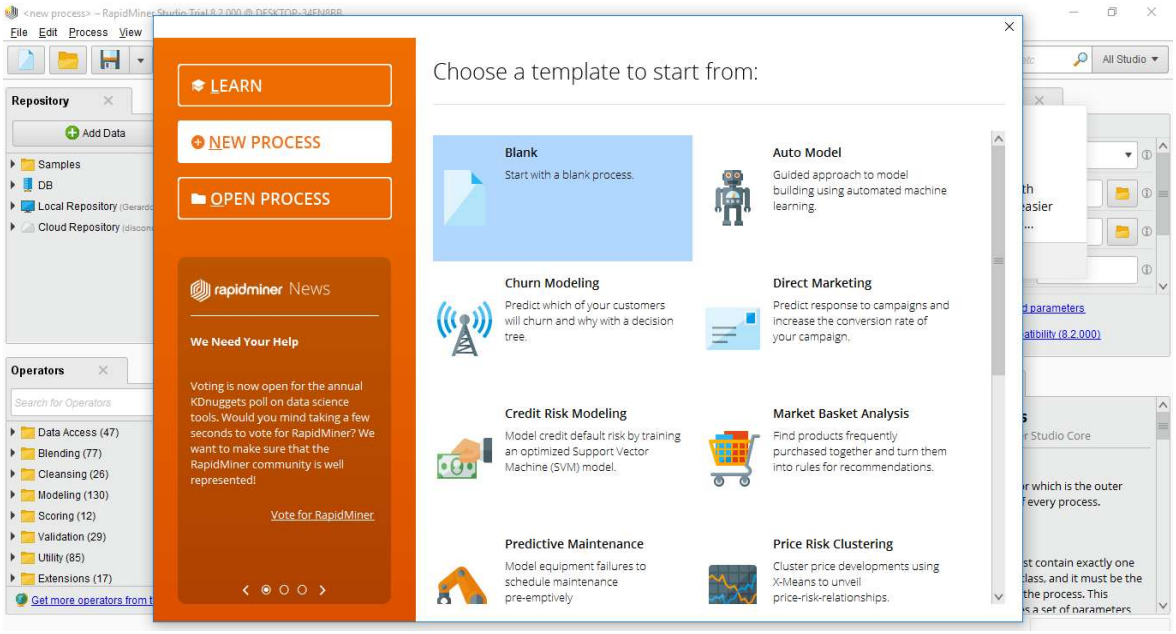
Diplomado en Minería de Datos

PEUVI, Facultad de Ciencias, UNAM

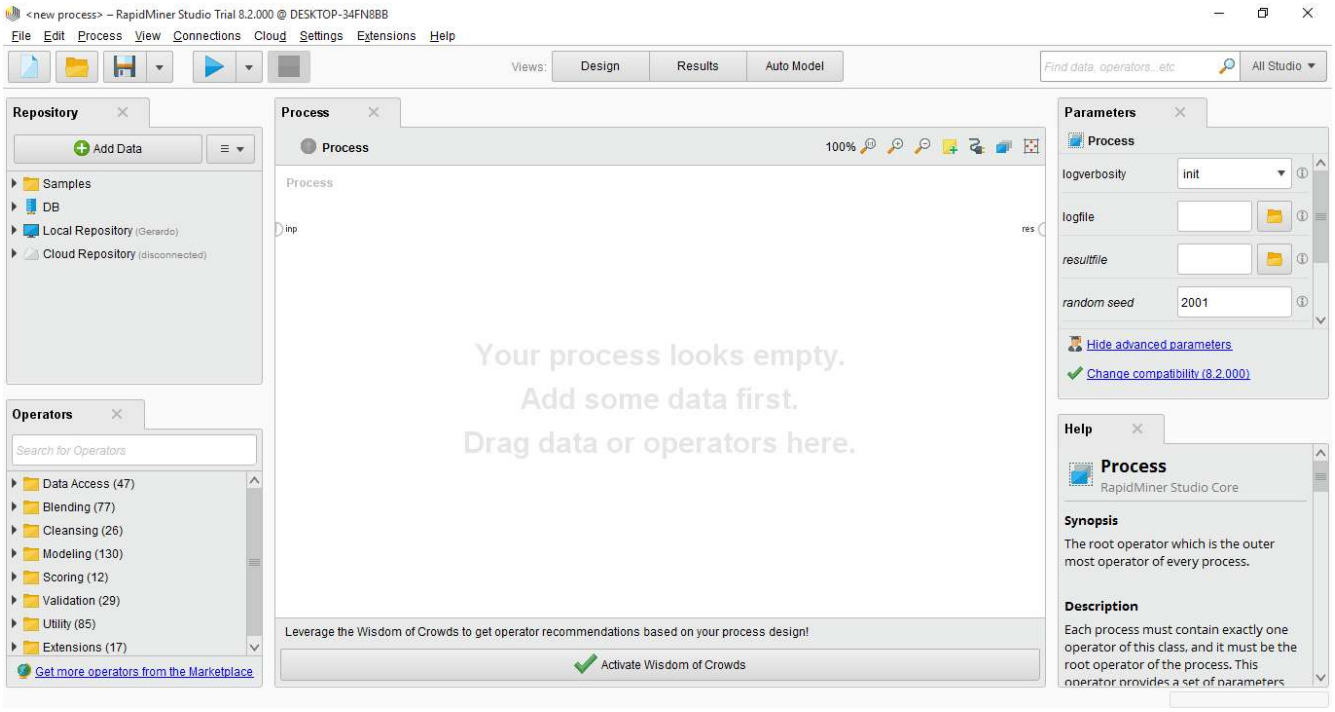
M.I. Gerardo Avilés Rosas
gar@ciencias.unam.mx

Módulo 5 **Minería de Datos**
Ejemplo de árboles de decisión con RapidMiner

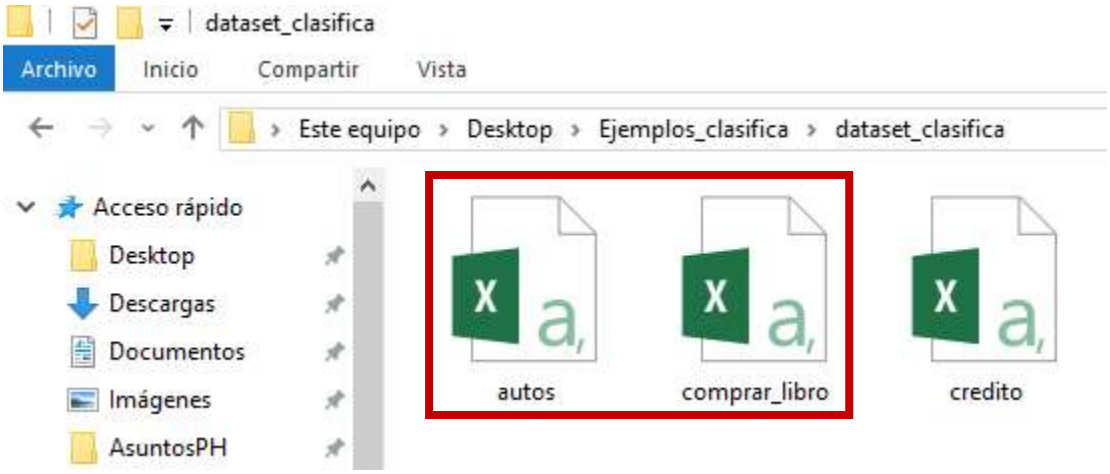
1. Iniciamos **RapidMiner** y seleccionamos **New process → Blank**:



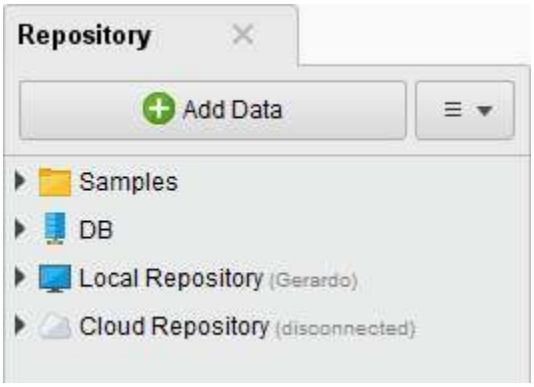
2. Se muestra la siguiente ventana. Como se puede observar, las opciones disponibles se encuentran en la sección **Operators**:

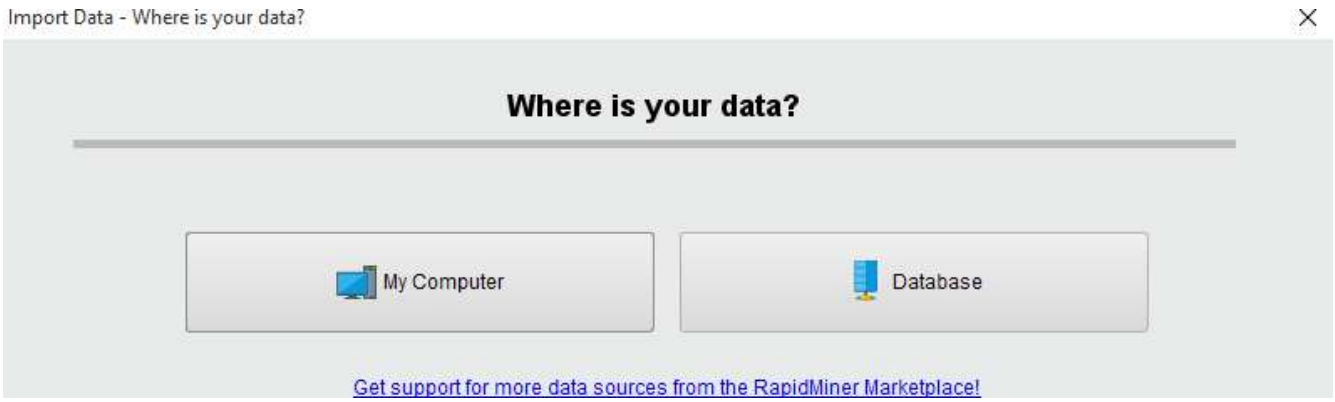


3. Lo primero que vamos a hacer es cargar los conjuntos de datos que se van a utilizar para este tutorial. Se trata de tres archivos **CSV**: **autos y comprar_libro**:

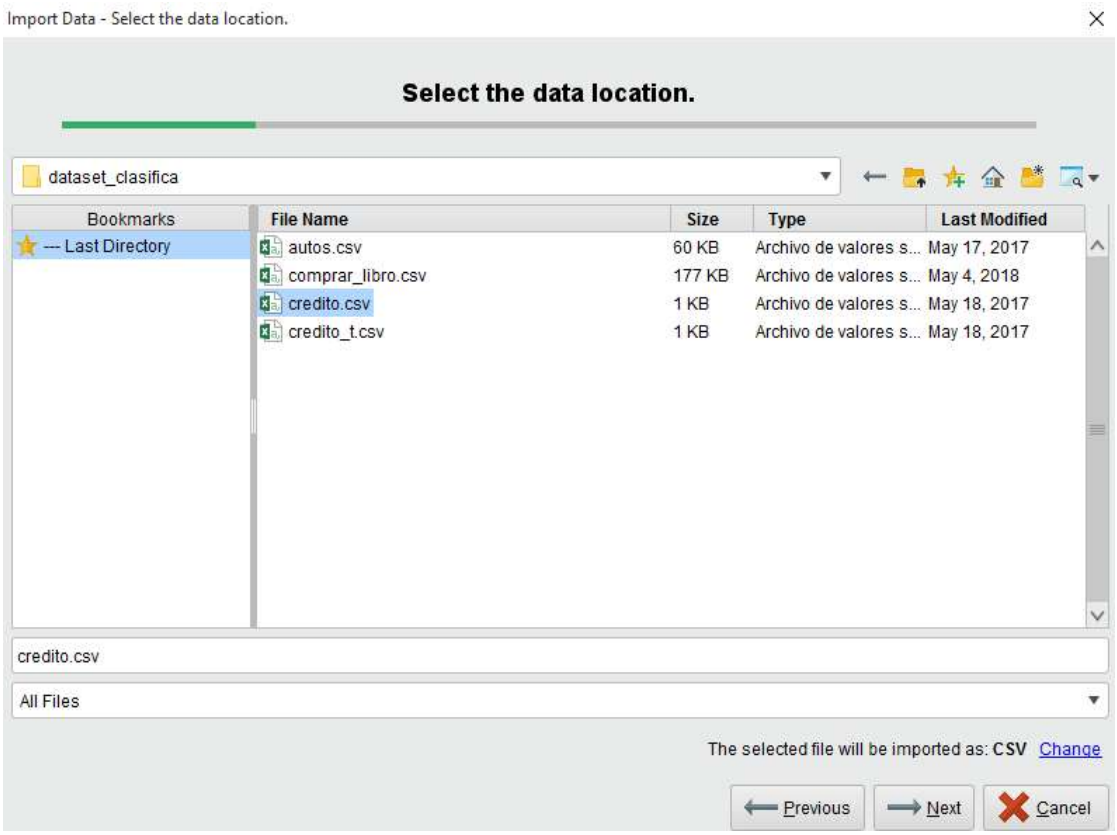


4. Para cargar los archivos, vamos a ir a la sección **Repository** y daremos clic en la opción **Add Data,** desplegándose la secuencia de ventanas que se muestra a continuación:

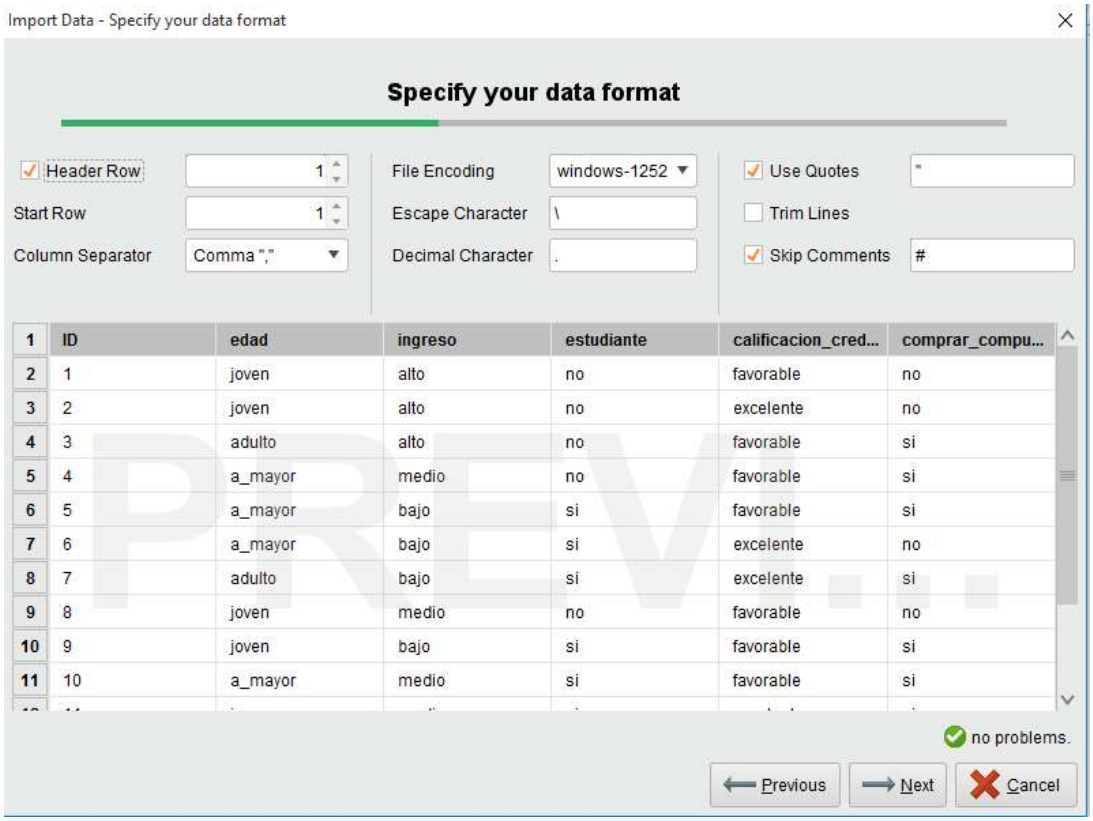




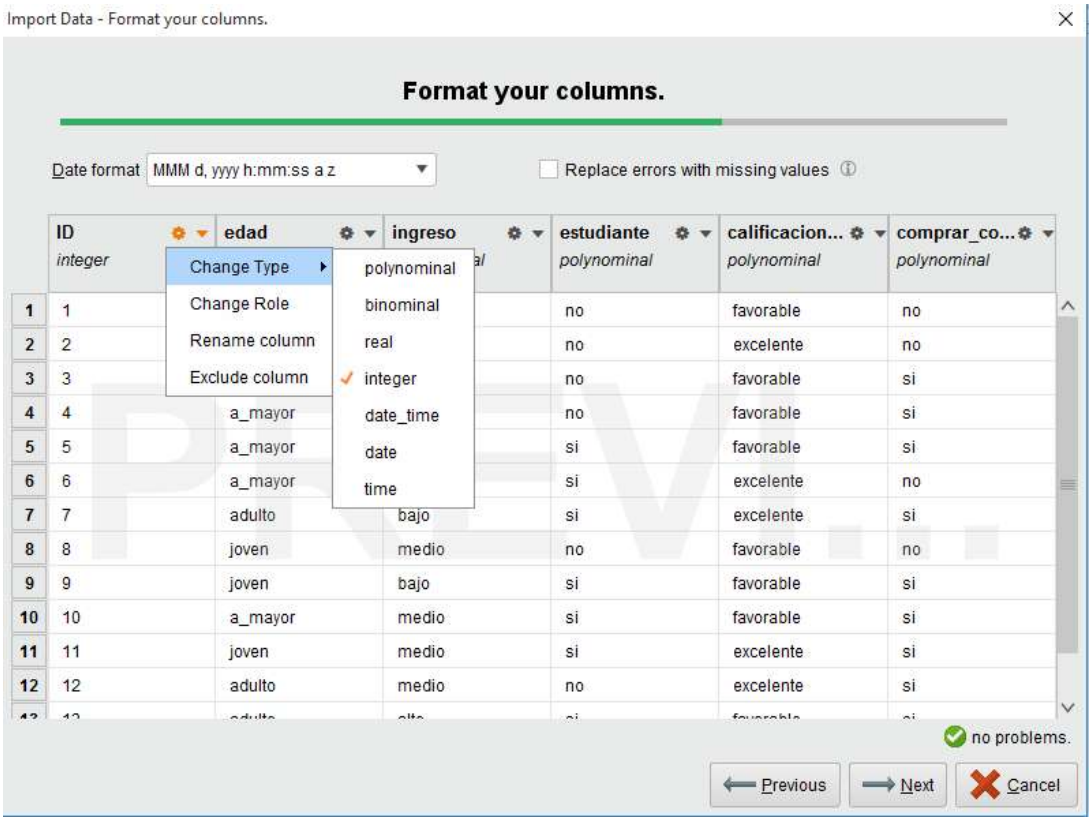
Elegimos la opción **My Computer** (el proceso que se hará para los tres datasets que se indicaron en el punto anterior, pero aquí solo se mostrará para uno de ellos). Buscamos el dataset que queremos cargar, en este caso, utilizaremos el dataset **crédito.csv** y damos clic en la opción **Next**:



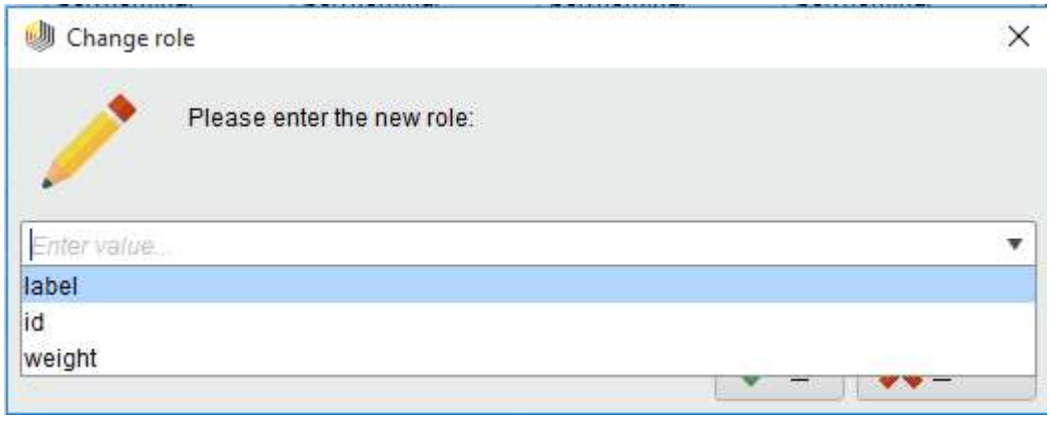
En la siguiente opción, se muestran varias opciones relacionadas con aspectos de **formato de los datos** que deseamos importar (tipo de datos, codificación, etc.). Vamos a dejar las opciones que se muestran por omisión:



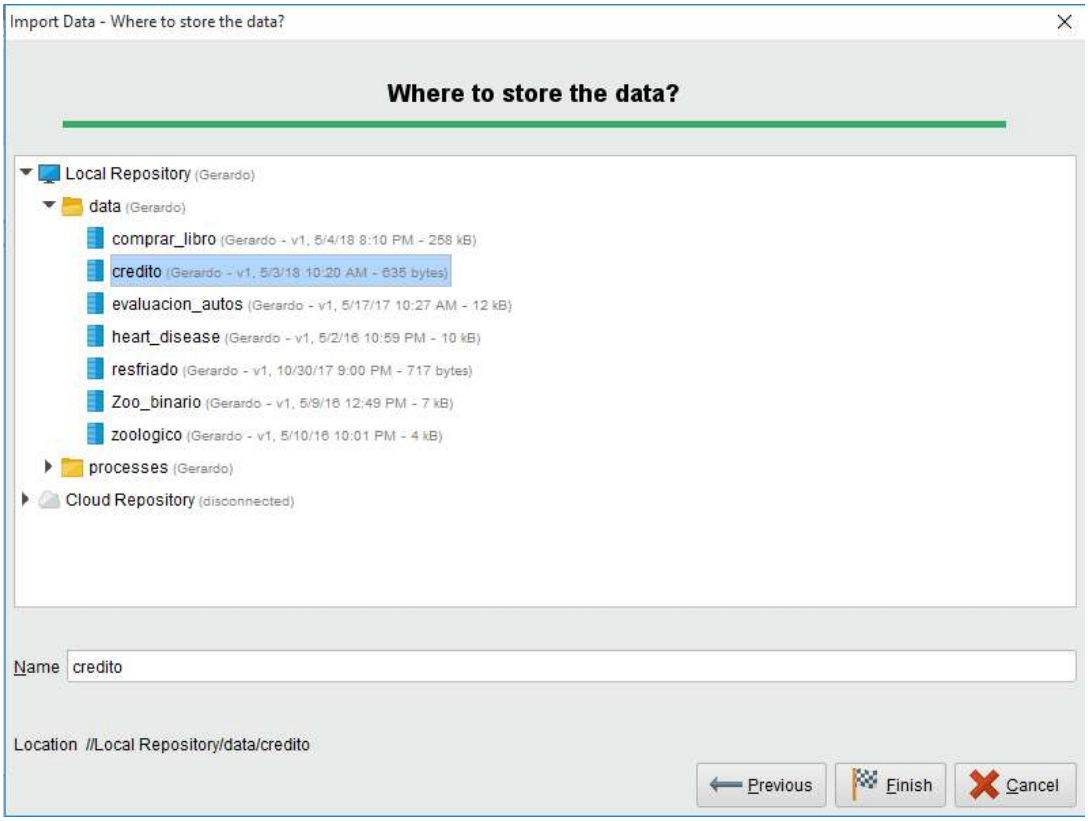
En la siguiente ventana, se muestran opciones del **formato de columnas**. Para cada una de ellas se indican las opciones disponibles: **Change type, Change Role, Rename column, Exclude column**. Inicialmente **RapidMiner** realiza una exploración de las mejores opciones. De igual forma, vamos a dejar las opciones que se indican por omisión.



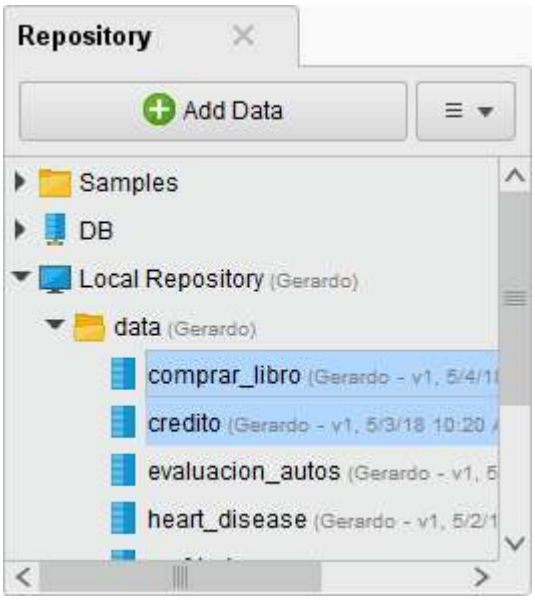
Un aspecto a tener en cuenta es sobre el tipo de tarea de Minería de Datos que se desea realizar. En el caso de la **Clasificación**, se requiere tener una variable que tenga la **etiqueta de clase**. Desde la ventana anterior, se puede controlar quién será la etiqueta de clase a partir de la opción **Change role** y desde aquí indicar **label** (en este caso, no vamos a elegir dicha opción):



Por último, vamos a elegir dónde se van a guardar los datos. En este caso, vamos a elegir la opción **data** y especificamos un nombre para este dataset (en este caso, RapidMiner elige por omisión el nombre del archivo que hayamos elegido para importar). Damos clic en **Finish**:



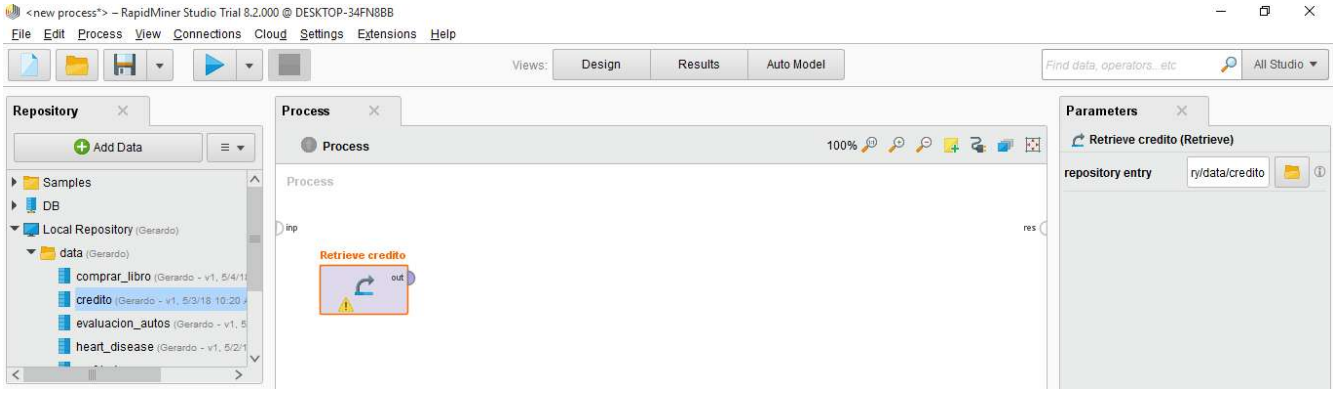
Una vez que hemos terminado el proceso de importación, deberemos de ver los datasets en la sección **Repository → Local repository → data**:



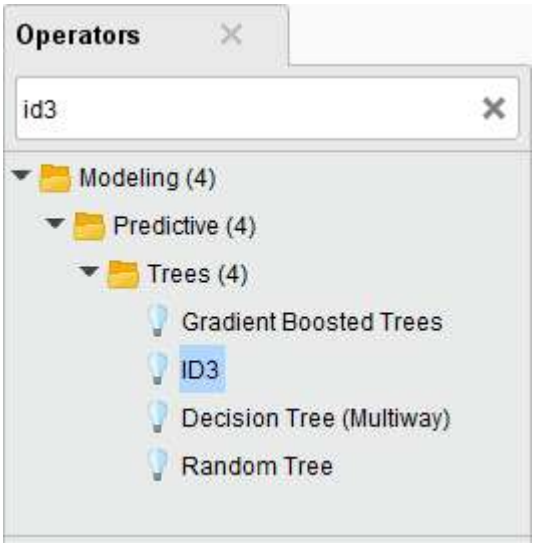
Una vez que ya están cargados los datos, vamos a proceder a ilustrar la generación de los árboles **ID3** (crédito) y **C4.5** (comprar_libro).

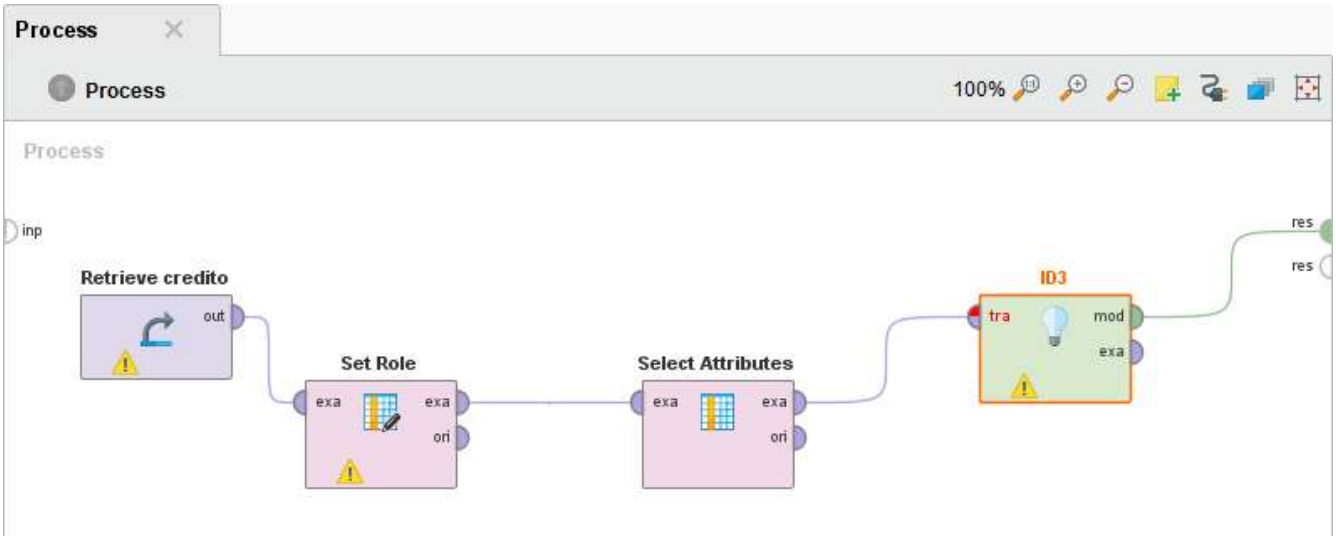
Dataset crédito → ID3

5. Vamos a comenzar seleccionando el dataset **crédito**, lo arrastramos al área de trabajo en blanco (se trata de un área drag&drop):

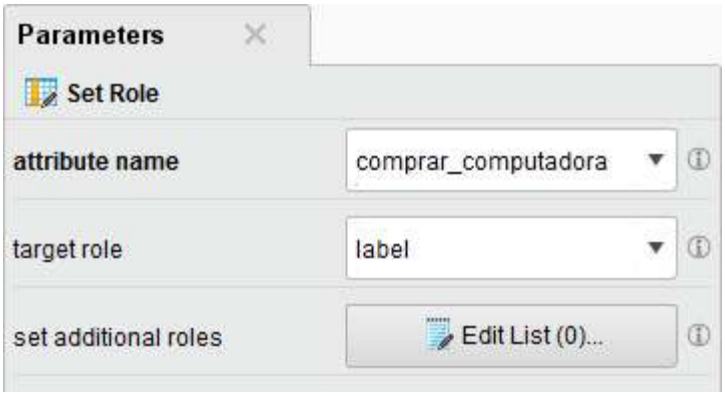


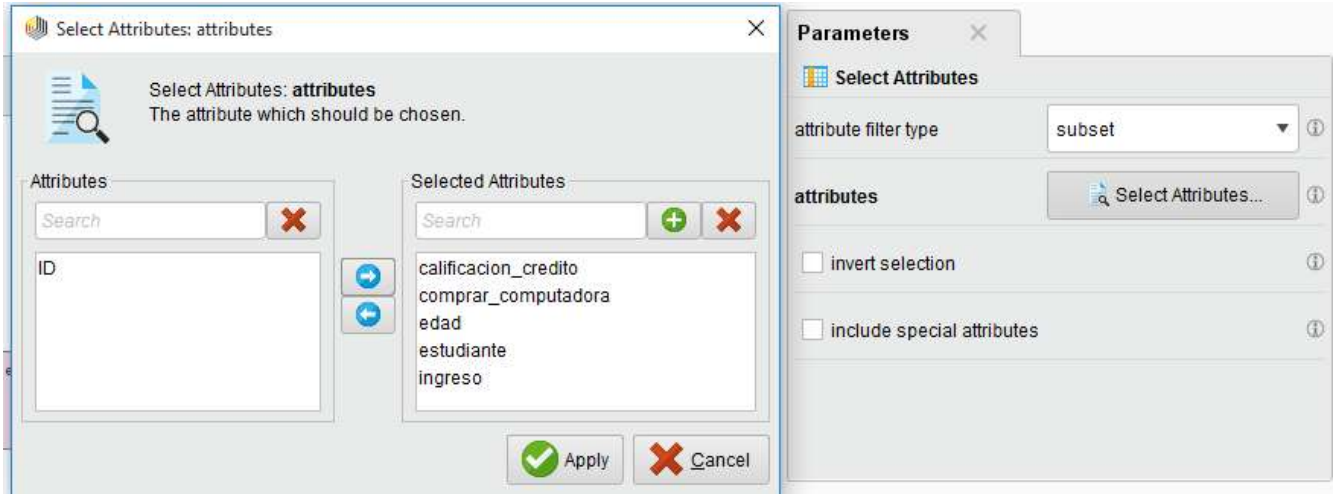
6. De la sección de **Operators**, vamos a seleccionar los siguientes (escribimos el nombre en el cuadro de búsqueda): **Set role, Select attributes, ID3**. Conectamos los nodos como se muestra:

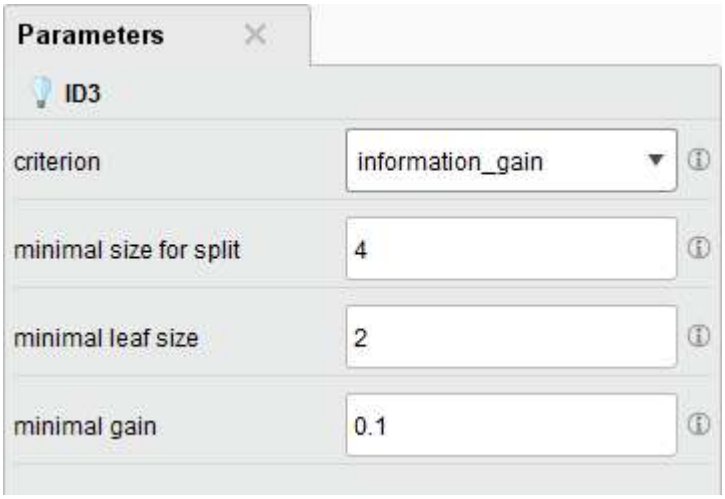




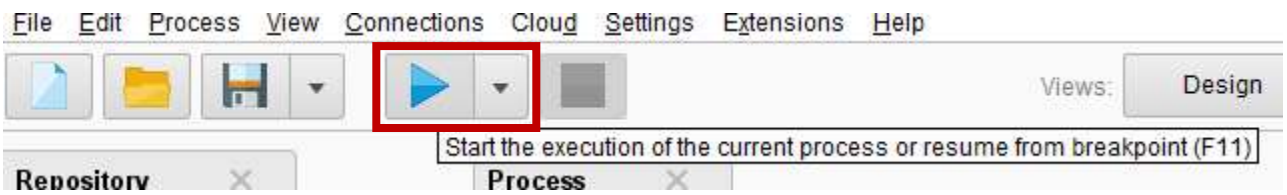
7. La configuración de cada uno de los nodos se muestra a continuación:

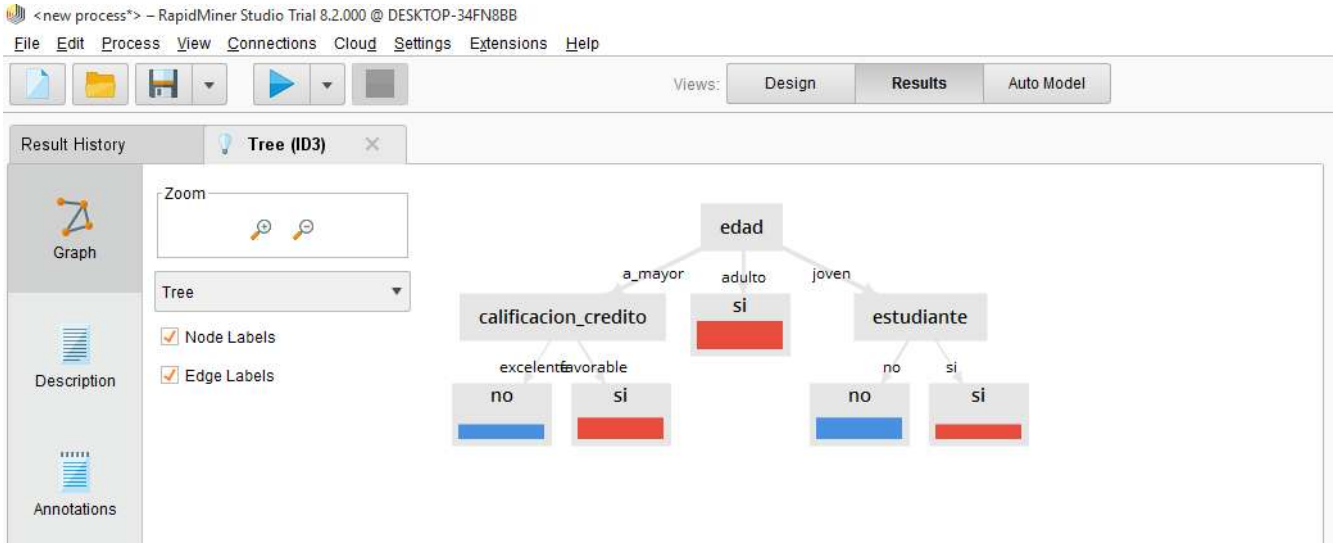


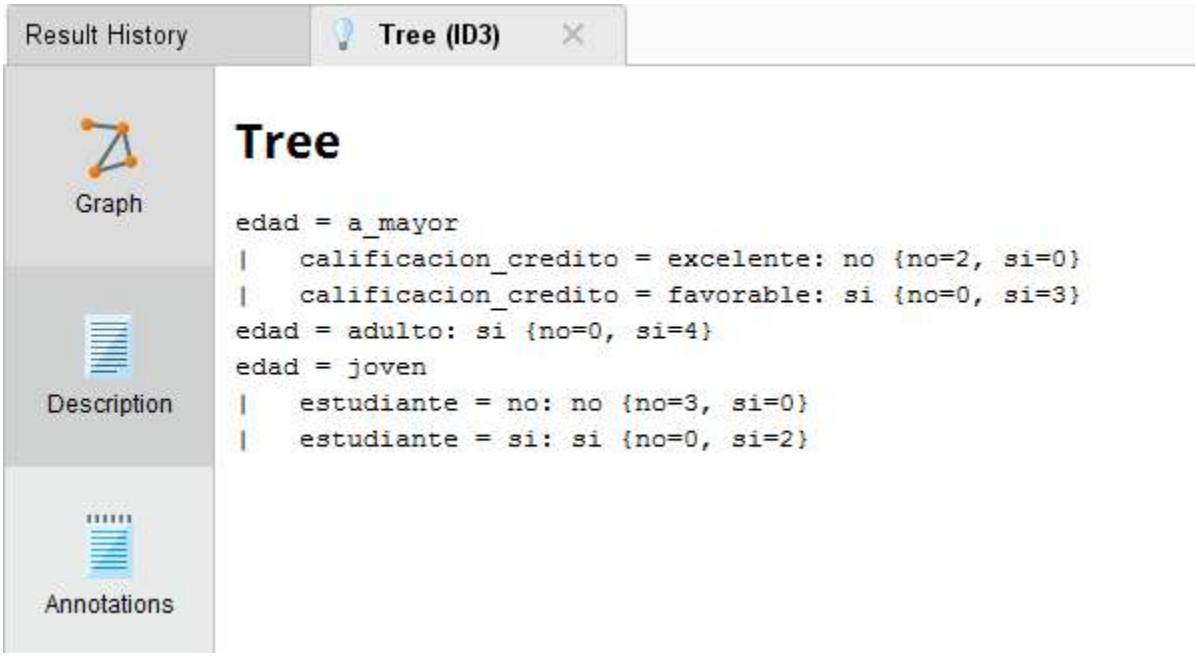




8. Una vez que se tiene configurado el proceso, vamos a ejecutarlo, como se muestra a continuación:

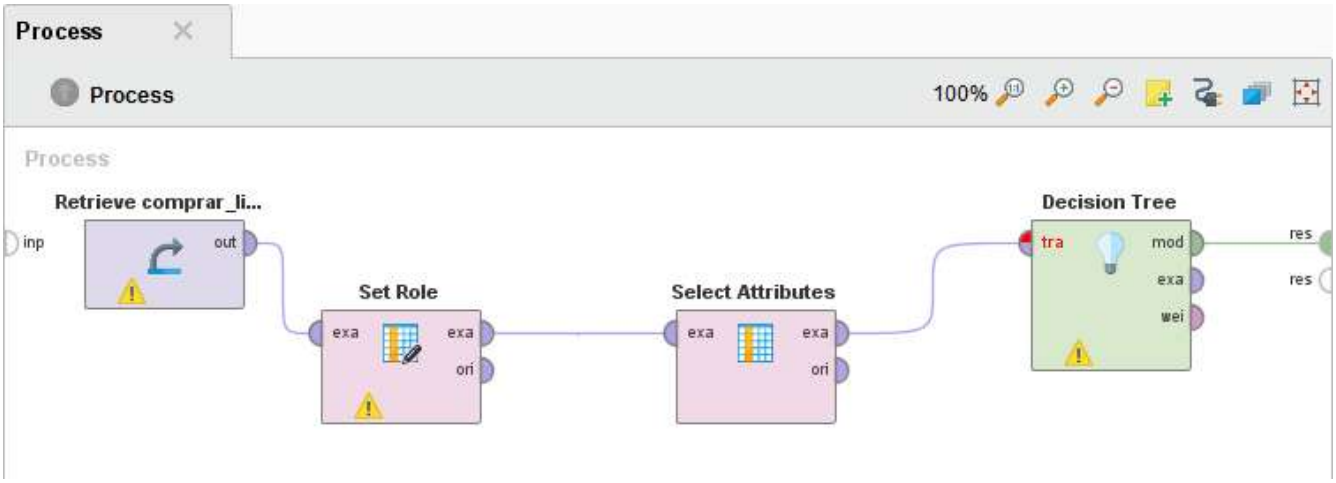



<new process*> – RapidMiner Studio Trial 8.2.000 @ DESKTOP-34FN8BB
File Edit Process View Connections Cloud Settings Extensions Help
Views: Design Results Auto Model
Result History
Tree (ID3)
Graph
Description
Annotations
Zoom
Tree
Node Labels
Edge Labels
edad
a_mayor
adulto
joven
calificacion_credito
si
estudiante
excelente
favorable
no
si
no
si

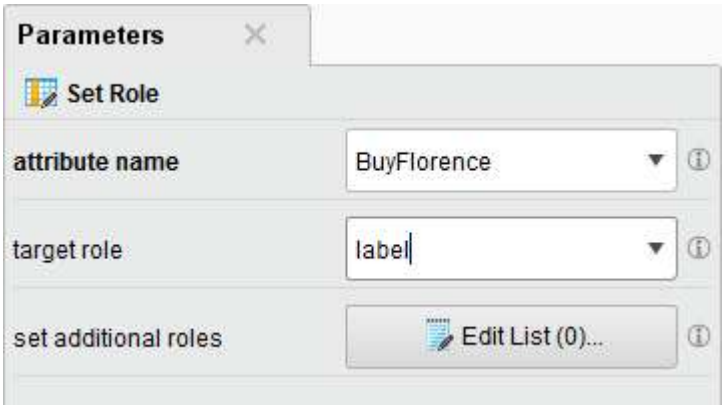

Result History
Tree (ID3)
Graph
Description
Annotations
Tree
edad = a_mayor
| calificacion_credito = excelente: no {no=2, si=0}
| calificacion_credito = favorable: si {no=0, si=3}
edad = adulto: si {no=0, si=4}
edad = joven
| estudiante = no: no {no=3, si=0}
| estudiante = si: si {no=0, si=2}

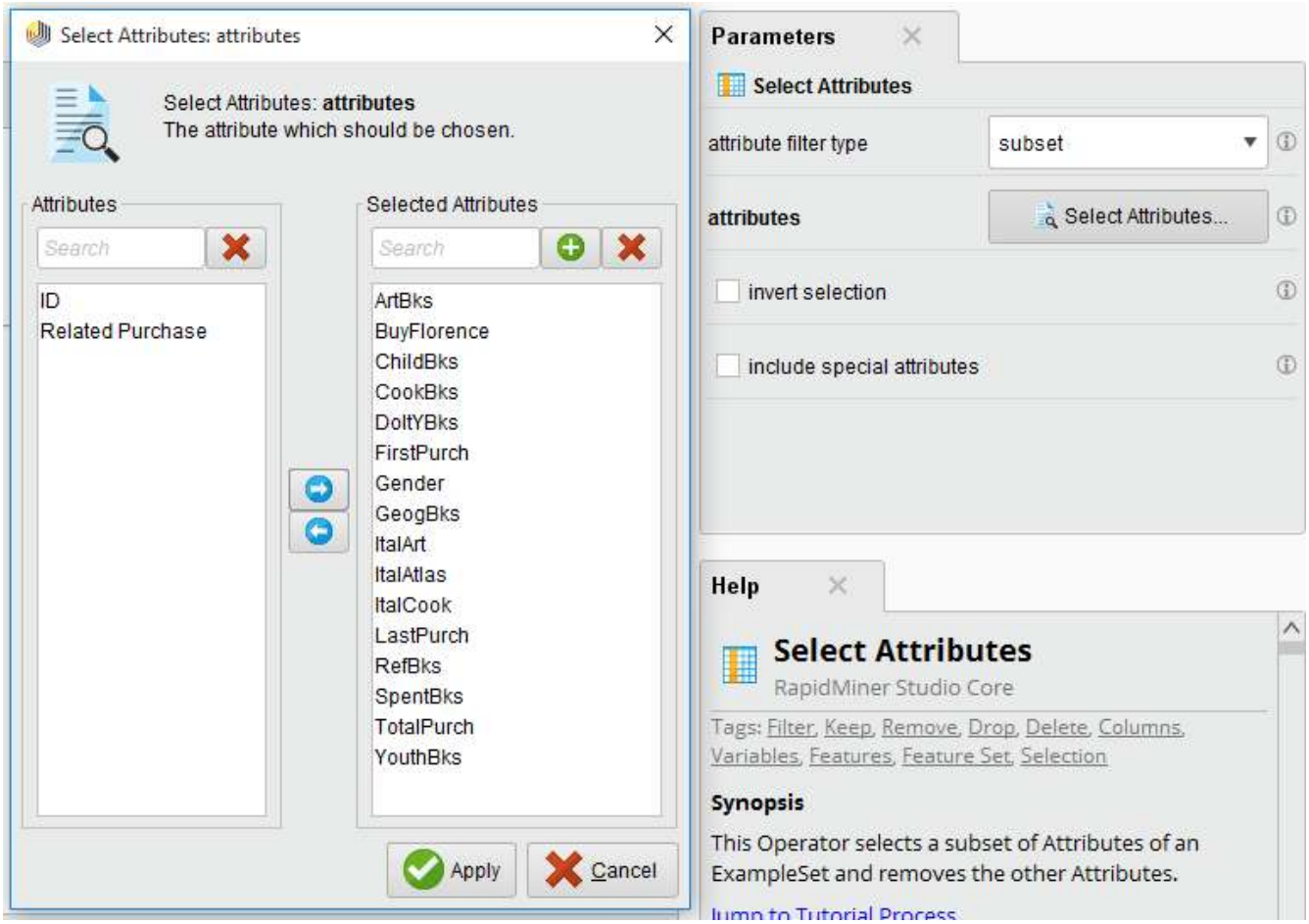
Dataset comprar_libro → C4.5

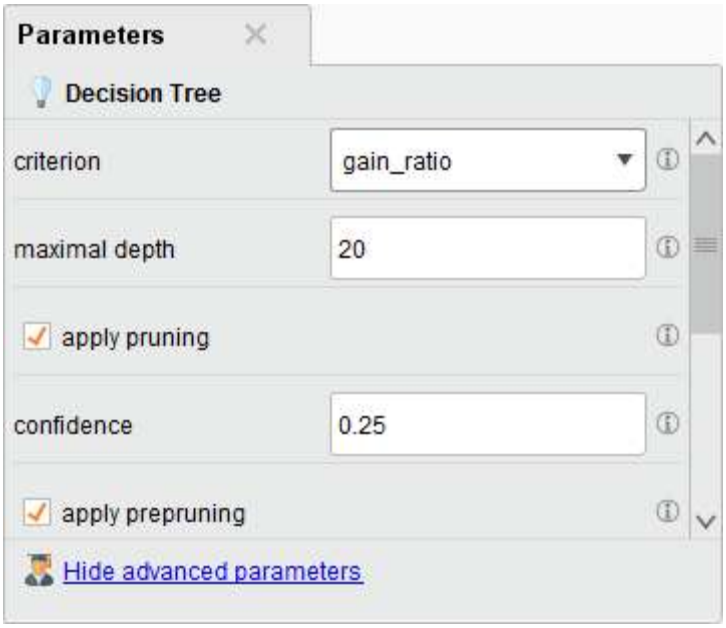
9. Vamos a comenzar seleccionando el dataset **comprar_libro**, lo arrastramos al área de trabajo en blanco (se trata de un área drag&drop). De la sección de **Operators**, vamos a seleccionar los siguientes (escribimos el nombre en el cuadro de búsqueda): **Set role, Select attributes, Decision Tree**. Conectamos los nodos como se muestra



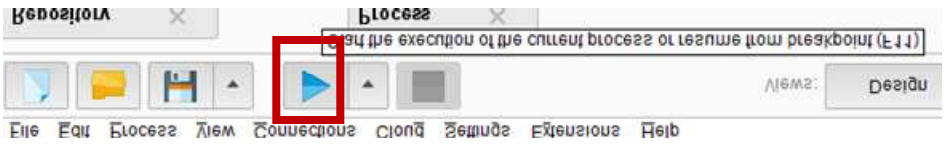
10. La configuración de cada uno de los nodos se muestra a continuación:





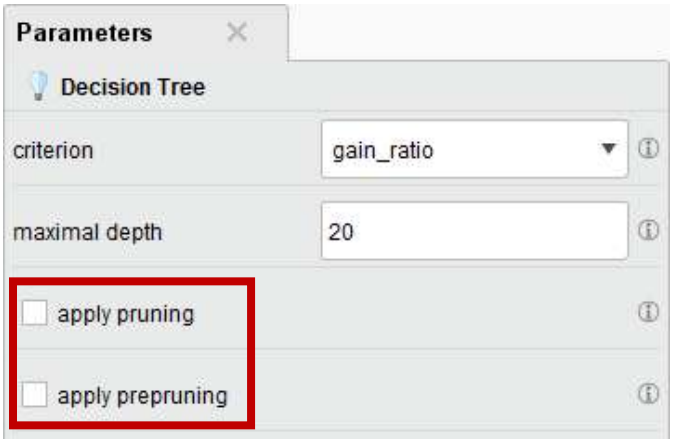


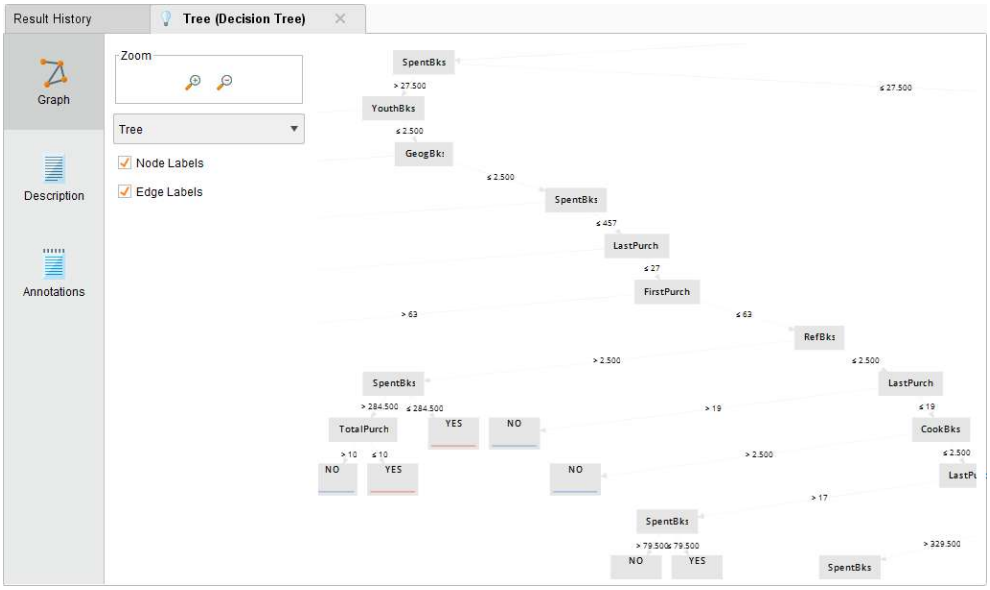
11. Una vez que se tiene configurado el proceso, vamos a ejecutarlo, como se muestra a continuación:



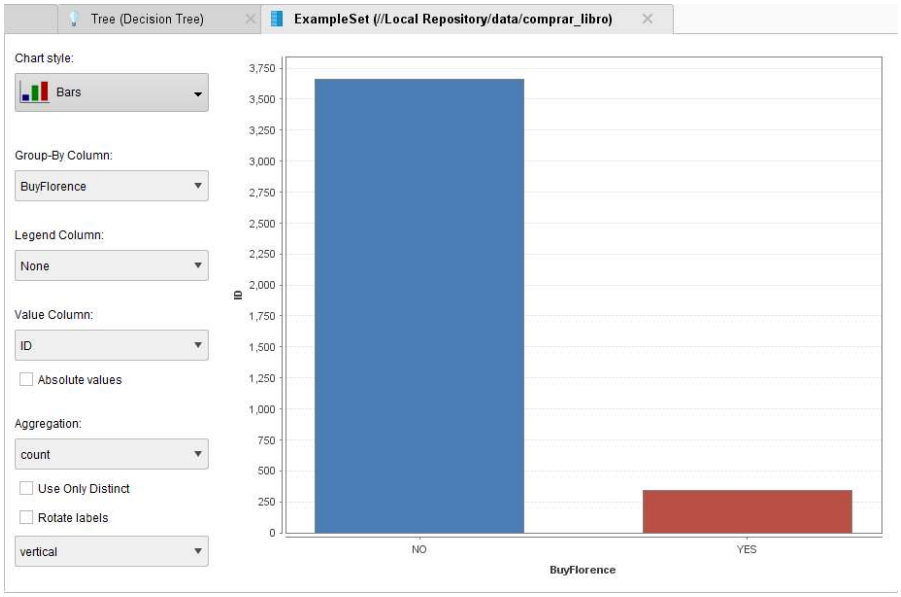
12. Como se puede observar **RapidMiner**, al aplicar aspectos de **"poda"**, decide que **no es necesario particionar más**, es decir, generar un árbol de decisión considera al conjunto de **4,000 tuplas** como **nodo hoja**. En este caso, como se observa, **no es una participación pura**. Si decidiéramos dejar el modelo, así como está, **el modelo clasificaría con NO**, cualquier tupla que se le presentara, independientemente de los atributos que tenga la tupla.

13. Podemos elegir **sobreajustar el árbol**, es decir, **obligarlo a particionar hasta obtener particiones puras**, en este caso, cambiamos las opciones de configuración del nodo que genera el árbol de decisión (desactivar las opciones marcadas). Ejecutamos de nuevo y veremos ahora si un árbol:

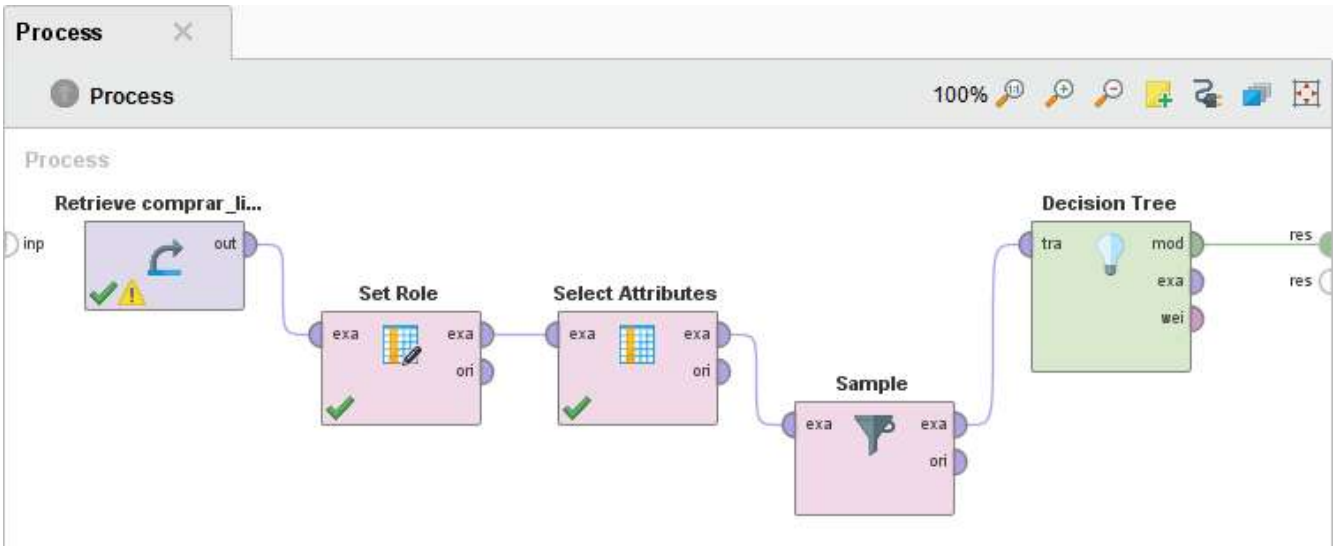


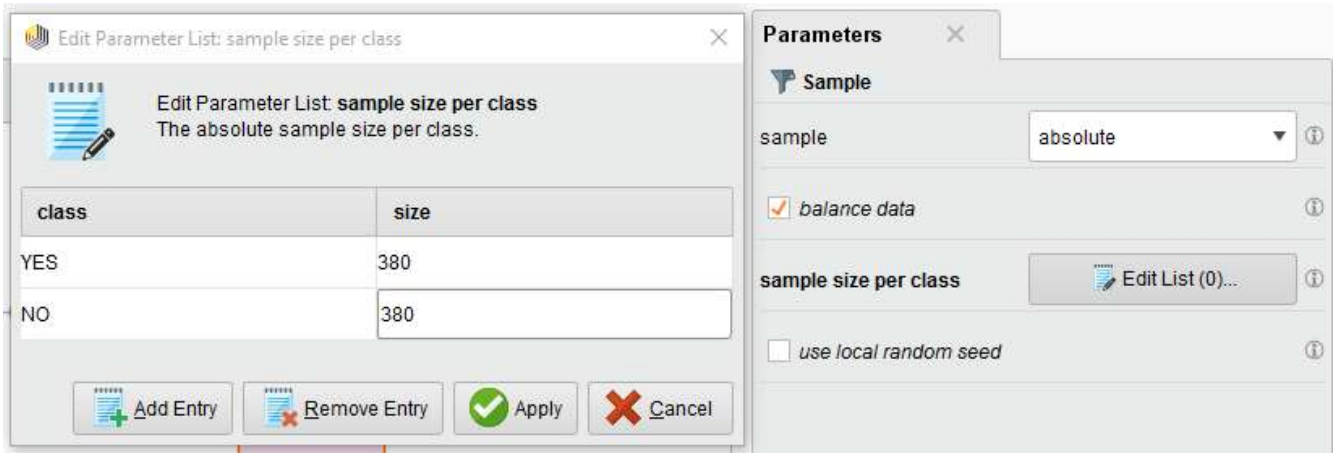


14. Por otro lado, si observamos la distribución de valores para el variable objetivo, nos damos cuenta que hay un problema de desbalance de los datos:

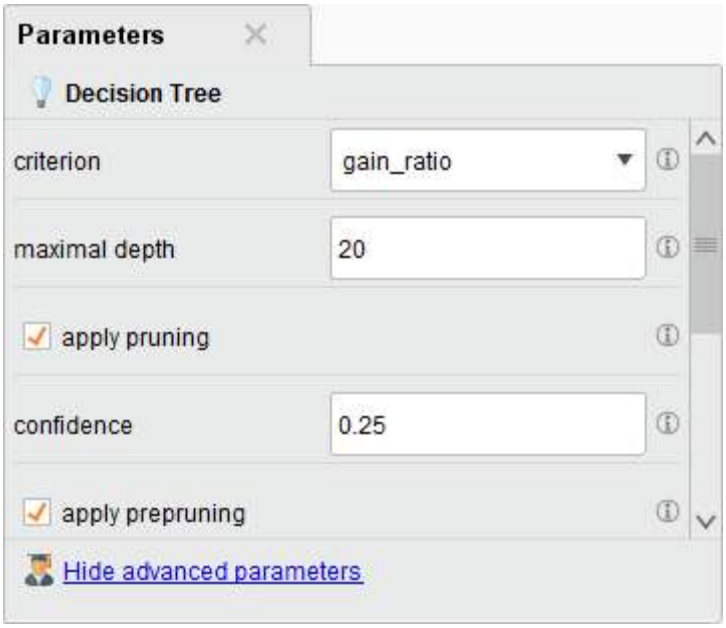


15. Hacemos un cambio en el proceso inicial y realizamos un muestreo estratificado:

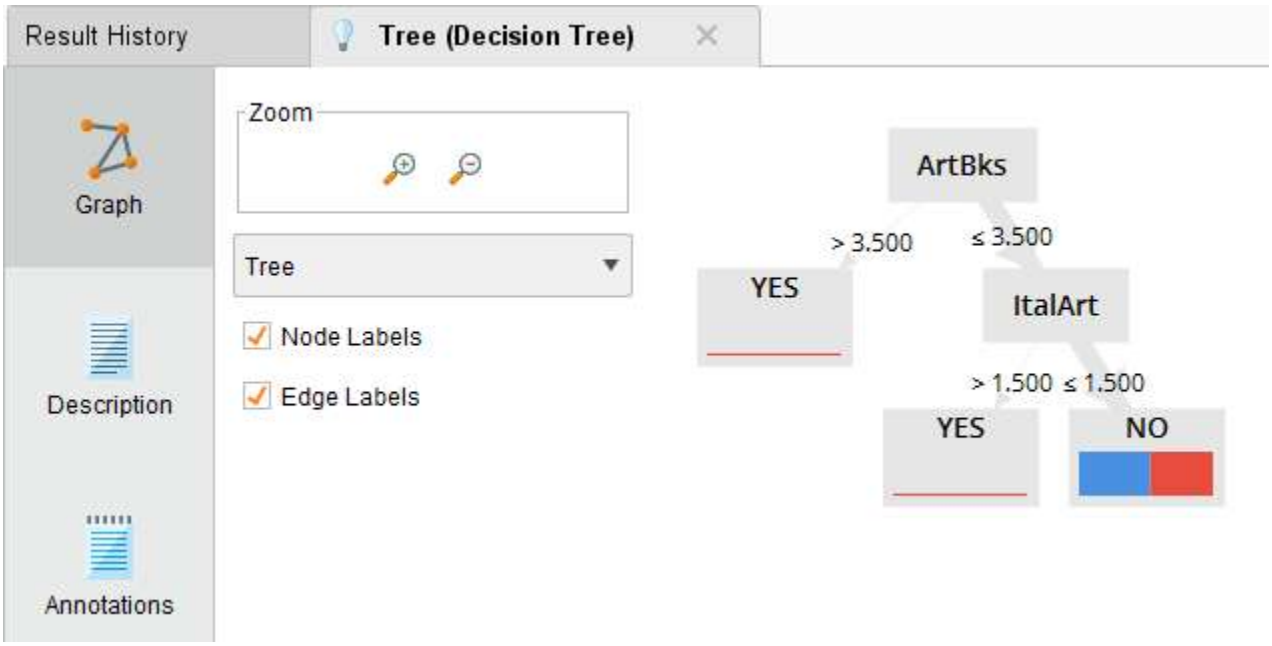




No debemos olvidar que se deben regresar las opciones iniciales del árbol de decisión:



16. Mostramos el resultado final:



17. Terminamos