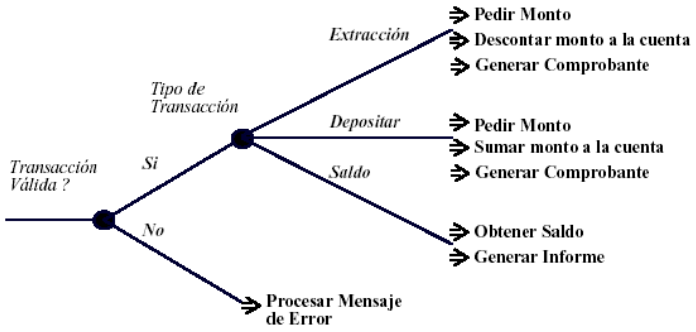
	UNIVERSIDAD DON BOSCO **FACULTAD DE INGENERIA ESCUELA DE COMPUTACION** **GUIA DE LABORATORIO Nº 11**
	Nombre de la práctica: **Árbol de decisión** Materia: **Data WareHouse y Minería de Datos**

Introducción

Aprendizaje basado en árboles de decisión utiliza un árbol de decisión como un modelo predictivo que mapea observaciones sobre un artículo a conclusiones sobre el valor objetivo del artículo. Es uno de los enfoques de modelado predictivo utilizadas en estadísticas, minería de datos y aprendizaje automático. Los modelos de árbol, donde la variable de destino puede tomar un conjunto finito de valores se denominan árboles de clasificación. En estas estructuras de árbol, las hojas representan etiquetas de clase y las ramas representan las conjunciones de características que conducen a esas etiquetas de clase. Los árboles de decisión, donde la variable de destino puede tomar valores continuos (por lo general números reales) se llaman árboles de regresión.

En análisis de decisión, un árbol de decisión se puede utilizar para representar visualmente y de forma explícita decisiones y toma de decisiones. En minería de datos, un árbol de decisión describe datos, pero no las decisiones; más bien el árbol de clasificación resultante puede ser un usado como entrada para la toma de decisiones. Esta página se ocupa de los árboles de decisión en la minería de datos.



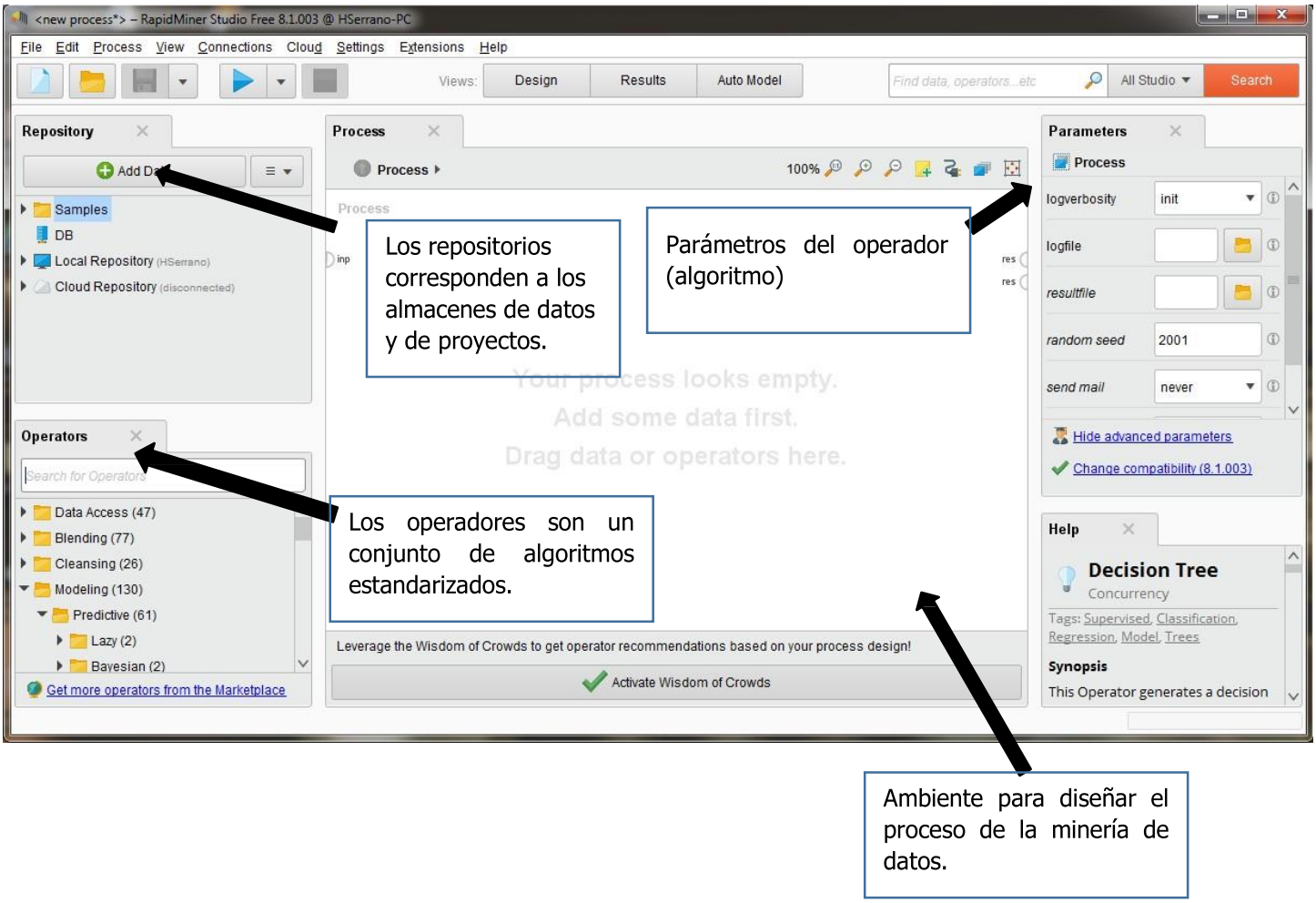
Equipo a utilizar:

- Computadora con rapidminer.
- Memoria USB.
- Guía proporcionada por el docente.

Procedimiento.

Lo primero que haremos es abrir rapidminer para conocer el entorno de trabajo.

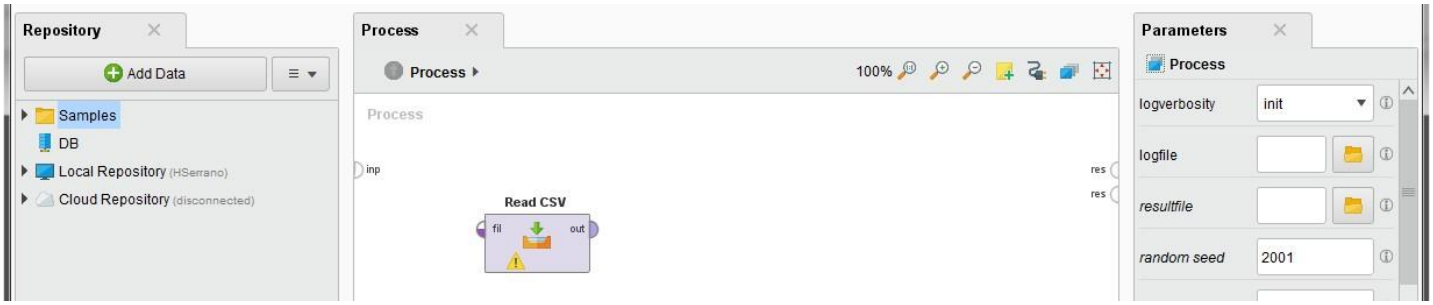
Entorno de rapidminer



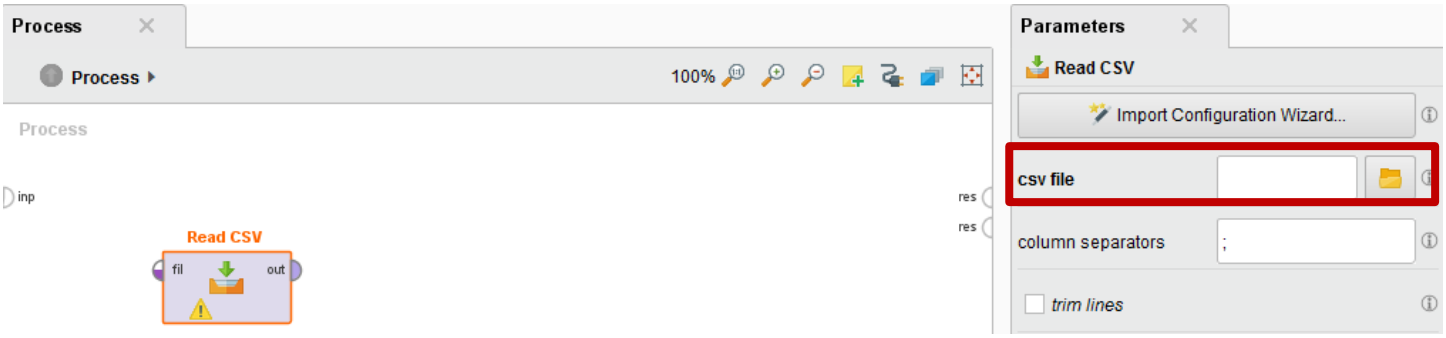
¿Cómo leer archivos en rapidminer?

Haremos un ejemplo utilizando una fuente de datos externa, la cual será un archivo csv (comma separated values).

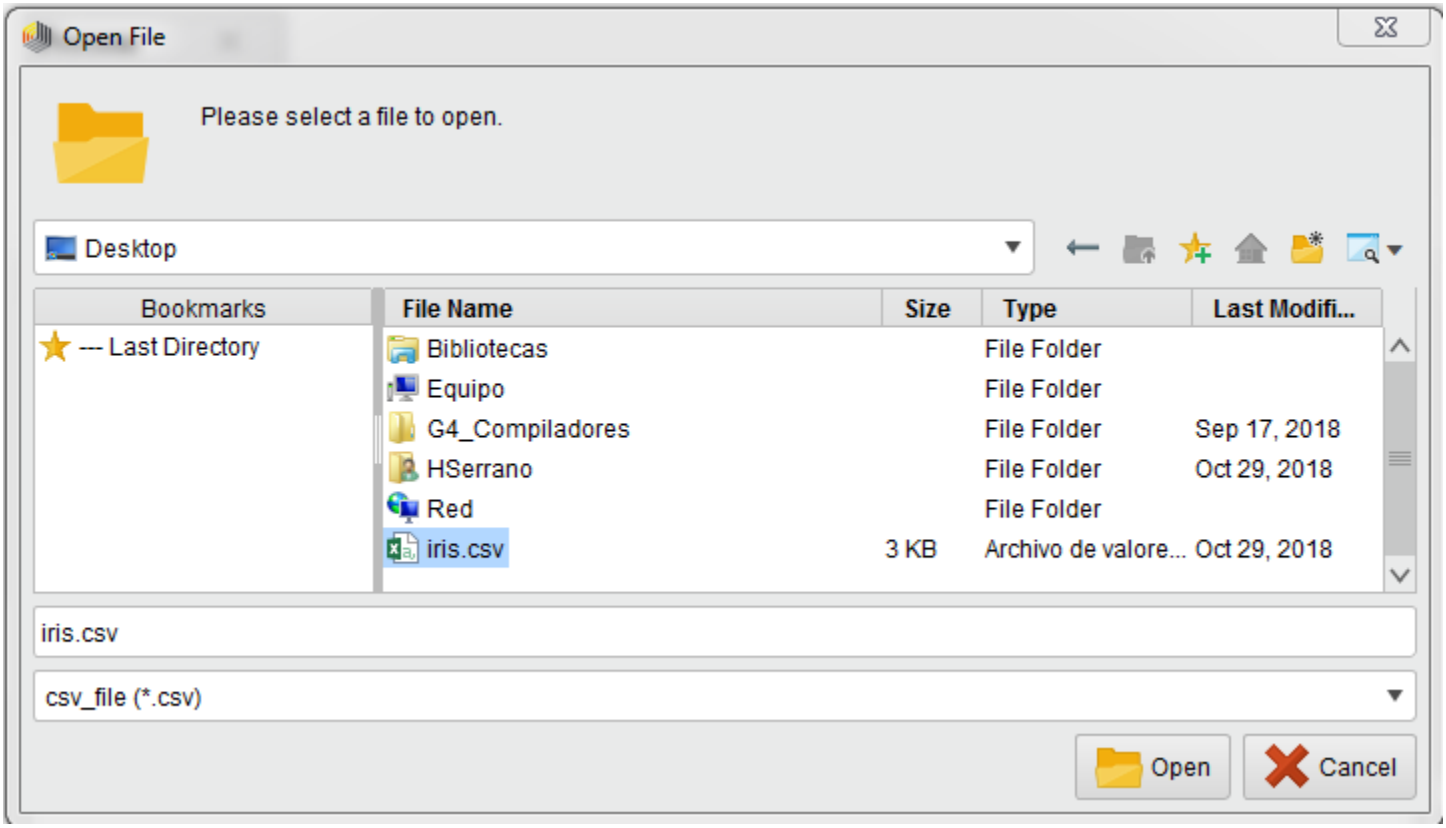
1. Vamos hacer uso de un repositorio de Inteligencia Artificial, el cual lo podemos descargar de la siguiente dirección (https://gist.github.com/curran/a08a1080b88344b0c8a7#file-iris-csv). Luego, vamos a utilizar bases de datos propias.
2. Una vez encontrado el código, se procede a guardar esa lista de datos en un archivo llamado "iris.csv".
3. Vamos a rapidminer y seleccionamos el operador "Read CSV" y lo arrastramos hasta el área de procesos de la aplicación.



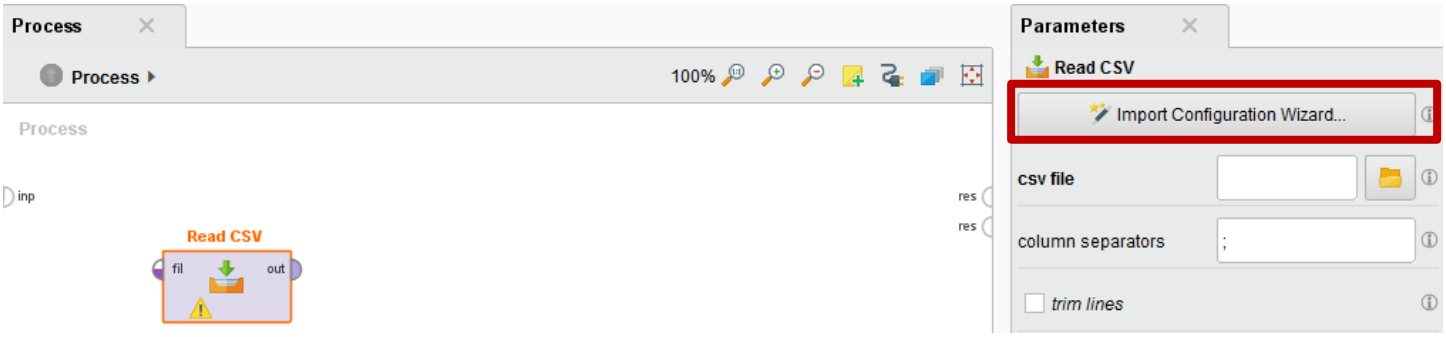
4. Ahora, cambiamos el parámetro, esto es definir dónde está ubicada la base de datos (en este caso nuestro archivo .csv). Para esto damos clic sobre el operador y buscamos la opción **csv file**.

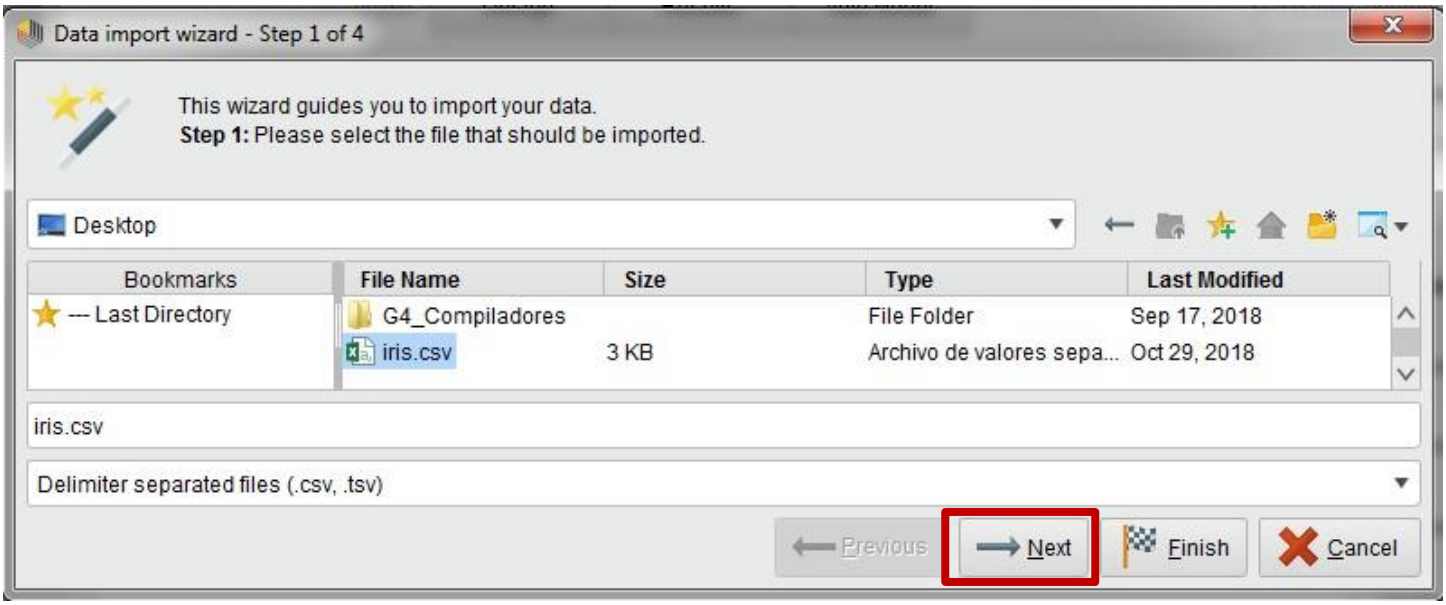


5. A continuación, se abre la siguiente ventana en la cual debemos buscar nuestra base de datos.

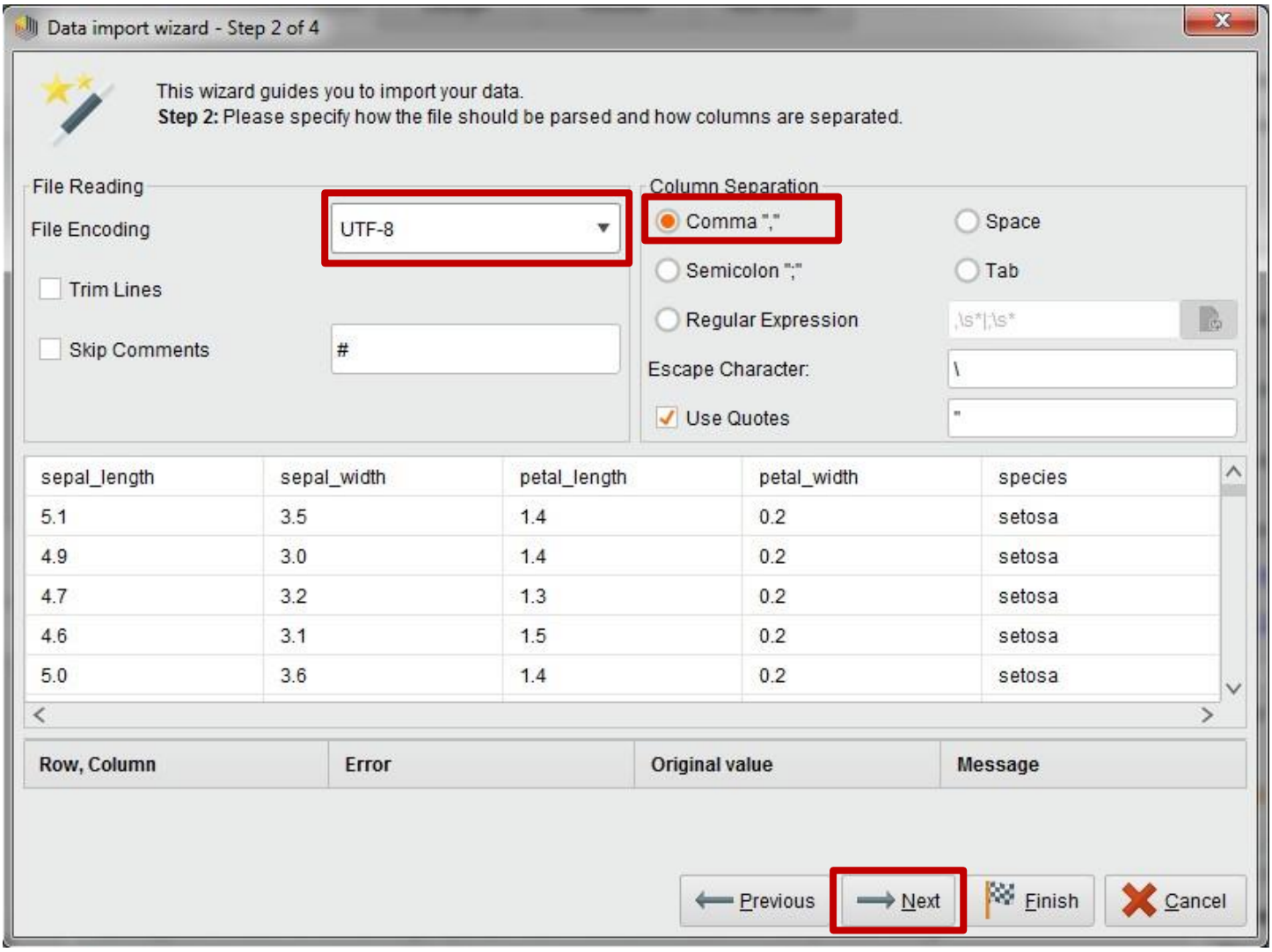


6. Ahora importamos la configuración haciendo clic en el siguiente botón:

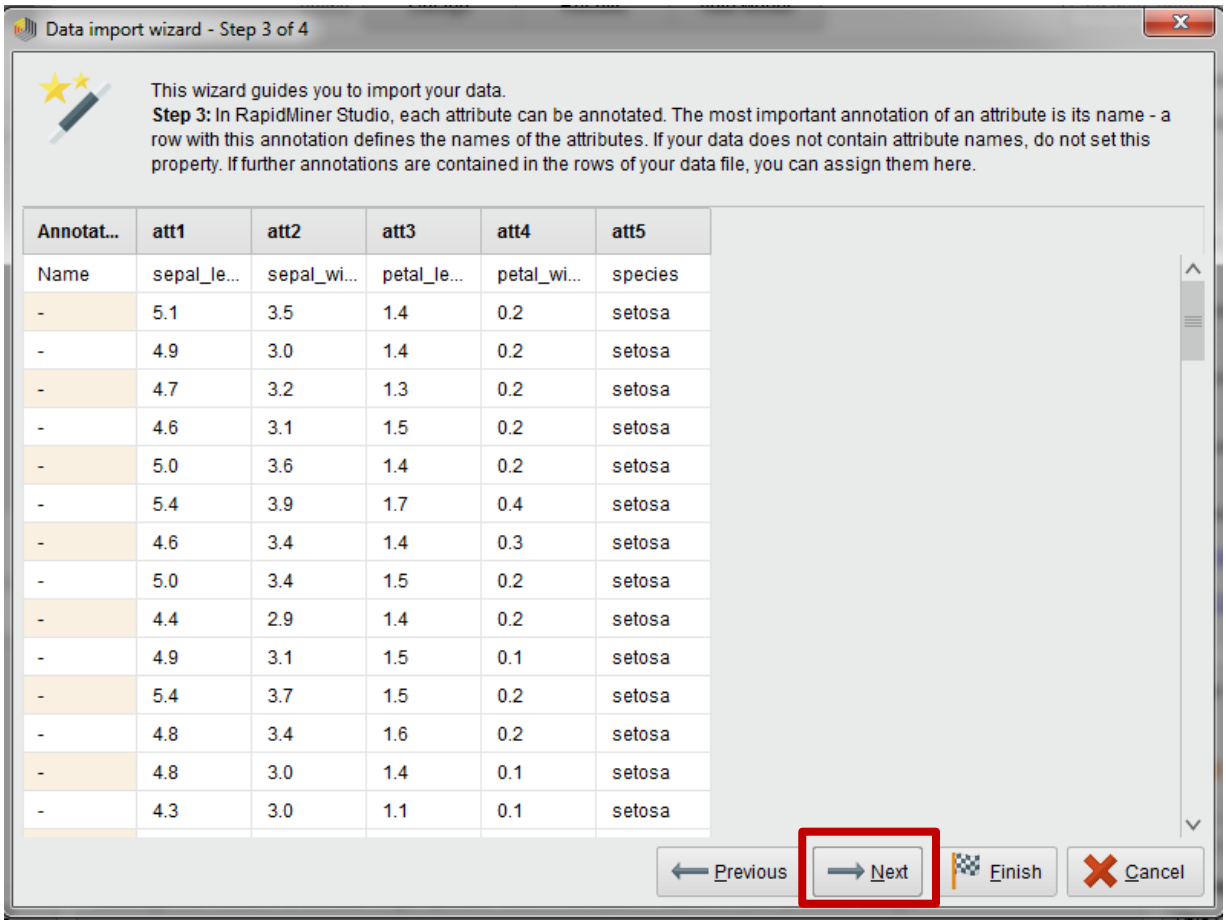




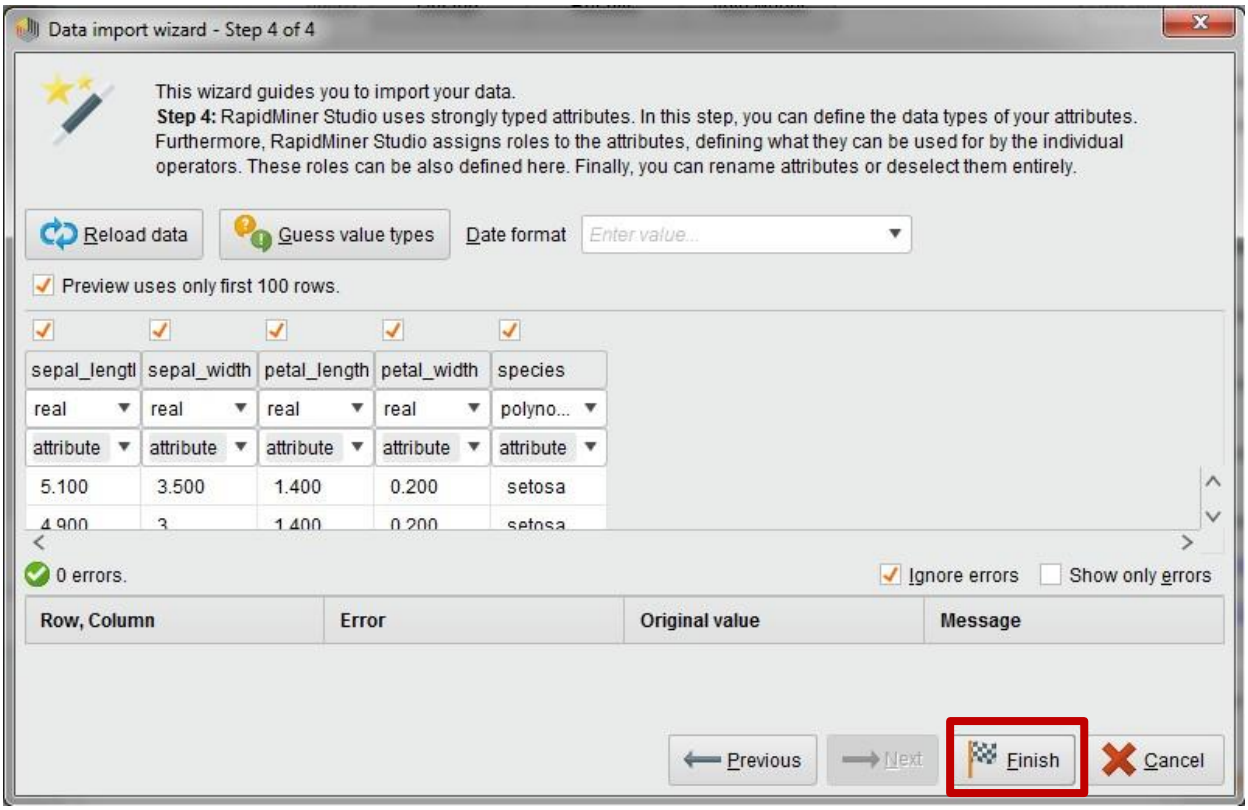
7. Seleccionamos algunas propiedades:



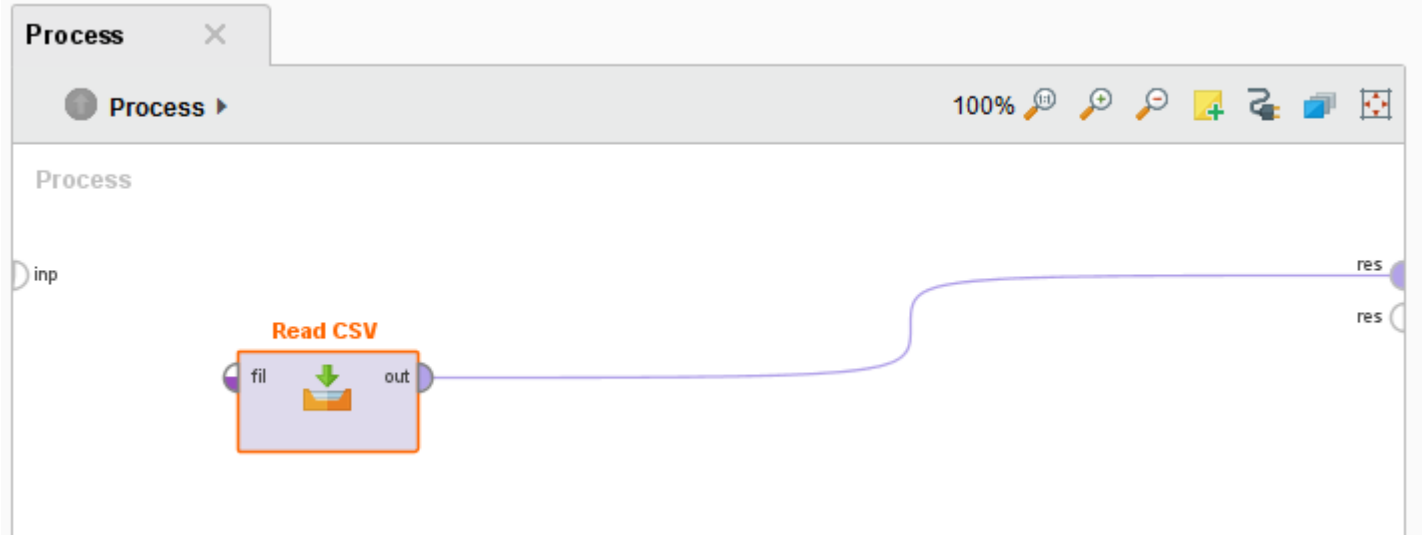
Debe aparecer la siguiente ventana:



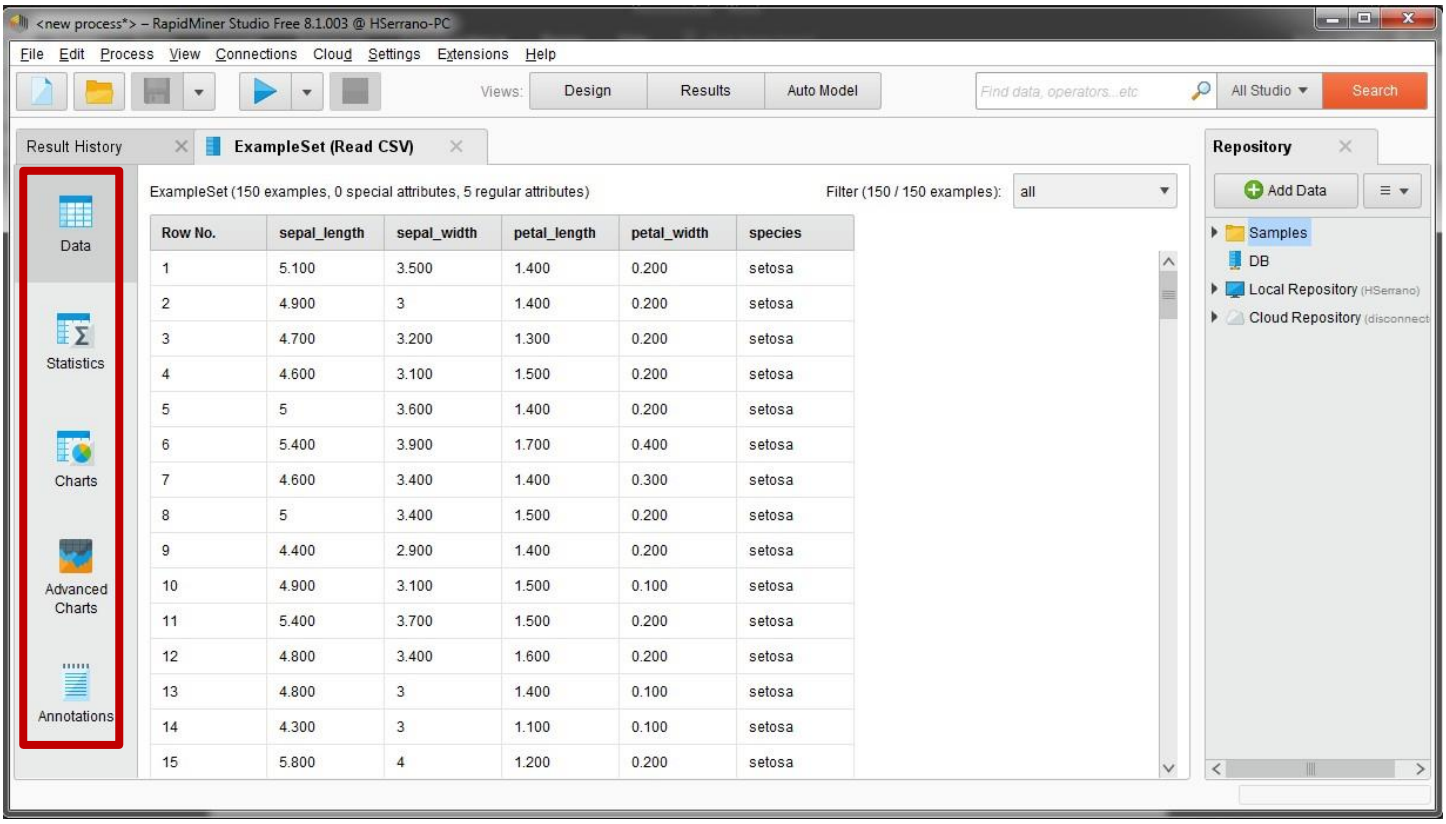
Deberá aparecer la siguiente ventana (ahí estarán los campos que vamos a utilizar)



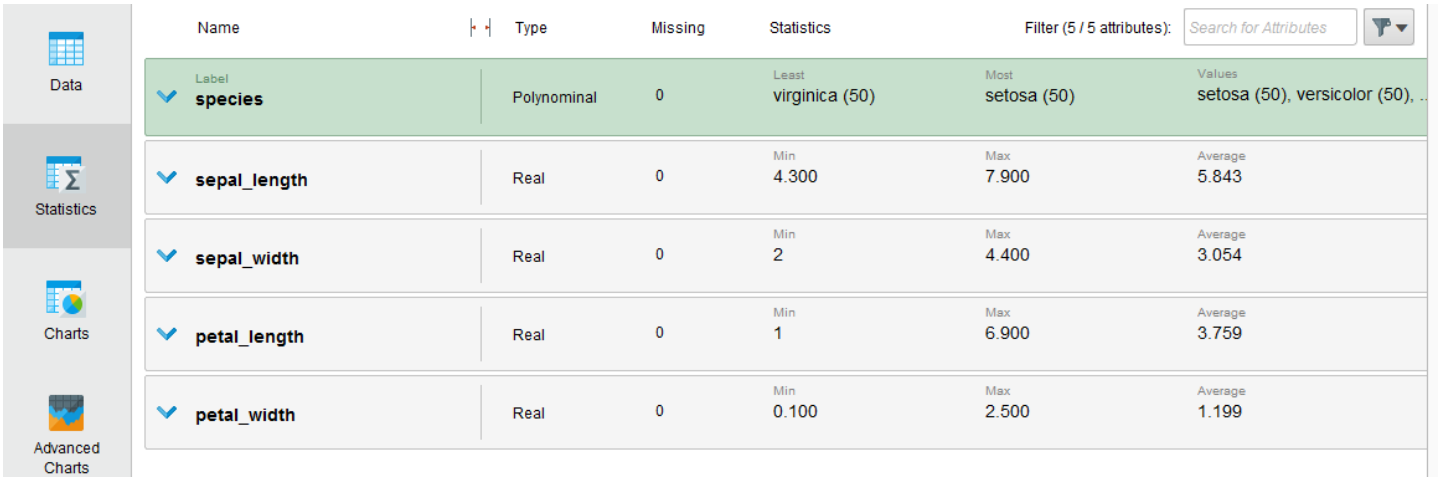
8. Realizamos la conexión:



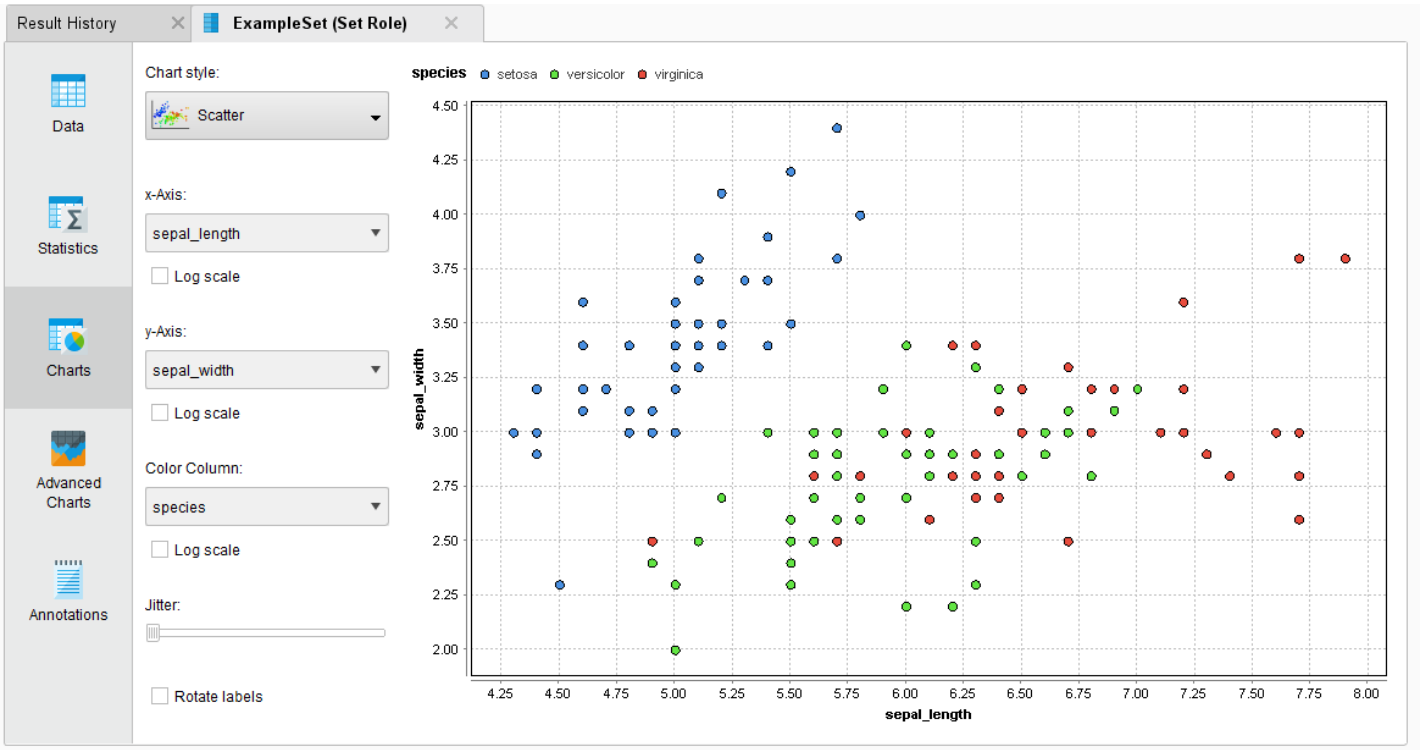
9. Ejecutamos y procedemos a verificar que los datos ahora están listos para ser manipulados y analizados en rapidminer:



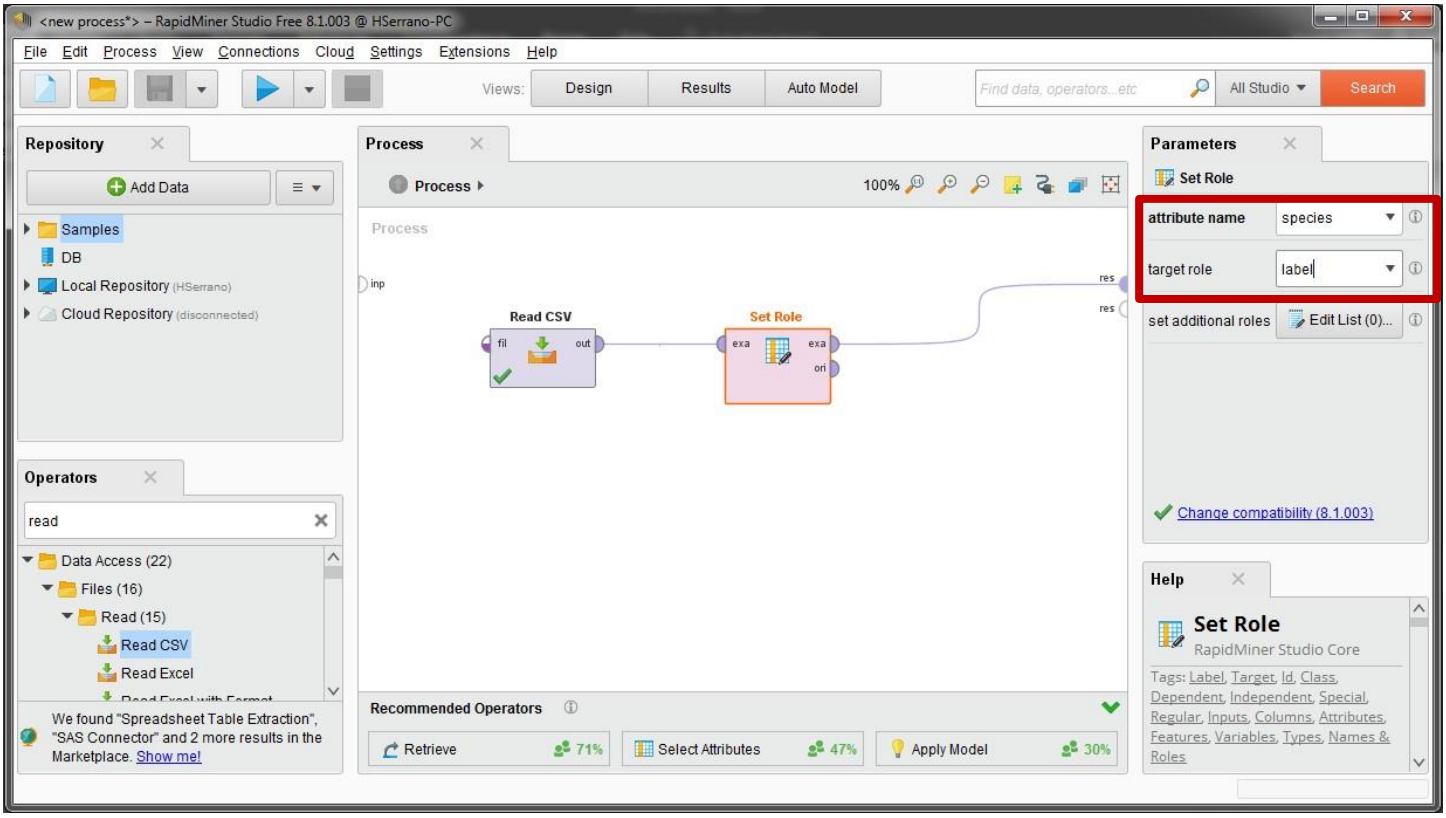
10. Usted puede verificar todas las opciones de rapidminer para representar la información. Para ello se seleccionan las opciones del panel izquierdo de la aplicación. Por ejemplo, veamos las estadísticas y gráficos (puede seleccionar diferentes tipos de gráficos).


Data
Statistics
Charts
Advanced Charts
Name
Type
Missing
Statistics
Filter (5 / 5 attributes):
Search for Attributes
Label
species
Polynominal
0
Least
virginica (50)
Most
setosa (50)
Values
setosa (50), versicolor (50), ..
sepal_length
Real
0
Min
4.300
Max
7.900
Average
5.843
sepal_width
Real
0
Min
2
Max
4.400
Average
3.054
petal_length
Real
0
Min
1
Max
6.900
Average
3.759
petal_width
Real
0
Min
0.100
Max
2.500
Average
1.199

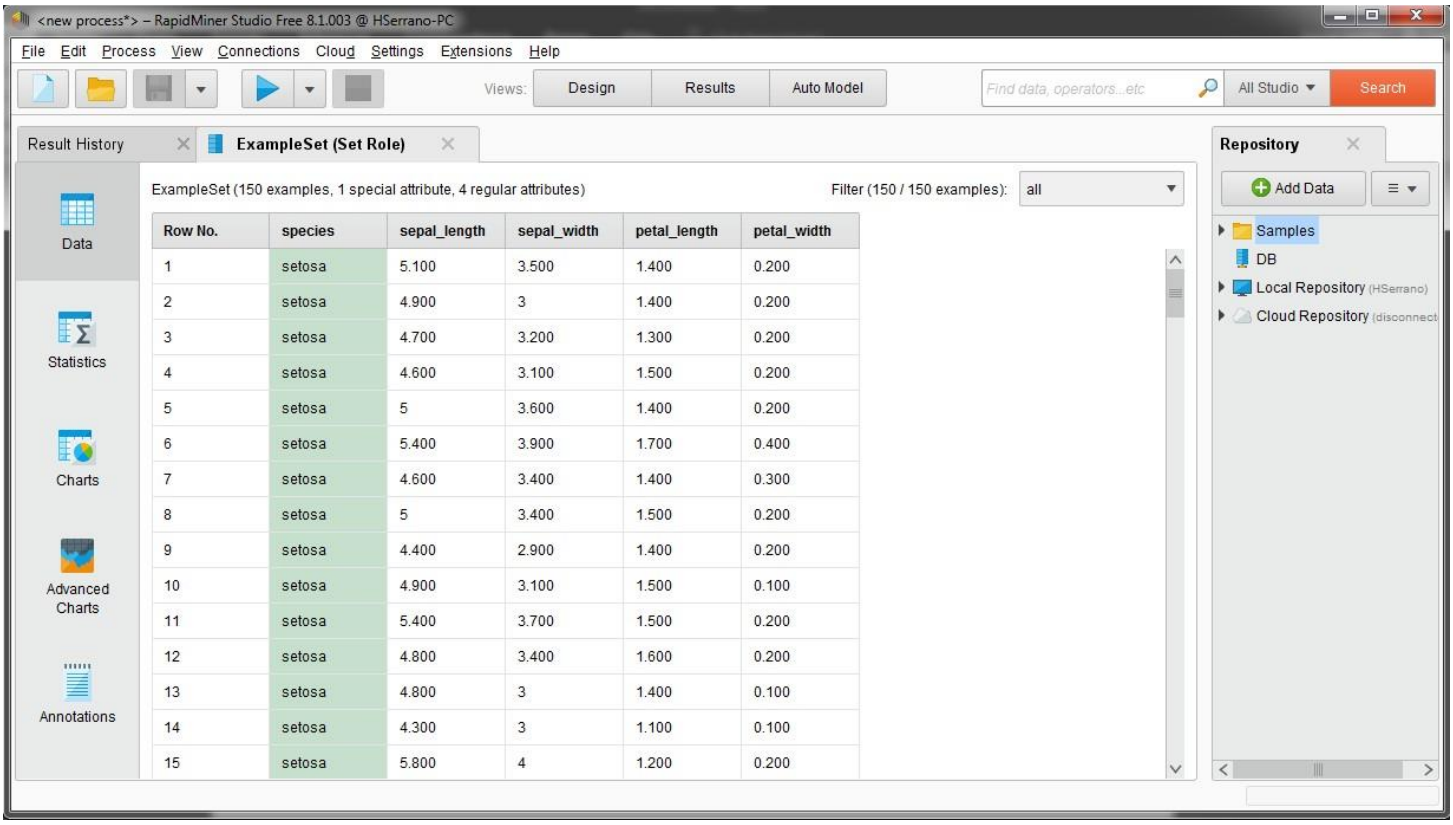
Gráficos:


Result History
ExampleSet (Set Role)
Data
Statistics
Charts
Advanced Charts
Annotations
Chart style:
Scatter
x-Axis:
sepal_length
Log scale
y-Axis:
sepal_width
Log scale
Color Column:
species
Log scale
Jitter:
Rotate labels
species
setosa
versicolor
virginica
sepal_width
sepal_length
4.50
4.25
4.00
3.75
3.50
3.25
3.00
2.75
2.50
2.25
2.00
4.25
4.50
4.75
5.00
5.25
5.50
5.75
6.00
6.25
6.50
6.75
7.00
7.25
7.50
7.75
8.00

11. Ahora, vamos a establecer un rol (Ser Role):

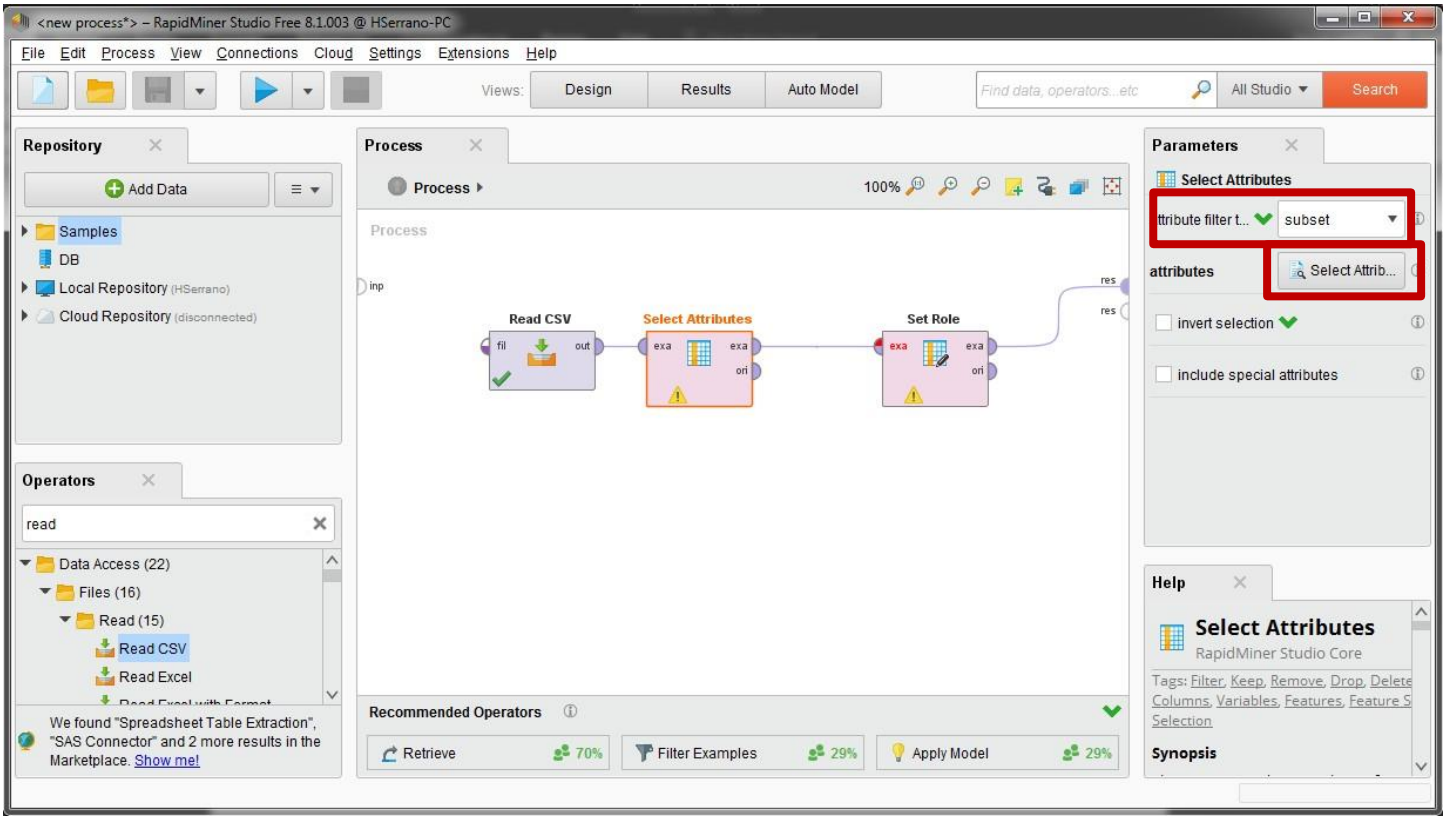


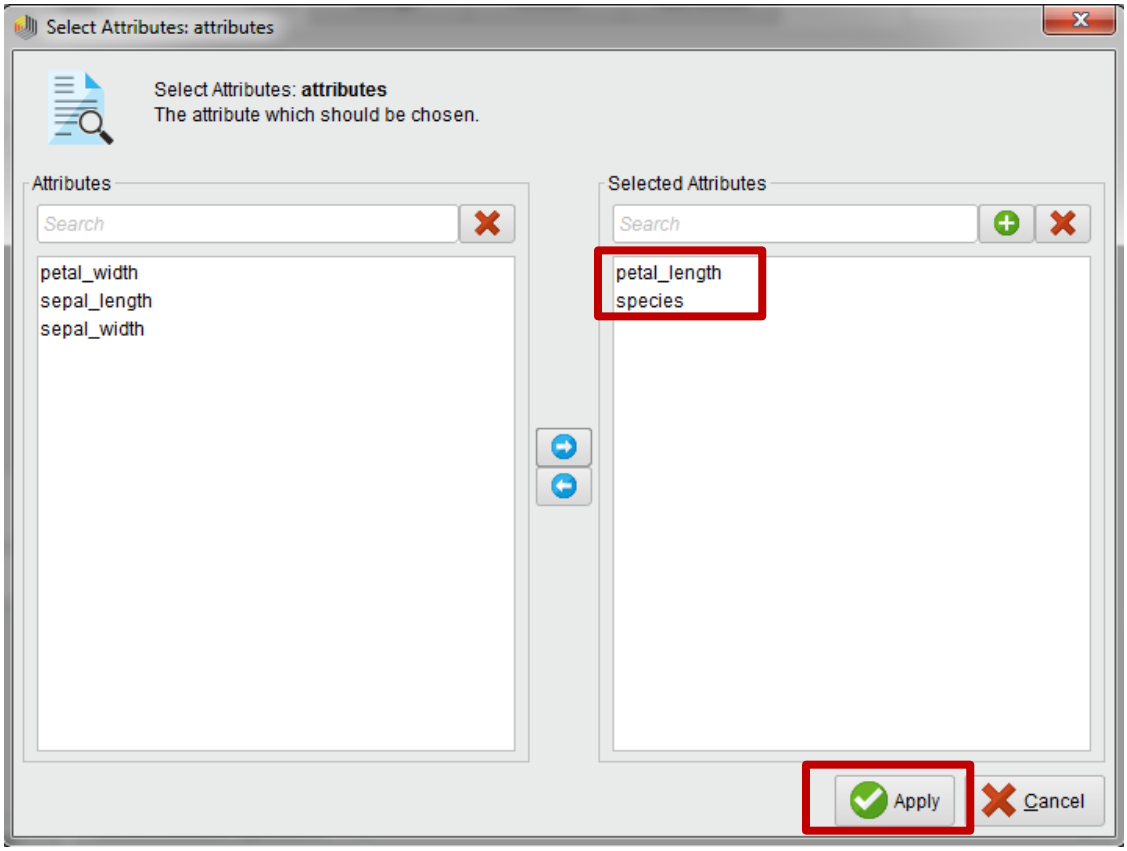
12. Al ejecutar se muestra lo siguiente:



Row No.	species	sepal_length	sepal_width	petal_length	petal_width
1	setosa	5.100	3.500	1.400	0.200
2	setosa	4.900	3	1.400	0.200
3	setosa	4.700	3.200	1.300	0.200
4	setosa	4.600	3.100	1.500	0.200
5	setosa	5	3.600	1.400	0.200
6	setosa	5.400	3.900	1.700	0.400
7	setosa	4.600	3.400	1.400	0.300
8	setosa	5	3.400	1.500	0.200
9	setosa	4.400	2.900	1.400	0.200
10	setosa	4.900	3.100	1.500	0.100
11	setosa	5.400	3.700	1.500	0.200
12	setosa	4.800	3.400	1.600	0.200
13	setosa	4.800	3	1.400	0.100
14	setosa	4.300	3	1.100	0.100
15	setosa	5.800	4	1.200	0.200

13. Y, por último, vamos a agregar un filtro a la información seleccionando subconjuntos de datos:





Solo van a aparecer estos registros:

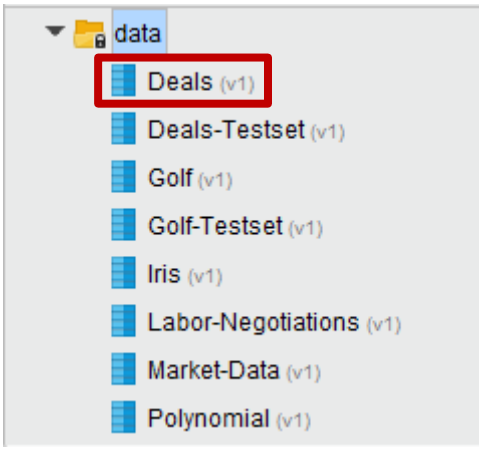
ExampleSet (150 examples, 1 special attribute, 1 regular attribute)

Row No.	species	petal_length
1	setosa	1.400
2	setosa	1.400
3	setosa	1.300
4	setosa	1.500
5	setosa	1.400
6	setosa	1.700
7	setosa	1.400
8	setosa	1.500
9	setosa	1.400
10	setosa	1.500
11	setosa	1.500
12	setosa	1.600
13	setosa	1.400
14	setosa	1.100
15	setosa	1.200

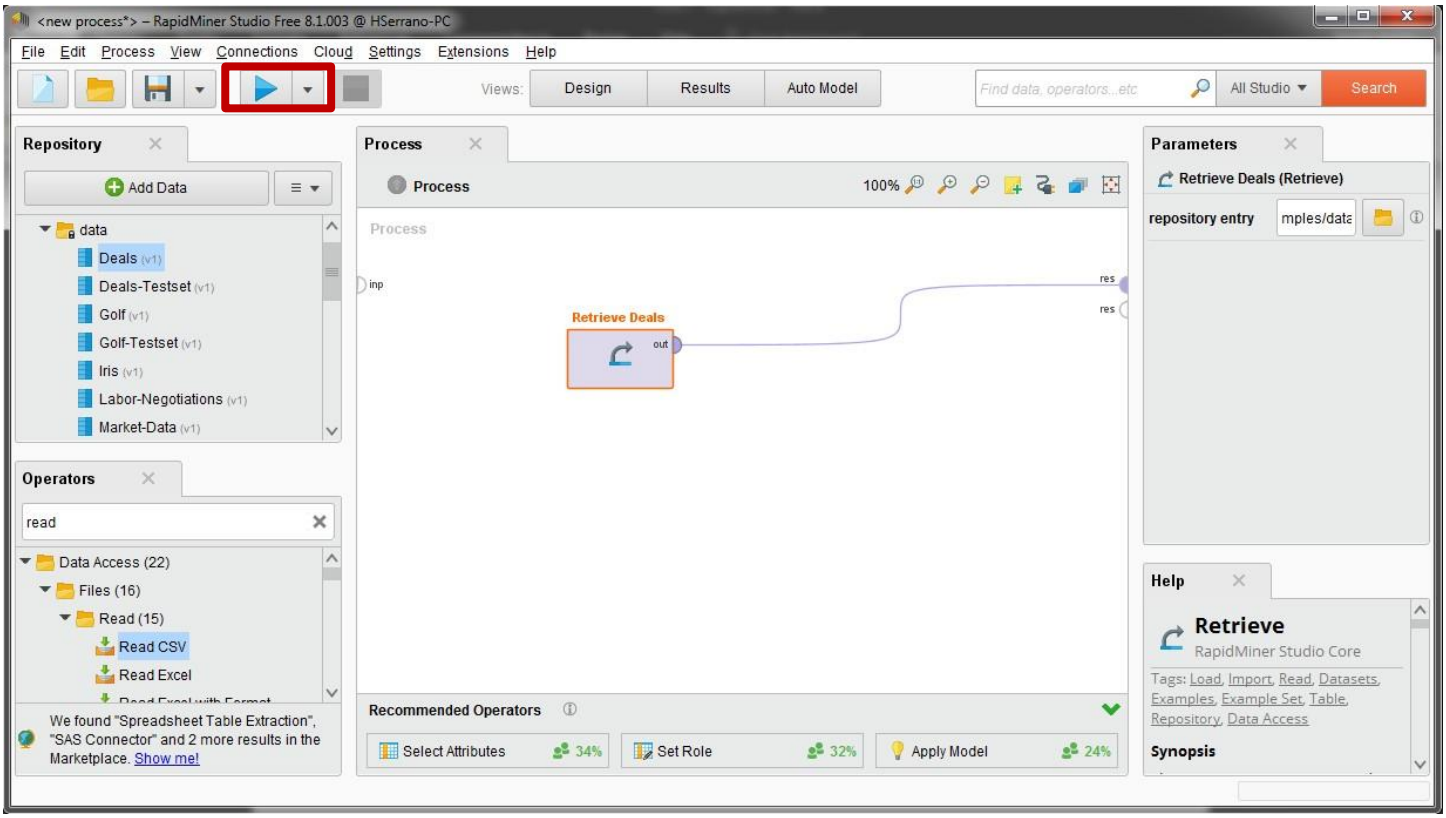
Árbol de decisión en RapidMiner

Haciendo uso de un repositorio que trae por defecto RapidMiner, Deals (ofertas) que muestra atributos para decidir si una persona puede ser un cliente futuro de acuerdo a sus atributos.

1. Acceder a los repositorios de RapidMiner y buscar "Deals".



2. Arrastramos hasta el área de trabajo, realiza la respectiva conexión y ejecute para verificar la correcta lectura de la base de datos.

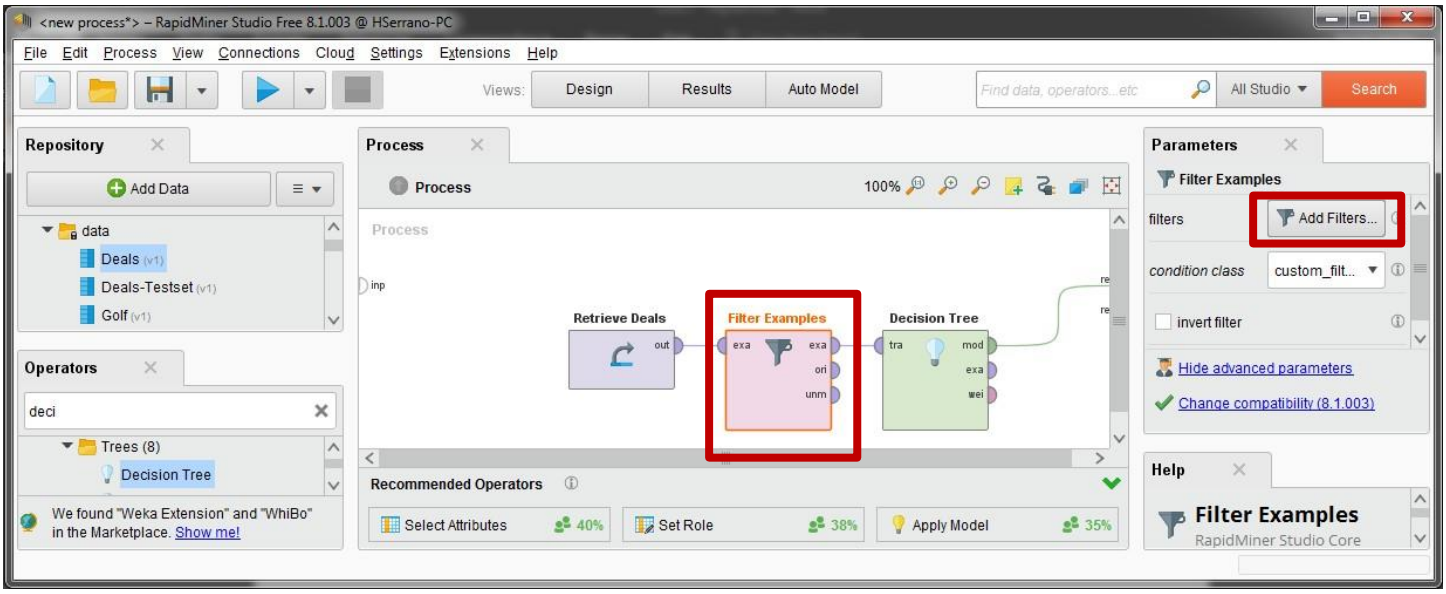


3. Al ejecutar le deberán aparecer los registros (en total son 1000 registros).

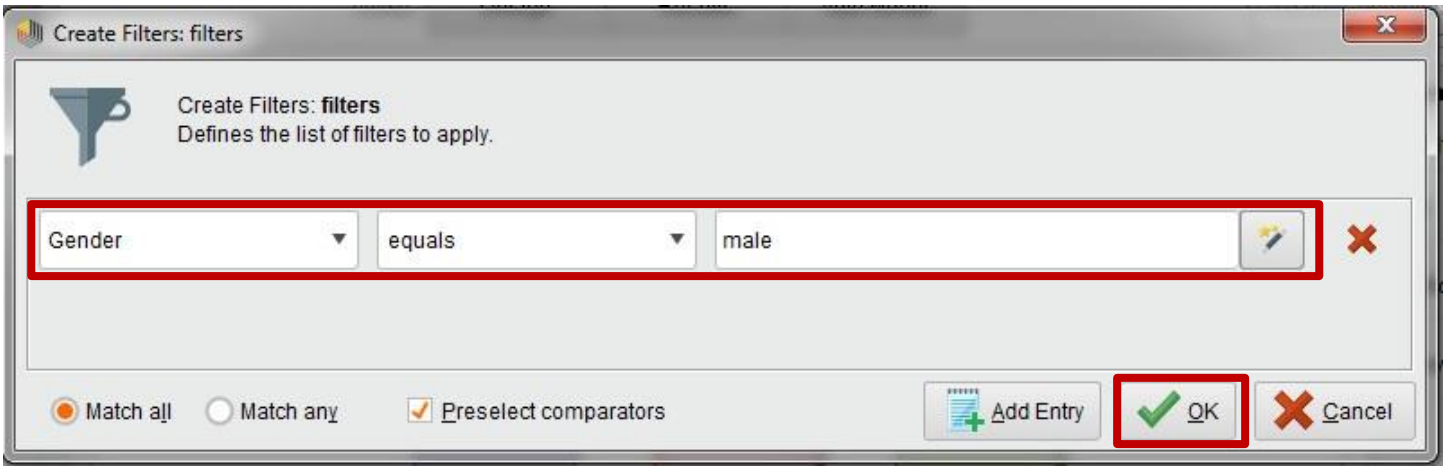
ExampleSet (1000 examples, 1 special attribute, 3 regular attributes)

Row No.	Future Cust...	Age	Gender	Payment Me...
1	yes	64	male	credit card
2	yes	35	male	cheque
3	yes	25	female	credit card
4	no	39	female	credit card
5	yes	39	male	credit card
6	no	28	female	cheque
7	yes	21	female	credit card
8	yes	48	male	credit card
9	no	70	female	credit card
10	yes	36	male	credit card
11	yes	22	male	credit card
12	no	53	female	cash
13	yes	27	male	cash
14	yes	40	male	credit card
15	yes	22	male	cash
16	no	49	female	credit card
17	no	24	female	cash

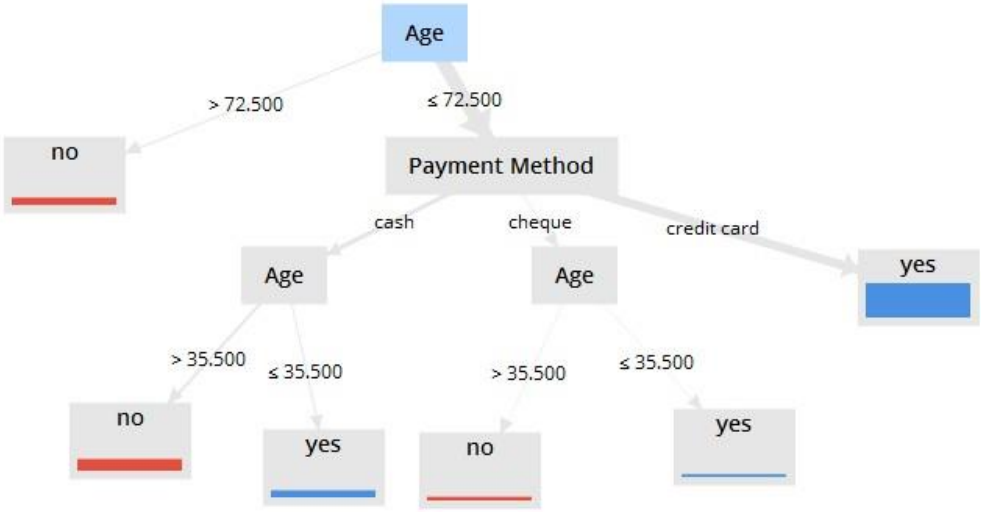
4. Ahora, conectamos el repositorio con un filtro para que solo tome a los del género masculino y posteriormente con el árbol de decisión. Hacemos clic sobre el filtro y luego en el botón "Add Filters" tal y como se muestra en la siguiente imagen:



5. Una vez seleccionada esa opción, vamos a elegir el género "masculino".



6. Al ejecutar (vea la pestaña "Results") se tiene lo siguiente:



Ejercicios:

1. Haciendo uso del archivo de Excel proporcionado por el docente:
   a) Cargar el archivo "whisky.xls" (proporcionado por el docente).
   b) Seleccionar los atributos "Calidad" y "Añejamiento", "Calidad" y "Precio".
   c) Generar las respectivas gráficas y visualizar las estadísticas.
   d) Aplicar un filtro y mostrar únicamente los precios de whisky mayores a 82.

2. Utilizando el repositorio "Deals" de RapidMiner:
   a) Cargar el archivo para visualizar los registros.
   b) Generar los respectivos arboles de decisión para conocer si puede ser un futuro cliente,
      - Aplicar filtro por tipo de pago cash y cheque
      - Aplicar filtro por edad

3. Genere el árbol de decisión correspondiente para la siguiente tabla, verificando si se le concederá un préstamo o no.

ID	CASA	ESTADO	INGRESOS	PRÉSTAMO
id_1	Propiedad	Soltero	125000	Conceder
id_2	Alquiler	Casado	100000	Conceder
id_3	Alquiler	Soltero	70000	Conceder
id_4	Propiedad	Casado	12000	Conceder
id_5	Alquiler	Divorciado	95000	Denegar
id_6	Alquiler	Casado	60000	Conceder
id_7	Propiedad	Divorciado	220000	Conceder
id_8	Alquiler	Soltero	85000	Denegar
id_9	Alquiler	Casado	75000	Conceder
id_{10}	Alquiler	Soltero	90000	Conceder

4. Utilizando el archivo de Excel estado_civil.xlsx, genere el árbol de decisión para determinar cual es el estado civil de la persona según los datos.
5. Utilizando el archivo de Excel primer_compra.xlsx, genere el árbol de decisión para determinar cual será la primera compra de la persona.