Manual Básico Knime

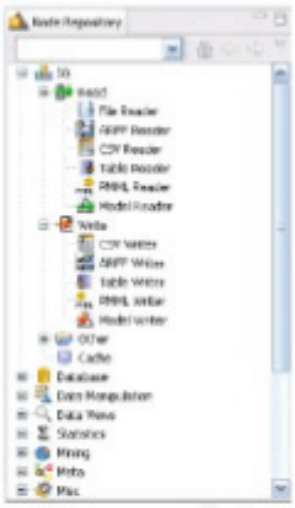
El presente manual tiene por objeto presentar un manual de uso básico del software de análisis datamining Knime. Desarrollado para programa Máster de IEDE- investigación de Mercado y Clientes.

Knime

KNIME es un entorno totalmente gratuito para el desarrollo y ejecución de técnicas de minería de datos. KNIME fue desarrollado originalmente en el departamento de bioinformática y minería de datos de la Universidad de Constanza, Alemania, bajo la supervisión del profesor Michael Berthold. En la actualidad, la empresa KNIME.com GmbH, radicada en Zúrich, Suiza, continúa su desarrollo, además de prestar servicios de formación y consultoría.

KNIME está desarrollado sobre la plataforma Eclipse y programado, esencialmente, en Java. Como otros entornos de este tipo, algunos de los cuales aparecen referenciados al final de este documento, su uso se basa en el diseño de un flujo de ejecución que plasme las distintas etapas de un proyecto de minería de datos.

Para ello, KNIME proporciona distintos nodos agrupados en fichas, como por ejemplo:



a) Entrada de datos [IO > Read].

b) Salida de datos [IO > Write].

c) Preprocesamiento [Data Manipulation], para filtrar,

discretizar, normalizar, filtrar, seleccionar variables…

d) Minería de datos [Mining], para construir modelos (reglas de asociación, clustering, clasificación, MDS, PCA…).

e) Salida de resultados [Data Views] para mostrar resultados en pantalla (ya sea de forma textual o gráfica).

Para crear un flujo de ejecución, las salidas de unos nodos se utilizan como entradas de otros. Por ejemplo, un flujo básico podría ser de la forma:

Instalación

El primer paso es ir a www.knime.org, dirigirse a la opción del menú superior "Downloads".

Seleccionar la opción "Knime Desktop" y luego presionar el botón , para iniciar descarga del software sin necesidad de registrarse en el sitio WEB.

Deberá seleccionar la opción de descarga que sea compatible con la versión del sistema operativo que maneja su PC, es decir, Knime Desktop para Windows (32 y 64 bits), Linux y para el sistema operativo MACOS de Apple.

NOTA:

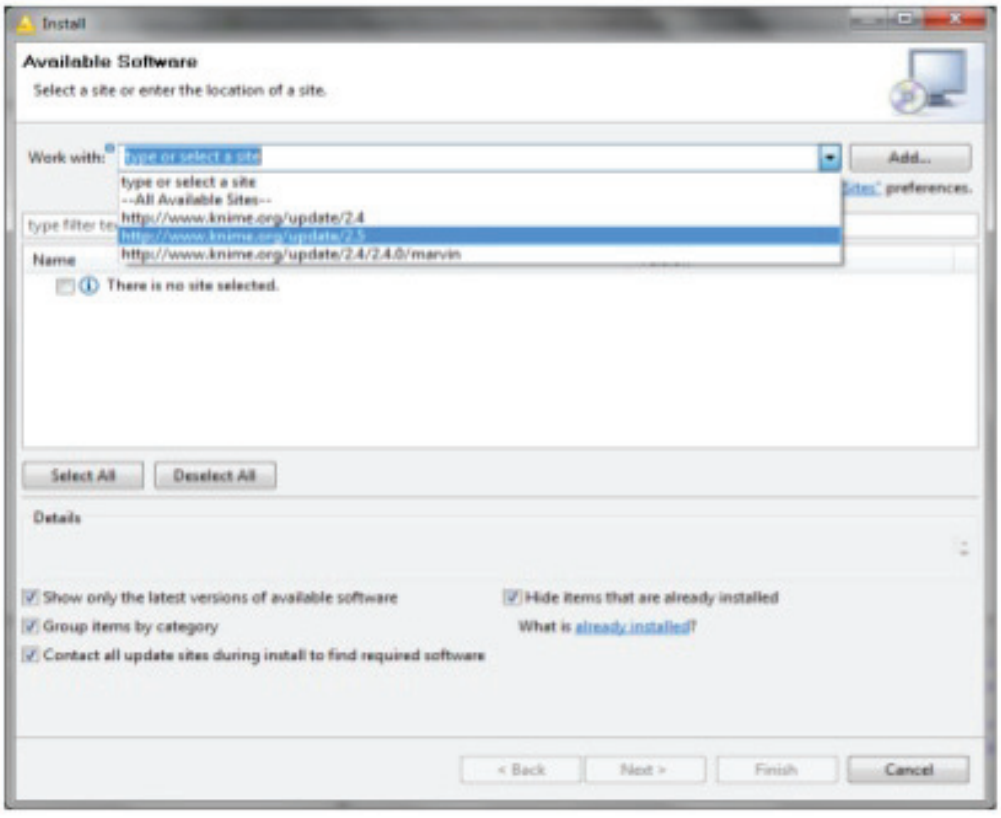
KNIME no requiere programa de instalación. El paquete se desinstala, simplemente, borrando la carpeta en la que lo hayamos descomprimido.

Ejecución

Para ejecutar KNIME, busque el fichero ***knime.exe*** en la carpeta en la que haya descomprimido el paquete de instalación y ejecútelo:

Antes de empezar a utilizar KNIME, nos aseguraremos de instalar los componentes de Weka, utilizando la opción "Get additional nodes" de la ventana de inicio de KNIME o accediendo a ellos a través del menú Help > Install New Software.

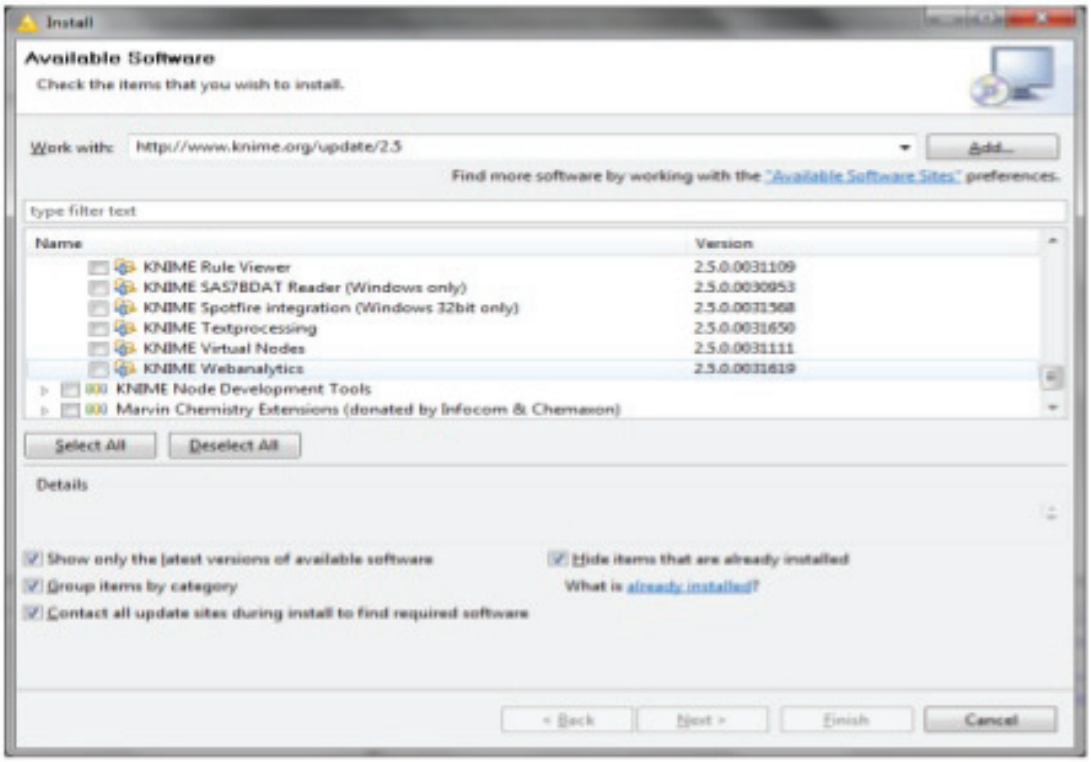
Luego seleccionar desde la lista http://www.knime.org/update/2.5 y luego en filtro seleccionar opcion WEKA Extensions (Waikato Environment for Knowledge Analysis)



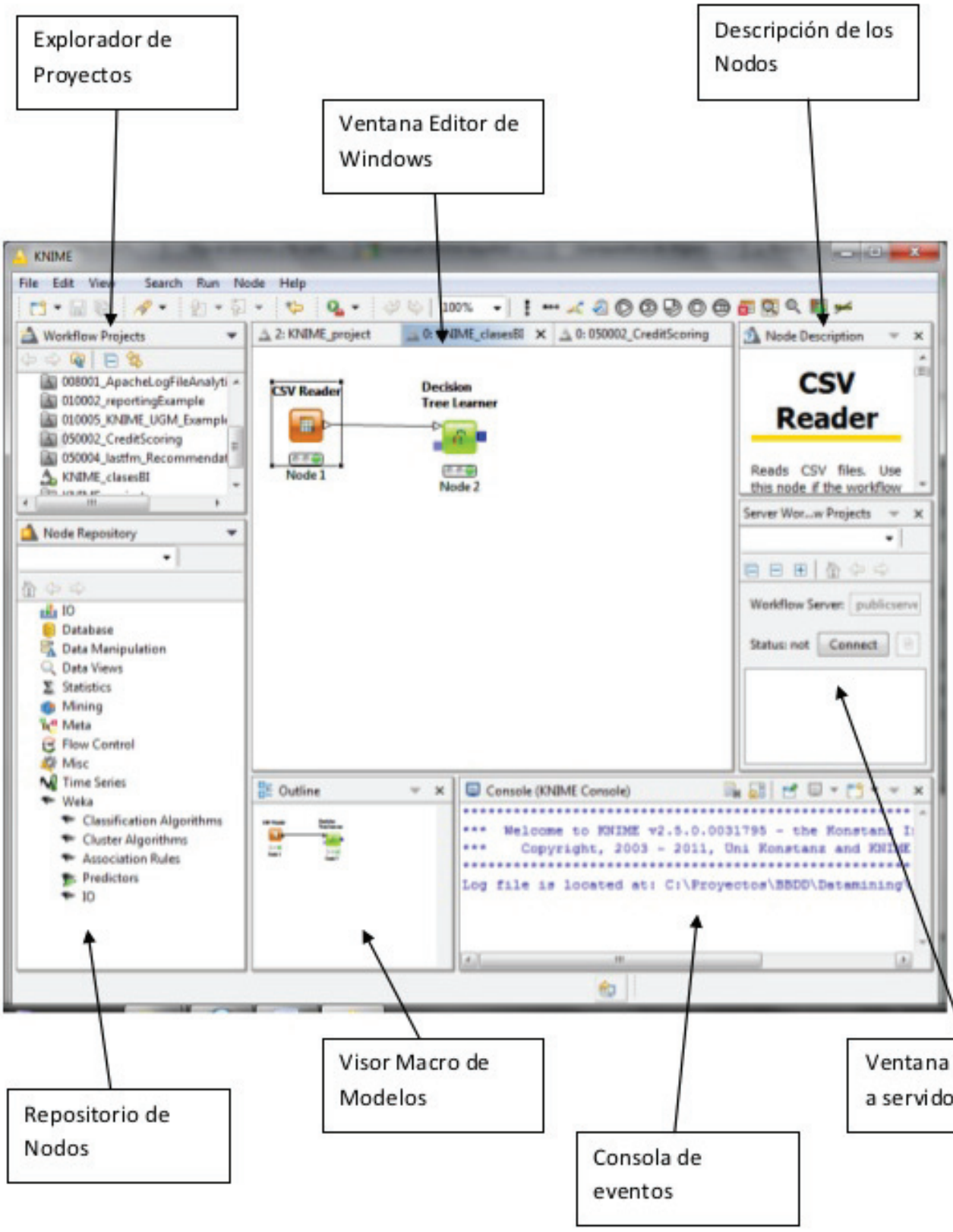
Más información de WEKA en http://www.cs.waikato.ac.nz/ml/weka/

Luego se instalará y deberá reiniciar el Knime.

Note que tiene otras opciones de análisis y extensiones que le permitirá explorar y analizar diferentes fuentes de información, por ejemplo WEB Analytics:



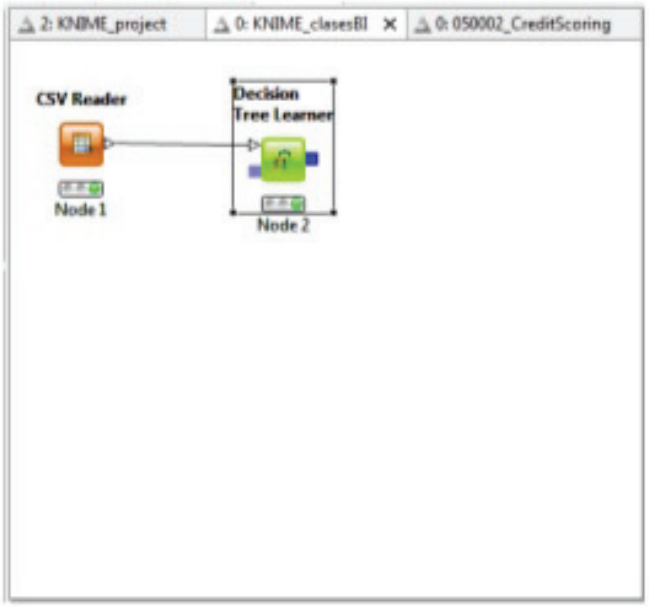
Partes de la herramienta



Ismael Tapia Vidal (Ismael.tapia.v@gmail.com)

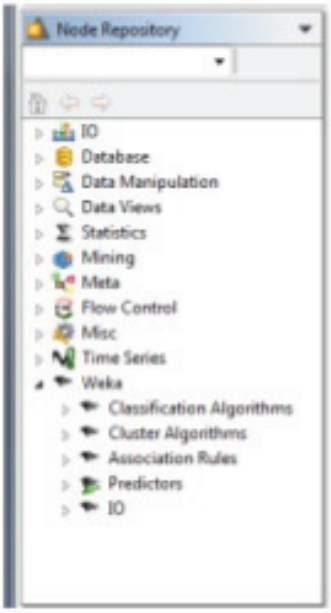
Venta Editor de Windows

Es la ventana donde se construyen los diagramas de análisis de datos



Repositorio de Nodos

Se encuentran todos los nodos que pueden agregarse a un flujo (están agrupados por categorías).

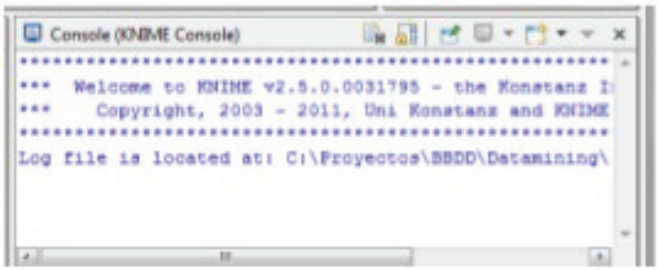


Categoría	Nodos
IO	File Reader, Table Reader, CSV Writter, Table Writter.
Manipulación de datos	Column, Row, Matrix.
Vistas de datos	Scatter Plot, Interactive Table, Box Plot, Histogram, Line Plot.
Mining	Clustering, Decision Tree, Association Rules.
Weka	Bayes, Trees, Rules.

Uso: Drag and Drop (Arrastrar y pegar) en el editor de Windows.

Consola de Eventos

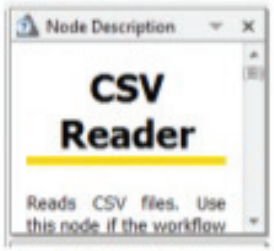
Muestra los mensajes de warning y error.



Esos mensajes también se almacenan en un archivo, llamado knime.log (stack trace), ubicado en la carpeta .metadata del proyecto.

Descripción de Nodos

Provee información sobre cualquier nodo seleccionado.

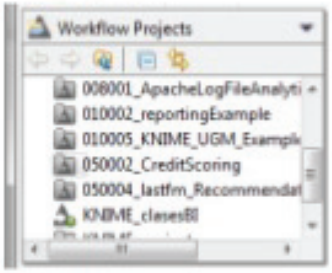


->Esta vacía si no seleccionamos ningún nodo en el flujo.

-> Si seleccionamos una categoría (que incluye varios nodos), entonces muestra los nodos que forman parte de esa categoría.

Explorador de Proyectos

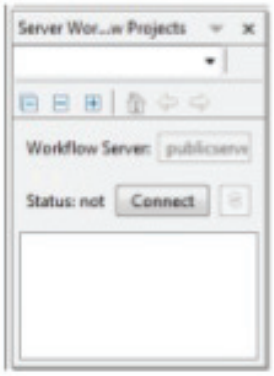
Muestra todo los proyectos existente en el workspace (espacio de trabajo, que se encuentra en el directorio donde descomprimimos el KNime).



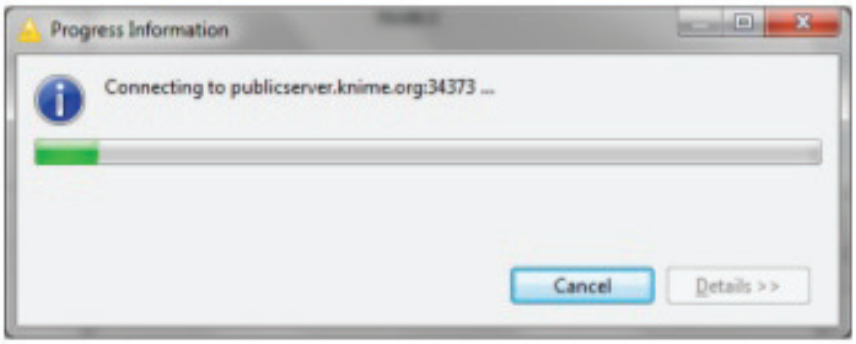
La herramienta permite crear nuevos proyectos, importar proyectos existentes o exportar nuestro proyecto, entre otras actividades.

Ventana de conexión a servidores públicos

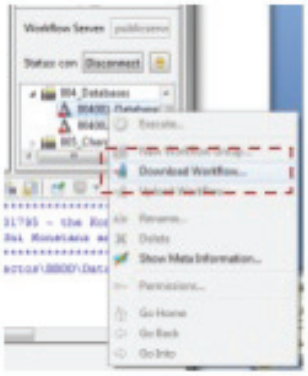
Es la ventana donde se pueden explorar los proyectos ejemplo desde los servidores públicos



Primero debe conectarse, presionando el botón "connect"



Una vez establecida la conexión, se podrá apreciar a modo de exploración los proyectos públicos para descargar y examinar. Se debe hacer clic sobre un workflow de ejemplo y seleccionar opción de descarga "Download Workflow".

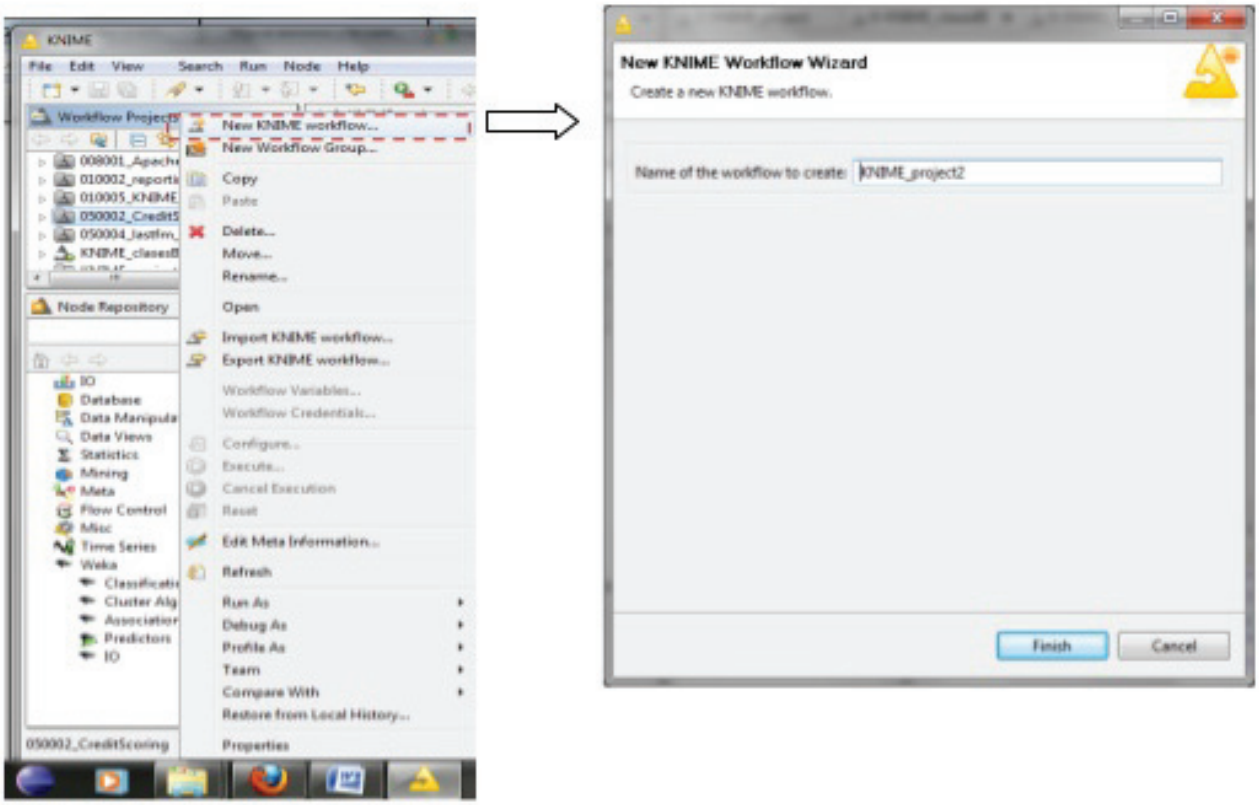


Acciones Básicas

- Crear un proyecto.
- Utilización de nodos.
- Construir un flujo de datos.
- Nodo Color Manager.
- Configuración de Nodos.
- Ejecución del flujo de datos.
- Resultados.
- HiLiting.

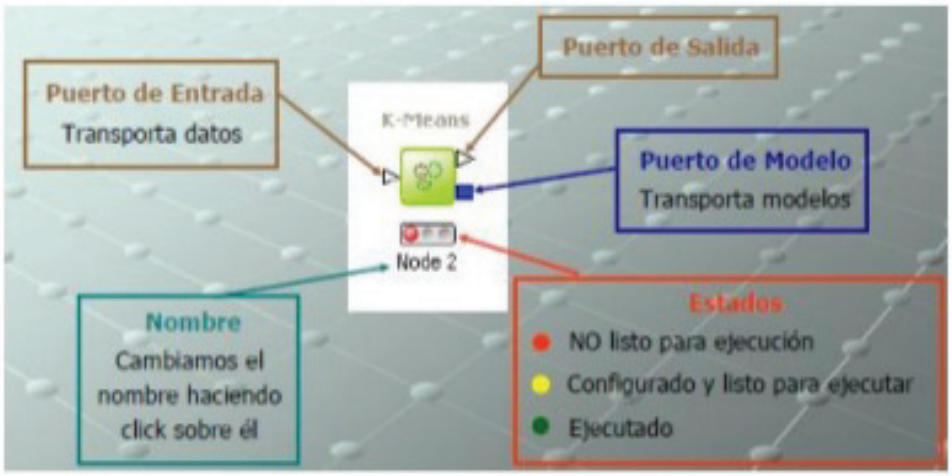
Crear un proyecto.

Clickeamos con botón derecho sobre el Workflow Proyect Navigator, elegimos la opción New KNIME workflow y después le ponemos un nombre al proyecto, luego asignamos el nombre y damos finalizar.



Utilización de nodos.

"Los nodos son básicamente unidades de procesamiento de un workflow"

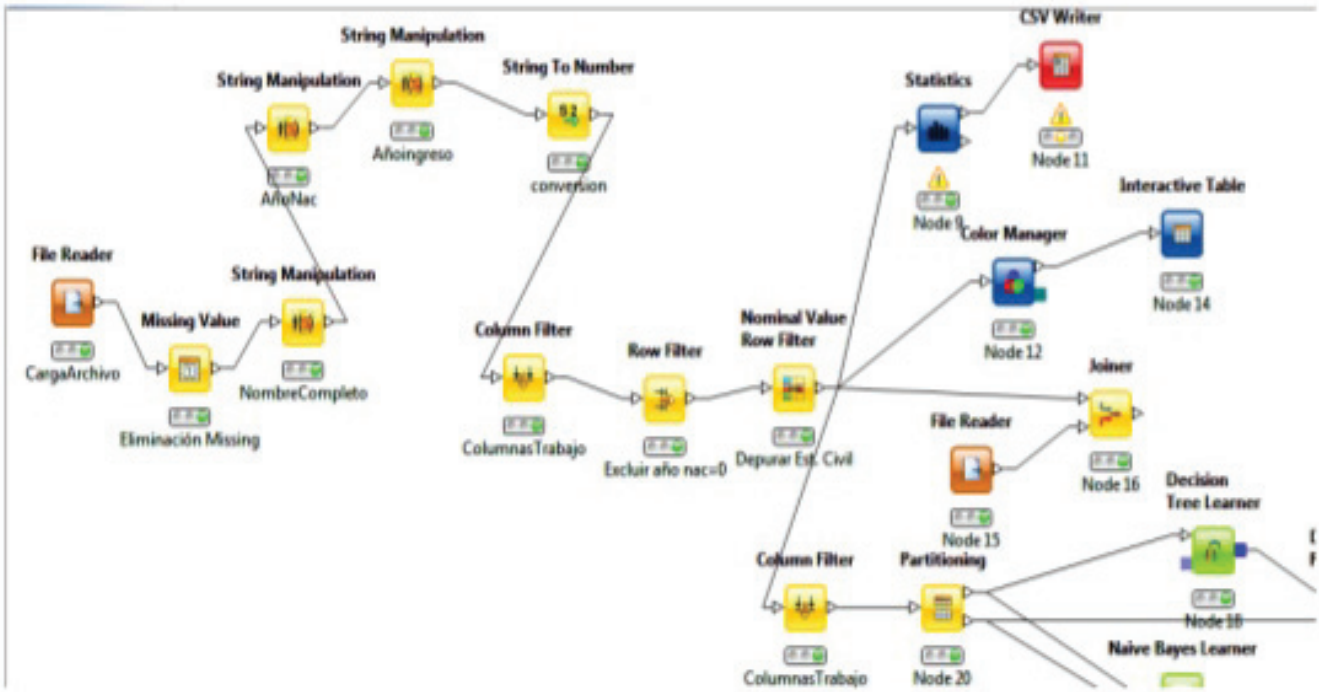


Construir un flujo de datos.

Se construye un flujo arrastrando y soltando (drag & drop) los nodos desde el repositorio de nodos al Editor Window y conectándolos entre ellos.

Los datos se transportan entre los nodos a través de los puertos de entrada y salida.

Después de colocar los nodos en el Editor Window, es necesario conectar la entrada de cada nodo con la salida del nodo predecesor.



Nodo Color Manager.

Este nodo se utiliza para colorear los resultados, generados a partir de los datos de entrada.

Los colores elegidos en el ColorManager afectan a muchas vistas y ayudan a diferenciar los datos.

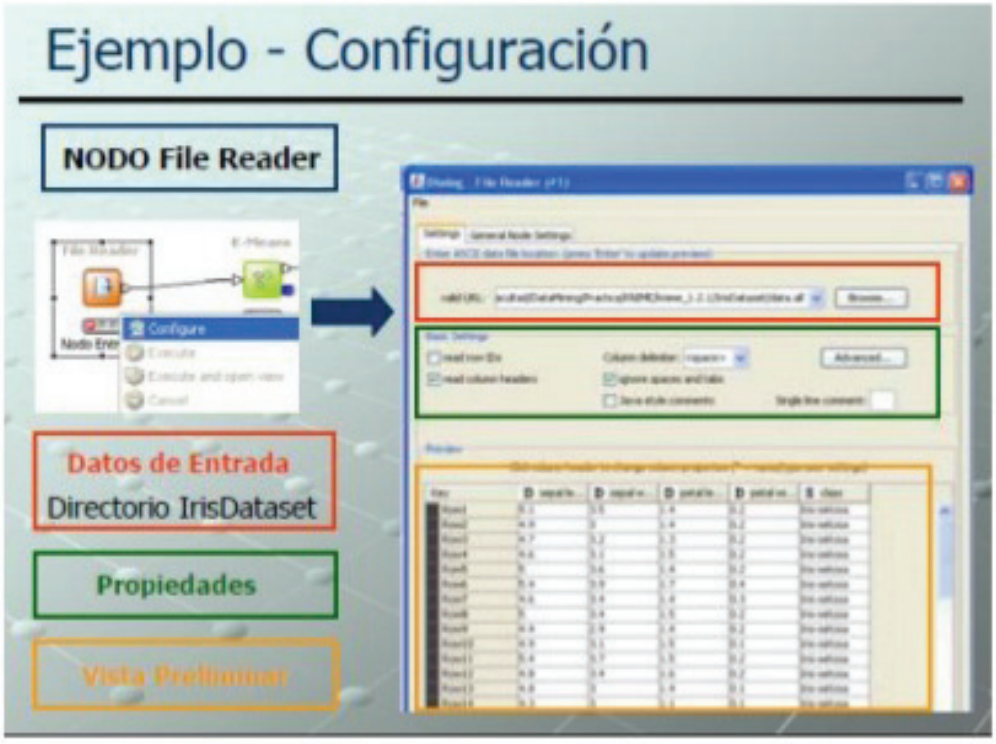
Si se inserta este nodo en el flujo, entonces los datos de entrada se codifican según un color específico, determinado por el ColorManager.

Si se inserta este nodo en el flujo, entonces los datos de entrada se codifican según un color específico, determinado por el ColorManager.

Configuración de Nodos.

Una vez que el flujo está totalmente conectado, si existen nodos que tienen su estado en color rojo, entonces deben ser configurados.

Si algún nodo presenta un color amarillo, después de que fueron conectados, significa que la configuración del mismo se realizó utilizando los valores por defecto.



Ejemplo - Configuración

NODO K-Means

Se configura automáticamente (una vez que se configura el File Reader). Su estado se encuentra en el color amarillo (esto significa que las propiedades por defecto del nodo fueron aplicadas, y el mismo esta listo para ejecutarse).

Número de Clusters

Iteraciones

Ejemplo - Configuración

NODO ColorManager

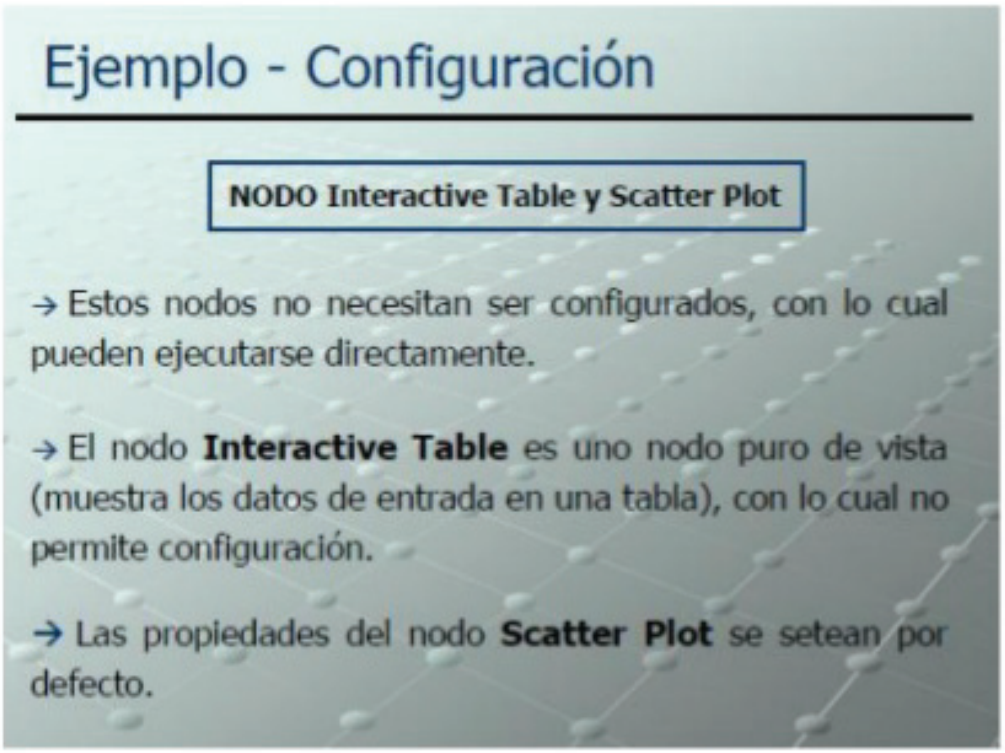
Es necesario configurar previamente el nodo K-Means, para determinar los valores a colorear.

Columna que contiene el atributo que queremos colorear (por defecto el cluster)

Columna con valores nominales (Cluster)

Columna con valores con rango (Sepal Lenght)

Colores para asignar a los datos



Ejecución del flujo de datos.

Cuando todos los nodos del flujo poseen un color amarillo, entonces el mismo puede ser ejecutado.

Los nodos se ejecutan de izquierda a derecha, es decir, que un nodo solo puede ejecutarse si todos sus nodos predecesores en el flujo han terminado su ejecución.

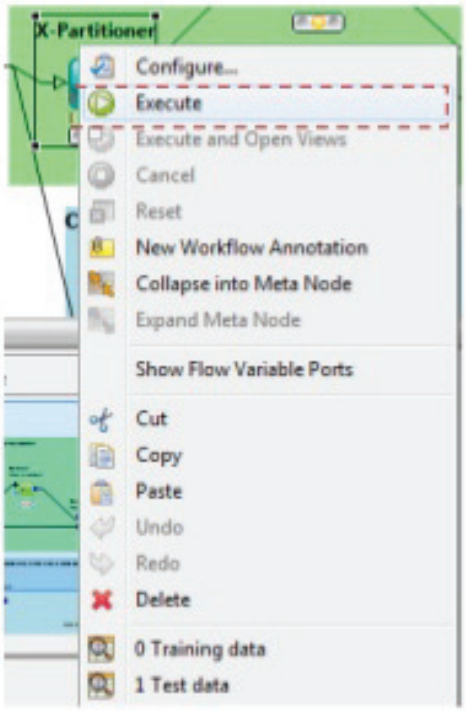
Formas de ejecución:

- Por nodo (eligiendo la opción Execute).
- Ejecutar el ultimo nodo del flujo (KNime ejecuta todos los nodos predecesores).
- Seleccionar varios nodos y disparar la ejecución (KNime determina el orden y ejecuta nodos en paralelo, si es posible).

: Nodo en Ejecución

:Nodo Ejecución con Error

Ejemplo de Ejecución de un nodo: Hacer clic con el botón derecho y seleccionar opción “Execute”.

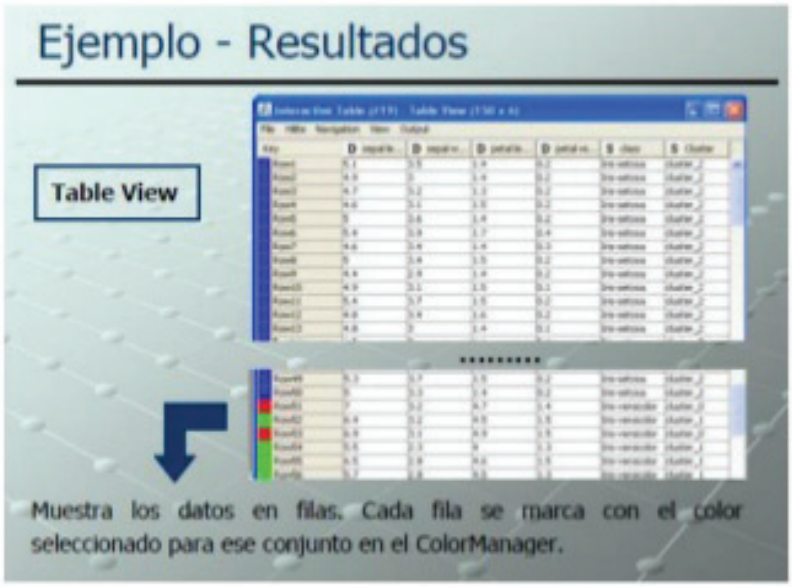


Resultados.

Algunos nodos proveen vistas que muestran sus resultados computacionales. Estas vistas se observan haciendo click con el botón derecho sobre el nodo y eligiendo la opción View.

Se pueden abrir varias instancias de una vista (de un nodo), lo cual permite tener diferentes vistas de los mismos datos.

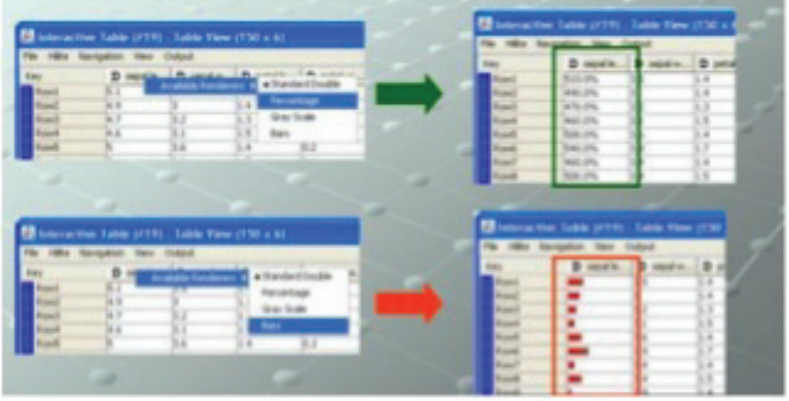
Algunas vistas, solo se muestran si el nodo fue ejecutado.



Ejemplo - Resultados

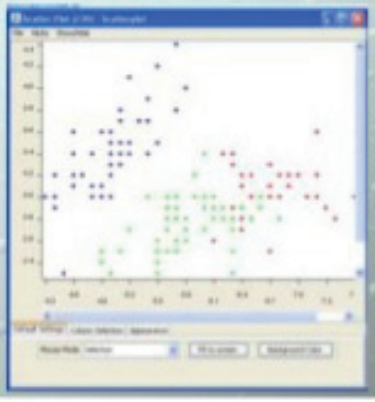
Table View

Renders para las columnas



Ejemplo - Resultados

Scatter Plot View

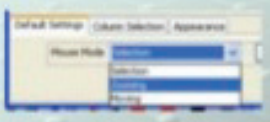


Cada uno de los puntos corresponde a un valor de los datos de entrada, y su color indica el cluster al cual pertenece (si en el ColorManager coloreamos los clusters).

Ejemplo - Resultados

Scatter Plot View

Operaciones sobre el gráfico



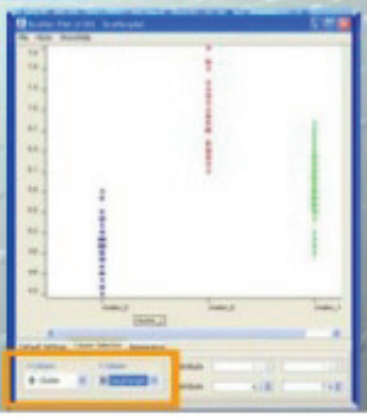
Zooming

Ismael Tapia Vidal (Ismael.tapia.v@gmail.com)

Ejemplo - Resultados

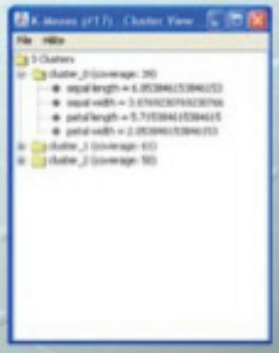
Scatter Plot View

Selección de columnas



Ejemplo - Resultados

Cluster View



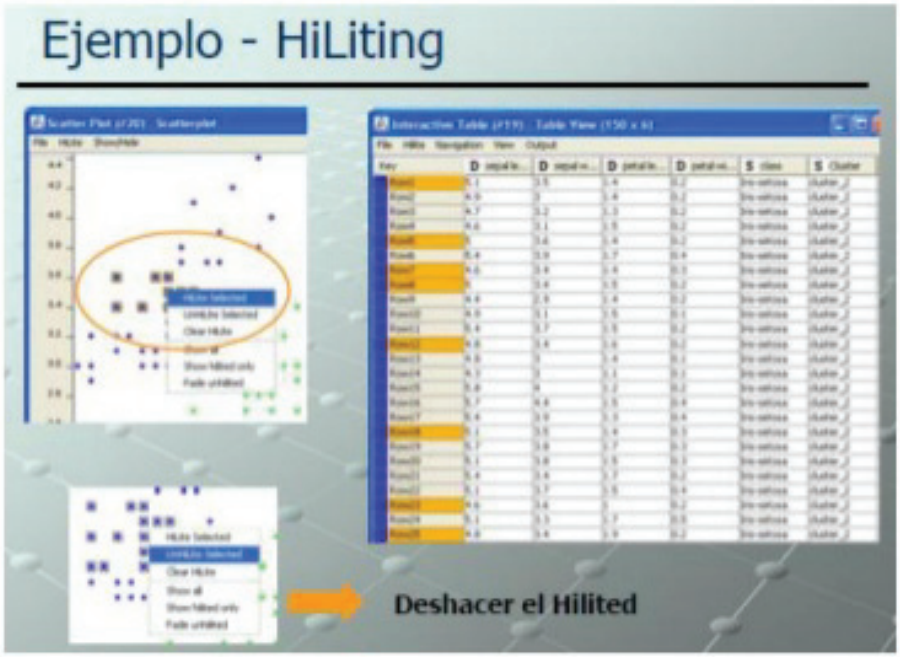
Muestra los cluster construidos por el algoritmo para los datos de entrada y los centroides correspondientes.

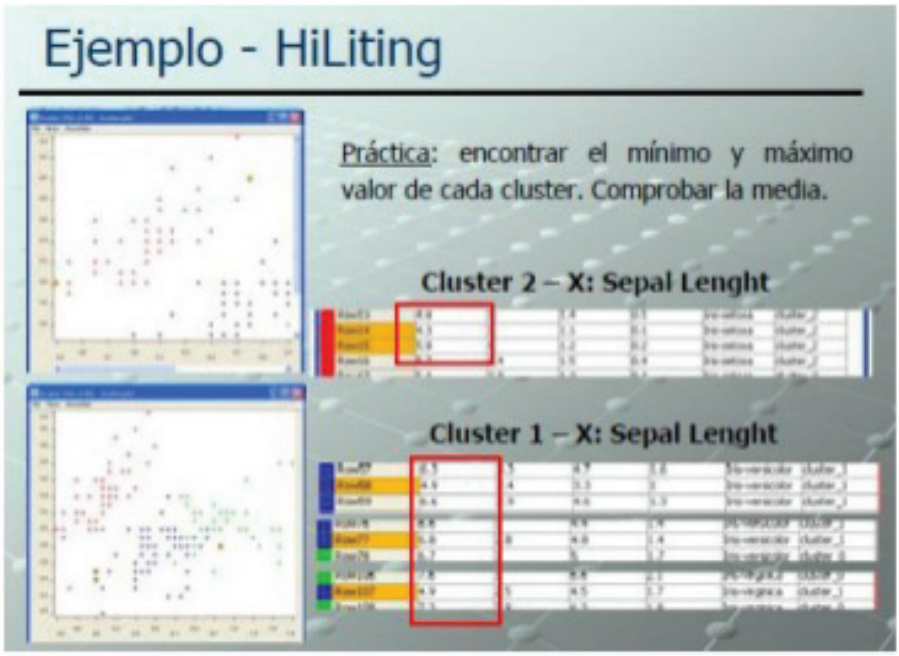
Ismael Tapia Vidal (Ismael.tapia.v@gmail.com)

HiLiting.

Si se seleccionan datos en una vista y se aplica un "hilite" sobre ellos, entonces se puede ver ese mismo efecto de hilite sobre los datos en todas las demás vistas disponibles del workflow que soporten esta función.

Los datos seleccionados y que recibieron el hilite se resaltan con un color naranja en todas las vistas mostrando los mismos datos.

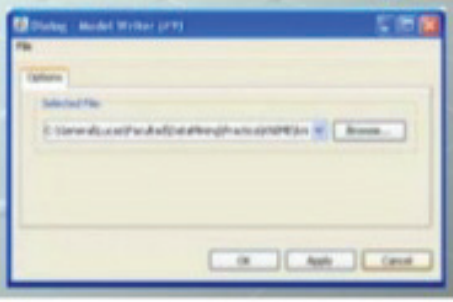




Ismael Tapia Vidal (Ismael.tapia.v@gmail.com)

Ejemplo 2: Nodos de Salida

Los resultados no se muestran, sino que se guardan en **nodos de salida**.

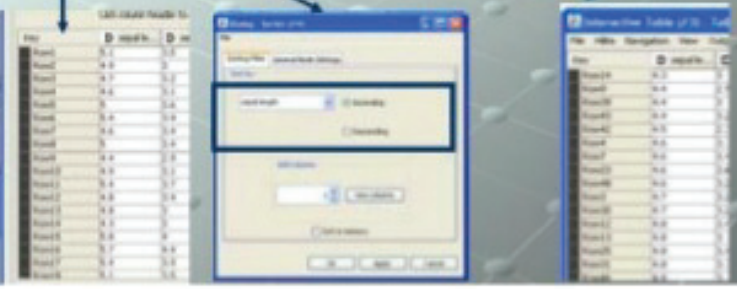


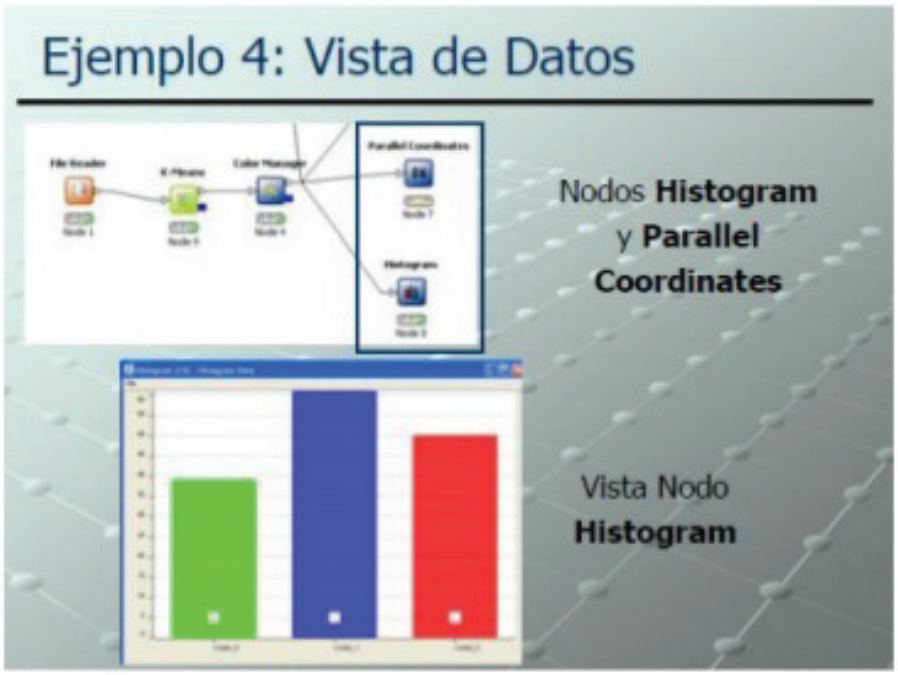
Configuración
Nombre para el archivo donde guardamos los datos

Ejemplo 3: Manipulación de Datos

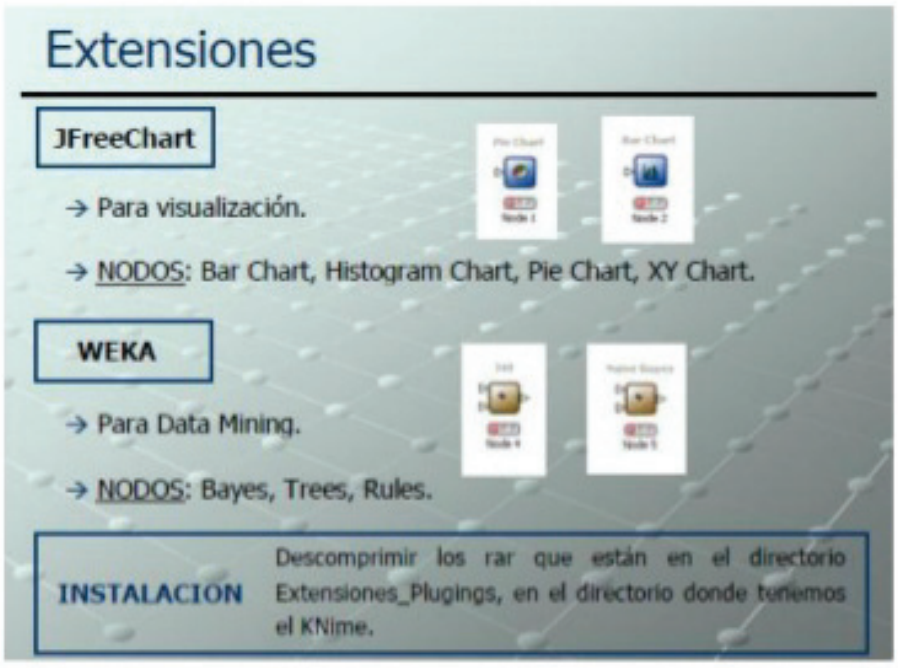
Podemos realizar operaciones sobre los datos. Se encuentran en la categoría **Data Manipulation.**

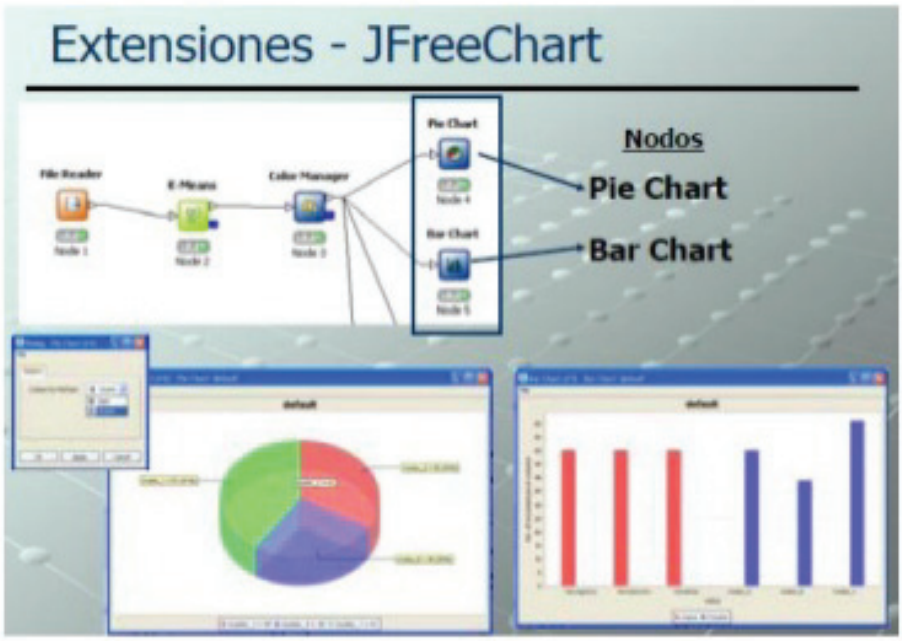
Nodo **Sorter**.
Categoría **Row**.

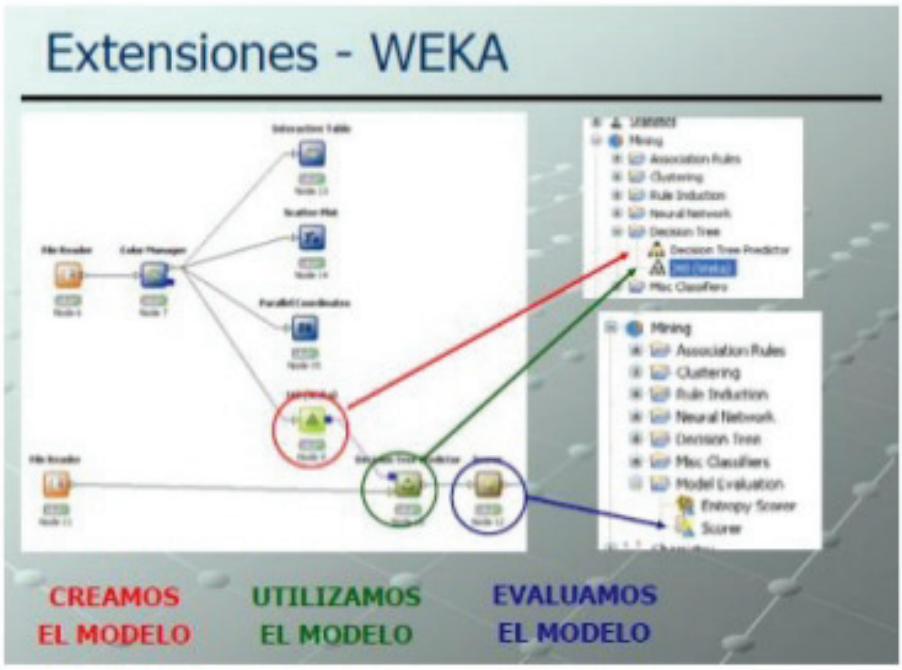




Mas Extensiones



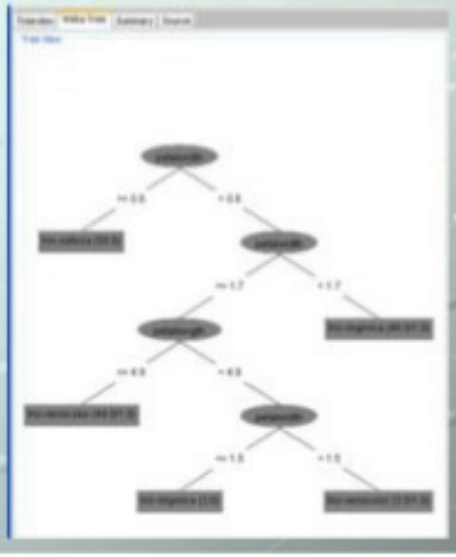

Extensiones - JFreeChart
Nodos
Pie Chart
Bar Chart


Extensiones - WEKA
CREAMOS EL MODELO
UTILIZAMOS EL MODELO
EVALUAMOS EL MODELO

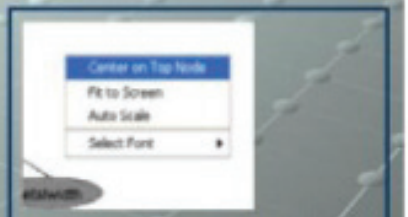
WEKA – J48

VISTA DE ARBOL

WEKA – J48

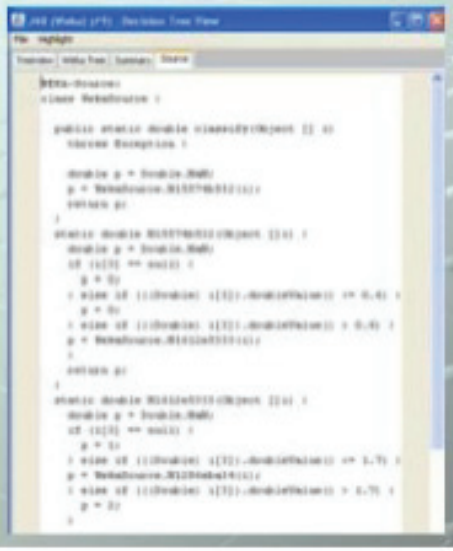


WEKA TREE



Para acomodar el árbol a la pantalla

WEKA – J48



CODIGO DEL ALGORITMO

WEKA – Scorer (Evaluación)

Clasificaciones correctas

Clasificaciones incorrectas