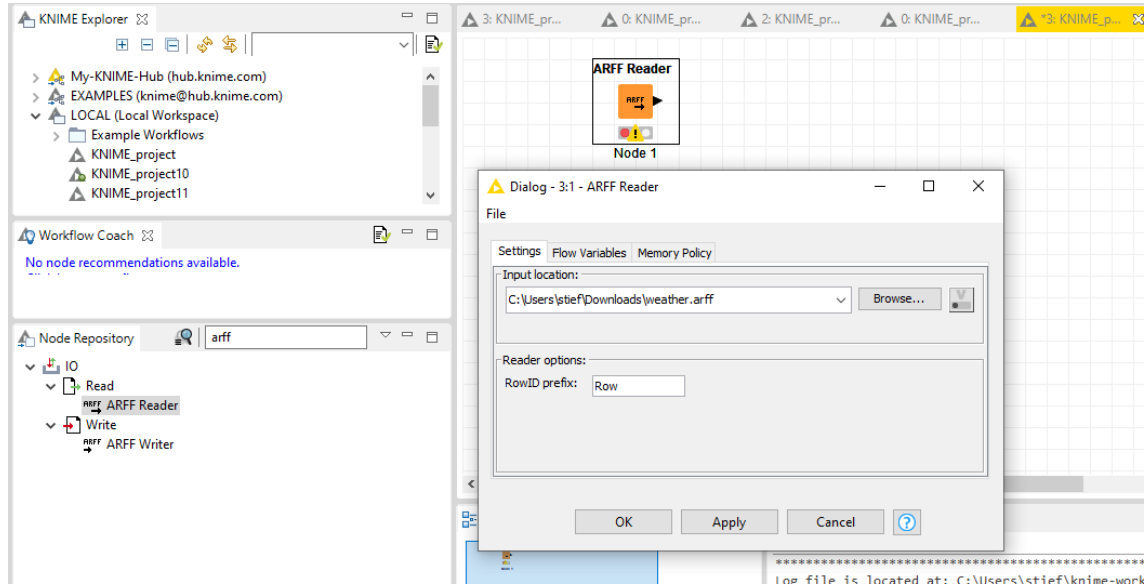
File Hilite Navigation View

| Table "spec_name" - Rows: 2 | | Spec - Columns: 2 | Properties | Flow Variables |
|---|---------------------------------|------------------------------------|------------|----------------|
| Row ID | <input type="checkbox"/> Quebró | <input type="checkbox"/> No que... | | |
| <input checked="" type="checkbox"/> Quebró | 4 | 1 | | |
| <input checked="" type="checkbox"/> No quebró | 1 | 4 | | |

Podemos observar que quedaron bien clasificadas un total de 8 (4 + 4) y mal clasificadas 2 (1 + 1).

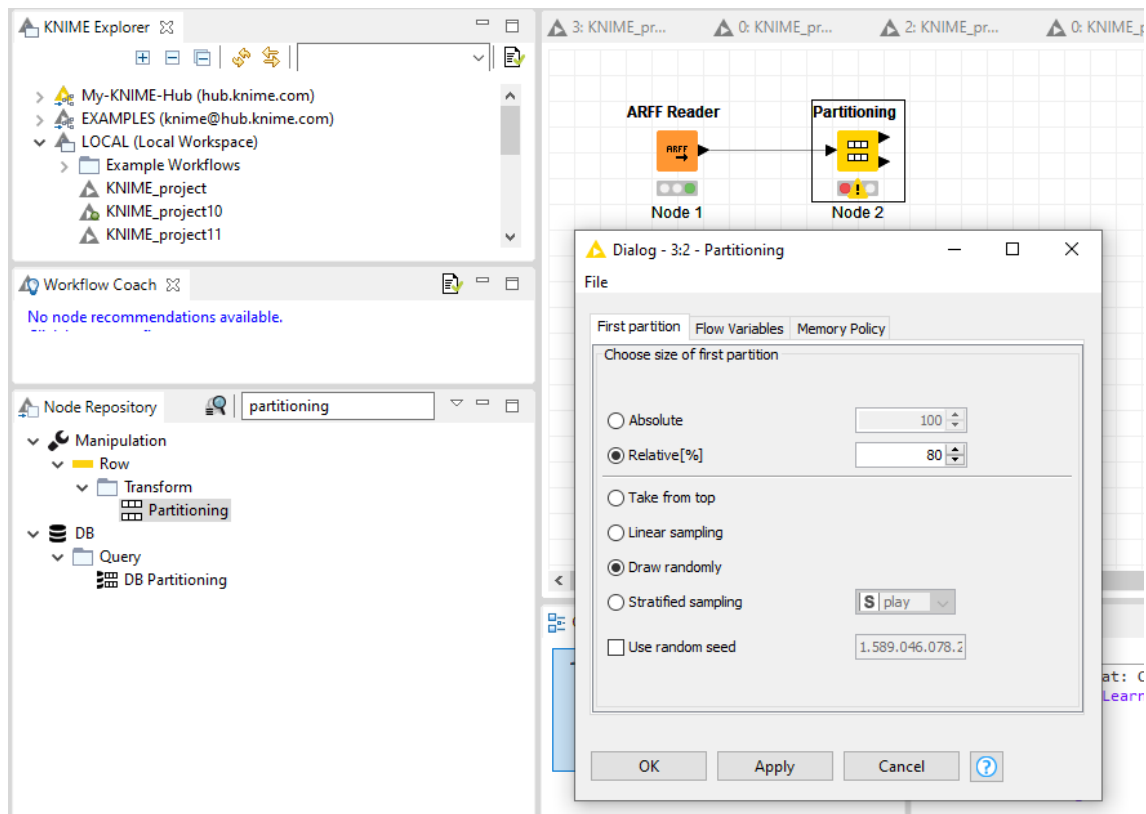
Vamos a hacer ahora un arbol de decisiones, utilizando el archivo weather.arff en esta ocasión, con la intención de predecir si un partido se jugará o no.

Procedemos a añadir un nodo “ARFF Reader” en el que seleccionaremos el archivo weather.arff



Pulsamos F7 para ejecutar. Seguidamente buscaremos el nodo “Partitioning” para dividir el numero de datos en el 80% y el 20%, que usaremos para crear y testear un modelo respectivamente.

Pulsamos F6 para confiurar el nodo y seleccionamos el 80%. Tras esto pulsamos F7 para ejecutar.



La salida superior de este nodo, lo conectaremos a un nodo “Decision Tree Learner” para crear el modelo con el 80% de los datos.

La salida inferior la conectaremos a la entrada inferior del nodo “Decision Tree Predictor”.

La entrada superior de este nodo será la salida del nodo “Decision Tree Learner”.

Para configurar el nodo “Decision Tree Learner” pulsamos F6 y nos aseguramos tener seleccionado el atributo que queremos explicado, en este caso “play”.

The screenshot displays the KNIME software interface. The main workspace shows a workflow with four nodes: 'ARFF Reader' (Node 1), 'Partitioning' (Node 2), 'Decision Tree Learner' (Node 3), and 'Decision Tree Predictor' (Node 4). The 'Decision Tree Learner' node is highlighted with a yellow warning icon. The 'Decision Tree Predictor' node is connected to the 'Decision Tree Learner' node. The 'Outline' pane on the left shows a hierarchical view of the workflow. The 'Console' pane at the bottom displays the following log messages:

```

KNIME Console
WARN Decision Tree Learner 0:
WARN Partitioning 0:4
WARN Scorer 0:6
WARN Scorer 0:6
WARN ARFF Reader 3:1
WARN Partitioning 3:2
WARN Decision Tree Learner 3:

```

The 'Dialog - 3:3 - Decision Tree Learner' window is open, showing the 'General' tab. The settings are as follows:

- Class column: **S** play
- Quality measure: Gini index
- Pruning method: No pruning
- ☒ Reduced Error Pruning
- Min number records per node: 2
- Number records to store for view: 10.000
- ☒ Average split point
- Number threads: 4
- ☒ Skip nominal columns without domain information

The 'Root split' section shows:

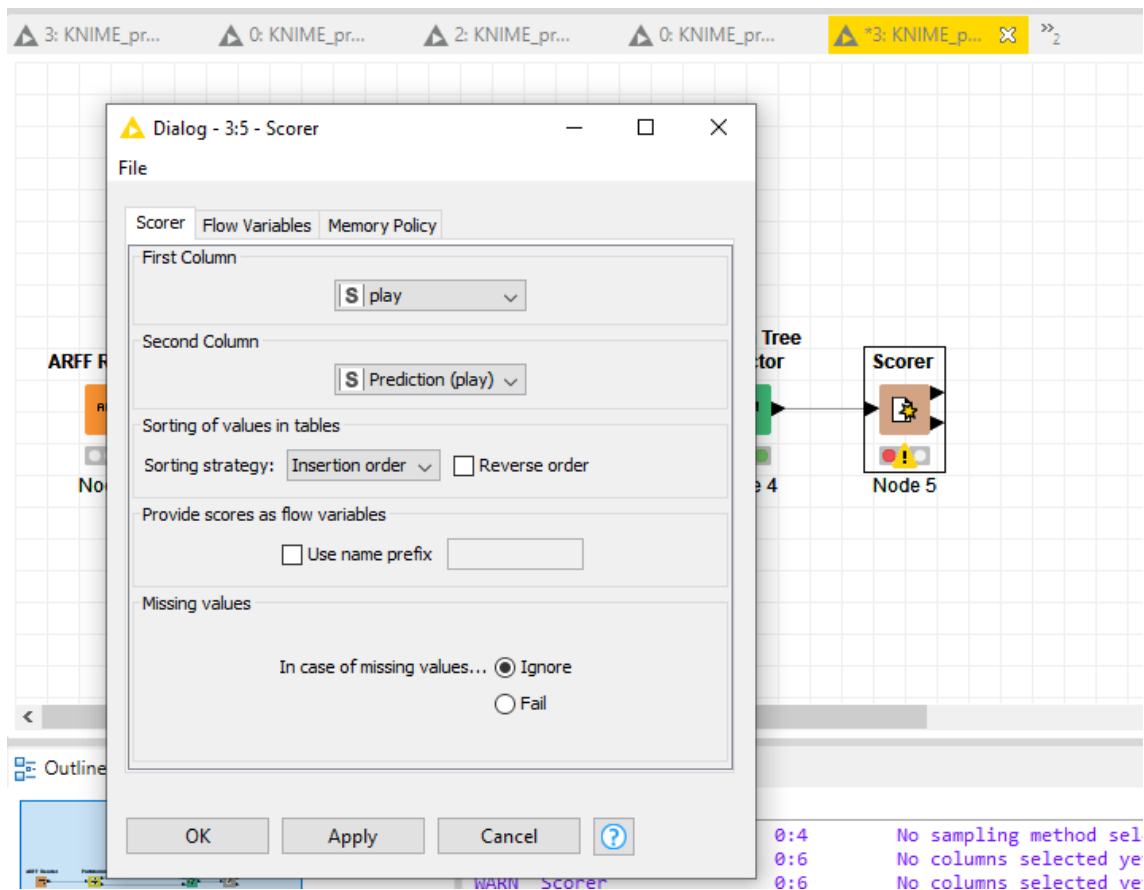
- ☐ Force root split column
- Root split column: **S** windy

The 'Binary nominal splits' section shows:

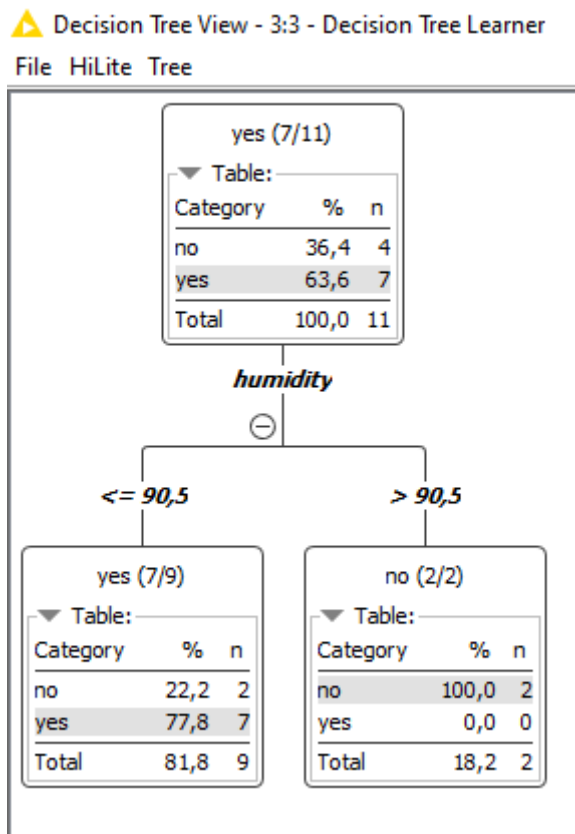
- ☐ Binary nominal splits
- Max #nominal: 10
- ☐ Filter invalid attribute values in child nodes

The dialog box has 'OK', 'Apply', and 'Cancel' buttons at the bottom.

Ejecutamos los nodos, y añadimos a la salida del “Decision Tree Predictor” el nodo Scorer para obtener la matriz de confusión. Recordamos que tenemos que tener la variable que queremos explicada en “First Column” y la variable predicha que queremos explicada en “Second Column”, que en nuestro caso será:

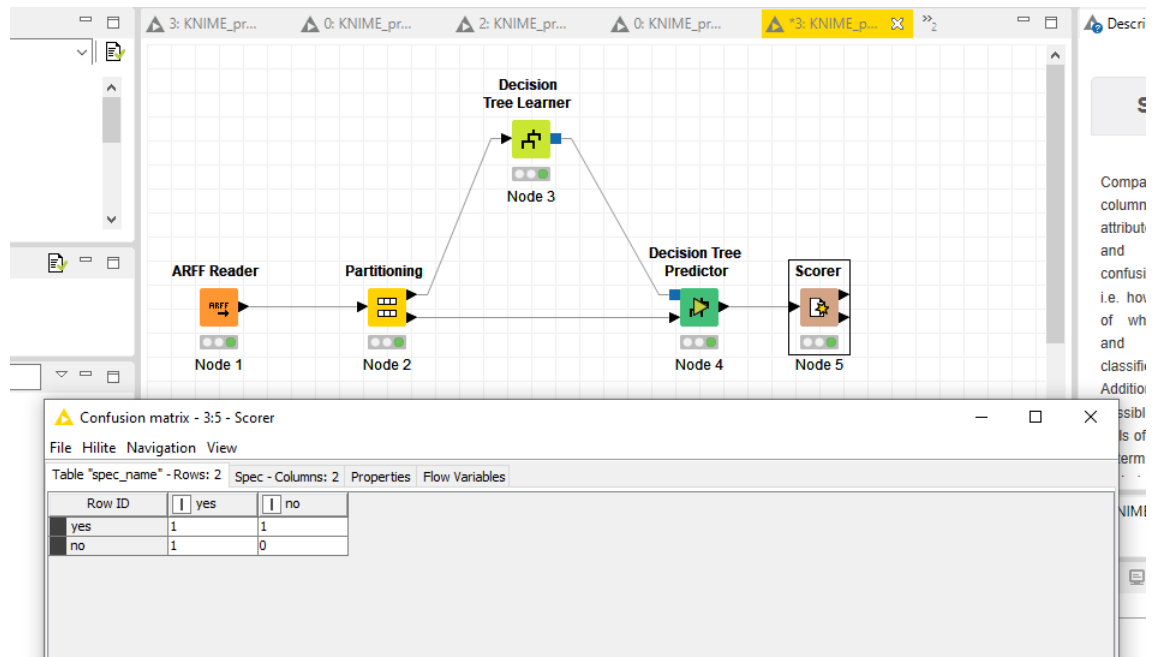


Haciendo click en el botón derecho podemos seleccionar la opción "Decision Tree View":



Vemos que la variable que más discrimina es la humedad.

Para ver la matriz de confusión, seleccionamos el nodo “Scorer”, pulsamos con el botón derecho y seleccionamos la opción “Confusion Matrix”:



Podemos observar que nos han quedado 2 mal clasificados (1 + 1) y 1 bien clasificado.