

**PLATAFORMA DE DESCUBRIMIENTO DE CONOCIMIENTO EN BASE A
VISUALIZACIÓN DE INFORMACIÓN, MINERÍA DE DATOS Y REDUCCIÓN DE
DIMENSIÓN**

VisMineDR

Manual de Usuario

CONTENIDO

1.	Preliminares	4
1.1	Proceso KDD	4
1.2	Notas, avisos, precauciones y símbolos.	6
1.2.1	Tipos de archivos soportados y /o generados por la herramienta.	6
2.	ETAPA DE DATOS (SELECCIÓN DE LOS DATOS A ANALIZAR)	8
2.1	Texto plano (<i>Plaint text</i>)	8
2.2	Conexión a base de datos (Connection DB)	10
3.	ETAPA DE LIMPIEZA (Data Cleaning)	14
3.1	Remover datos vacíos (Remove Missing)	14
3.2	Cambiar datos vacíos (Udate Missing)	17
3.3	Rangos (Muestra)	18
3.4	Reduction (Reducción)	20
3.5	Replace Value (Reemplazar Valor)	23
3.6	Numeric Range (Rangos Numéricos)	25
3.7	Discretize (Categorización de variables numéricas)	27
3.8	Codification (Codificación)	28
3.9	Selection (Selección)	31
3.10	Standarize (Estandarizar)	33
4.	ETAPA DE MINERIA DE DATOS (DATA MINING)	36
4.1	Asociación (ASSOCIATION)	36
4.1.1	Apriori.	36
4.1.2	FP-Growth	36
4.1.3	EquipAsso	37
4.2	Clasificación (CLASIFICATION)	39
4.2.1	MATE	40
4.2.2	C4.5	40
4.2.3	SLIQ	41
4.3	Cluster	46
4.3.1	Algoritmo K-Means	47
4.3.2	BIRCH	47

1. Preliminares

La herramienta de minería de datos que se describe en el presente documento está basada en el Proceso KDD (Knowledge Discovery in Databases), por lo cual es importante puntualizar su definición.

1.1 Proceso KDD

KDD o descubrimiento de conocimiento en bases de datos [14], se lo ha definido como un proceso iterativo e interactivo y no trivial para la identificación en los datos de patrones validos, nuevos, potencialmente útiles y comprensibles para descubrir conocimiento [49].

Las tareas comunes en KDD son: la inducción de reglas, la clasificación, el clustering, el reconocimiento de patrones, el modelado predictivo, la detección de dependencias, entre otras.

En este proceso se contemplan distintas etapas [36], en primera instancia se establece la conexión al conjunto de datos, es posible realizarla de distintas formas, dependiendo del tipo de acoplamiento al conjunto de datos (Fuerte, débil y medianamente acoplados) [16] [41] y también del tipo de la fuente de datos, como archivos planos, conexión a gestores de bases de datos o repositorios remotos.

Una vez establecida la conexión se procede a seleccionar los datos objeto, que serán los elementos de estudio, en esta etapa se debe excluir los datos que no son relevantes o que no generarían conocimiento, es decir, atributos que no sean pertinentes de selección como aquellos que contengan valores particulares, de los cuales no se puedan establecer reglas generales, por ejemplo, si estamos observando un conjunto de datos de estudiantes en una universidad, con el propósito de analizar la deserción estudiantil, los datos a seleccionar de éste conjunto podrían ser: el programa al que pertenece, calificación promedio, estrato social, lugar de procedencia, entre otros, sin embargo no sería adecuado seleccionar los atributos de identificación o nombre, ya que sobre estos datos no se podrían generalizar resultados.

Sobre los datos objeto, es necesario aplicar un proceso de filtrado, pre-procesamiento o también llamado *data cleaning*, el cual es una limpieza de valores que permitan adecuar el conjunto a los requerimientos de los analistas, ya que dichos datos pueden estar contaminados o corruptos, por ejemplo, datos nulos o que nunca fueron capturados, datos que no correspondan a la realidad como la cesaría en hombres, datos atípicos o que están significativamente fuera de un rango, como por ejemplo que una persona tenga edad 700, además de los datos corruptos también es necesario adecuarlos a un esquema acorde al proceso de descubrimiento, en donde es necesario aplicar otros tipos de filtrados como la discretización, muestreo, codificación, generación de nuevos atributos a partir de otros, y existen muchos filtros más [31]. Con lo anterior obtenemos datos limpios, depurados y homogéneos los cuales son aptos para aplicarles la etapa siguiente que es precisamente el núcleo del proceso KDD [18] [34], y donde intervienen los algoritmos

inteligentes de aprendizaje automático y analítica visual de forma sinérgica, interactiva e iterativa, tema que se trata en secciones posteriores.

Después de aplicar el núcleo KDD, es posible establecer patrones los cuales describen eventos periódicos o tendencias predecibles, mediante expresiones que describen un subconjunto de los datos, su valor real reside en la información que podamos extraer de ellos para comprender los fenómenos que nos rodean, muchos autores han denominado a esto la pepita de oro [26], es decir, se ha extraído la información que estaba escondida en esas montañas gigantescas de datos, y que ahora se han convertido en información, palabra que proviene del verbo latino “informare” dar forma a una idea, por lo tanto, en esta etapa se ha tomado los datos y se les ha dado forma de patrón, sobre los cuales se pueden generar modelos que son una visión simplificada de la compleja realidad, dicho modelo puede ser simulado, visualizado y manipulado con la intención de generar hipótesis que son explicaciones sugeridas a un fenómeno observable, relaciona las posibles causas con los efectos.

Finalmente, el experto o analista ha generado nuevo conocimiento a partir de las experiencias y competencias adquiridas después de aplicar el proceso KDD. En la Figura 1 se esquematiza las etapas del proceso KDD [46] [5].

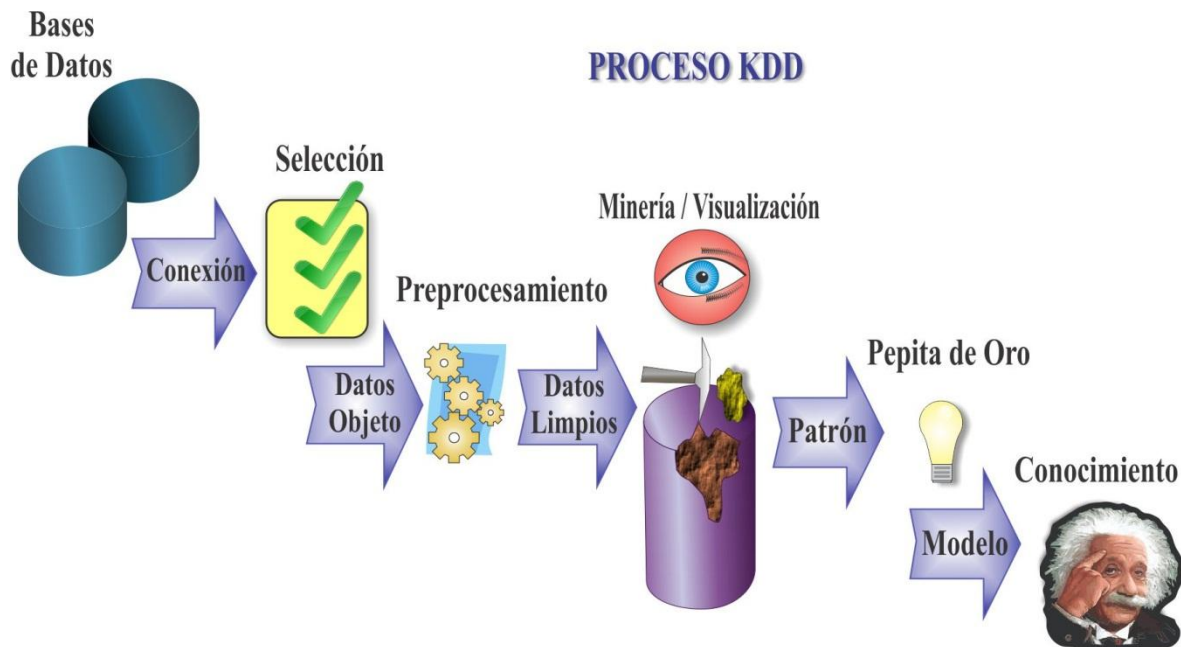


Figura 1. Etapas del proceso KDD

1.2 Notas, avisos, precauciones y símbolos.

1.2.1 Tipos de archivos soportados y /o generados por la herramienta.

Archivos de texto plano csv (comma separated value):

Los archivos de texto plano **CSV** son un tipo de datos en formato sencillo para representar información en forma de tablas, en donde los nombres de los atributos se ubican en la primera fila y las columnas son separadas por comas (,) o en algunos casos por punto y coma (;). Las filas son representadas por saltos de línea. El formato **CSV** es muy sencillo y no indica un tipo de datos concreto. Es importante que el número de columnas sea el mismo en cada fila, si no se dispone de un dato no hay que omitirlo, hay que dejar la separación del dato con la coma correspondiente lo que dará lugar a una doble coma por ej.: “,,”

Ejemplo:

```
ID,NOMBRE,EDAD,DIRECCION
987,juan,28,10 norte 342
876,pedro,42,8 oriente 342
123,jorge,22,av. libertad 23
69,vicente,61,valencia nº183
18,lorenzo,24,sol nº18
19,lucía,38,luna nº8
```

Archivos de texto tipo arff (attribute relation file format):

Los archivos de este tipo se han diseñado especialmente para el trabajo con minería de datos, presenta una definición más clara de la estructura de datos contenidos. Esta se basa en tres áreas. La primera, es el área de definición del encabezado que inicia por el indicador *@RELATION* seguida por el nombre de la relación que se quiere dar a la estructura de datos. Si el nombre contiene espacios se debe colocarlo entre comillas.

Ej.: @RELATION cliente ->Nombre de relación sin espacios

@RELATION “clientes tienda” ->Nombre de relación con espacios

El área No. 2 corresponde al segmento de definición de los atributos que tendrá el archivo de datos. Cada atributo contará con una línea para su definición. Para su construcción se debe colocar el indicador *@ATTRIBUTE* seguido de un espacio más el nombre del atributo que debe empezar por una letra (no se permiten atributos que comiencen con números), si el nombre del atributo contiene espacios se debe entrecomillar. Luego del nombre se debe colocar el tipo de datos que contendrá el atributo, para esto se tienen 3 tipos:

Numéric: numérico

String: Cadena de texto

Date: fecha. El tipo de fecha por defecto es `yyyy-MM-dd HH:mm:ss`

Nominal-specification: tipos de datos definidos por nosotros mismos, en general se refiere a categorías que expresamos de forma explícita.

La tercera sección es exclusiva para los datos propiamente dichos. Inicia por el indicador `@DATA` en una sola línea y los datos debajo de este indicador. Separaremos cada columna por comas y todas las filas deberán tener el mismo número de columnas, número que coincide con el de las declaraciones `@ATTRIBUTE` que añadimos en la sección anterior.

Si no disponemos de algún dato, colocaremos un signo de interrogación (?) en su lugar. El separador de decimales tiene que ser obligatoriamente el punto y las cadenas de tipo string tienen que estar entre comillas simples.

Por tanto, en un archivo de tipo `arff` tendremos las áreas `@relation`, `@attribute`, y `@data`.

Ej:

```
@RELATION clientes
```

```
@ATTRIBUTE nombre string
```

```
@ATTRIBUTE cedula numeric
```

```
@ATTRIBUTE "fecha de nacimiento" date "yyyy-MM-dd HH:mm:ss"
```

```
@ATTRIBUTE género {M, F}
```

```
@ATTRIBUTE "tipo de pago" {CREDITO, CONTADO, DONACION}
```

```
@ATTRIBUTE "valor de pago realizado" real
```

```
@DATA
```

```
Carlos, 1234545, "1959-01-22 12:12:45",M,CREDITO, 1435.55
```

```
María, 45366445, "1980-05-25 09:05:10",F,CONTADO, 2000
```

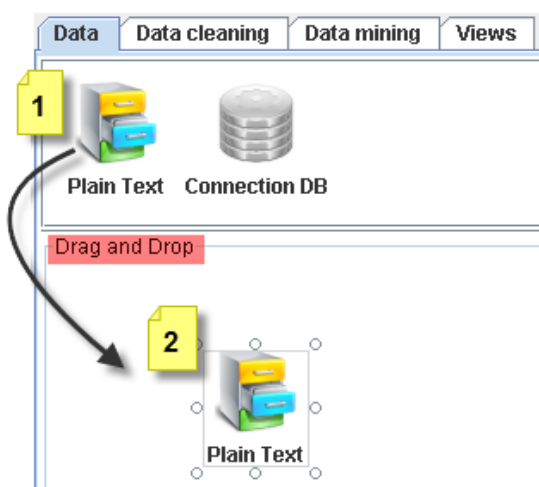
```
Pedro, 89789879, ?,M,DONACION,2750.3
```

```
Gloria,2342342, "1977-03-25 09:05:10",F,CONTADO,1987
```

```
Pablo,?,?,M,CREDITO,3950003.540
```

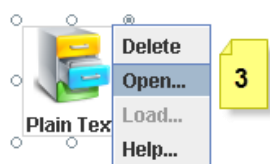
2. ETAPA DE DATOS (SELECCIÓN DE LOS DATOS A ANALIZAR)

2.1 Texto plano (*Plaint text*)

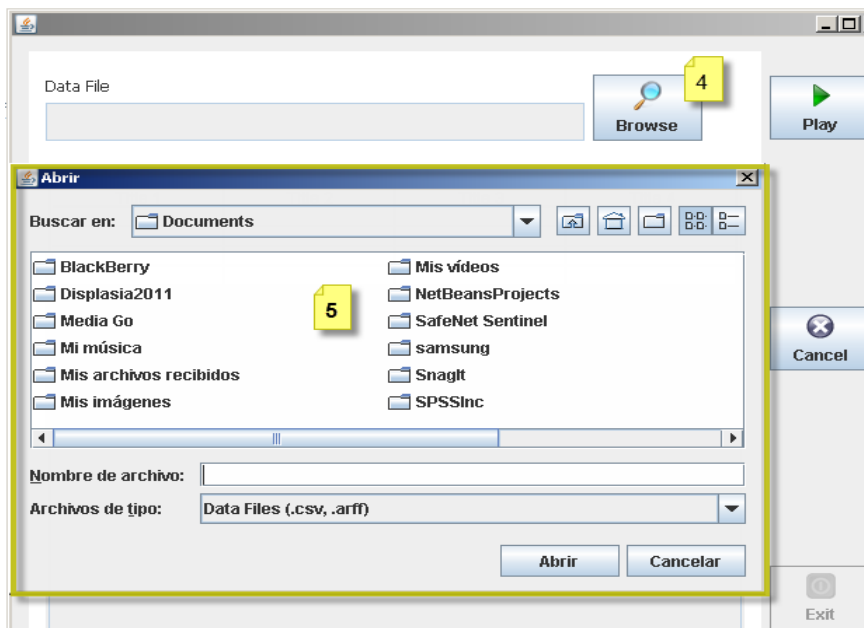



1. Desde la pestaña datos (**Data**) se debe tomar el ícono **Plain Text** haciendo clic sostenido sobre el ícono que representa los archivos de datos de texto plano que se cargará en la herramienta.

2. Una vez tomado se debe llevar al área de **Drag and Drop** con clic sostenido y soltándolo en ella.

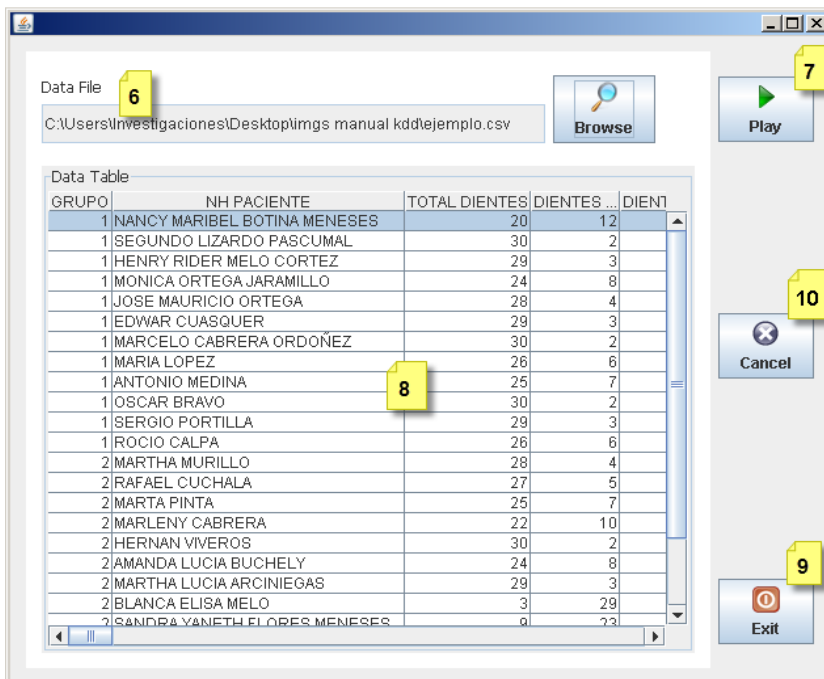


3. Una vez colocado el objeto, se debe dar clic derecho sobre él para desplegar las opciones que permitirá el objeto. Una vez realizado el procedimiento se desplegarán cuatro opciones, de las cuales debemos seleccionar **Open...** Con esta acción se abrirá una ventana para seleccionar el archivo de datos.




4. Se debe seleccionar el botón **Browse**  para abrir la ventana de búsqueda y selección de archivos.


5. En la ventana que se despliega se debe buscar el archivo de datos a cargar (sólo se permiten los tipos de archivo **.csv**: del inglés comma-separated values y **.arff**: attribute-relation file format), una vez encontrado se selecciona y se da clic al botón **Abrir**. Con esto queda seleccionado el archivo y se vuelve a la ventana anterior.




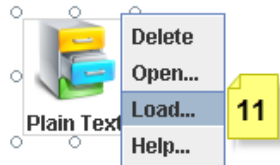
6. Cuando se realiza esta acción se vuelve a la ventana de selección de archivos y en la caja de texto **Data File** aparecerá la ruta del archivo seleccionado.

7. Para completar este procedimiento se debe dar clic en el botón **Play**  de la parte derecha de la ventana, con lo cual se confirma la carga y aceptación de datos.

8. Realizado lo anterior en la regilla central de la ventana aparece una vista preliminar de los datos.

9. Para salir se debe dar clic en el botón **Exit**. 

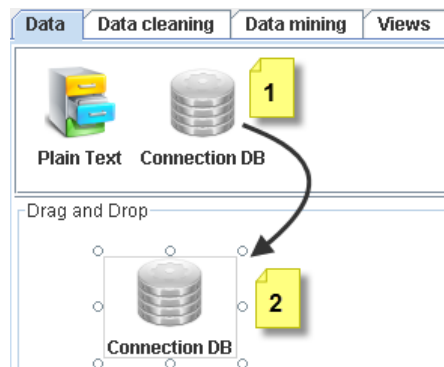
10. Si no se aceptan los datos existe el botón de **Cancel**  que cancelará la acción.



11. Una vez completados los pasos 7-9 debe ejecutarse la carga de datos, para esto se debe nuevamente dar clic derecho sobre el objeto **Plain text** en el área de **Drag and drop** y del menú que aparece seleccionar la opción **Load**, que ejecutará el proceso y nos informará mediante un mensaje en la barra de estado de la página el estado de la carga. Con esto ha finalizado el proceso de carga de datos por medio de archivos planos.

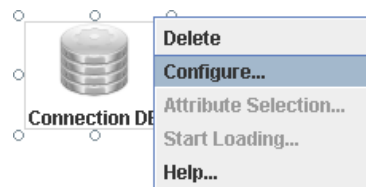
2.2 Conexión a base de datos (Connection DB)

Otra opción que se tiene para la carga de datos es mediante la conexión a base de datos mediante el comando **Connection DB** del área de datos.

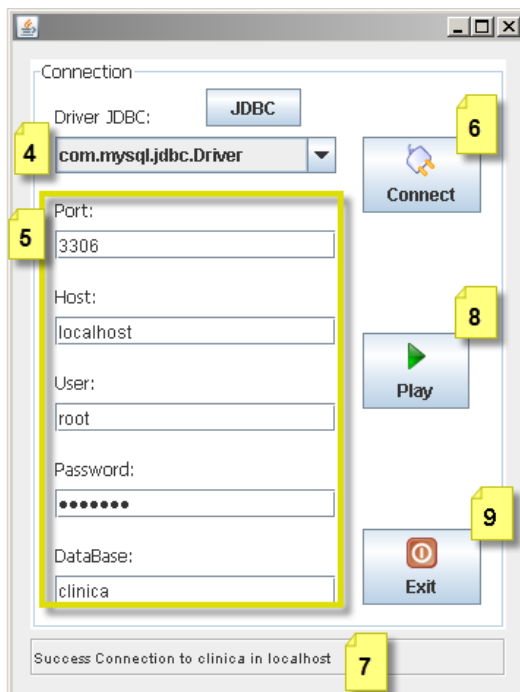


1. Para realizar este procedimiento se debe seleccionar el icono **Connection DB** de la pestaña datos (**Data**).

2. Con clic sostenido se lleva el objeto hasta el área de **Drag and Drop** soltándolo en ella.




3. Se continúa haciendo clic derecho sobre el ícono **Connection DB** en el área **Drag and Drop** para seleccionar la opción **Configure...** que desplegará la ventana para conexión a base de datos.




4. En esta ventana se debe seleccionar el tipo de motor de base de datos del cual se realizará la conexión. Existen tres tipos de motores soportados: PostgreSQL, MySQL y Oracle.

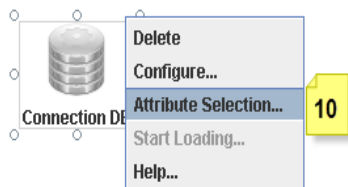
5. Seleccionado el motor se debe diligenciar los datos de conexión en los campos correspondientes: *Port* (puerto), *Host* (huésped), *User* (usuario de la base de datos), *Password* (contraseña) y *DataBase* (base de datos).

6. Una vez diligenciados los campos se debe dar clic al botón **Connect**. 

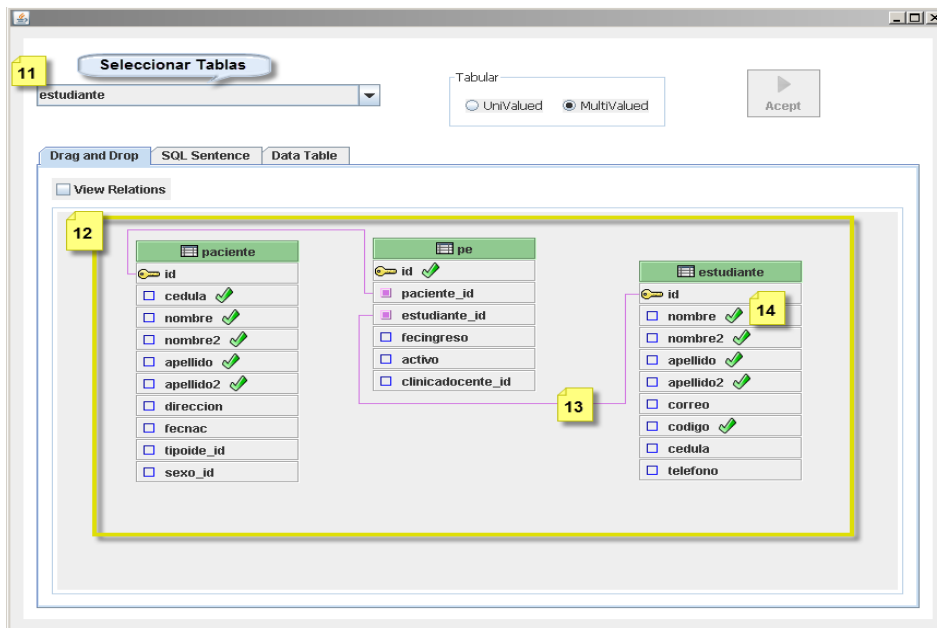
7. Mediante un mensaje en el pie de la ventana informará si la conexión fue satisfactoria o hubo algún inconveniente (tipos de inconvenientes ¿?).

8. Si la conexión fue exitosa (*Success*) se debe dar clic al botón  **Play**.

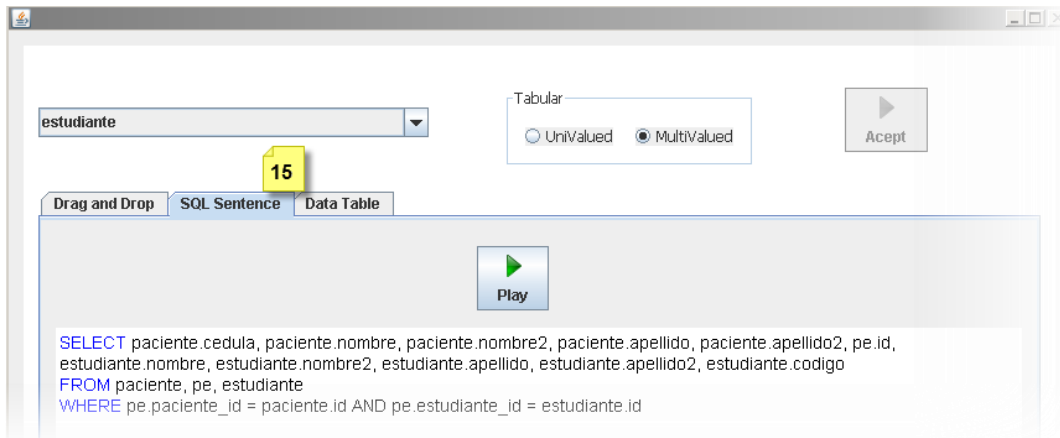
9. Posteriormente salir mediante el botón **Exit**  que devolverá el foco a la ventana principal.



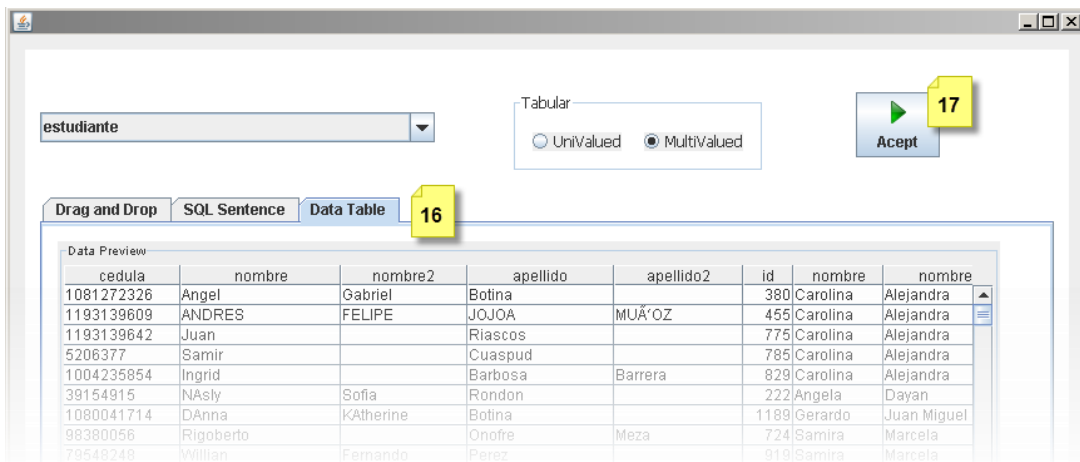
10. Posteriormente en el área **Drag and Drop** se debe nuevamente dar clic derecho en objeto **Connection DB** y seleccionar la opción **Attribute Selection**. A continuación se desplegará la ventana de selección de atributos.




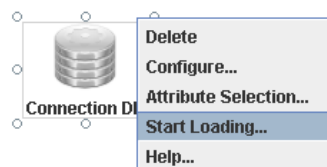
11. En la ventana de selección de atributos se cargarán las tablas de datos de la conexión realizada en los pasos anteriores. En primer lugar el usuario tiene la posibilidad de seleccionar con cuales tablas de la base de datos quiere trabajar, simplemente seleccionando las tablas desde la lista desplegable.
12. Las tablas que vaya seleccionando se irán acomodando en el área de color gris de la pestaña *Drag and Drop* detallando nombre de la tabla y sus respectivos campos.
13. En esta área el usuario tiene la posibilidad de establecer conexiones entre los diferentes campos de las tablas seleccionadas, simplemente haciendo clic sostenido en alguno de los atributos en el cuadro ☐ azul y llevándolo hasta algún atributo de otra tabla, con lo cual se formará una línea de color fucsia que identificará la relación e irá construyendo la consulta a la base de datos.
14. En esta misma área el usuario tiene la posibilidad de ir seleccionando los campos que harán parte de la consulta a la base de datos para el análisis simplemente dando clic sobre los nombres de los atributos de las tablas seleccionadas. Los atributos seleccionados tendrán al lado del nombre el ícono de *seleccionado*. Igual, si se equivoca haciendo clic nuevamente sobre el nombre, el atributo dejará de ser seleccionado.



15. Una vez el usuario ha seleccionado los atributos y establecido las relaciones según su criterio, ha de pasar a la siguiente pestaña, **SQL Sentence** en la cual se presentará el texto de la consulta a realizarse en la base de datos y donde el usuario simplemente tiene que ejecutar la consulta haciendo clic en el botón **Play**



16. Concluido el paso anterior el sistema mostrará al usuario los resultados de la consulta en una grilla denominada **Data Preview** en la pestaña **Data Table**.
17. A continuación el usuario debe dar clic en el botón aceptar  para cargar los datos seleccionados y volver a la ventana principal.



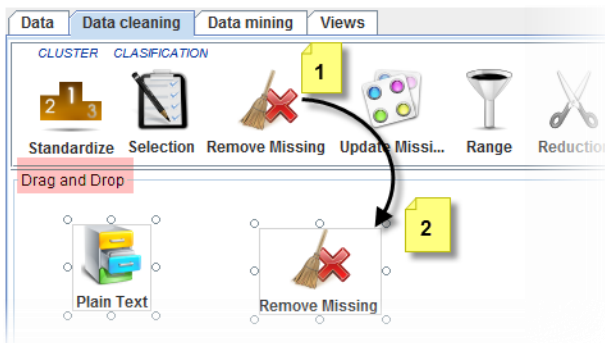
18. Finalmente, en el área de **Drag and Drop** de la ventana principal se debe dar nuevamente clic derecho sobre el icono **Connection DB** y seleccionar la opción **Start Loading...** con lo cual el sistema cargará la selección de datos de la base de datos conectada e indicará las instancias cargadas en el área de notificación de la barra de estado de la aplicación.

Load 25 instances.

3. ETAPA DE LIMPIEZA (Data Cleaning)

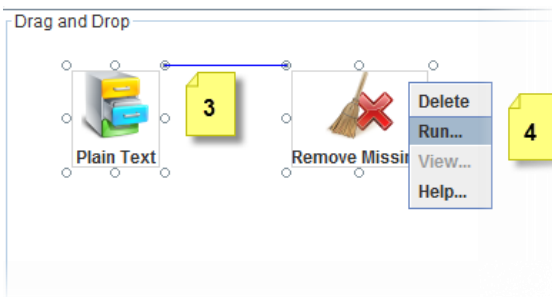
Para este apartado la entrada de datos puede ser un archivo de texto (*Plain text*), una base de datos (*Connection DB*) u otro filtro.

3.1 Remover datos vacíos (Remove Missing)



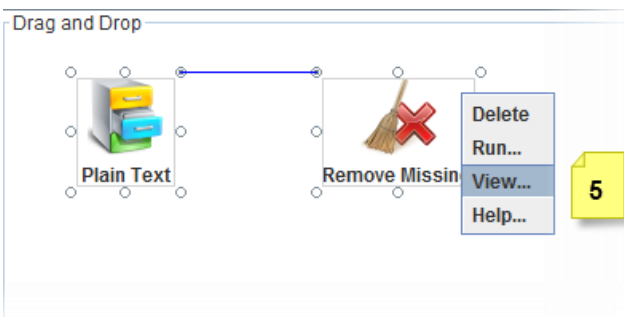
1. Desde la pestaña **Data cleaning** se debe tomar el ícono correspondiente a **Remove Missing** haciendo clic sostenido sobre él.

2. Una vez tomado se debe llevar al área **Drag and Drop** con clic sostenido y soltándolo en ella.



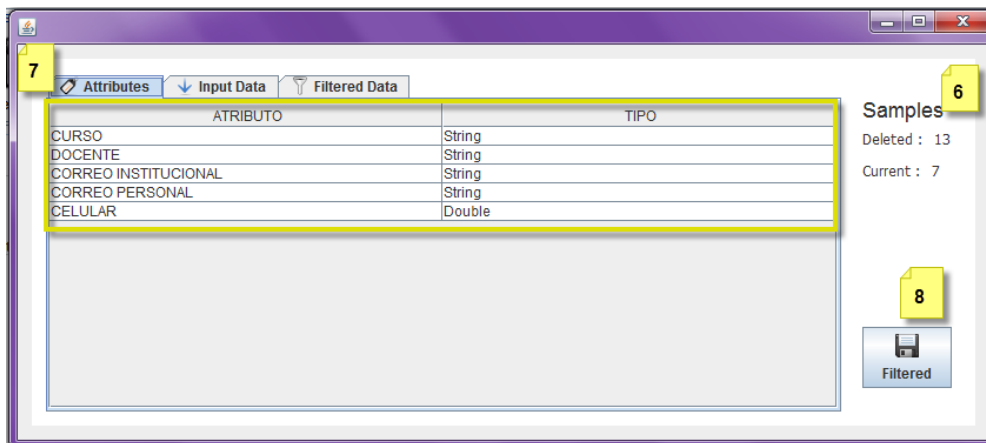
3. Una vez colocado el objeto, se debe establecer la conexión con el objeto de datos (*Plain text*, *Connection DB* u otro filtro), para esto se toma cualquiera de los puntos de conexión que rodean al objeto de datos con clic izquierdo sostenido se lo lleva al objeto **Remove Missing** en cualquiera de los puntos de conexión que rodea al objeto, una vez realizado esto se establecerá una línea de conexión entre ambos objetos, esto quiere decir que ya se encuentran enlazados los datos al objeto que realizará la limpieza.

4. Una vez realizado esto se debe dar clic derecho sobre el objeto **Remove Missing** y del menú de opciones que se despliegan dar clic sobre la opción **Run...** Con esta acción se cargará el proceso correspondiente para remover datos vacíos, el filtro estará ejecutado.



5. Una vez ejecutado el proceso se debe dar nuevamente clic derecho sobre el objeto **Remove Missing** y del menú de

opciones que se despliegan seleccionar la opción **View...** Realizado este proceso se cargará la ventana de filtrado de datos. El contenido de la ventana se describe a continuación.



6. En primer lugar la ventana en la parte derecha muestra información correspondiente a cuántos datos vacíos fueron filtrados (*Deleted*) y cuántos datos quedan después de aplicado el filtro (*Current*).
7. En la parte central se despliegan 3 pestañas, la primera de ellas **Attributes** muestra una tabla con dos columnas con la información de los nombres de los atributos o tipos de datos y su tipo correspondiente, dependiendo de lo que se encuentre en el archivo de datos, la herramienta identificará a qué tipo corresponde cada columna de datos, así:
 - a. Tipo de datos *String*: si contiene caracteres, datos vacíos, nulos o con la palabra NULL.
 - b. Tipo de datos *Double*: si contiene números con separador de decimales
 - c. Tipo de datos *Integer*: si contiene números enteros.
8. Así mismo en la ventana aparece el botón **Filtered** con el cual es posible guardar los datos filtrados, al hacer clic sobre él aparecerá la ventana para guardar los datos en la ubicación que se desee dentro del equipo en el cual se esté trabajando, es importante recordar que estos datos filtrados serán guardados en el formato de archivos **.csv** que es uno de los tipo de datos con los cuales trabaja la herramienta y que puede ver su descripción en el área de [Notas preliminares](#) al inicio de este manual.

9

CURSO	DOCENTE	CORREO INSTITUCI...	CORREO PERSONAL	CELULAR
BIOQUIMICA I	CLAUDIA GUEVARA	Claudia.Guevara@ca...	quimicasofi@hotmail...	3.013.935.735
BIOFISICA	ALVARO VILLOTA	Alvaro.Villota@camp...		3.013.626.114
BIOLOGIA MOLECUL...	CAROL CASTILLO		carol.castillop@gmail...	3.166.046.739
BIOLOGIA MOLECUL...	ARMANDO FOLLECO		wolffy115@hotmail.com	3.006.785.721
EMBRIOLOGIA	LORENA LIMA	Lorena.Lima@camp...	lalitalimamd@hotmail...	3.007.754.564
ANATOMIA I				
HISTOLOGIA	TERESITA FAJARDO	Teresita.Fajardo@ca...	teresita_fajardomedi...	3.128.660.105
BIOQUIMICA II	GABRIEL OBANDO		obandoga@gmail.com	3.017.542.762
ANATOMIA II	ALVARO HERNANDEZ	Alvaro.Hernandez@c...	alvarojhz@yahoo.com	3.117.492.537
MIGUEL DARIO MAR...	Miguel.Martinez@ca...	emedell1@hotmail.c...	3.006203157E9	
NEUROANATOMIA	JHON PABLO MEZA	John.Meza@campus...		3.108.491.518
FISIOLOGIA	JORGE RAMOS	Jorge.Ramos@camp...	jorcolos1@hotmail.c...	3.146.317.832
BIOQUIMICA III	MILENA GUERRERO		3.17705234E9	
PATOLOGIA	RONALD BASTIDAS		ronalgbg@yahoo.com	3.006.195.524
OSCAR MEJIA	Oscar.Mejia@campu...	oscarandresmejia@...	3.154763098E9	

Samples
Deleted : 13
Current : 7

Filtered

9. Otra de las pestañas que contiene la ventana es **Input Data** en la cual se muestra una tabla con la información de los datos de entrada al filtro (datos presentados en forma original), en la cual podemos darnos cuenta de aquellas casillas que se encuentren vacías y que posteriormente serán filtradas por la herramienta.

10

CURSO	DOCENTE	CORREO INSTITUCI...	CORREO PERSONAL	CELULAR
BIOQUIMICA I	CLAUDIA GUEVARA	Claudia.Guevara@ca...	quimicasofi@hotmail...	3.013.935.735
EMBRIOLOGIA	LORENA LIMA	Lorena.Lima@campu...	lalitalimamd@hotmail...	3.007.754.564
HISTOLOGIA	TERESITA FAJARDO	Teresita.Fajardo@ca...	teresita_fajardomedi...	3.128.660.105
ANATOMIA II	ALVARO HERNANDEZ	Alvaro.Hernandez@ca...	alvarojhz@yahoo.com	3.117.492.537
FISIOLOGIA	JORGE RAMOS	Jorge.Ramos@camp...	jorcolos1@hotmail.com	3.146.317.832
MICROBIOLOGIA Y LA...	MONICA GUERRERO	monica.guerrero@ca...	moncall2004@gmail...	3.122.583.494
FARMACOLOGIA	NORBERTO LOPEZ	Norberto.Lopez@cam...	norberto-lo-m@hotma...	3.127.760.676

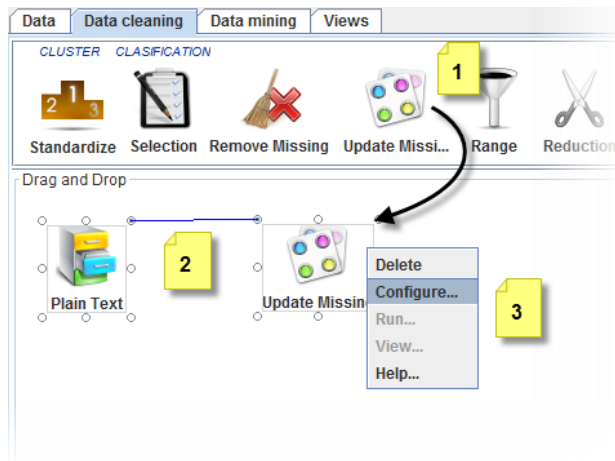
Samples
Deleted : 13
Current : 7

Filtered

10. La última pestaña de esta ventana es **Filtered Data**, en la cual en una tabla se nos muestra la información ya aplicado el filtro de remover los datos vacíos. Una vez ejecutado el filtro este se convierte en un nuevo flujo de entrada que puede ser utilizado para otros procedimientos.

NOTA: Cada flujo conserva su estado (Datos de entrada y salida)

3.2 Cambiar datos vacíos (Update Missing)

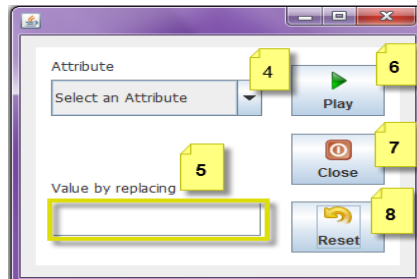


1. Desde la pestaña **Data cleaning** se debe seleccionar el ícono correspondiente a la opción **Update Missing** y con clic sostenido llevarlo al área de **Drag and Drop**.

2. Una vez realizado el paso anterior se debe establecer la conexión con el objeto contenedor de los datos (*Plain text*, *Connection DB* u otro filtro). Para esto se debe tomar con clic sostenido cualquiera de los 8 puntos que se encuentran alrededor del objeto de datos y llevarlo a cualquiera de

los puntos de conexión del objeto **Update Missing**.

3. Ahora se debe dar clic derecho sobre el objeto **Update Missing** y del menú de opciones que se despliega seleccionar la opción **Configure...** Con la cual se desplegará una ventana de opciones.



4. En esta ventana se debe seleccionar el atributo o columna de datos que contiene valores vacíos.

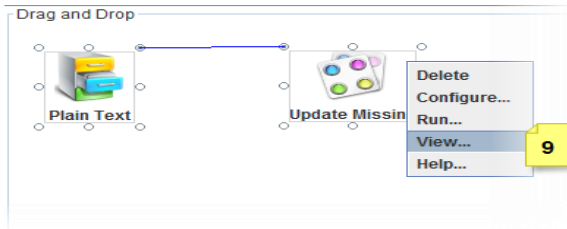
5. Luego en la caja de texto se colocará el valor que reemplazará al dato vacío dentro del atributo.

6. Posteriormente se ejecutará el reemplazó mediante el botón **Play**.

Nota: Los pasos 4 a 6 se pueden repetir con los atributos en que se desea reemplazar los datos vacíos.

7. Una vez realizado el procedimiento se da clic en el botón **Close** para cerrar la ventana.

8. Existe la posibilidad de revertir el proceso si se ha cometido alguna equivocación dando clic en el botón **Reset**.



9. Luego se debe nuevamente dar clic derecho sobre el objeto **Update Missing** y del menú que se despliega seleccionar la opción **View...** la cual desplegará una ventana con la información correspondiente al filtro.

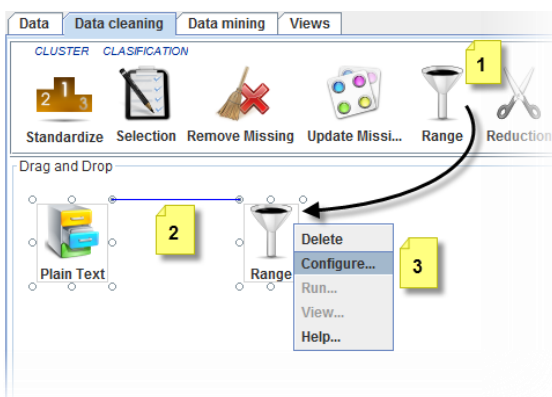
CURSO	DOCENTE	CORREO INSTITUCIONAL	CORREO PERSONAL	CELULAR
BIOQUIMICA I	CLAUDIA GUEVARA	Claudia.Guevara@ca...	micasoft@hotmail...	3.013.935.735
BIOFISICA	ALVARO VILLOTA	Alvaro.Villota@campu...		3.013.626.114
BIOLOGIA MOLECUL...	CAROL CASTILLO	n/a	l.castillop@gmail...	3.166.046.739
BIOLOGIA MOLECUL...	ARMANDO FOLLECO	n/a	wolff115@hotmail.com	3.006.785.721
EMBRIOLOGIA	LORENA LIMA	Lorena.Lima@camp...	talitalimamd@hotmail...	3.007.754.564
ANATOMIA I		n/a		
HISTOLOGIA	TERESITA FAJARDO	Teresita.Fajardo@ca...	teresita_fajardomedic...	3.128.660.105
BIOQUIMICA II	GABRIEL OBANDO	n/a	obandoga@gmail.com	3.017.542.762
ANATOMIA II	ALVARO HERNANDEZ	Alvaro.Hernandez@c...	alvarojhz@yahoo.com	3.117.492.537
MIGUEL DARIO MAR...	Miguel.Martinez@ca...	emedell1@hotmail.c...	3.006203157E9	
NEUROANATOMIA	JHON PABLO MEZA	John.Meza@campus...		3.108.491.518
FISIOLOGIA	JORGE RAMOS	Jorge.Ramos@camp...	jorcolos1@hotmail.co...	3.146.317.832
BIOQUIMICA III	MILENA GUERRERO	n/a	3.17705234E9	
PATOLOGIA	RONALD BASTIDAS	n/a	ronalgbg@yahoo.com	3.006.195.524
OSCAR MEJIA	Oscar.Mejia@campu...	oscarandresmejia@...	3.154763098E9	
MICROBIOLOGIA Y L...	MONICA GUERRERO	monica.guerrero@ca...	moncali2004@gmail...	3.122.583.494
DORIS MARTINEZ	doris.martinez@cam...	dmartinezjurado@ya...	3.155689221E9	
BIOQUIMICA IV	JAIME NARVAEZ	n/a	jaimenarvaez@gmail...	3.006.102.083
INMUNOGENETICA	IVAN HERNANDEZ	Ivan.Hernandez@ca...		3.104.981.031
FARMACOLOGIA	NORBERTO LOPEZ	Norberto.Lopez@ca...	norberto-lo-m@hotm...	3.127.760.676

10. En esta ventana es importante observar la pestaña **Filtered Data** que es en donde se puede observar cómo fueron reemplazados los datos vacíos con los que se colocaron en las opciones de filtrado. Para el ejemplo los datos vacíos fueron reemplazados con el texto "n/a".

11. Este filtro también puede ser guardado mediante el botón **Filtered**.

3.3 Rangos (Muestra)

Esta opción permite partir los datos con el propósito de hacer un conjunto control y un conjunto experimental.

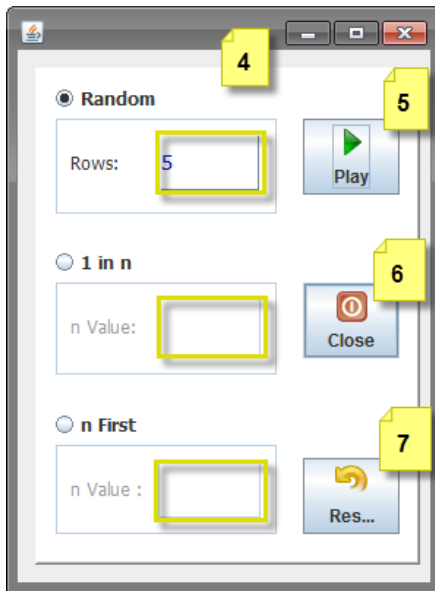


1. Desde la pestaña **Data cleaning** se debe seleccionar el ícono correspondiente a la opción **Range** y con clic sostenido llevarlo al área de **Draga and Drop**.

2. Una vez realizado el paso anterior se debe establecer la conexión con el objeto contenedor de los datos (*Plain text, Connection DB* u otro filtro). Para esto se debe tomar con clic

sostenido cualquiera de los 8 puntos que se encuentran alrededor del objeto de datos y llevarlo a cualquiera de los puntos de conexión del objeto **Range**.

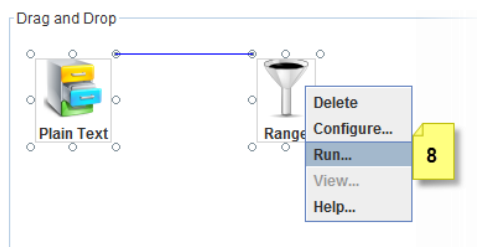
- Ahora se debe dar clic derecho sobre el objeto **Range** y del menú de opciones que se despliega seleccionar la opción **Configure...** Con la cual se desplegará la ventana de opciones que se describe a continuación.



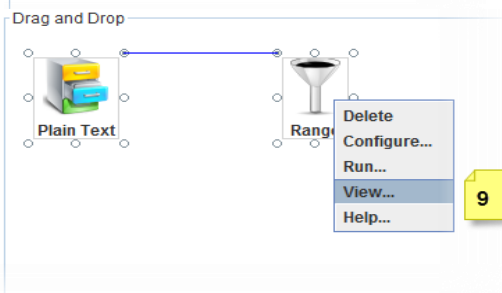
4. En esta ventana se tiene 3 opciones para seleccionar datos, en primer lugar la opción **Random** permite seleccionar cierto número de filas aleatoriamente ingresando el número deseado en el apartado **Rows**; la segunda opción permite seleccionar filas en saltos de 1 en n partiendo desde la fila 1, para esto se debe digitar el valor del salto en el apartado **n Value**; y en tercer lugar está la opción de seleccionar las n primeras filas, en donde sólo debemos digitar el valor de las filas a seleccionar en el apartado **n Value**.

5. Una vez realizado el paso anterior en cualquiera de las opciones correspondientes, se dará clic al botón **Play** para ejecutar el procedimiento.

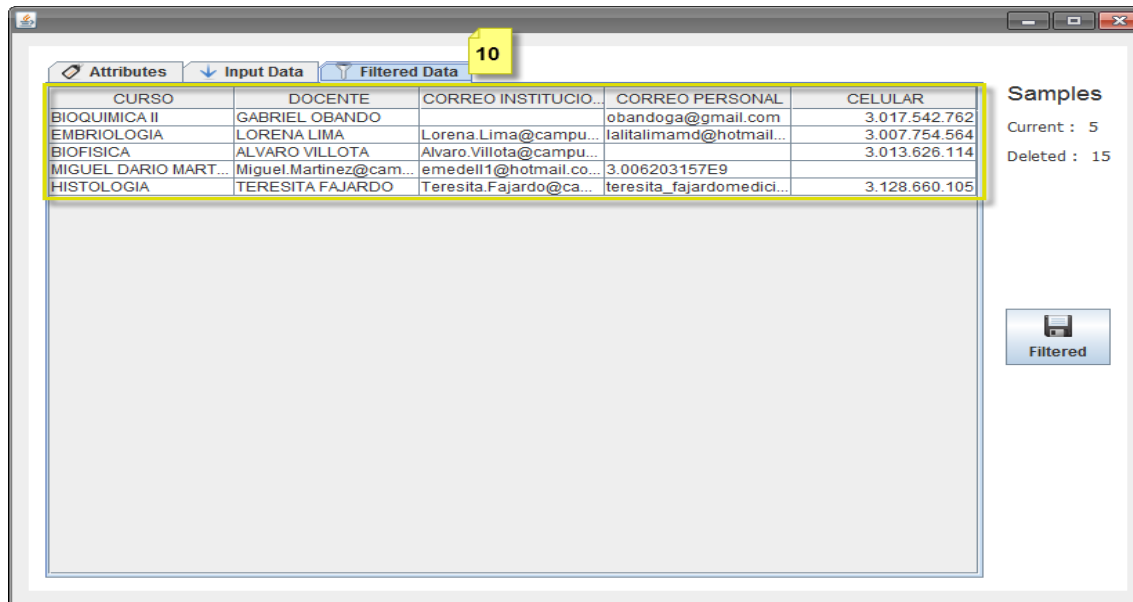
- Una vez ejecutado el procedimiento podremos salir de la configuración a través del botón **Close**.
- Si se desea restablecer algún campo de la ventana de configuración, simplemente se debe dar clic en el botón **Reset**.



8. Una vez configurado, se debe dar clic derecho en el ícono correspondiente a **Range** de la ventana **Drag and Drop** y de las opciones que se despliegan se debe seleccionar **Run...** con lo cual ejecutaremos el procedimiento que acabamos de configurar.



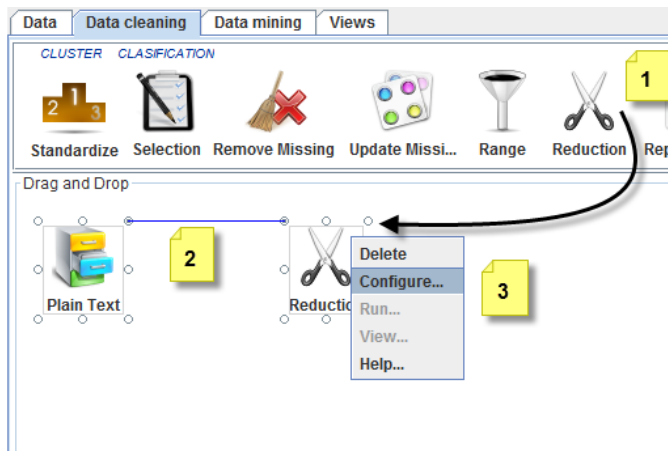
9. Luego se debe nuevamente dar clic derecho sobre el objeto **Range** y del menú que se despliega seleccionar la opción **View...** la cual desplegará una ventana con la información correspondiente al filtro.



10. En esta ventana es importante observar la pestaña **Filtered Data** que es donde se puede observar cómo fueron seleccionados los datos de acuerdo a la configuración dada al filtro.
11. Este filtro también puede ser guardado mediante el botón **Filtered**.

3.4 Reduction (Reducción)

La reducción tiene como propósito seleccionar o eliminar un conjunto de datos de acuerdo a ciertas opciones.

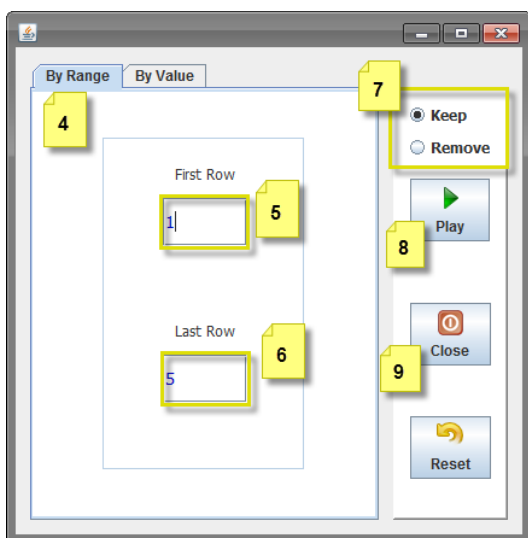


1. Desde la pestaña **Data cleaning** se debe seleccionar el ícono correspondiente a la opción **Reduction** y con clic sostenido llevarlo al área **Drag and Drop**.

2. Una vez realizado el paso anterior se debe establecer la conexión con el objeto contenedor de los datos (*Plain text*, *Connection DB* u otro filtro). Para esto se debe tomar con clic sostenido cualquiera

de los 8 puntos que se encuentran alrededor del objeto de datos y llevarlo a cualquiera de los puntos de conexión del objeto **Reduction**.

3. Ahora se debe dar clic derecho sobre el objeto **Reduction** y del menú de opciones que se despliega seleccionar la opción **Configure...** Con la cual se desplegará una ventana de opciones.



4. El filtro **Reduction** tiene la posibilidad de realizar la reducción de datos por dos métodos: por rangos (pestaña **By Range**) y por valores (pestaña **By Value**).

5. En la pestaña **By Range** se debe digitar la fila inicial en el apartado **First Row**.

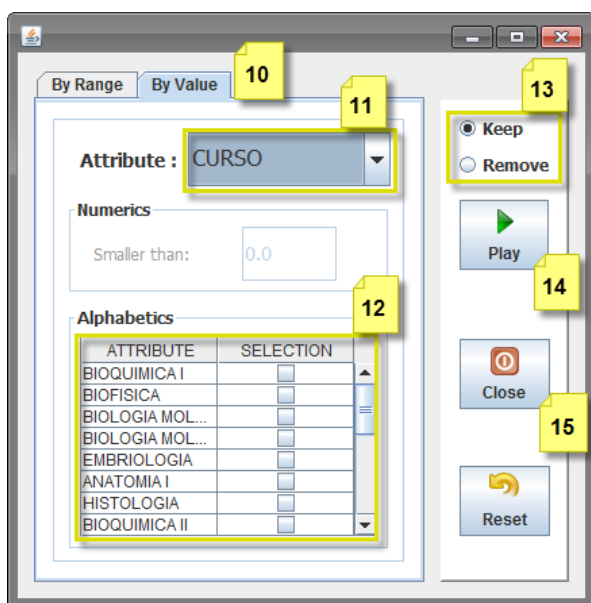
6. En la misma pestaña en el apartado **Last Row** se debe digitar el valor de la última fila a seleccionar.

7. Para el filtrado se tiene dos opciones: Si lo que se desea es utilizar el filtro para la conservación o selección de datos se debe seleccionar **Keep** de este apartado; en cambio si lo que se desea es utilizar el

filtro para la remoción de datos se debe seleccionar **Remove**.

8. Para ejecutar el procedimiento se debe dar clic en el botón **Play**.

9. Para salir del procedimiento se debe dar clic en el botón **Close**.



10. La otra opción que permite este filtro es seleccionar datos por valor desde la pestaña **By Value**.

11. En esta pestaña se debe en primer lugar seleccionar del apartado **Attribute** la variable a filtrar del conjunto de datos.

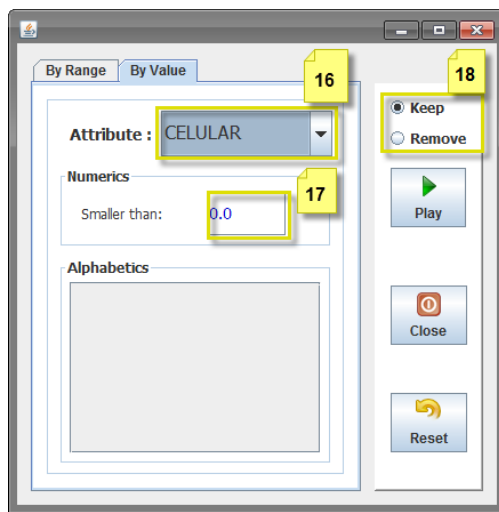
12. Si el tipo de datos del atributo seleccionado es de tipo carácter o cadena (**String**) se activará el apartado **Alphabetic** discriminando en una tabla los diferentes datos categorizados del atributo seleccionado, en este apartado al frente de cada dato se encuentra una caja de selección que se debe activar si el dato será

seleccionado para realizar el filtro.

13. Una vez seleccionados los datos del apartado anterior sólo queda decidir si los datos seleccionados serán conservados (seleccionar **Keep**) o eliminados (seleccionar **Remove**).

14. Una vez efectuada la selección para ejecutar el procedimiento se debe dar clic en el botón **Play**.

15. Para salir de la ventana se debe dar clic en el botón **Close**.

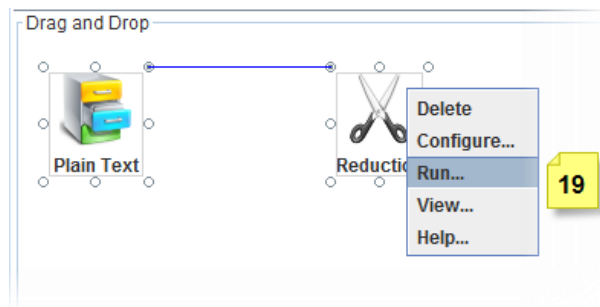


16. En cambio si el tipo de datos seleccionado en **Attribute** es de tipo numérico,

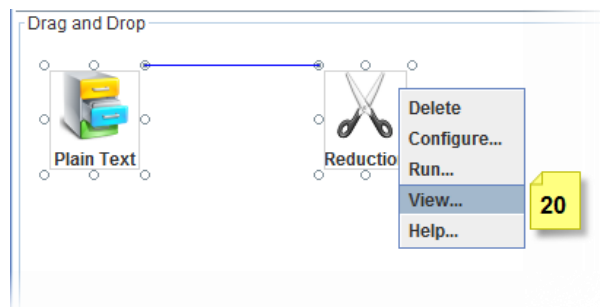
17. Se activará el apartado **Numerics** en donde se debe digitar el valor numérico de tope máximo para la selección de datos de acuerdo al atributo.

18. Luego se debe seleccionar qué se desea realizar con los datos (**Keep o Remove**).

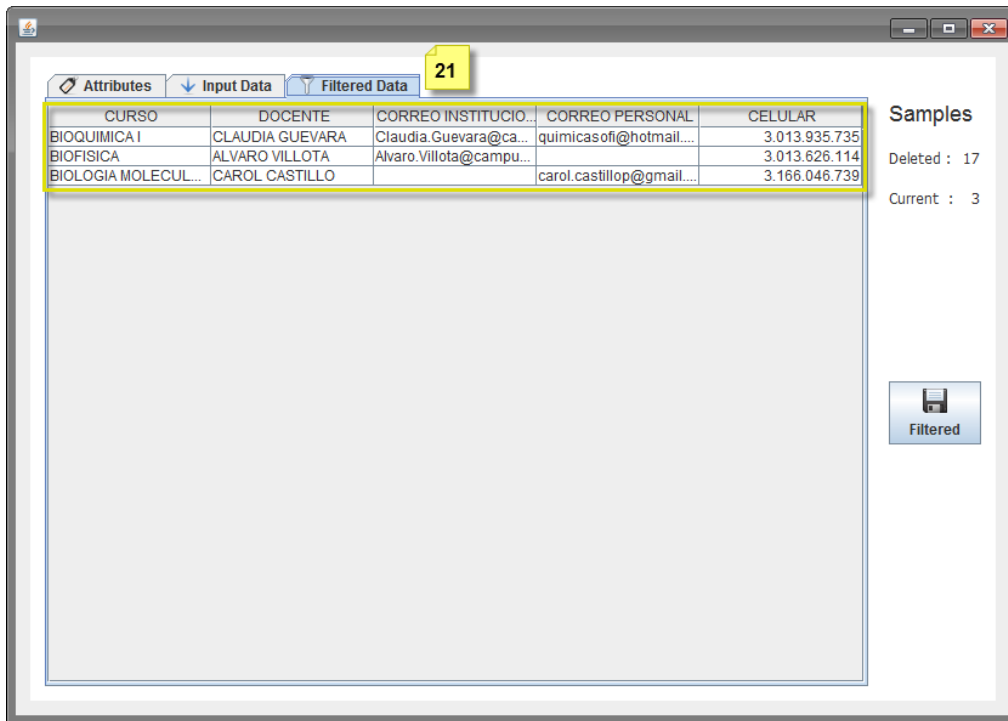
Realizado los pasos anteriores se debe seleccionar el botón **Play** para ejecutar el procedimiento y para salir de la ventana el botón **Close**.



19. Nuevamente en el área **Drag and Drop** se debe dar clic derecho al objeto **Reduction** y seleccionar **Run...** del menú de opciones para ejecutar el procedimiento.



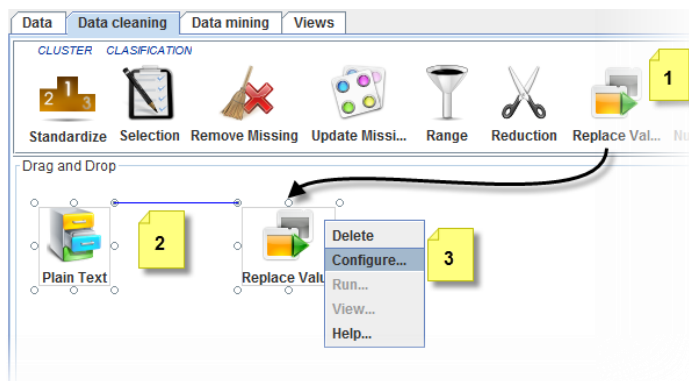
20. Posteriormente se debe dar clic derecho y seleccionar **View...** para observar la ejecución del filtro.



21. Realizado el anterior paso se abrirá la ventana de resultados del filtro en donde observaremos la pestaña correspondiente a **Filtered Data** que presentará los datos que fueron seleccionados de acuerdo a la configuración del filtro. En el lado derecho también observaremos un resumen de los datos seleccionados (filtrados: *Current*) y los datos eliminados (no seleccionados: *Deleted*).

3.5 Replace Value (Reemplazar Valor)

Este filtro aplica para atributos de tipo categórico nada más, es decir aquellos atributos del conjunto de datos que contengan información de tipo cadena (*String*).

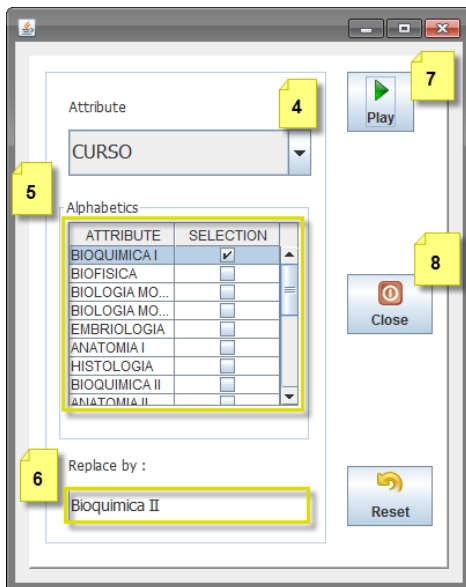


1. De la pestaña **Data cleaning** se selecciona el ícono correspondiente a **Replace Value** y con clic sostenido se lo lleva al área **Drag and Drop**.

2. Una vez realizado el paso anterior se debe establecer la conexión con el objeto contenedor de los datos (*Plain text*, *Connection DB* u otro filtro). Para esto se debe tomar con clic sostenido

cualquiera de los 8 puntos que se encuentran alrededor del objeto de datos y llevarlo a cualquiera de los puntos de conexión del objeto **Replace Value**.

- Ahora se debe dar clic derecho sobre el objeto **Replace Value** y del menú de opciones que se despliega seleccionar la opción **Configure...** Con la cual se desplegará una ventana de opciones

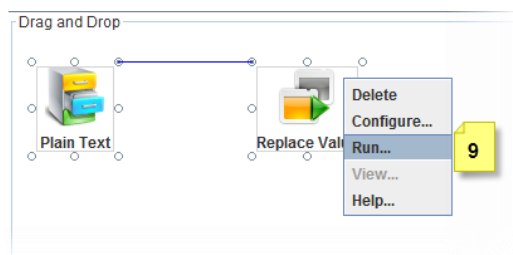


4. En la ventana de opciones se debe seleccionar el atributo que será filtrado (sólo se seleccionarán atributos de tipo cadena).

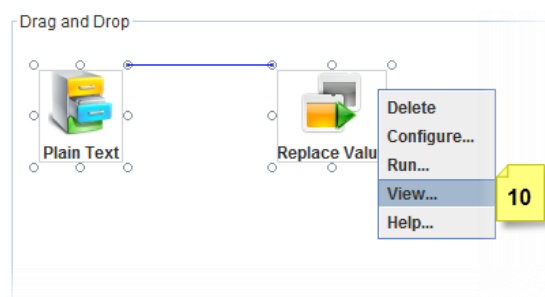
5. Una vez seleccionado el atributo en el área **Alphabetic** aparecerán las categorías de datos del atributo seleccionado, donde se debe seleccionar las categorías de datos que serán filtradas, esto se hace marcando la casilla de verificación que se encuentra al frente de cada dato.

6. En el apartado **Replace by** se debe digitar el valor de cadena que remplazará a los atributos seleccionados en el apartado anterior.

- Una vez realizado el paso anterior queda simplemente dar clic en el botón **Play** para ejecutar el filtro de remplazo.
- Luego de ejecutado el paso anterior se debe dar clic en el botón **Close** para cerrar la ventana y volver al área de datos (**Drag and Drop**)



9. Después de realizado el paso anterior se debe dar clic derecho al objeto **Replace Value** y del menú que se despliega seleccionar la opción **Run...**



10. Nuevamente se debe hacer clic derecho sobre el objeto **Replace Value** y seleccionar la opción **View...** para ir a la ventana de resultados.

Attributes

Input Data

Filtered Data

CURSO	DOCENTE	CORREO INSTITUCI...	CORREO PERSONAL
BIOQUIMICA I	CLAUDIA GUEVARA	Claudia.Guevara@ca...	quimicasofi@hotmail...
BIOFISICA	ALVARO VILLOTA	Alvaro.Villota@campu...	
BIOLOGIA MOLECUL...	CAROL CASTILLO		carol.castillo@gmail...
BIOLOGIA MOLECUL...	ARMANDO FOLLECO		wolffy115@hotmail.com
EMBRIOLOGIA	LORENA LIMA	Lorena.Lima@campu...	lalitalimamd@hotmail...
ANATOMIA I			
HISTOLOGIA	TERESITA FAJARDO	Teresita.Fajardo@ca...	teresita_fajardomedi...

11

Attributes

Input Data

Filtered Data

CURSO	DOCENTE	CORREO INSTITUCI...	CORREO PERSONAL	CELULAR
Bioquimica II	CLAUDIA GUEVARA	Claudia.Guevara@ca...	quimicasofi@hotmail...	3.013.935.735
BIOFISICA	ALVARO VILLOTA	Alvaro.Villota@campu...		3.013.626.114
BIOLOGIA MOLECUL...	CAROL CASTILLO		carol.castillo@gmail...	3.166.046.739
BIOLOGIA MOLECUL...	ARMANDO FOLLECO		wolffy115@hotmail.com	3.006.785.721
EMBRIOLOGIA	LORENA LIMA	Lorena.Lima@campu...	lalitalimamd@hotmail...	3.007.754.564
ANATOMIA I				
HISTOLOGIA	TERESITA FAJARDO	Teresita.Fajardo@ca...	teresita_fajardomedi...	3.166.046.735

12

Samples

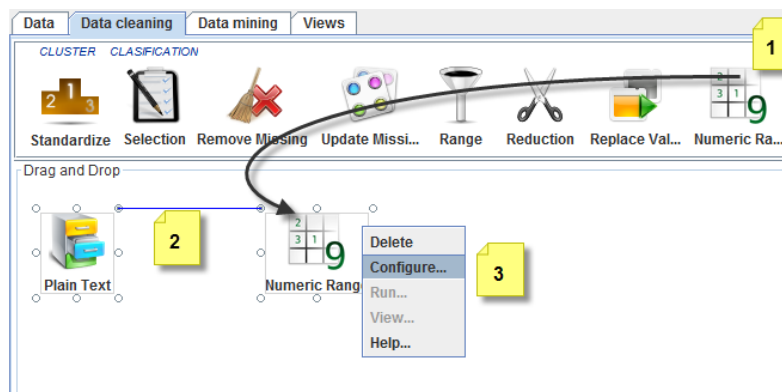
Currents : 20

Replaced : 1

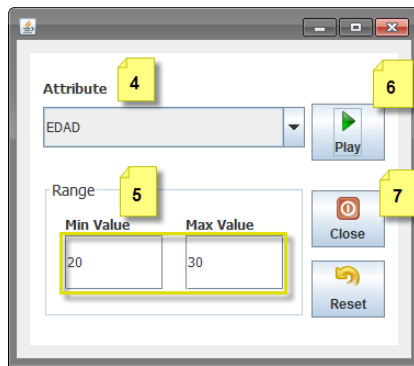
- En la ventana de resultados es preciso centrar la atención en las pestañas **Input Data** y **Filtered Data**. En la primera, para mirar los atributos que se seleccionaron para filtrar en su estado anterior y,
- En la ventana **Filtered Data** se observará cómo fue realizado el cambio por el filtro **Replace Value**.
- Al lado derecho de la ventana de resultados se puede observar bajo el título **Samples** el resumen de los valores filtrados (**Replaced**) y los que quedaron por fuera del filtro (**Currents**).

3.6 Numeric Range (Rangos Numéricos)

Este filtro permitirá el filtrado de datos por atributos de tipo numérico (*int*, *double*)



- De la pestaña **Data cleaning** se selecciona el ícono correspondiente a **Numeric Range** y con clic sostenido se lo lleva al área **Drag and Drop**.
- Una vez realizado el paso anterior se debe establecer la conexión con el objeto contenedor de los datos (*Plain text*, *Connection DB* u otro filtro). Para esto se debe tomar con clic sostenido cualquiera de los 8 puntos que se encuentran alrededor del objeto de datos y llevarlo a cualquiera de los puntos de conexión del objeto **Numeric Range**.
- Ahora se debe dar clic derecho sobre el objeto **Numeric Range** y del menú de opciones que se despliega seleccionar la opción **Configure...** Con la cual se desplegará una ventana de opciones

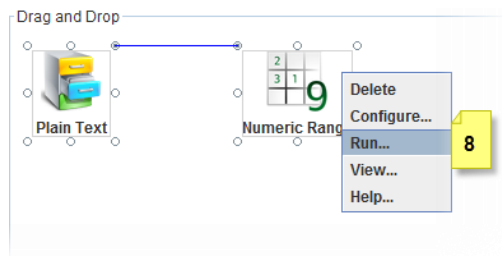


4. En la ventana se seleccionará el atributo de tipo numérico al cual se le aplicará el filtro.

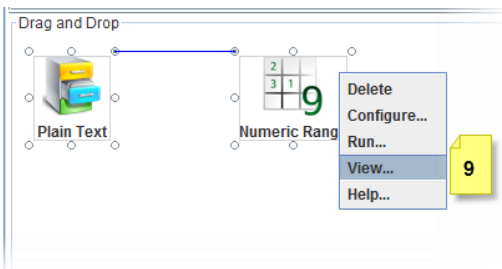
5. Posteriormente en el apartado **Range** se establecerá el valor mínimo (**Min Value**) y valor máximo (**Max Value**) para el rango.

6. Una vez realizado el paso anterior se ejecutará el procedimiento haciendo clic en el botón **Play**.

7. Para cerrar la ventana y volver a la principal se da clic en el botón **Close**.



8. Posteriormente se debe dar clic derecho al objeto **Numeric Range** y del menú de opciones que se presenta seleccionar **Run...** para ejecutar el filtro.



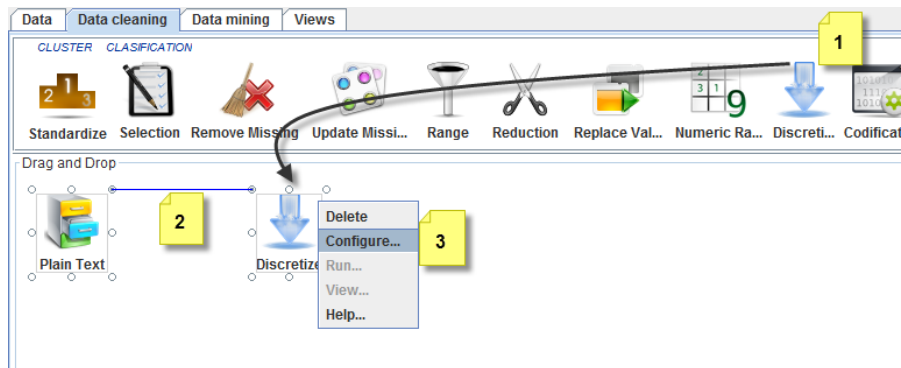
9. Luego se debe dar clic derecho sobre el objeto nuevamente para seleccionar la opción **View...** del menú que se despliega con el objetivo de mirar los resultados del filtro.

CURSO	DOCENTE	CORREO INSTIT.	CORREO PERS.	CELULAR	EDAD
BIOQUIMICA I	CLAUDIA GUEVA...	Claudia Guevara...	quimicasof@hot...	3.013.935.735	35
EMBRIOLOGIA	LORENA LIMA	Lorena Lima@ca...	lalitalimamd@ho...	3.007.754.564	34
QUIMICA	OSCAR MEJIA	Oscar Mejia@ca...	oscarandresmej...	3.154.763.098	38
MICROBIOLOGIA	MONICA GUERRR...	monica.guerrero...	moncal2004@g...	3.122.583.494	40
FISICA	DORIS MARTINEZ	doris.martinez@c...	dmartinezjurado...	3.155.689.221	37

10. En la ventana de resultados que se presenta se debe observar la pestaña **Filtered Data** la cual presentará los atributos numéricos que fueron trabajados con el filtro y de acuerdo a los valores mínimos y máximos que fueron colocados se observará el filtrado de los datos.

3.7 Discretize (Categorización de variables numéricas)

Este filtro presenta la posibilidad de realizar categorización de atributos de carácter numérico tanto valores continuos como valores discretos.

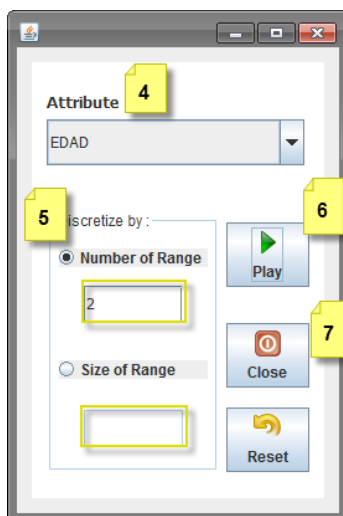


1. De la pestaña **Data cleaning** se selecciona el ícono correspondiente a **Discretize** y con clic sostenido se lo lleva al área **Drag and Drop**.

2. Una vez realizado el paso anterior se debe

establecer la conexión con el objeto contenedor de los datos (*Plain text*, *Connection DB* u otro filtro). Para esto se debe tomar con clic sostenido cualquiera de los 8 puntos que se encuentran alrededor del objeto de datos y llevarlo a cualquiera de los puntos de conexión del objeto **Discretize**.

3. Ahora se debe dar clic derecho sobre el objeto **Discretize** y del menú de opciones que se despliega seleccionar la opción **Configure...** Con la cual se desplegará una ventana de configuración.

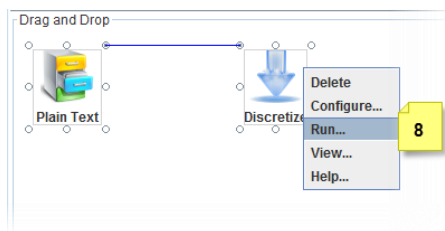


4. En la ventana de configuración se debe seleccionar el atributo de tipo numérico que será filtrado.

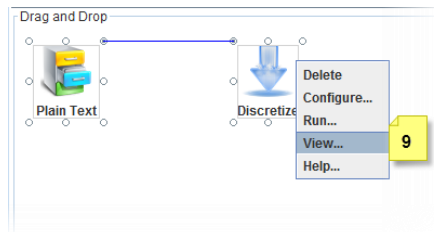
5. En el apartado **Discretize by** se tiene de dos opciones para realizar la categorización. En primer lugar por número de rangos (**Number of Range**) en donde se debe digitar el número de rangos que se desea obtener del filtrado. La segunda opción permite realizar rangos por tamaño de rango, para lo cual se tiene que digitar el número de datos que abarcará cada categoría como máximo (**Size of Range**).

6. Una vez realizado el paso anterior por cualquiera de las dos opciones se debe dar clic al botón **Play** para ejecutar el filtrado.

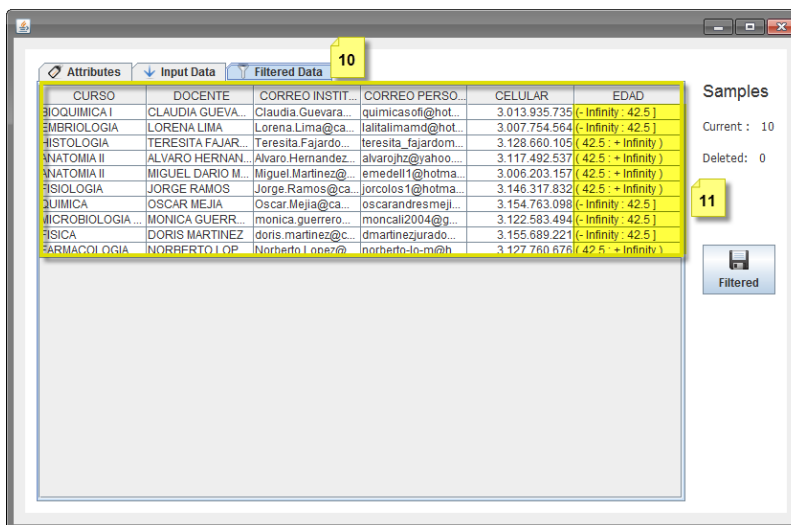
7. Luego se debe dar clic al botón **Close** para volver a la ventana principal.



8. En la ventana principal se debe dar clic derecho al objeto **Discretize** y seleccionar la opción **Run...** del menú que se despliega con el propósito de ejecutar la función del objeto.



9. Luego nuevamente con dando clic derecho sobre el objeto, se selecciona la opción **View...** del menú que se despliega con el objetivo de mirar los resultados de aplicar el filtro.

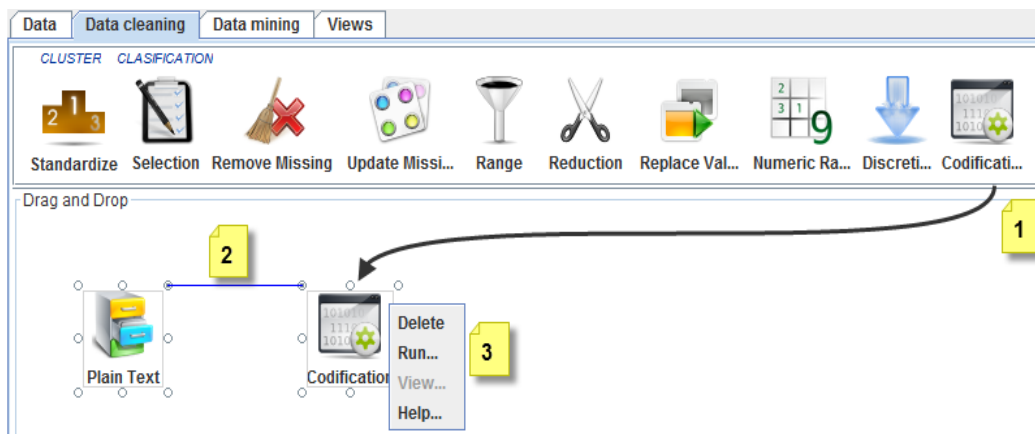


10. En la ventana de resultados se debe centrar la atención en el apartado **Filtered Data** en el cual se presenta los resultados de aplicación del filtro.

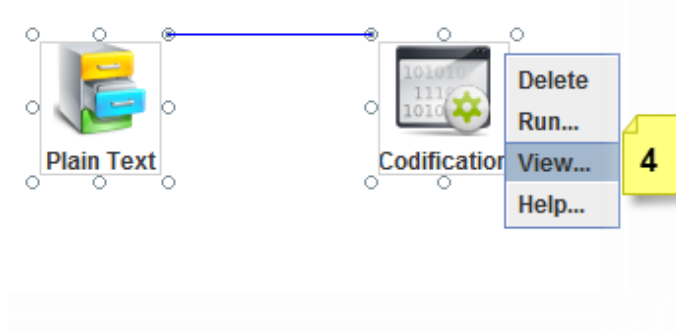
11. La variable a la cual se le aplicó el filtro mostrará los datos de acuerdo a la categorización obtenida por el filtro.

3.8 Codification (Codificación)

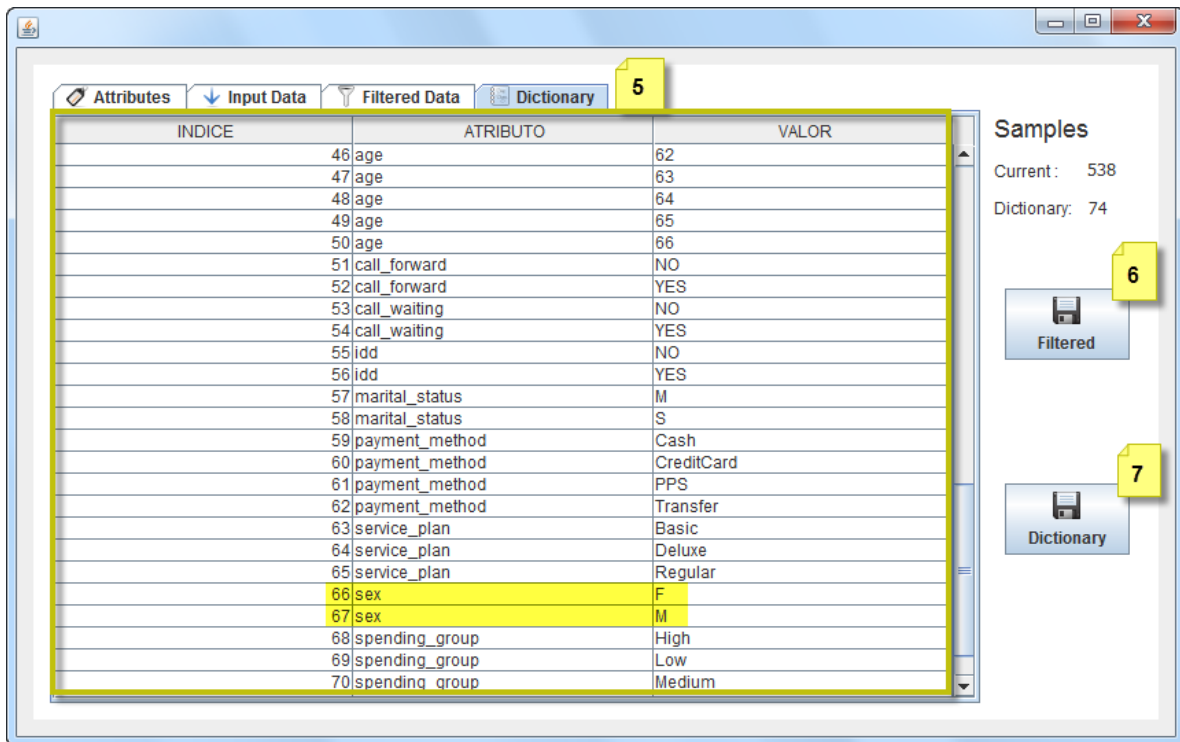
Este filtro es muy apto para trabajar con muchos datos. Cuando se aplica este filtro sobre un conjunto de datos no necesita configurarlo simplemente correrlo y mirar los resultados. En la ventana de resultados aparece una pestaña importante que se denomina **Dictionary** la cual presenta el diccionario de datos de la codificación.



1. De la pestaña **Data cleaning** se selecciona el ícono correspondiente a **Codification** y con clic sostenido se lo lleva al área **Drag and Drop**.
2. Una vez realizado el paso anterior se debe establecer la conexión con el objeto contenedor de los datos (*Plain text*, *Connection DB* u otro filtro). Para esto se debe tomar con clic sostenido cualquiera de los 8 puntos que se encuentran alrededor del objeto de datos y llevarlo a cualquiera de los puntos de conexión del objeto **Codification**.
3. Ahora se debe dar clic derecho sobre el objeto **Codification** y del menú de opciones que se despliega seleccionar la opción **Run...** con la cual se ejecutará el filtro. Una vez realizado este procedimiento aparecerá el mensaje: *Filter Codification Loaded* en la barra de estado de la aplicación.



4. Posteriormente, damos clic derecho sobre el objeto **Codification** y del menú de opciones que se despliega seleccionar la opción **View...** con la cual se desplegará la ventana de resultados que se describirá a continuación.



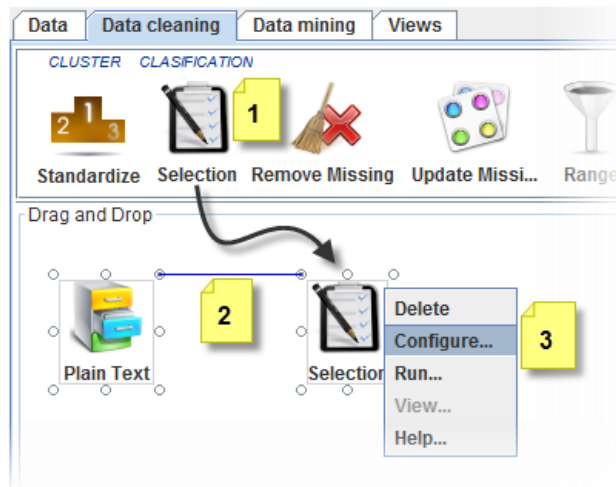
- En la ventana de resultados se debe prestar atención a la pestaña **Dictionary** pues ésta nos presenta el diccionario de datos de la codificación aplicada a los datos. Esta se encuentra distribuida en 3 columnas, la primera muestra el INDICE que es el código aplicado al dato, en la segunda se muestra el ATRIBUTO al cual pertenece ese dato y en la tercer columna se muestra el VALOR que es el dato que contiene el atributo y que fue codificado con el dato que aparece en el INDICE. Así, se puede observar en la imagen de ejemplo que para el atributo **sex** se asignó el código **66** para **F** y el código **67** para **M**.

Es importante anotar que el filtro **Codification** se aplica tomando los atributos en orden alfabético.

- En este filtro se presenta la posibilidad de guardar los datos filtrados (datos que están codificados) mediante el botón **Filtered**.
- En el mismo sentido, el filtro presenta la posibilidad de guardar el diccionario de datos, el cual contempla cómo están codificados los datos. Esto se realiza a través del botón **Dictionary**.

3.9 Selection (Selección)

En la mayor parte de los algoritmos de minería de datos es necesario utilizar este filtro pues permite focalizarse en los atributos útiles para ser incorporados al modelo. Reducir el número de columnas y atributos puede mejorar el rendimiento y la calidad del modelo.

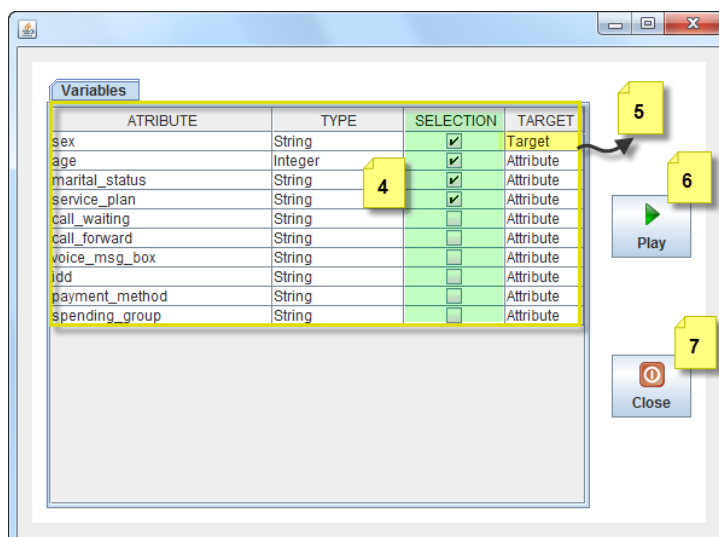


1. De la pestaña **Data cleaning** se selecciona el ícono correspondiente a **Selection** y con clic sostenido se lo lleva al área **Drag and Drop**.

2. Una vez realizado el paso anterior se debe establecer la conexión con el objeto contenedor de los datos (*Plain text*, *Connection DB* u otro filtro). Para esto se debe tomar con clic sostenido cualquiera de los 8 puntos que se encuentran alrededor

del objeto de datos y llevarlo a cualquiera de los puntos de conexión del objeto **Selection**.

3. Ahora se debe dar clic derecho sobre el objeto **Selection** y del menú de opciones que se despliega seleccionar la opción **Configure...** Con la cual se desplegará una ventana de opciones.

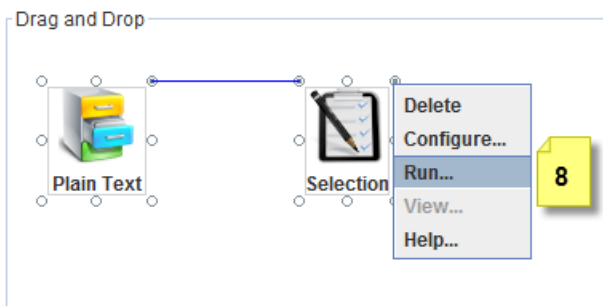


4. En la ventana de configuración se debe seleccionar los atributos que entraran a ser parte del modelo, esto se realiza en la columna **SELECTION** donde se debe activar o desactivar la casilla de verificación para cada atributo.

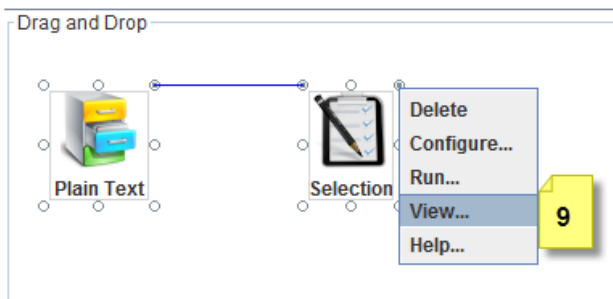
5. Si el objetivo de aplicar minería de datos es la **Clasificación** en la casilla **TARGET** se debe escoger el atributo objetivo (*Target*).

6. Para ejecutar el procedimiento se debe hacer clic sobre el botón **play**.

7. Finalmente se cierra la ventana dando clic en el botón **close**.



8. Una vez realizado el anterior procedimiento, en la ventana principal se debe dar clic derecho sobre el objeto **Selection** y seleccionar la opción **Run...** del menú que se despliega para que la selección configurada anteriormente sea ejecutada.



9. Luego nuevamente con dando clic derecho sobre el objeto, se selecciona la opción **View...** del menú que se despliega con el objetivo de mirar los resultados de aplicar el filtro.

age	marital_status	service_plan	sex
61	M	Basic	M
27	M	Regular	M
29	M	Basic	F
49	M	Basic	M
29	S	Basic	M
60	S	Basic	M
53	S	Regular	F
63	S	Basic	F
46	M	Regular	M
25	S	Basic	F
40	M	Regular	F
29	S	Basic	M
49	S	Basic	F
41	M	Regular	M
47	S	Regular	F
43	M	Basic	F
63	M	Basic	M
37	M	Regular	F
32	S	Basic	M
48	M	Basic	M
57	M	Basic	F
36	M	Basic	M
55	S	Basic	F
50	S	Basic	F
48	S	Basic	F
33	S	Regular	F

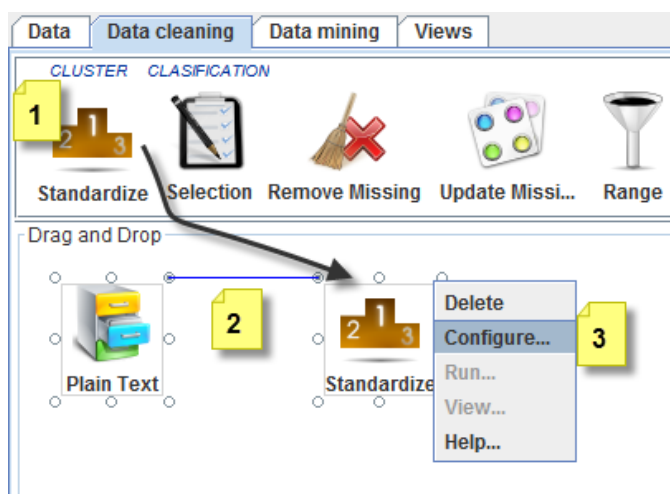
10. En la ventana de resultados se observará en la pestaña **Filtered Data** los atributos seleccionados con los datos correspondientes.

11. El apartado **Samples** muestra el resumen del filtro informando el número de atributos borrados y el número de atributos seleccionados.

12. Este filtro nos da la oportunidad de guardar los resultados del filtro. Para ello simplemente se debe dar clic sobre el botón **Filtered** y luego seleccionar el lugar donde guardaremos el filtro.

3.10 Standarize (Estandarizar)

Este filtro está realizado para trabajara sólo con atributos de tipo numérico (*integer o double*) para realizar correcciones a la información que contengan con el propósito de evitar los valores extremos y/o atípicos.

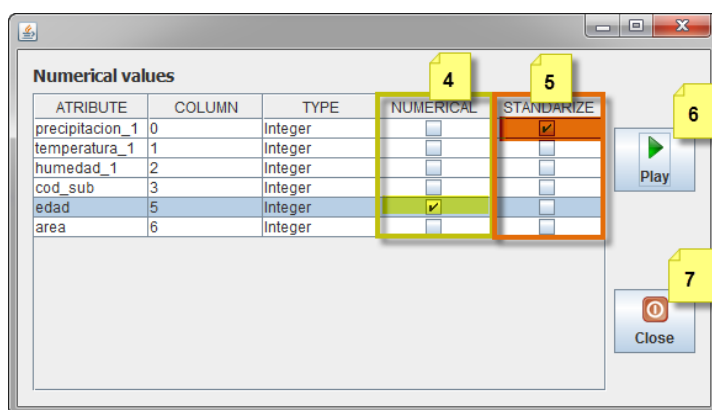


1. De la pestaña **Data cleaning** se selecciona el ícono correspondiente a **Standarize** y con clic sostenido se lo lleva al área **Drag and Drop**.

2. Una vez realizado el paso anterior se debe establecer la conexión con el objeto contenedor de los datos (*Plain text, Connection DB* u otro filtro). Para esto se debe tomar con clic sostenido cualquiera de los 8 puntos que se encuentran alrededor del objeto de datos y llevarlo a cualquiera de los

puntos de conexión del objeto **Standardize**.

3. Ahora se debe dar clic derecho sobre el objeto **Standardize** y del menú de opciones que se despliega seleccionar la opción **Configure...** Con la cual se desplegará una ventana de configuración.

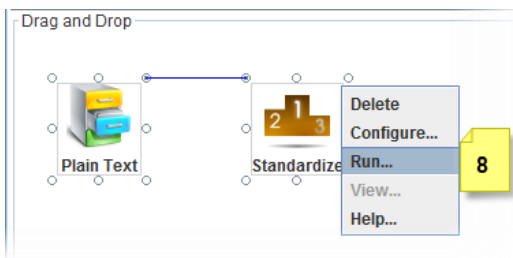


4. En la ventana de configuración se presenta una tabla donde se muestra la información de los atributos de tipo numérico. Una de las opciones que permite el filtro es la realización de una normalización de atributos marcando la casilla de verificación NUMERICAL.

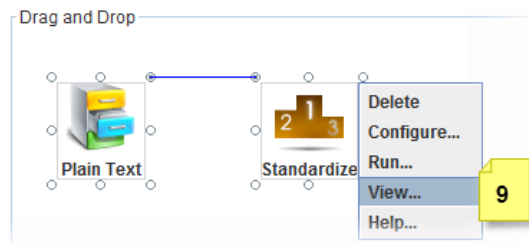
5. Si lo que se desea es estandarizar el atributo, se debe marcar la casilla de verificación en la columna STANDARIZE.

6. Una vez establecidos los métodos (NUMERICAL O STANDARIZE) en los atributos que se desea filtrar se debe dar clic al botón **play** para que la configuración quede establecida.

7. Finalmente, para cerrar la ventana de configuración se debe dar clic en el botón **Close** que permitirá volver a la ventana principal.



8. Nuevamente en la ventana principal se debe dar clic derecho al elemento **Standardize** y seleccionar la opción **Run...** de las opciones que se despliegan. Hecho esto el filtro será ejecutado.



9. Realizado el procedimiento anterior, se debe dar clic derecho nuevamente en el objeto **Standardize** y seleccionar la opción **View...** del menú de opciones para poder visualizar los resultados de aplicar el filtro a los atributos seleccionados en el paso número 4 y 5.

FIELD	TYPE	MAXIMUM	MINIMUM	MEAN	VARIANCE	STANDARD D.
precipitacion...	Numerical	3.60268781...	-1.27210074...	-2.64649925...	1.00000000...	1.00000000...
temperatura_1	Numerical	27.0	24.0	26.0369213...	0.49963085...	0.70684570...
humedad_1	Numerical	88.0	78.0	83.0188866...	4.27726328...	2.06815456...
cod_sub	Numerical	93.0	0.0	4.47202499...	244.385258...	15.6328263...
tip_id	Categorical	0.03436523...	1.42005112...	0.02040477...	8.45363151...	0.00919436...
edad	Numerical	0.03436523...	1.42005112...	0.02040477...	8.45363151...	0.00919436...
area	Numerical	3.0	1.0	2.07242260...	0.80248004...	0.89581250...

10. En la ventana de resultados se observa en la pestaña **Attributes** los atributos numéricos del conjunto de datos y algunos estadísticos descriptivos (máximo, mínimo, media, varianza, desviación estándar) que son los datos que utiliza el filtro para su procedimiento.

precipitacion_1	temperatura_1	humedad_1	cod_sub	tip_id	edad	area
0.5073115094	26	83	15	AS	0.0254189150	3
3.6026878110	26	86	0	AS	0.0208747514	3
1.2740036062	26	84	0	MS	0.0183186594	3
1.2740036062	26	84	0	AS	0.0153365521	3
0.728550378	25	83	0	AS	0.0149105367	3
0.728550378	25	83	16	AS	0.0112184038	3
0.728550378	25	83	0	AS	0.0133484805	3
0.728550378	25	83	0	AS	0.0278330019	3
0.728550378	25	83	8	AS	0.0278330019	2
0.516851814	25	80	8	TI	0.0320931553	2
0.413863323	26	83	1	AS	0.0244248792	3
0.5073115094	26	83	40	AS	0.0278330019	3
0.413863323	26	83	0	AS	0.0069582504	2
0.476800734	25	79	1	CC	0.0018460664	2
0.980300021	26	83	8	RC	0.0343652371	2
0.413863323	26	83	0	AS	0.0129224652	3
3.6026878110	26	86	0	AS	0.0308151093	3
3.6026878110	26	86	0	AS	0.0308151093	3
0.041960440	27	85	0	AS	0.0018460664	3
0.041960440	27	85	1	AS	0.0153365521	3
0.413863323	26	83	0	AS	0.0217267821	3
0.413863323	26	83	52	MS	0.0186026696	3
0.5073115094	26	83	52	MS	0.0336552115	3
0.5073115094	26	83	0	MS	0.0231468332	3
0.6617942453	27	83	0	AS	0.0142005112	3
0.6617942453	27	83	0	MS	0.0186026696	3

11. En la ventana de resultados, en la pestaña **Filtered Data** se puede observar el resultado del filtro en los atributos que fueron seleccionados para ello, así en el ejemplo se puede ver el atributo **precipitación_1** al cual se le aplicó el método **STANDARDIZE**, el cual organiza los datos en base a su media

y desviación estándar.

12. Así mismo, el atributo *edad* tiene aplicado el método NUMERICAL, el cual organiza los datos en base a la normalización de los mismos tomando una media de 0 y una desviación estándar de 1.
13. También esta disponible como en los anteriores métodos la posibilidad de guardar los datos filtrados, para esto se debe dar clic en el botón ***Filtered*** de la ventana de resultados.

4. ETAPA DE MINERIA DE DATOS (DATA MINING)

4.1 Asociación (ASSOCIATION)

Las asociaciones buscan patrones en los que la presencia de algo implica la presencia de algo más. Aquí se presenta a los ítems y una colección de transacciones que son subconjuntos de esos ítems. La tarea es encontrar relaciones entre los ítems de esos subconjuntos para descubrir reglas de asociación que cumplan unas especificaciones mínimas dadas por el usuario, expresadas en forma de **soporte** y **confianza**. Estas dos medidas son las que dan validez al modelo de asociación.

Una regla o patrón de asociación es una implicación de la forma $X \Rightarrow Y$, que significa que si X está presente en una transacción entonces Y también está presente. El soporte para una regla de asociación del tipo $X \Rightarrow Y$ es la proporción (frecuencia) de transacciones en el conjunto de datos que contienen tanto a X como a Y (ref: Larose, 2004).

La confianza de una regla de asociación $X \Rightarrow Y$ es una medida de exactitud (fuerza de implicación) de la regla determinada por el porcentaje de transacciones en el conjunto de datos que contienen a A y B (ref: Larose, 2004).

Los algoritmos que contempla la herramienta para el caso de la asociación son: **Apriori**, **FPGrowth** y **EquipAsso**.

4.1.1 Apriori.

Este algoritmo ejecuta todas las combinaciones posibles en el conjunto de datos, basándose en el conocimiento previo o “a priori” de los conjuntos frecuentes, esto conduce a reducir el espacio de búsqueda y aumentar la eficiencia.

4.1.2 FP-Growth

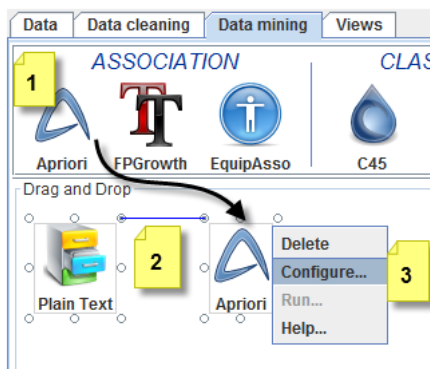
En [ref](#) se propone un interesante método para minar conjuntos de datos (*itemsets*) frecuentes sin la generación de candidatos denominado *Frequent Pattern growth (FP-growth)*, basado en la construcción de un árbol denominado *Frequent Pattern tree (FP-tree)*, el cual es una extensión de una estructura de árbol prefija donde se almacena de manera compacta, toda la información cuantitativa acerca de los patrones frecuentes. La eficiencia de la minería se apoya en las siguientes estrategias: primero, la base de datos se comprime en una estructura de datos mucho más pequeña y altamente condensada, que evita los costos de recorrer la base de datos repetidamente. Segundo, la minería basada en *FP-tree* adopta un método denominado *pattern fragment growth* para evitar la costosa generación de un gran número de *itemsets* candidatos, y por último se adopta la estrategia de divide y vencerás para descomponer la tarea de minería de datos en un conjunto de tareas más pequeñas, lo cual reduce drásticamente el espacio de búsqueda.

Importante: Los resultados de aplicar los algoritmos de minería de datos se podrán observar a través de las diferentes vistas que se encuentran en la pestaña **Views** de la ventana principal de la herramienta.

4.1.3 EquipAsso

EquipAsso es un algoritmo que se basa en la combinación de dos primitivas del algebra relacional implementadas en las primitivas del SQL, pero que su fundamento es la matemática matricial.

En el presente manual se explicará el procedimiento para aplicar el algoritmo de asociación **A priori**, anotando que para los otros algoritmos de asociación (**FP-Growth** y **EquipAsso**) el procedimiento es similar.

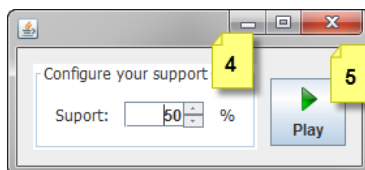


1. De la pestaña **Data mining** se selecciona el ícono correspondiente a **Apriori** y con clic sostenido se lo lleva al área **Drag and Drop**.

2. Una vez realizado el paso anterior se debe establecer la conexión con el objeto contenedor de los datos (*Plain text*, *Connection DB* u otro filtro). Para esto se debe tomar con clic sostenido cualquiera de los 8 puntos que se encuentran alrededor del objeto de datos y llevarlo a cualquiera de los

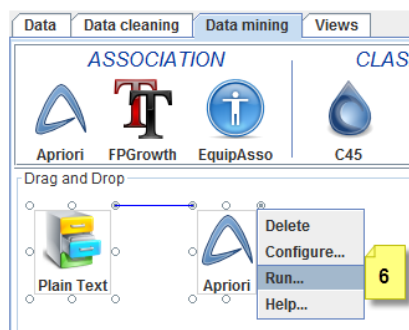
puntos de conexión del objeto **Apriori**.

3. Ahora se debe dar clic derecho sobre el objeto **Apriori** y del menú de opciones que se despliega seleccionar la opción **Configure...** Con la cual se desplegará una ventana de opciones.

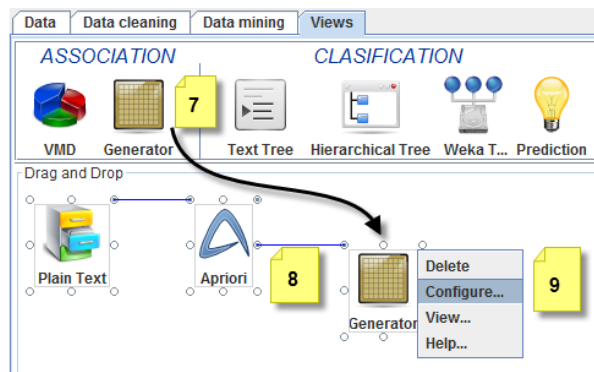


4. En la ventana de configuración se debe establecer el nivel de soporte que tendrá la ejecución del algoritmo. Este es un valor entre 0 y 100%.

5. Luego de colocar el valor del soporte se dará clic en el botón **play** para dejarlo establecido y volver a la ventana principal.

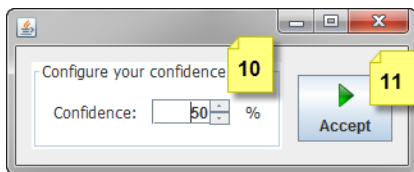


6. En la ventana principal, se tiene que hacer clic derecho sobre el elemento **Apriori** y del menú de opciones seleccionar **Run...** con lo cual se ejecutará el procedimiento de Asociación Apriori.



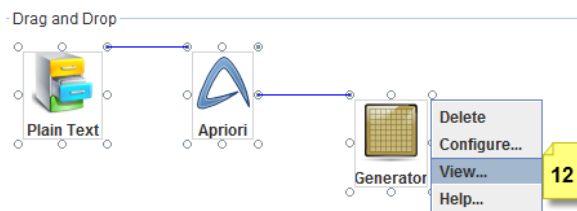
7. Luego del paso anterior se debe cambiar a la pestaña **Views** que nos permitirá a través de uno de sus elementos observar los resultados del algoritmo aplicado. Para esto, de esta pestaña tomamos el ícono correspondiente a **Generator** y con un clic sostenido se lo lleva al área de **Drag and Drop**.

8. Una vez realizado el paso anterior se debe establecer la conexión con el objeto **Apriori**. Para esto se debe tomar con clic sostenido cualquiera de los 8 puntos que se encuentran alrededor del objeto de datos y llevarlo a cualquiera de los puntos de conexión del elemento **Generator**.
9. Ahora se debe dar clic derecho sobre el objeto **Generator** y del menú de opciones que se despliega seleccionar la opción **Configure...** con la cual se desplegará una ventana de opciones.

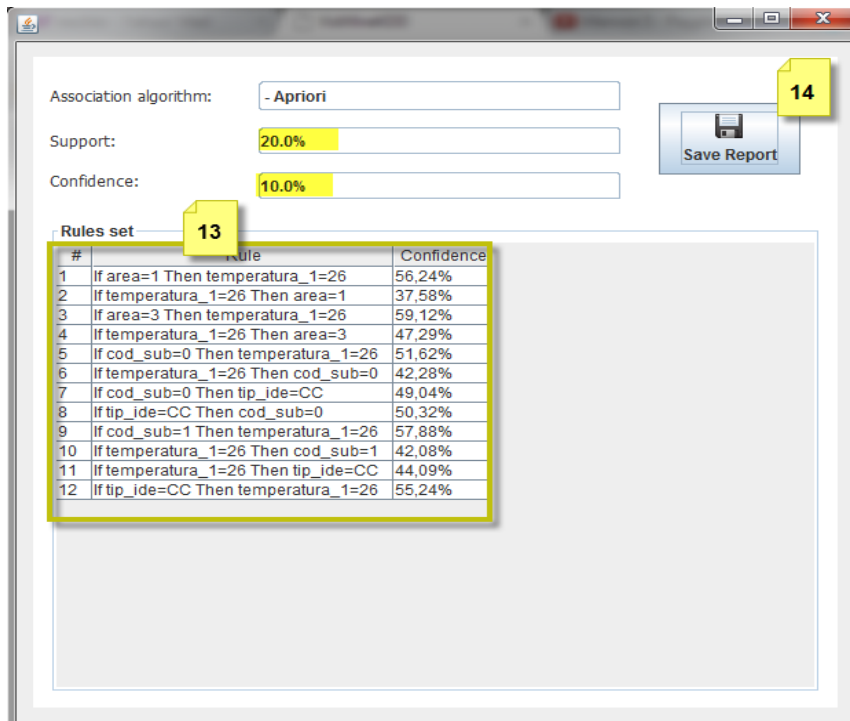


10. Esta vez en la ventana de opciones que se presenta se debe seleccionar el porcentaje de Confianza que se aplicará al modelo. El rango para este nivel es de 0 a 100%.

11. Luego de lo anterior se debe dar clic en el botón **play** de la ventana para que quede establecida la confianza y volver a la ventana principal.



12. Nuevamente en la ventana principal se debe dar clic derecho sobre el elemento **Generator** y seleccionar la opción **View...** del menú que se despliega.



13. En la ventana que se despliega se informa los resultados obtenidos. Se muestra el algoritmo de asociación que se aplicó, el nivel de soporte y confianza que estableció y más abajo en una tabla se presentan el conjunto de reglas que se encontraron en el conjunto de datos luego de aplicado el algoritmo. Al lado de cada regla encontrada se informa el porcentaje de confianza que tiene la regla a nivel individual.

14. Mediante el botón **Save Report** es posible guardar el conjunto de reglas encontrado.

4.2 Clasificación (CLASIFICATION)

La predicción tiene como objetivo estimar el posible valor o comportamiento de una variable o un conjunto de variables a partir de un conjunto de datos, utilizando distintos tipos de métodos como estadísticos de regresión lineal y no lineal [23], dependiendo del comportamiento de los datos, en el que se pretende aproximar la instancia de los datos a una figura estandarizada como una línea recta o curva, en algunos casos es necesario manipular los datos de origen para poderlos ajustar a los modelos de regresión. Otro tipo de métodos que es aplicable en ésta técnica es el empleo de árboles de decisión, los cuales utilizan una medida estadística conocida como Entropía, la cual sirve para medir el grado de desorden de un sistema, permitiendo diferenciar los datos útiles, es decir los que generan información, de los datos inútiles o aquellos que producen desorden o desconocimiento del sistema. En el árbol de predicción la entropía permite establecer cual es el atributo que contribuye a generar mas información, dicho atributo se establece como nodo principal del árbol y a partir de ese nodo el proceso continua hasta tener una comprensión general

del sistema, por lo tanto, la entropía aplicada al descubrimiento de conocimiento pretende aplacar el desconocimiento reduciendo la incertidumbre en el sistema, organizando los datos en información útil, como un modelo que toma la forma de un árbol jerarquizado.

Un ejemplo típico de la aplicación de éste tipo de técnicas es el marketing dirigido, sobre la tendencia de mercados, prediciendo el comportamiento de compra de los usuarios según sus gustos, cultura, economía, entre otras variables.

El proceso consiste en agrupar conjuntos de datos mutuamente excluyentes según una clase objetivo, estableciendo una distancia entre ellas. Si la clase objetivo es numérica, se toma la media y si la clase objetivo es nominal o categórica se toma la moda [12].

Hay muchos métodos que sirven para realizar clasificación [24], como tablas y arboles de decisión, inducción de reglas, clasificadores bayesianos, clasificadores basados en casos o ejemplos, algoritmos genéticos, lógica difusa y redes neuronales, sin embargo los arboles de decisión se presentan como un modelo de fácil abstracción, en donde los atributos son los nodos y los valores de dichos atributos son las ramas, los cuales se organizan en base a la clase objetivo, así pues, las reglas de clasificación resultan de recorrer el árbol descendentemente, sin embargo, el árbol puede crecer mucho haciéndose complejo e incomprensible, por lo cual se hace necesario aplicar técnicas de poda, que permiten acotar el árbol suprimiendo hojas y reduciendo el modelo. Como ejemplo de algoritmo de esta tarea podemos referenciar a Sliq, IDE3 [25] [26] C4.5 [27] [5] y Mate [28].

4.2.1 MATE

Este algoritmo genera, por cada una de las tuplas de una relación, todas las posibles combinaciones formadas por los valores no nulos de los atributos pertenecientes a una lista de atributos denominados Atributos Condición, y el valor no nulo del atributo denominado Atributo Clase.

Mate empareja en cada partición todos los atributos condición con el atributo clase, lo que facilita el conteo y el posterior cálculo de las medidas de entropía y ganancia de información. Mate genera estas combinaciones, en una sola pasada sobre la tabla de entrenamiento (lo que redundará en la eficiencia del proceso de construcción del árbol de decisión).

4.2.2 C4.5

El algoritmo C4.5 genera un árbol de decisión a partir de los datos mediante particiones realizadas recursivamente. El árbol se construye mediante la estrategia de profundidad-primero. El algoritmo considera todas las pruebas posibles que pueden dividir el conjunto de datos y selecciona la prueba que resulta en la mayor ganancia de información en base a la métrica de la entropía.

4.2.3 SLIQ

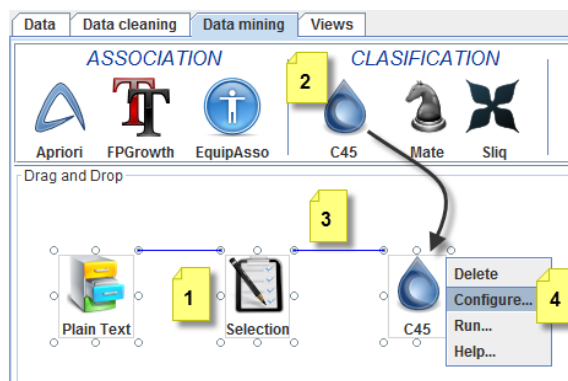
SLIQ es un clasificador que usa árboles de decisión y que puede manejar tanto atributos numéricos como categóricos. También usa una técnica de pre-clasificación en la etapa de construcción del árbol para reducir el coste de la evaluación de particiones por atributos numéricos

SLIQ puede clasificar grandes bases de datos residentes en disco si utilizar memoria. Otra característica interesante es el uso de un nuevo algoritmo de poda que es poco costoso y produce árboles compactos y eficaces.

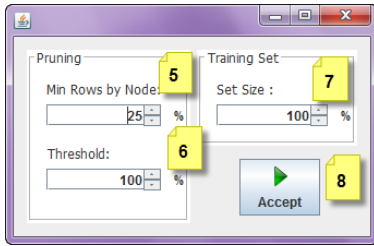
Poda del árbol:

El conjunto de datos puede generar mucho ruido el cual implicaría un crecimiento desmedido del árbol. Este hecho nos puede llevar a errores en la clasificación es decir, clasifica muy bien los datos de entrenamiento pero luego no sabe generalizar al conjunto de prueba. Este efecto indeseado. Es posible controlarlo configurando el número de transacciones por nodo en la caja de texto Min rows by node y un umbral o límite de conocimiento de crecimiento threshold, es decir hasta que porcentaje se considera que el árbol debe crecer para alcanzar un modelo de conocimiento adecuado.

En el presente manual se explicará el procedimiento para aplicar el algoritmo de clasificación **C4.5**, anotando que para los otros algoritmos de clasificación (**Mate y Sliq**) el procedimiento es similar.



1. Para utilizar el algoritmo de clasificación debe haberse realizado el proceso de selección, el cual se explicó en la etapa de [Data cleaning \(limpieza de datos\)](#).
2. De la pestaña **Data mining** se selecciona el ícono correspondiente a **C45** y con clic sostenido se lo lleva al área **Drag and Drop**.
3. Una vez realizado el paso anterior se debe establecer la conexión con el objeto **Selection** que se encuentra en el área de **Drag and drop**. Para esto se debe tomar con clic sostenido cualquiera de los 8 puntos que se encuentran alrededor del objeto **Selection** y llevarlo a cualquiera de los puntos de conexión del objeto **C45**.
4. Ahora se debe dar clic derecho sobre el objeto **Selection** y del menú de opciones que se despliega seleccionar la opción **Configure...** Con la cual se desplegará una ventana de opciones.

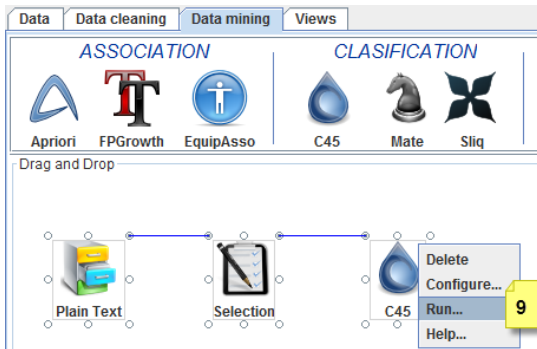


5. En la ventana de configuración se debe establecer para la poda (*Pruning*) del árbol el porcentaje mínimo de filas por nodo (*Min Rows by Node*).

6. Así mismo, se debe establecer el porcentaje para el umbral (*Threshold*) de la poda.

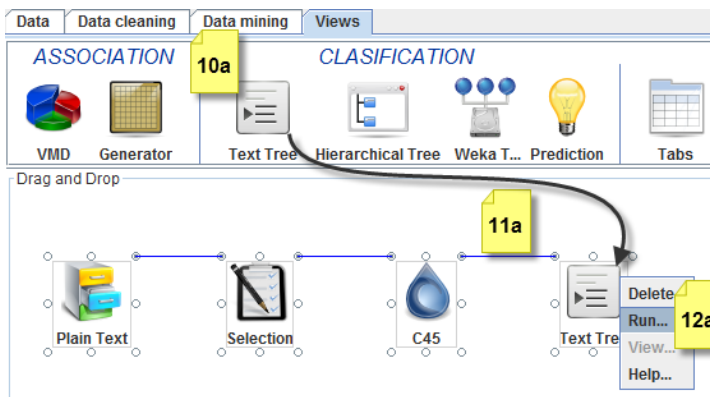
7. Igualmente, en esta ventana se debe configurar el porcentaje de entrenamiento para el modelo (*Set size*).

8. Una vez configurados los anteriores parámetros se puede dar clic al botón **Accept** que dejará listo el algoritmo para ser ejecutado.



9. Posteriormente a la configuración, se tiene que hacer clic derecho sobre el elemento **C45** y del menú de opciones **Run...** con lo cual se ejecutará el procedimiento de Clasificación C45.

Luego de ejecutado el procedimiento se procede a visualizar los resultados, para lo cual se ha establecido mostrar las diferentes opciones de visualización que tienen los algoritmos de clasificación, por cuanto, observaremos en los pasos una vocal en cada número de paso que identificará al visualizador seleccionado.

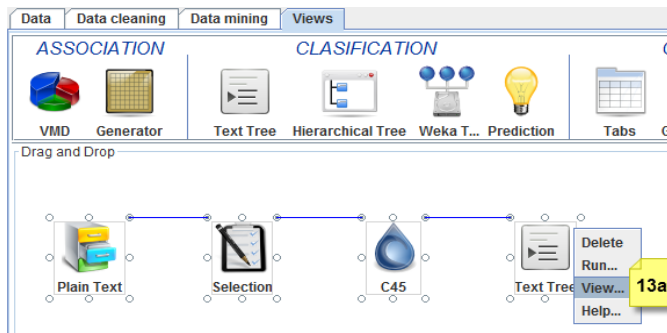


10a. Luego de haber ejecutado el algoritmo de clasificación, se debe cambiar a la pestaña **Views**. De esta, tomamos el ícono correspondiente a **Text Tree** y con clic sostenido se lo lleva al área de **Drag and Drop**.

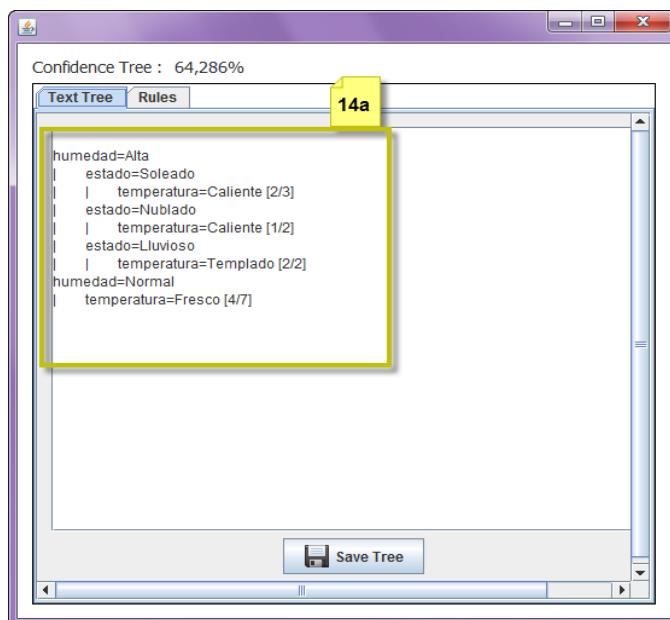
11a. Una vez realizado el paso anterior se debe establecer la conexión con el objeto **C45**. Para esto se debe tomar con clic

sostenido cualquiera de los 8 puntos que se encuentran alrededor del objeto de datos y llevarlo a cualquiera de los puntos de conexión del elemento **Text Tree**.

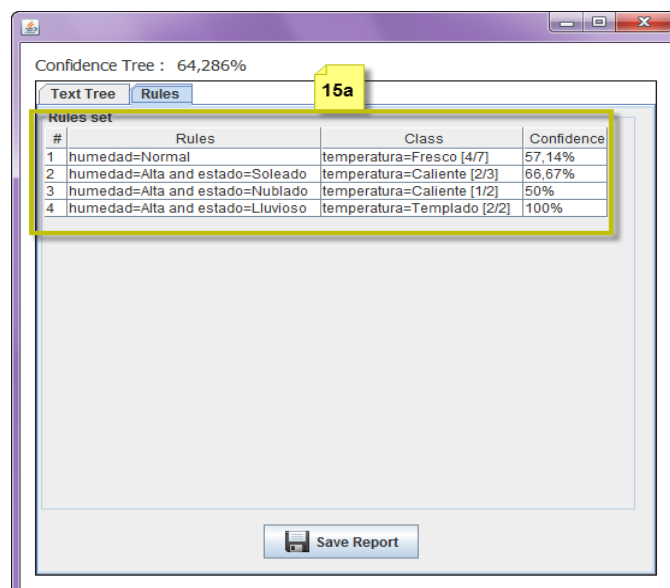
12a. Luego damos clic derecho sobre el objeto **Text Tree** y de las opciones seleccionamos **Run...** que ejecutará el procedimiento de visualización.



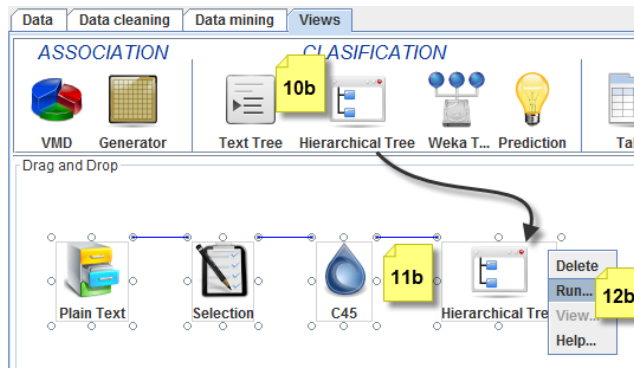
13a. Luego volvemos a dar clic derecho sobre el objeto **Text Tree** y de las opciones seleccionamos **View...** que desplegará la ventana de resultados.



14a. En la ventana de resultados se nos presentan la información correspondiente al porcentaje de confiabilidad del árbol en la parte superior (*Confidence Tree*). Por debajo de ésta, se encuentran dos pestañas, la primera denominada **Text Tree** que muestra el árbol de clasificación organizado en formato texto. Aquí se pueden ver las reglas de clasificación que generó el algoritmo. Este resultado lo podemos guardar a través del botón **Save Tree**.



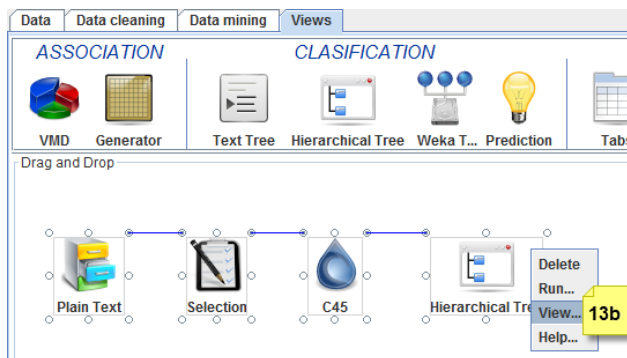
15a. De igual manera, en la misma ventana en la pestaña **Rules** se puede observar una grilla que muestra el conjunto de reglas que se encuentran en el árbol junto con información de las clases que comprende la regla y el porcentaje de confianza de la misma.



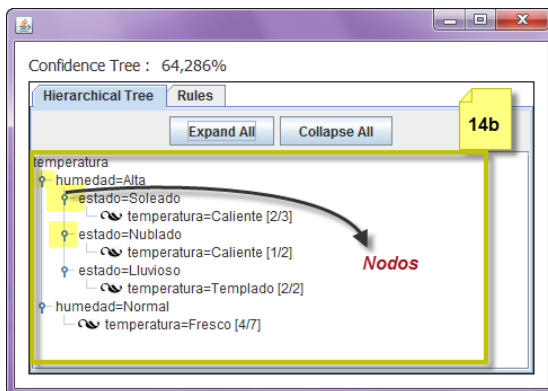
10b. Ahora veremos el procedimiento para visualizar los resultados de los algoritmos de clasificación mediante el visualizador **Hierarchical Tree**. De esta manera tomamos el ícono correspondiente a **Hierarchical Tree** y con clic sostenido se lo lleva al área de **Drag and Drop**.

C45. Para esto se debe tomar con clic sostenido cualquiera de los 8 puntos que se encuentran alrededor del objeto de datos y llevarlo a cualquiera de los puntos de conexión del elemento **Hierarchical Tree**.

12b. Luego damos clic derecho sobre el objeto **Hierarchical Tree** y de las opciones seleccionamos **Run...** que ejecutará el procedimiento de visualización.

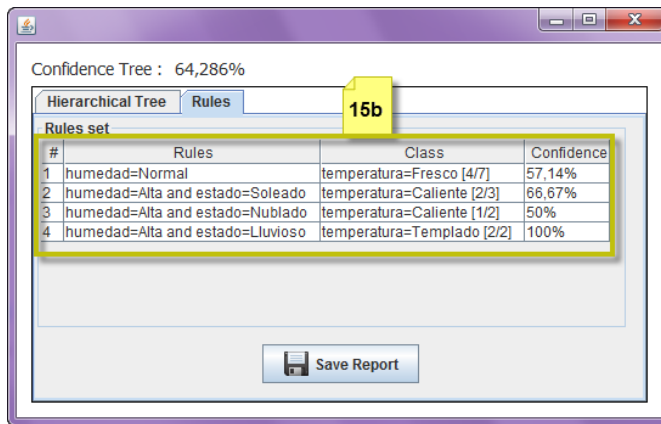


13b. Luego volvemos a dar clic derecho sobre el objeto **Hierarchical Tree** y de las opciones seleccionamos **View...** que desplegará la ventana de resultados.

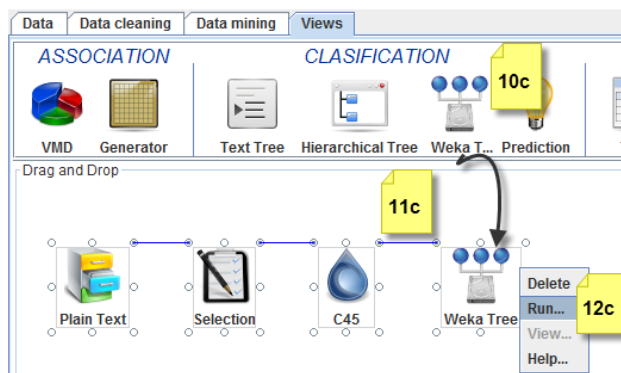


All y Collapse All respectivamente.

14b. En la ventana de resultados se nos presentan la información correspondiente al porcentaje de confiabilidad del árbol en la parte superior (**Confidence Tree**). Por debajo de ésta, se encuentran dos pestañas, la primera denominada **Hierarchical Tree** que muestra el árbol de clasificación organizado en jerárquica. Aquí se pueden ver las reglas de clasificación que generó el algoritmo. Además, se puede expandir o colapsar las reglas mediante los botones **Expand**



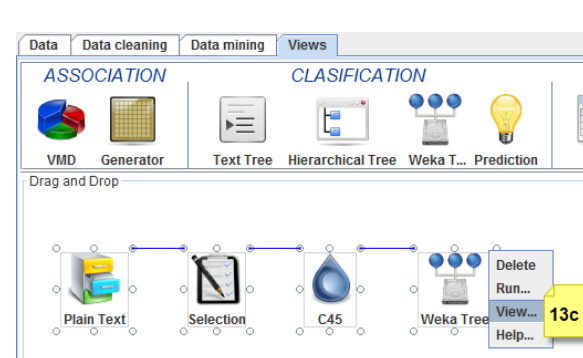
15b. De igual manera, en la misma ventana en la pestaña **Rules** se puede observar una grilla que muestra el conjunto de reglas que se encuentran en el árbol junto con información de las clases que comprende la regla y el porcentaje de confianza de la misma.



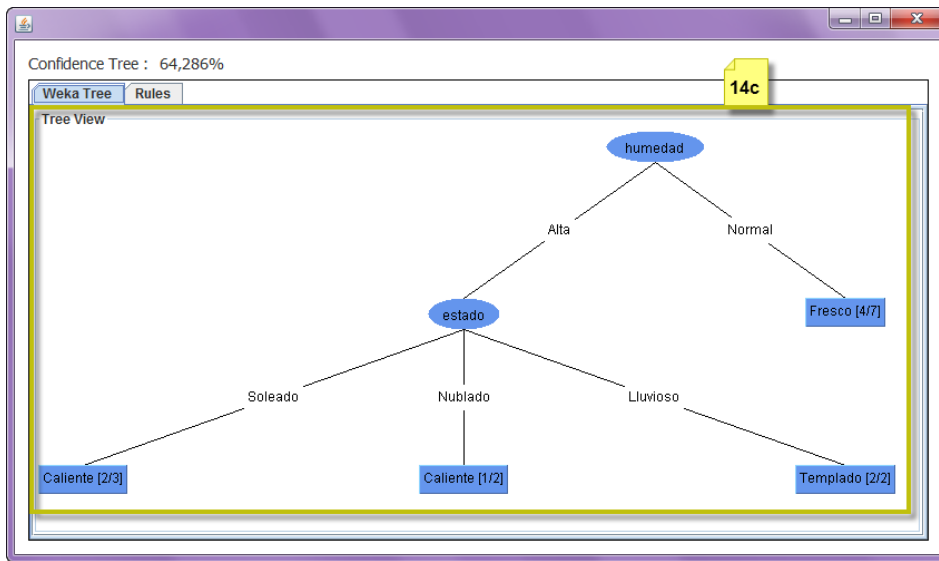
10c. Ahora veremos el procedimiento para visualizar los resultados de los algoritmos de clasificación mediante el visualizador **Weka Tree**. De esta manera tomamos el ícono correspondiente a **Weka Tree** y con clic sostenido se lo lleva al área de **Drag and Drop**.

C45. Para esto se debe tomar con clic sostenido cualquiera de los 8 puntos que se encuentran alrededor del objeto de datos y llevarlo a cualquiera de los puntos de conexión del elemento **Weka Tree**.

12c. Luego damos clic derecho sobre el objeto **Hierarchical Tree** y de las opciones seleccionamos **Run...** que ejecutará el procedimiento de visualización.



13c. Luego volvemos a dar clic derecho sobre el objeto **Weka Tree** y de las opciones seleccionamos **View...** que desplegará la ventana de resultados.



14c. En la ventana de resultados se nos presentan la información correspondiente al porcentaje de confiabilidad del árbol en la parte superior (*Confidence Tree*). Por debajo de ésta, se encuentran dos pestañas, la primera denominada **Weka Tree** que muestra el árbol de clasificación organizado en modo gráfico similar a un árbol.

#	Rules	Class	Confidence
1	humedad=Normal	temperatura=Fresco [4/7]	57.14%
2	humedad=Alta and estado=Soleado	temperatura=Caliente [2/3]	66.67%
3	humedad=Alta and estado=Nublado	temperatura=Caliente [1/2]	50%
4	humedad=Alta and estado=Lluvioso	temperatura=Templado [2/2]	100%

15c. De igual manera, en la misma ventana en la pestaña **Rules** se puede observar una grilla que muestra el conjunto de reglas que se encuentran en el árbol junto con información de las clases que comprende la regla y el porcentaje de confianza de la misma.

4.3 Cluster

También es conocida como la técnica de agrupamiento, permite segmentar el conjunto de datos en grupos que presentan dos características indispensables: que los elementos de un grupo presenten alta similitud entre ellos y muchas diferencias con los elementos de otros grupos, según medidas de atracción y repulsión que dependen del método utilizado [21], el cual subdivide ésta técnica en métodos numérico, conceptual y probabilístico, como ejemplos respectivos de algoritmos desarrollados en dichos métodos tenemos los siguientes: k-medias, Cobweb, EM [9].

Por ejemplo, los animales que presenten las características: presencia de pelo, reproducción vivípara y lactancia, serán segmentados en un grupo diferencial de otro que tenga como atributos:

presencia de plumas, reproducción ovípara, no lactante, ya que la medida de su similitud intragrupo y diferencias extragrupo es alta.

Esta técnica es empleada en muchas y diversas áreas, sin embargo ha tenido gran importancia en la bioinformática, especialmente la enfocada a la genética con el proyecto GENOME [21] [22] analizando las interacciones de los genes y su repercusión en posibles enfermedades.

4.3.1 Algoritmo K-Means

Es uno de los algoritmos más conocidos de agrupamiento, sigue una forma fácil y simple para dividir una base de datos en k grupos fijados a priori definiendo k centroides, uno para cada grupo, posteriormente toma cada punto de la base de datos y lo sitúa en la clase del centroide más cercano. La formación de los grupos (clúster) se basa en un criterio de cercanía. El criterio de cercanía generalmente se define como una función de distancia, entre las que se destacan la euclidiana (más utilizada), Manhattan y Minkowski, que son las que utiliza la herramienta.

El proceso se repite hasta que ya no es posible generar cambios en los grupos de un paso al siguiente.

4.3.2 BIRCH

Es un método que descompone de forma jerárquica un conjunto de datos, creando un dendrograma o árbol que divide la base de datos recursivamente en conjuntos cada vez más pequeños, tratando de minimizar la distancia total entre los registros y sus conglomerados, El algoritmo realiza dos pasos independientes: primero, ordena los registros de entrada en un árbol de característica de conglomerado de modo que los registros similares pasan a formar parte de los nodos del mismo árbol; a continuación, agrupa las hojas de este árbol en la memoria para generar el resultado de conglomerado definitivo

4.3.3 CLARANS

Es un algoritmo de clúster particional, evoluciona a partir de los algoritmos PAM Y CLARA, éste algoritmo no realiza particiones geométricas para identificar centroides, si no que selecciona un elemento representativo del grupo denominado medoide. A partir de los medoides se asigna cada objeto al clúster representado por el medoide más cercano y se computa la función de calidad como la media de las distancias de cada objeto a su medoide correspondiente del clúster.

