



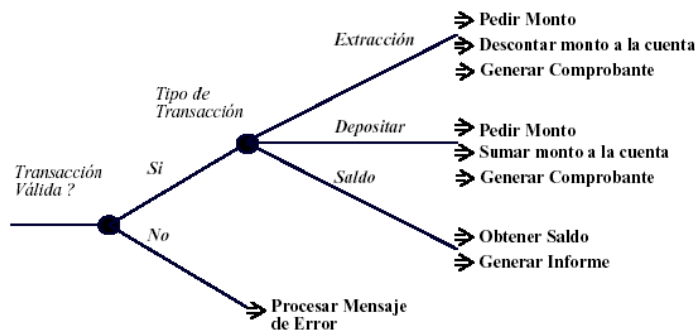
**UNIVERSIDAD DON BOSCO**  
**FACULTAD DE INGENIERIA ESCUELA DE COMPUTACION**  
**GUIA DE LABORATORIO Nº 11**

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 uso como entrada para la toma de decisiones. Esta página se ocupa de los árboles de decisión en la minería de datos.



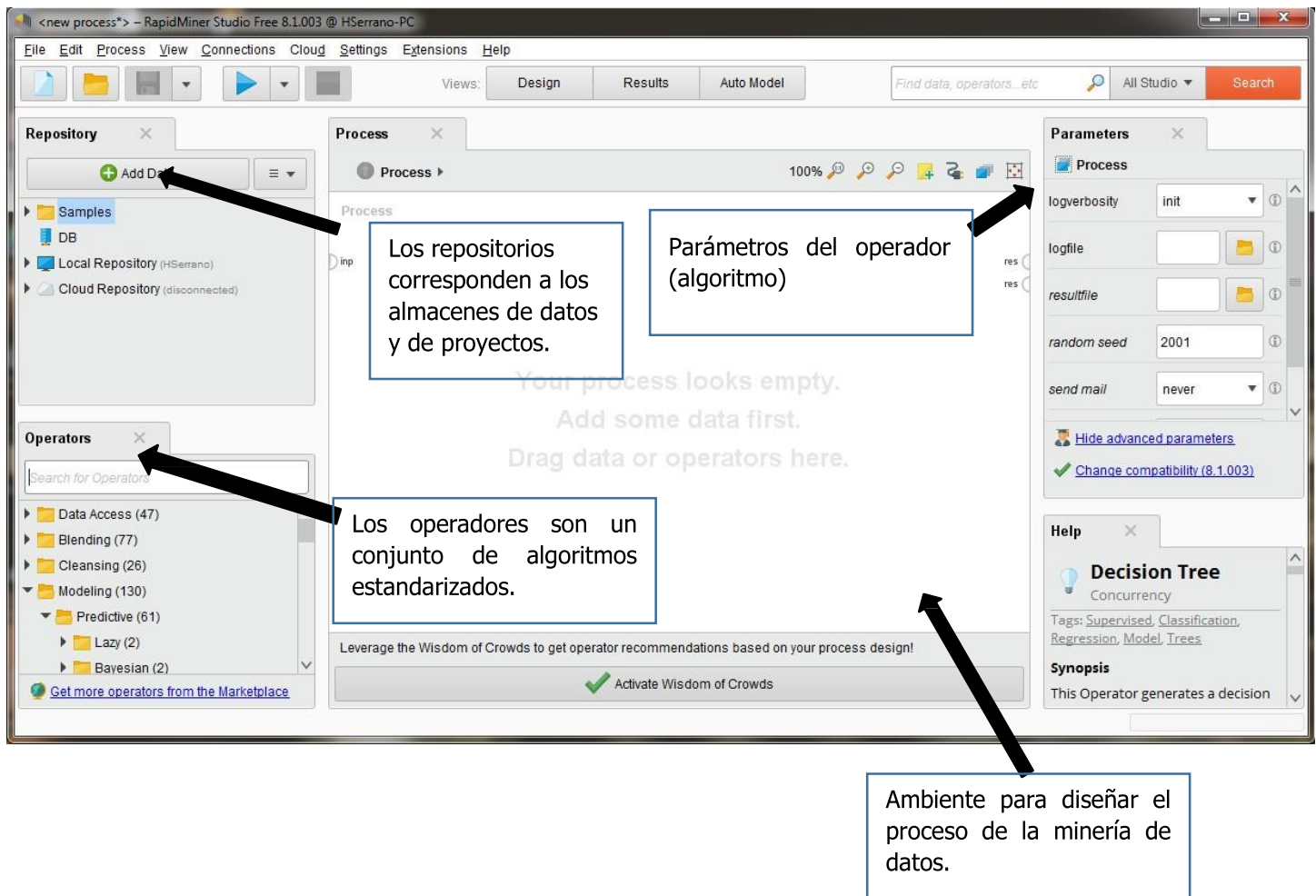
## Equipo a utilizar:

- 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



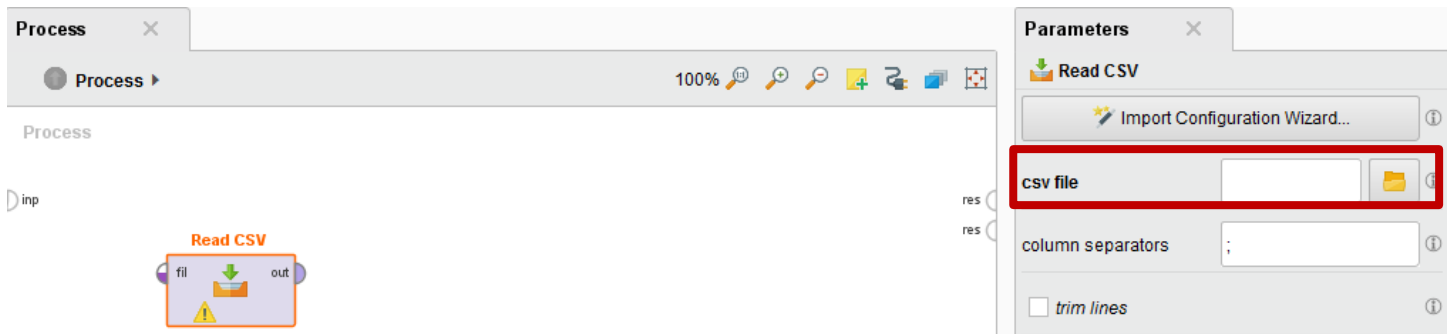
## ¿Cómo leer archivos en rapidminer?

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

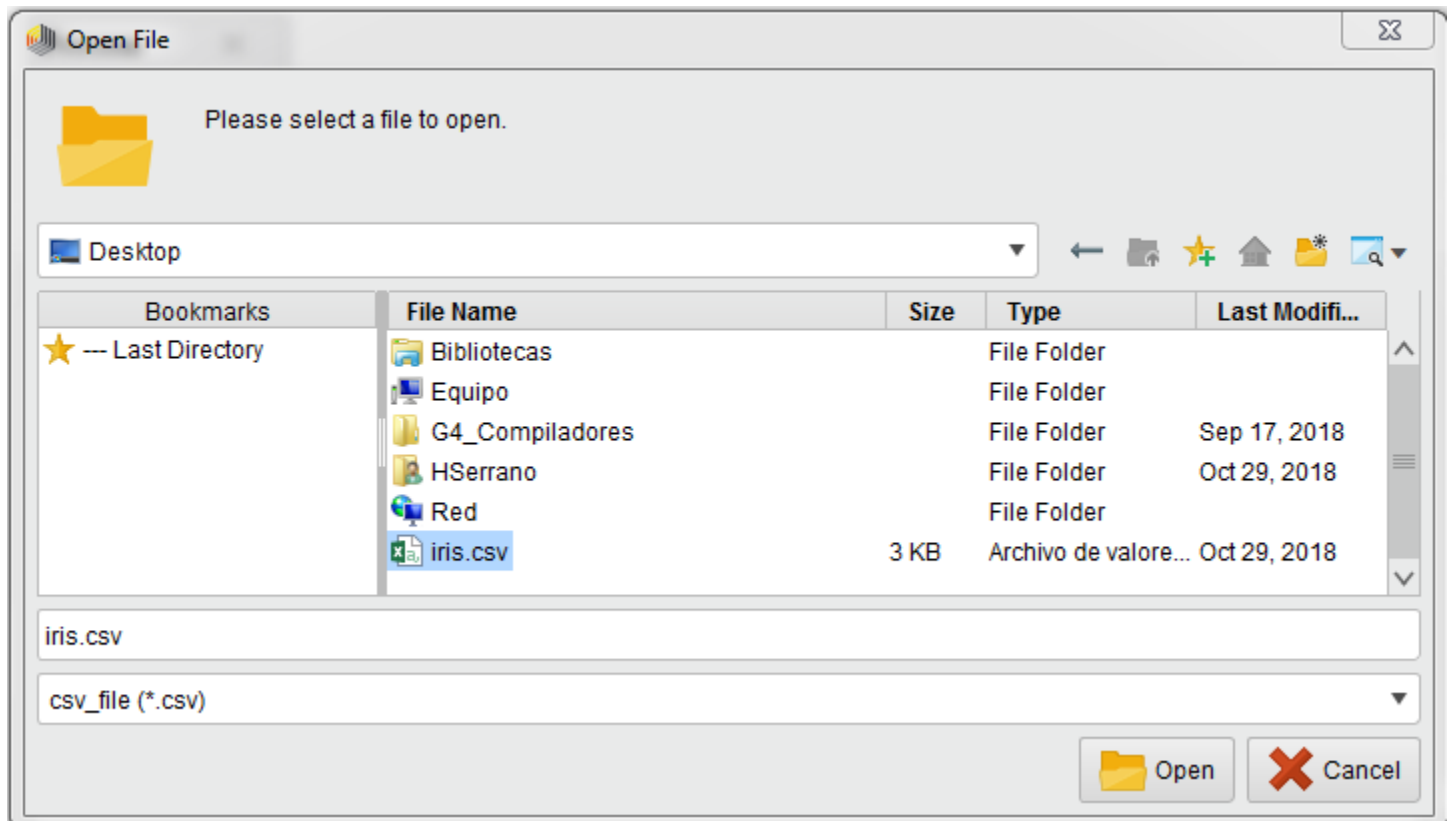
1. Vamos hacer uso de un repositorio de Inteligencia Artificial, el cual lo podemos descargar de la siguiente dirección (<https://gist.github.com/curran/a08a1080b88344b0c8a7#file-iris-csv>). Luego, vamos a utilizar bases de datos propias.
2. Una vez encontrado el código, se procede a guardar esa lista 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.



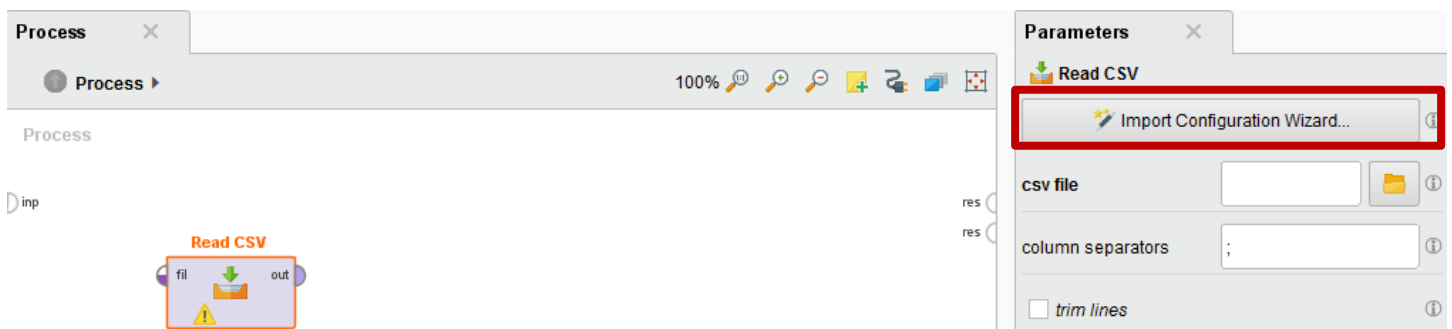
4. Ahora, cambiamos el parámetro, esto es definir dónde está ubicada la base de datos (en este caso nuestro archivo .csv). Para esto damos clic sobre el operador y buscamos la opción **csv file**.



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:



**Data import wizard - Step 1 of 4**

This wizard guides you to import your data.  
**Step 1:** Please select the file that should be imported.

Desktop

| Bookmarks           | File Name       | Size | Type                       | Last Modified |
|---------------------|-----------------|------|----------------------------|---------------|
| ★ -- Last Directory | G4_Compiladores |      | File Folder                | Sep 17, 2018  |
|                     | iris.csv        | 3 KB | Archivo de valores sepa... | Oct 29, 2018  |

iris.csv

Delimiter separated files (.csv, .tsv)

Previous **Next** Finish Cancel

7. Seleccionamos algunas propiedades:

**Data import wizard - Step 2 of 4**

This wizard guides you to import your data.  
**Step 2:** Please specify how the file should be parsed and how columns are separated.

**File Reading**

File Encoding: **UTF-8**

☐ Trim Lines

☐ Skip Comments #

**Column Separation**

☒ Comma ","

☐ Space

☐ Semicolon ";"

☐ Tab

☐ Regular Expression `/s*/s*`

Escape Character: \

☒ Use Quotes "

| sepal_length | sepal_width | petal_length | petal_width | species |
|--------------|-------------|--------------|-------------|---------|
| 5.1          | 3.5         | 1.4          | 0.2         | setosa  |
| 4.9          | 3.0         | 1.4          | 0.2         | setosa  |
| 4.7          | 3.2         | 1.3          | 0.2         | setosa  |
| 4.6          | 3.1         | 1.5          | 0.2         | setosa  |
| 5.0          | 3.6         | 1.4          | 0.2         | setosa  |

| Row, Column | Error | Original value | Message |
|-------------|-------|----------------|---------|
|-------------|-------|----------------|---------|

Previous **Next** Finish Cancel

Debe aparecer la siguiente ventana:

**Data import wizard - Step 3 of 4**

This wizard guides you to import your data.  
**Step 3:** In RapidMiner Studio, each attribute can be annotated. The most important annotation of an attribute is its name - a row with this annotation defines the names of the attributes. If your data does not contain attribute names, do not set this property. If further annotations are contained in the rows of your data file, you can assign them here.

| Annotat... | att1        | att2        | att3        | att4        | att5    |
|------------|-------------|-------------|-------------|-------------|---------|
| Name       | sepal_le... | sepal_wi... | petal_le... | petal_wi... | species |
| -          | 5.1         | 3.5         | 1.4         | 0.2         | setosa  |
| -          | 4.9         | 3.0         | 1.4         | 0.2         | setosa  |
| -          | 4.7         | 3.2         | 1.3         | 0.2         | setosa  |
| -          | 4.6         | 3.1         | 1.5         | 0.2         | setosa  |
| -          | 5.0         | 3.6         | 1.4         | 0.2         | setosa  |
| -          | 5.4         | 3.9         | 1.7         | 0.4         | setosa  |
| -          | 4.6         | 3.4         | 1.4         | 0.3         | setosa  |
| -          | 5.0         | 3.4         | 1.5         | 0.2         | setosa  |
| -          | 4.4         | 2.9         | 1.4         | 0.2         | setosa  |
| -          | 4.9         | 3.1         | 1.5         | 0.1         | setosa  |
| -          | 5.4         | 3.7         | 1.5         | 0.2         | setosa  |
| -          | 4.8         | 3.4         | 1.6         | 0.2         | setosa  |
| -          | 4.8         | 3.0         | 1.4         | 0.1         | setosa  |
| -          | 4.3         | 3.0         | 1.1         | 0.1         | setosa  |

Previous **Next** Finish Cancel

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

**Data import wizard - Step 4 of 4**

This wizard guides you to import your data.  
**Step 4:** RapidMiner Studio uses strongly typed attributes. In this step, you can define the data types of your attributes. Furthermore, RapidMiner Studio assigns roles to the attributes, defining what they can be used for by the individual operators. These roles can be also defined here. Finally, you can rename attributes or deselect them entirely.

Reload data Guess value types Date format Enter value...

☒ Preview uses only first 100 rows.

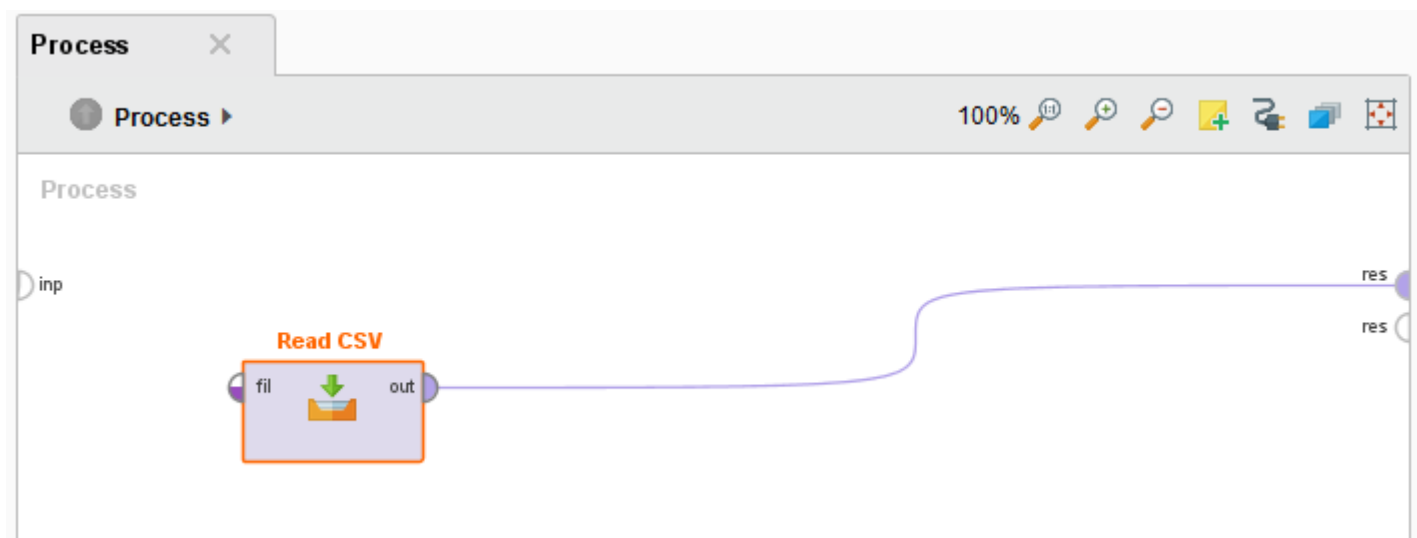
| sepal_length | sepal_width | petal_length | petal_width | species   |
|--------------|-------------|--------------|-------------|-----------|
| real         | real        | real         | real        | polyno... |
| attribute    | attribute   | attribute    | attribute   | attribute |
| 5.100        | 3.500       | 1.400        | 0.200       | setosa    |
| 4.900        | 3.000       | 1.400        | 0.200       | setosa    |

0 errors. ☒ Ignore errors ☐ Show only errors

| Row, Column | Error | Original value | Message |
|-------------|-------|----------------|---------|
|-------------|-------|----------------|---------|

Previous Next **Finish** Cancel

8. Realizamos la conexión:



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

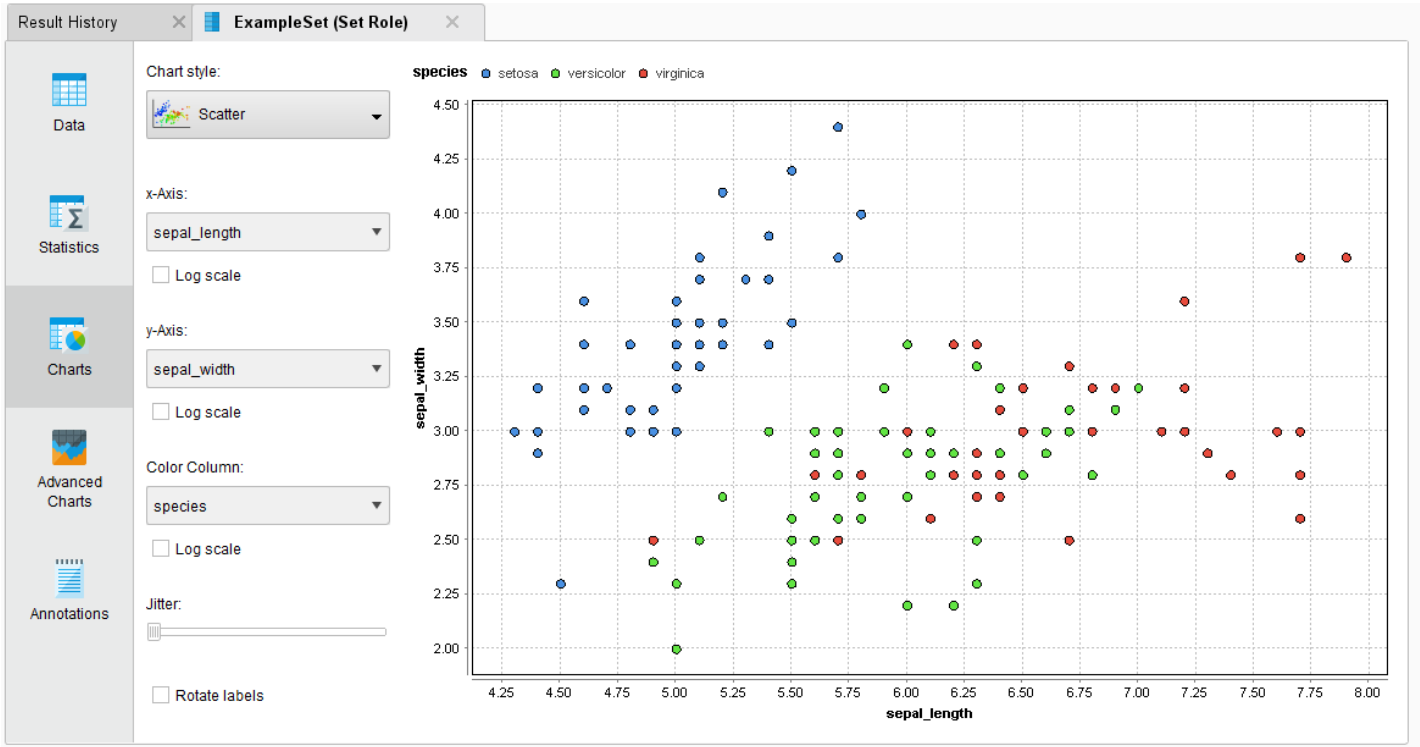
The screenshot shows the RapidMiner Studio Free 8.1.003 interface. The 'Result History' panel on the left is highlighted with a red box, showing options for Data, Statistics, Charts, Advanced Charts, and Annotations. The main window displays the 'ExampleSet (Read CSV)' results, which are 150 examples with 5 regular attributes. The data is presented in a table with columns: Row No., sepal\_length, sepal\_width, petal\_length, petal\_width, and species. The 'Repository' panel on the right shows the 'Add Data' button and a list of data sources: Samples, DB, Local Repository (HSerrano), and Cloud Repository (disconnect).

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

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).

|                 | Name         | Type       | Missing | Statistics           | Filter (5 / 5 attributes): |  |
|-----------------|--------------|------------|---------|----------------------|----------------------------|--|
| Data            | Label        |            |         |                      |                            |  |
|                 | species      | Polynomial | 0       | Least virginica (50) | Most setosa (50)           | Values setosa (50), versicolor (50), ... |
| Statistics      | 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                            |
| Charts          | petal_length | Real       | 0       | Min 1                | Max 6.900                  | Average 3.759                            |
| Advanced Charts | petal_width  | Real       | 0       | Min 0.100            | Max 2.500                  | Average 1.199                            |

Gráficos:





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

The screenshot shows the RapidMiner Studio interface. In the center, the 'Process' canvas displays a workflow with two operators: 'Read CSV' and 'Set Role'. The 'Set Role' operator is highlighted, and its parameters are shown in the right-hand 'Parameters' panel. The 'attribute name' is set to 'species' and the 'target role' is set to 'label'. The 'Repository' panel on the left shows the 'Samples' folder. The 'Operators' panel shows the 'Read CSV' operator selected. The 'Recommended Operators' panel at the bottom suggests 'Retrieve', 'Select Attributes', and 'Apply Model'.

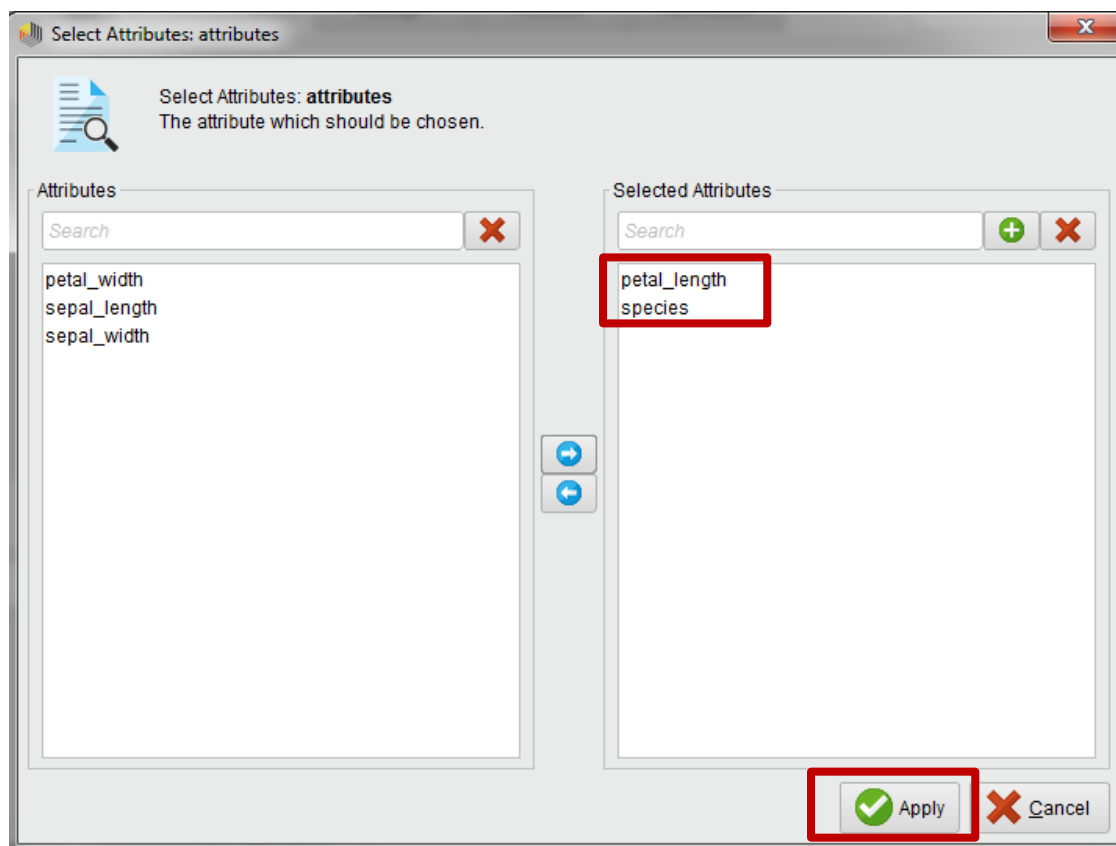
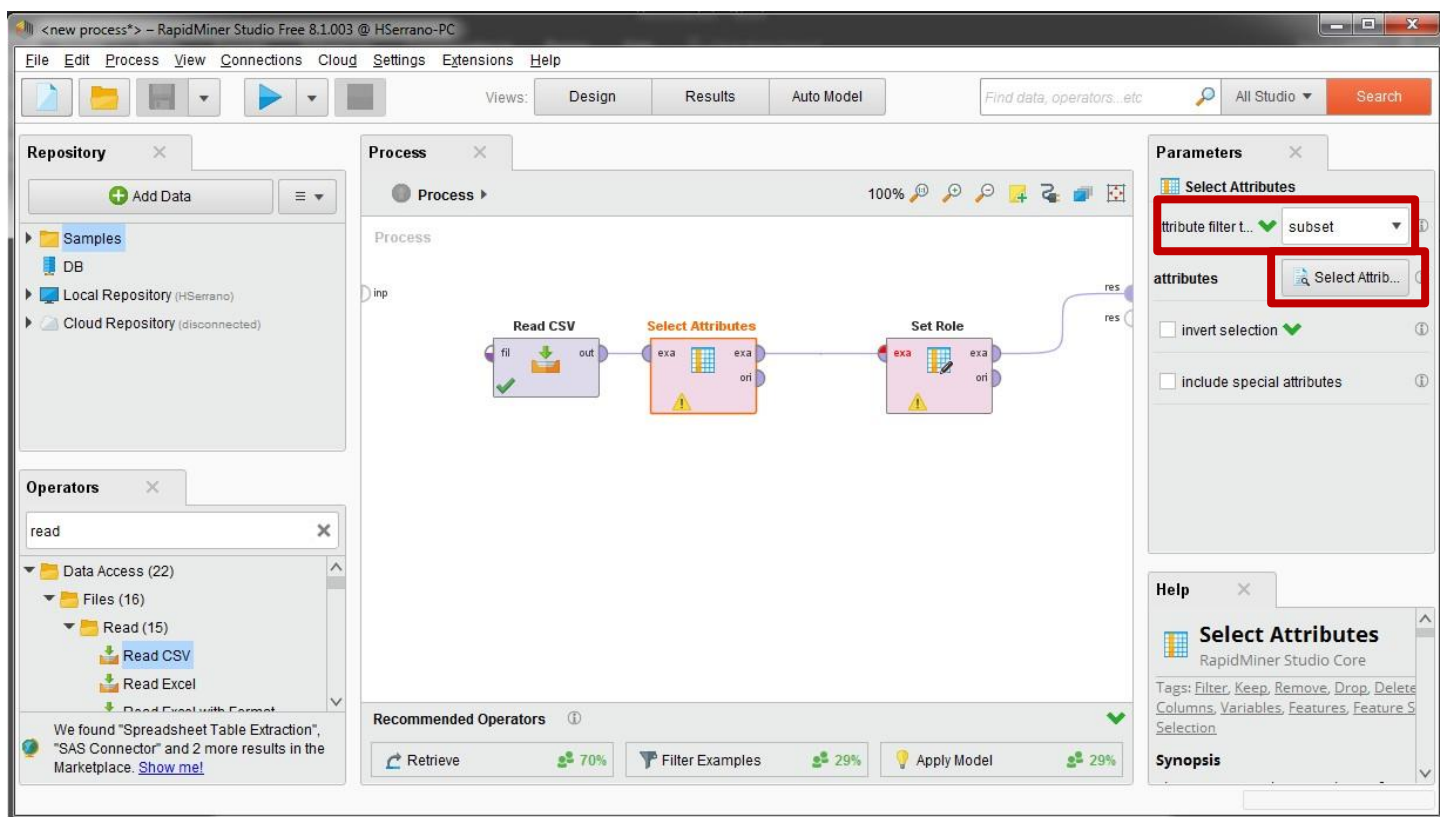
12. Al ejecutar se muestra lo siguiente:

The screenshot shows the 'Result History' panel in RapidMiner Studio. The 'ExampleSet (Set Role)' result is displayed, showing a table with 15 rows of data. The 'species' column is highlighted in green, indicating it is the target role. The table contains the following data:

| 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:

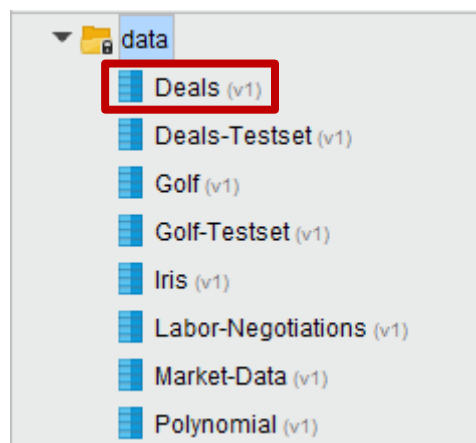
ExampleSet (150 examples, 1 special attribute, 1 regular attribute)

| 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        |

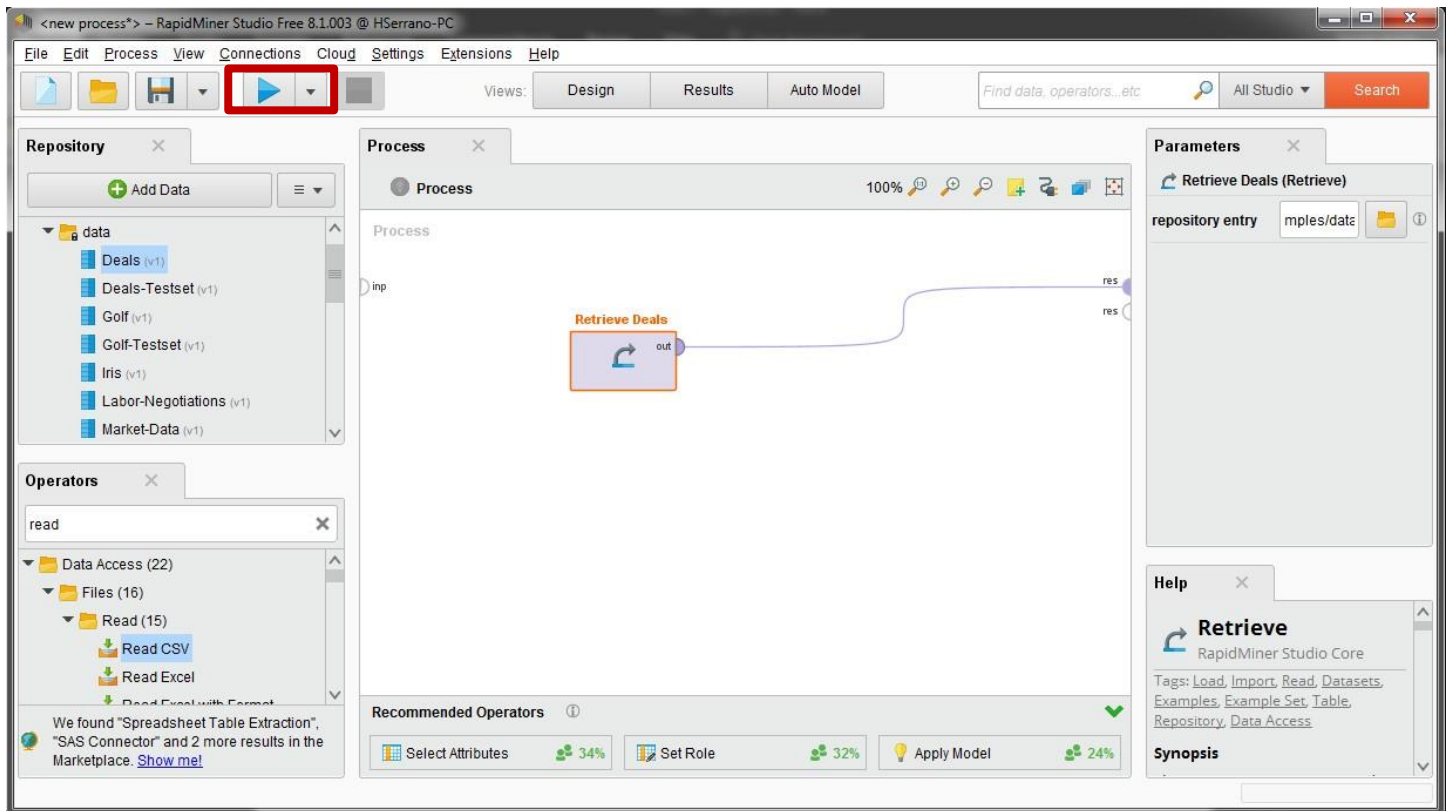
## Árbol de decisión en RapidMiner

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.

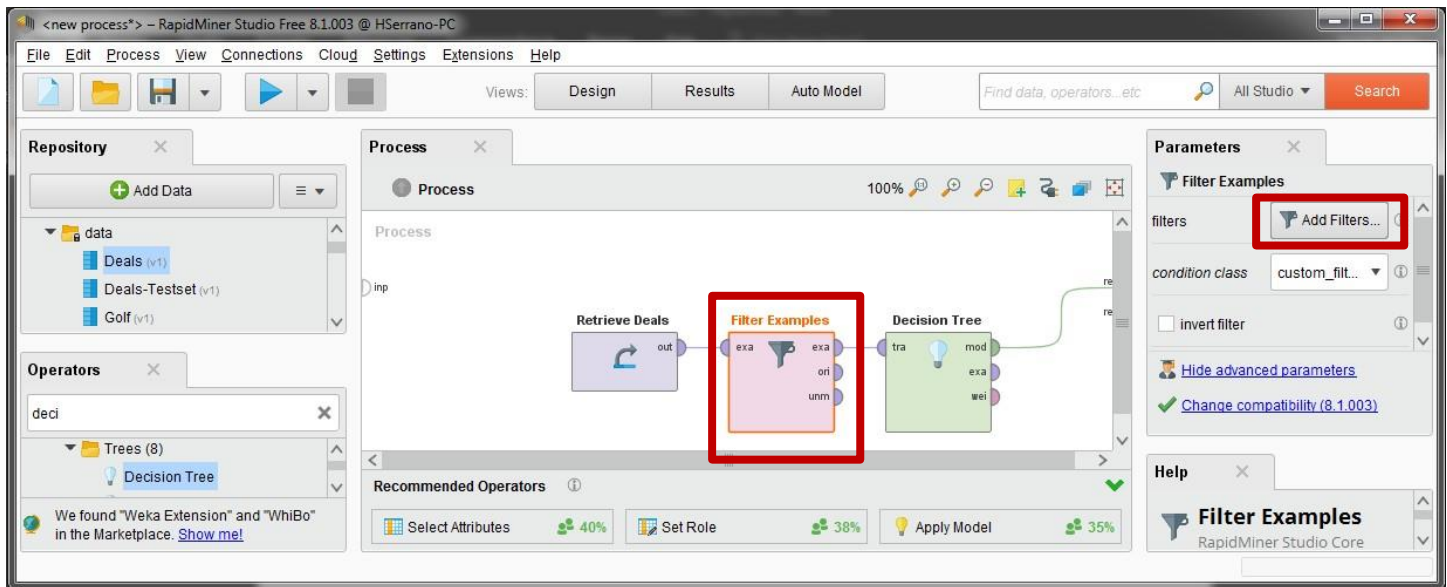


3. Al ejecutar le deberán aparecer los registros (en total son 1000 registros).

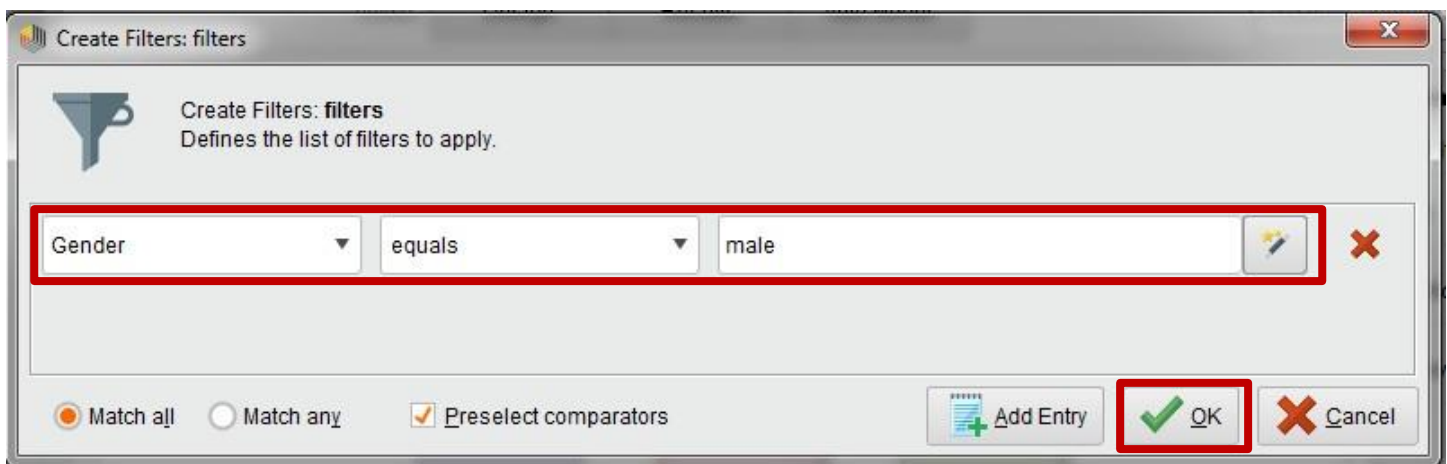
ExampleSet (1000 examples, 1 special attribute, 3 regular attributes)

| 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          |

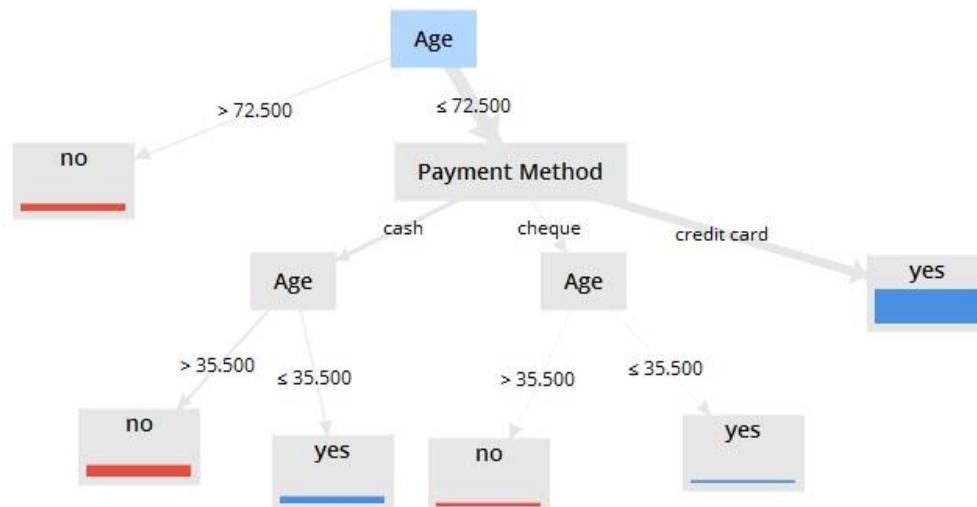
4. Ahora, conectamos el repositorio con un filtro para que solo tome a los del género masculino y posteriormente con el árbol de decisión. Hacemos clic sobre el filtro y luego en el botón "Add Filters" tal y como se muestra en la siguiente imagen:



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 árboles 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 |
|------------------------|-----------|------------|----------|----------|
| <i>id<sub>1</sub></i>  | Propiedad | Soltero    | 125000   | Conceder |
| <i>id<sub>2</sub></i>  | Alquiler  | Casado     | 100000   | Conceder |
| <i>id<sub>3</sub></i>  | Alquiler  | Soltero    | 70000    | Conceder |
| <i>id<sub>4</sub></i>  | Propiedad | Casado     | 12000    | Conceder |
| <i>id<sub>5</sub></i>  | Alquiler  | Divorciado | 95000    | Denegar  |
| <i>id<sub>6</sub></i>  | Alquiler  | Casado     | 60000    | Conceder |
| <i>id<sub>7</sub></i>  | Propiedad | Divorciado | 220000   | Conceder |
| <i>id<sub>8</sub></i>  | Alquiler  | Soltero    | 85000    | Denegar  |
| <i>id<sub>9</sub></i>  | Alquiler  | Casado     | 75000    | Conceder |
| <i>id<sub>10</sub></i> | 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.