MANUAL DEL USUARIO

MATE-KDD: Una Herramienta Genérica para el Descubrimiento de Reglas de Clasificación Medianamente Acoplada al SGBD PostgreSQL

Desarrolladores

Claudia Milena Castro Rodríguez
Mari Aleyda Cabrera Cabrera

Director

Dr. Ricardo Timarán Pereira, Ph.D.

Universidad de Nariño
Facultad de Ingeniería
Departamento de Sistemas
Grupo de Investigación
GRIASKDD

2007

TABLA DE CONTENIDO

INTRODUCCIÓN

En este manual se encontrará con la información necesaria acerca MATE-KDD: Una Herramienta Genérica para el Descubrimiento de Reglas de Clasificación Medianamente Acoplada al SGBD PostgreSQL, instalación en su sistema.

DE QUÉ TRATA EL PROGRAMA

El programa MATE-KDD, es una herramienta creada para contribuir con el desarrollo de actividades de minería de datos, es decir que con esta herramienta se pretende poner a disposición de los usuarios tres algoritmos para obtener reglas de clasificación a partir de una base de datos diseñada en PostgresSQL, con lo cual puede comparar resultados y rendimiento de cada uno de los algoritmos.

1. INSTALACIÓN DEL PROGRAMA

Para la instalación de MATE-TREE se debe tener en cuenta los pasos que se describen a continuación.

- Se ingresa como usuario root y se desplaza al directorio </usr/local/src/> donde se ubicaran las fuentes de postgresql

  $ su
  # cd /usr/local/src/
  # cp /cdrom/postgresql-KDD.tar.gz.

- Se desempaqueta el archivo y se ingresa en el directorio generado.

  # tar xvfz postgresql-KDD.tar.gz
  # cd postgresql-7.3.4

- Antes de iniciar la instalación se configura el código fuente de la siguiente forma:

  # ./configure

- Para compilar e instalar, se digitan las siguientes secuencias de comandos:

  # gmake
  # gmake install

- Se crea el súper-usuario postgres.

  # adduser postgres

- Se crea el directorio de datos y de generación de archivos de seguimiento, y se le asigna permisos de propietario al usuario postgres.

  # mkdir /usr/local/pgsql/data
  # chown postgres:postgres /usr/local/pgsql/data

- Se inician las bases de datos básicas para el correcto funcionamiento de PostgreSQL y se ejecuta el servidor de PostgreSQL. Para esto, es necesario identificarse como usuario postgres.

  ```
  # su – postgres
  $ /usr/local/pgsql/bin/initdb -D /usr/local/pgsql/data
  $ /usr/local/pgsql/bin/postmaster –i –s -D /usr/local/pgsql/data >logfile 2>&1 &
  ```

- Una vez tenemos instalado y configurado nuestro Postgres, procedemos a instalar La herramienta en el directorio **<</opt/>>**

  ```
  # cd /opt/
  # cp /cdrom/Matekdd.tar.gz.
  ```

- Se desempaqueta el archivo y se ingresa en el directorio generado.

  ```
  # tar –xvfz Matekdd.tar.gz.
  ```

- Antes de ejecutar por primera vez la herramienta se debe cambiar los permisos del directorio **/models** y su contenido.

  ```
  # chmod –777 /opt/Matekdd/models
  ```

Para ejecutar la herramienta se ingresa al directorio **<<.../dist>>**

```
# cd /opt/Matekdd/dist
# java –jar Matekdd.jar
```

2. CONSIDERACIONES DEL SISTEMA OPERATIVO

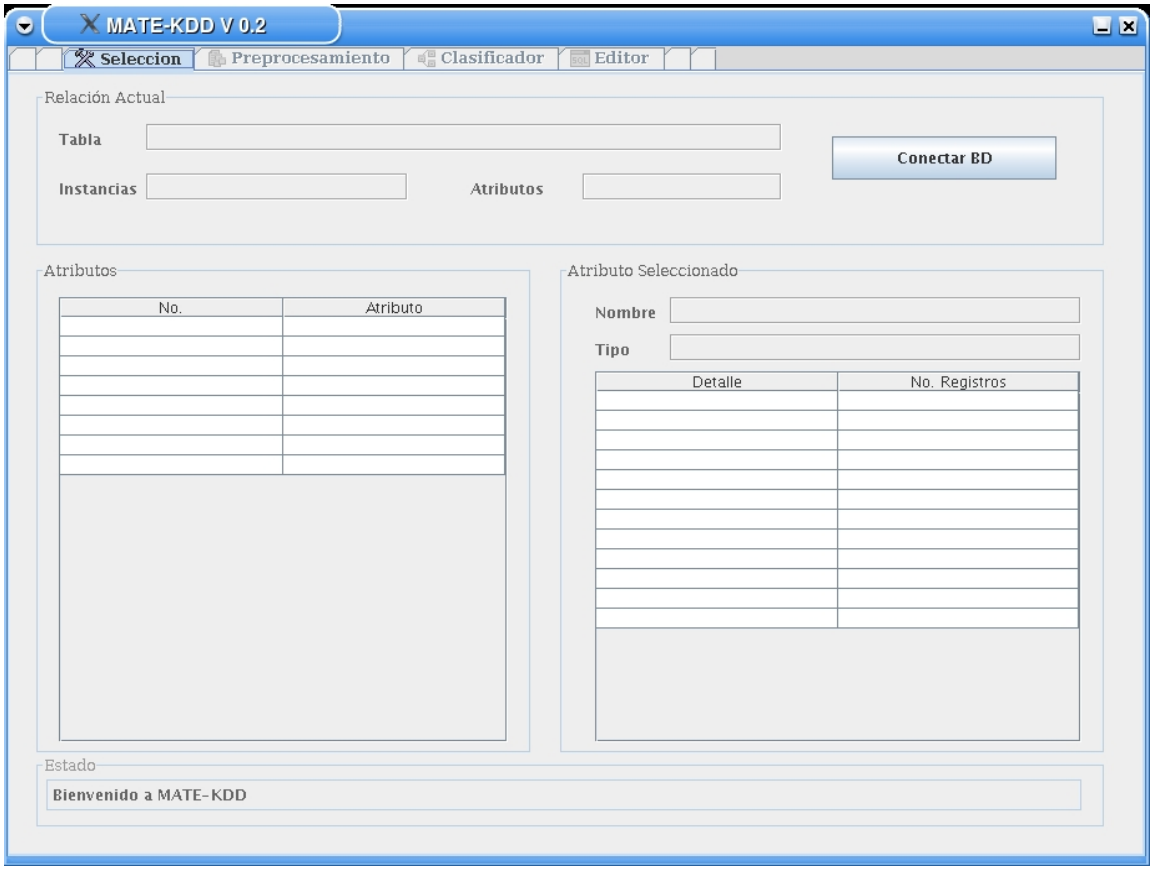
El proyecto se desarrollo bajo filosofía de Software Libre, la herramienta funciona sobre una plataforma Linux, la distribución que se utilizo fue Slakware 10.

Se debe tener instalado Postgresql-7.3.4 que se encuentra incluida en el CD de instalación. En caso de preferir utilizar otra versión de Postgresql es necesario copiar el directorio **"kdd"** al directorio **"contrib"** de la otra versión instalada de Postgresql.

3. AMBIENTE DEL PROGRAMA

Esta herramienta está diseñada para facilitar el proceso de la Minería de datos mediante algoritmos de clasificación como son SLIQ, C4.5 y MATE-TREE.

El diseño, es bastante amigable y predecible para una persona conocedora del proceso de Minería de Datos, ya que la aplicación organiza cada uno de los procesos que se siguen secuencialmente para minar una base de datos, para ello encontramos en una ventana diferente la aplicación de cada fase (selección de la BD, preprocesamiento, modelación con determinado algoritmo y resultados), interfaces fáciles de manejar y perfectamente validadas de tal forma que se puede llevar a cabo el proceso de minería de principio a fin sin dificultad.



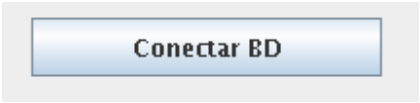
3.1 PESTAÑAS DEL PROGRAMA

En esta parte de la pantalla se visualiza las funciones de la aplicación: Selección, Preprocesamiento, Clasificador y Editor. Cada pestaña permite que naveguemos por la aplicación y en cada item encontramos los pasos principales para minar, empezando por seleccionar la tabla que contiene los datos para procesar, manipulación de la misma, según criterio del analista con elfin de obtener resultados validos, aplicación del modelo para finalmente visualizar las reglas obtenidas y el árbol del modelo. Se implemento una opción importante que permite la interacción directa del analista con el SGBD PostgreSQL, este es el editor.

3.1.1 Opción de Selección

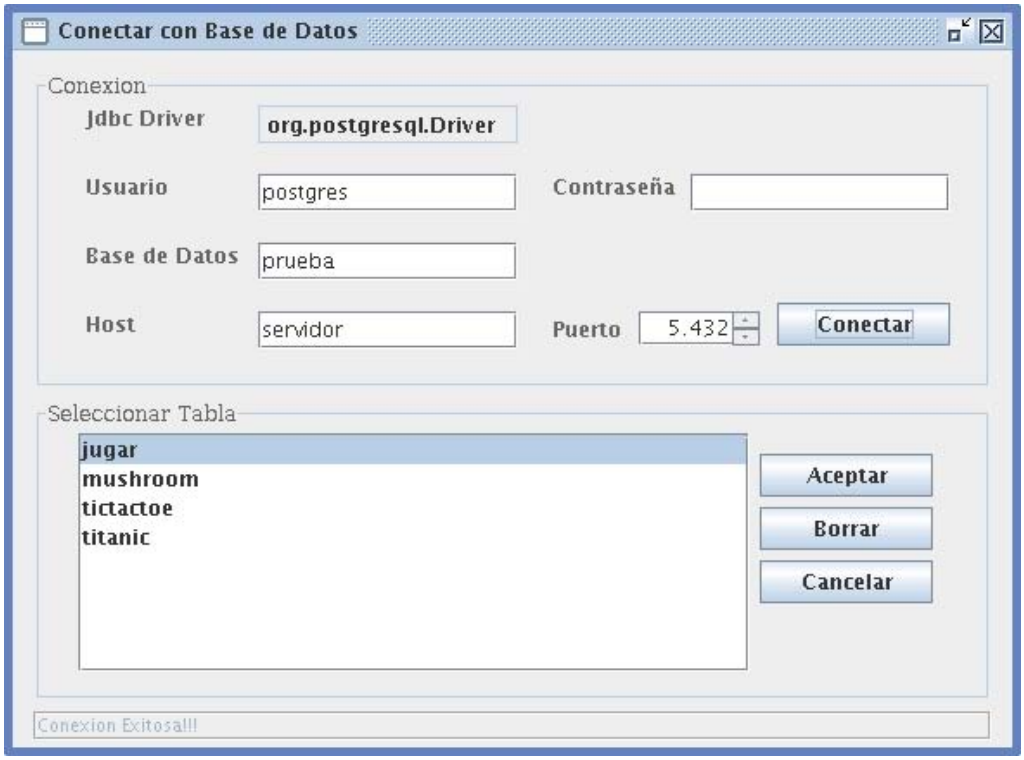
En la opción de Selección podemos conectarnos con el SGBD para ubicar la tabla que vamos a minar.

3.1.1.1 Conexión a la BD



Con este Botón podemos desplegar en la que podemos elegir la Base de Datos que se encuentre Creada en Postgres a la cual accedemos con un nombre de Usuario y una contraseña, el nombre de la base de datos y Host, una vez seguidos estos pasos en el recuadro blanco Rotulado como “Seleccionar Tabla” se visualizan la tablas que tiene creada la base de datos y seleccionamos la

deseada para trabajar. También permite borrar tablas que no queramos o ya no necesitemos de nuestra Base de Datos.

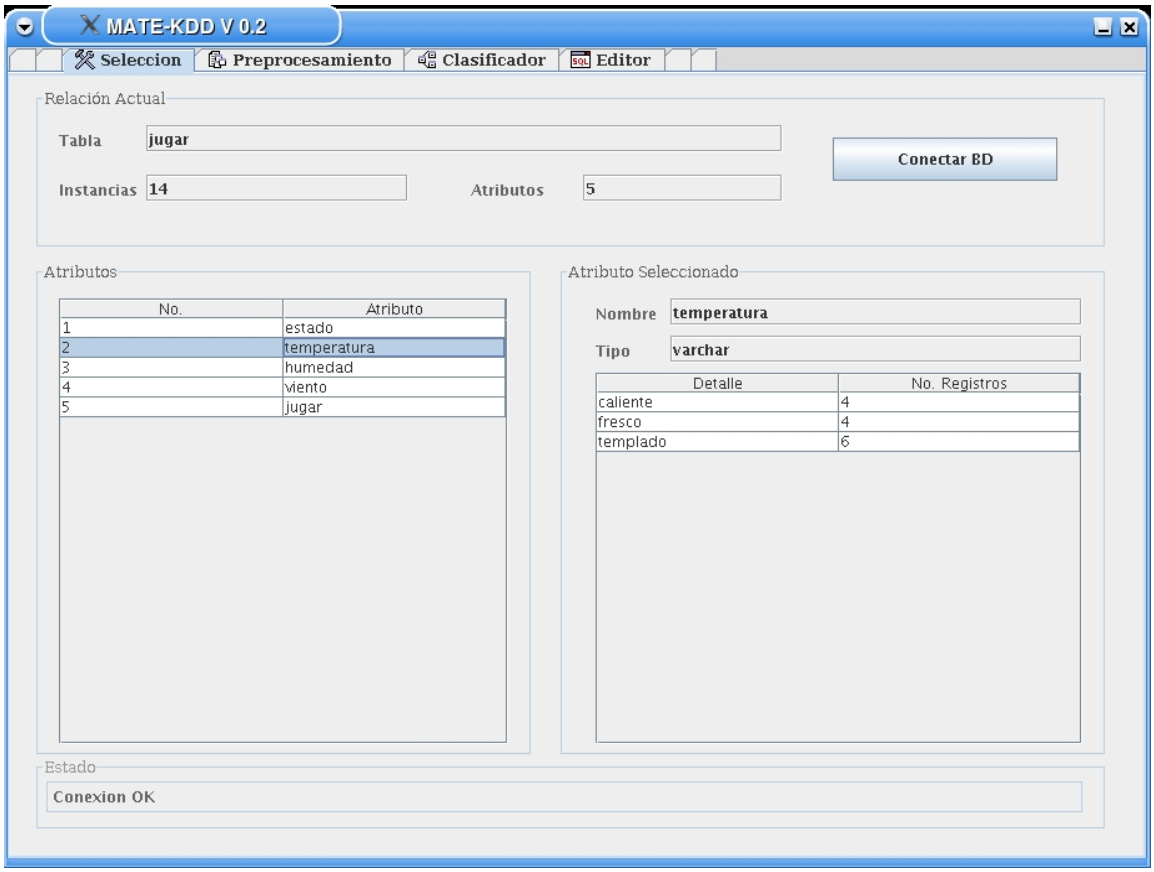


Simplemente haciendo clic sobre el botón *"Aceptar"* obtendremos una descripción detallada de la tabla, sus atributos, tipos de datos. Tenemos dos secciones en la ventana principal en la que podemos visualizar la información de los datos con el fin de conocer sus características para empezar a trabajar y determinar si vamos a aplicar algún cambio en la opción de preprocesamiento.

En el área de trabajo podemos diferenciar la siguiente información:

Relación Actual	Muestra el nombre de la tabla, así como el número de atributos e instancias que la componen. También esta los botones para abrir o guardar una tabla.
Atributos	Aquí se muestran los nombres de los atributos de la tabla.
Atributo Seleccionado	Al seleccionar con un clic un atributo de *"Atributos"* en esta parte se muestra las características de ese atributo.

Podemos observar que en la ventana se detalla la tabla seleccionada, sus atributos, tipos de datos, cantidades.

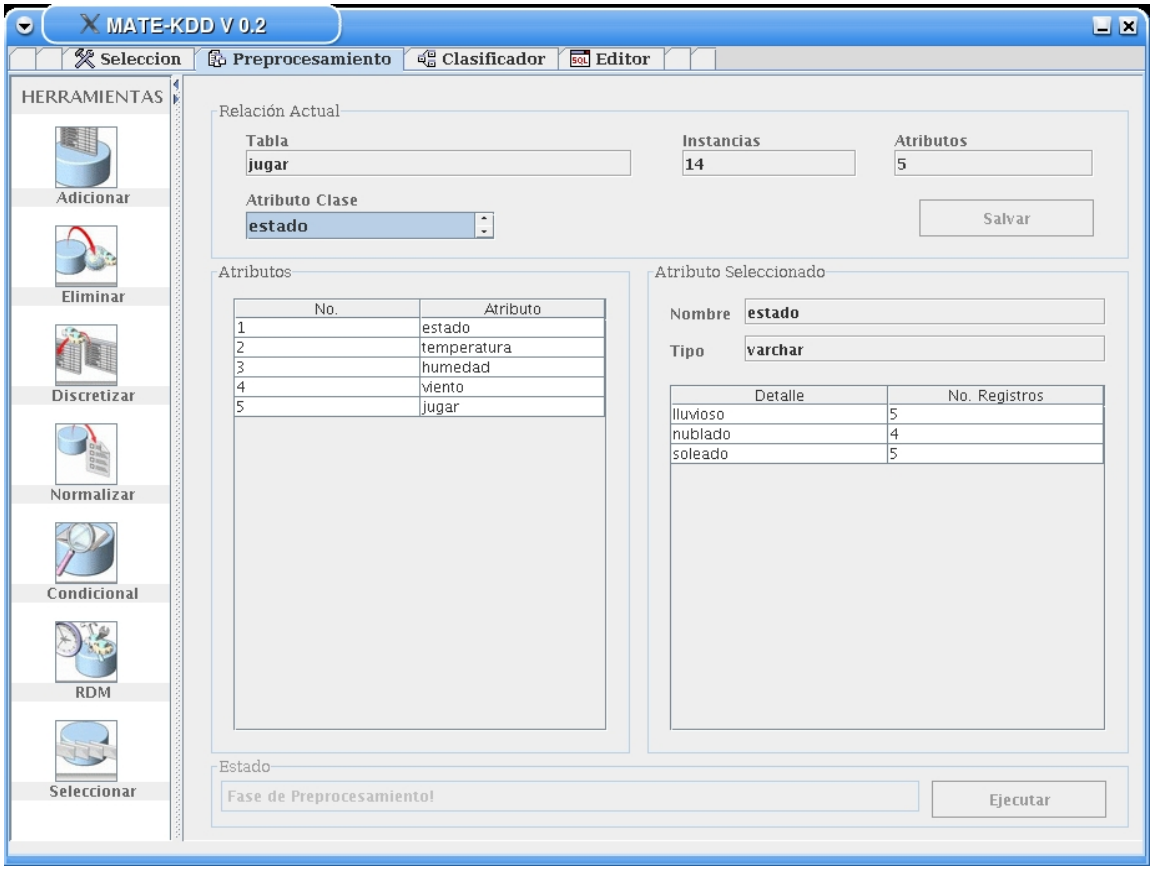


3.1.2 Opción Preprocesamiento

Mate-kdd permite aplicar en la etapa de preprocesamiento algunas técnicas de depuración y discretización que son indispensables para poder procesar los datos y de esta forma obtener unos patrones confiables.

Se cuenta con algunos filtros que permiten manipular la tabla a criterio del analista para depurar datos ruidosos, Discretizar atributos numéricos que es muy importante según el algoritmo que deseemos aplicar, esto tonel fin de realizar transformaciones a partir de la tabla original de la forma en que el analista vea pertinente, una vez se aplica cualquiera de los filtros se crea una tabla con la

nueva configuración a la que podemos dar un nombre. Es en esta fase en donde se debe seleccionar el atributo clase.



Adicionar: Esta herramienta permite adicionar un atributo a la tabla, éste puede ser el resultado de una operación matemática entre dos atributos de la misma o un atributo con un número.

Normalizar: Esta herramienta permite normalizar los datos numéricos que se encuentran en base 10 a rangos entre 0 y 1.

Discretizar: Con esta Herramienta podemos Discretizar un conjunto de atributos numéricos en rangos de datos.

Condicional: mediante esta herramienta podemos filtrar la tabla con una condición booleana (***or*** o ***and***).

RDM: Nos permite seleccionar ramdomicamente un número de datos de la tabla original, el analista especifica cuantos datos quiere

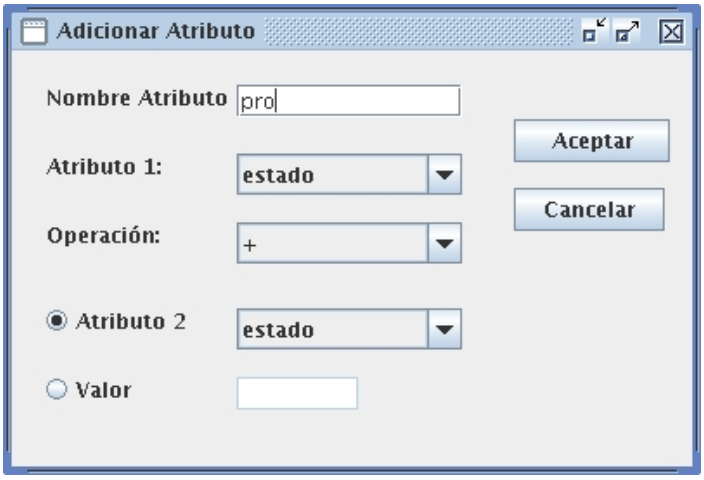
Eliminar: Podemos también eliminar un atributo de la tabla que vamos a minar cuando seleccionamos esta Herramienta.

Seleccionar: Permite seleccionar determinados atributos de una tabla para trabajar con ella.

Cuando tenemos seleccionada la tabla para minar podemos entonces aplicar alguno de los filtros antes mencionados dando clic en el botón “Ejecutar”.

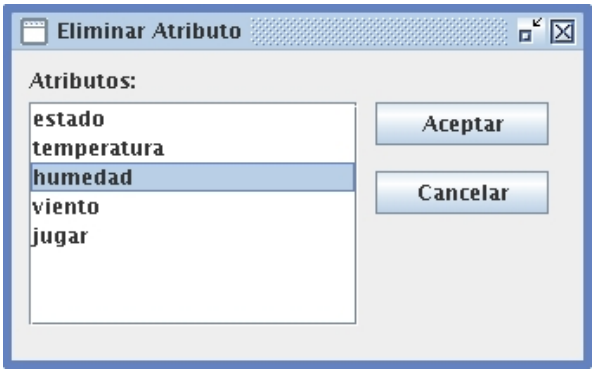
3.1.2.1 Adicionar

En esta ventana podemos dar un nombre al atributo que vamos a adicionar, seleccionamos el primer atributo la operación matemática que aplicaremos y el segundo atributo o un valor para efectuar el proceso. Al dar clic en el botón “Aceptar”, se construye la instrucción SQL para su posterior ejecución.



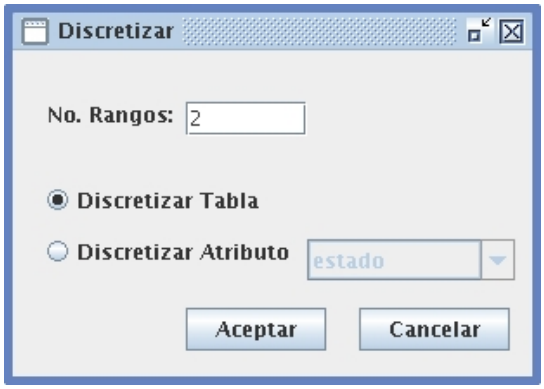
3.1.2.2 Eliminar

En la ventana de Eliminar atributo podemos seleccionar el atributo que deseamos eliminar de la tabla Al dar clic en el botón “Aceptar”, se construye la instrucción SQL para su posterior ejecución.



3.1.2.3 Discretizar

Para Discretizar una tabla o un atributo debemos tener en cuanta que los atributos deben ser numéricos para discretizarse; introducimos el número de rangos que deseamos para formar los grupos de datos que no puede ser mayor que 10, seleccionamos “Discretizar Tabla” o “Discretizar Atributo” según necesitemos y aplicamos el proceso.

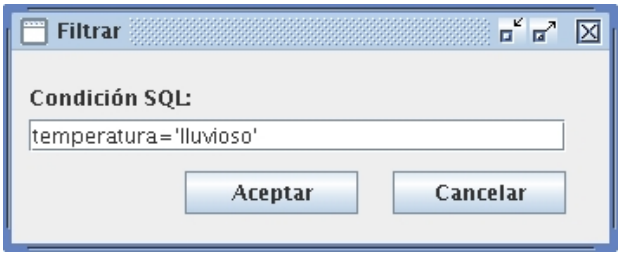


3.1.2.4 Normalizar

Para Normalizar una tabla o un atributo procedemos igual que para Discretizar, seleccionamos si el proceso se va a realizar sobre la tabla o sobre un atributo para seleccionar.

3.1.2.5 Condicional

Esta opción nos permite hacer un filtro muy especial, podemos condicionar un atributo mediante una condición lógica.



3.1.2.6 Seleccionar

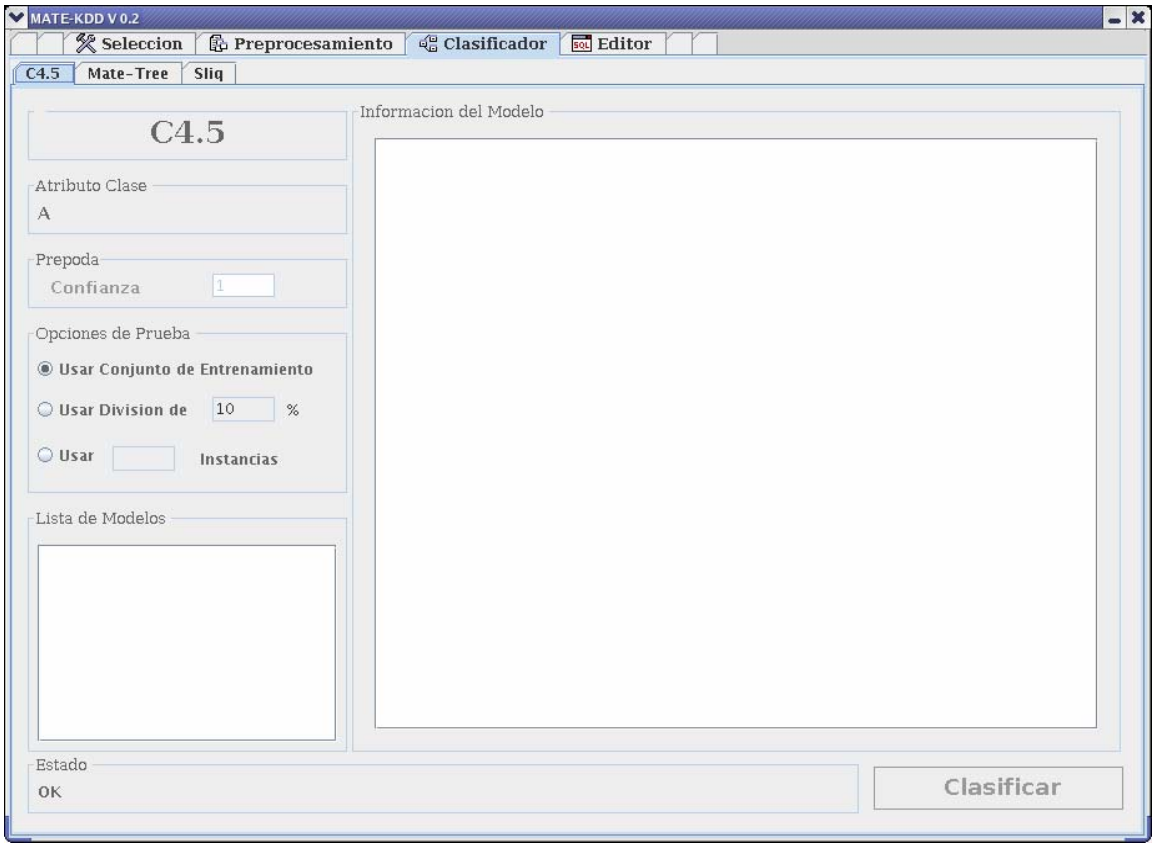
Esta opción permite escoger los atributos que el analista desee de una tabla para clasificar solo con ellos.

3.1.3 Opción Clasificador

Aquí tenemos la posibilidad de seleccionar el algoritmo C4.5, Mate-Tree o Sliq para aplicarlo.

3.1.3.1 C4.5

Al seleccionar la pestaña C4.5 aparece esta pantalla en donde se presenta el proceso de clasificación utilizando el algoritmo C4.5.



Cuando nos disponemos a modelar con cualquiera de los tres algoritmos implementados podemos configurar el los conjuntos para entrenamiento y prueba del modelo:

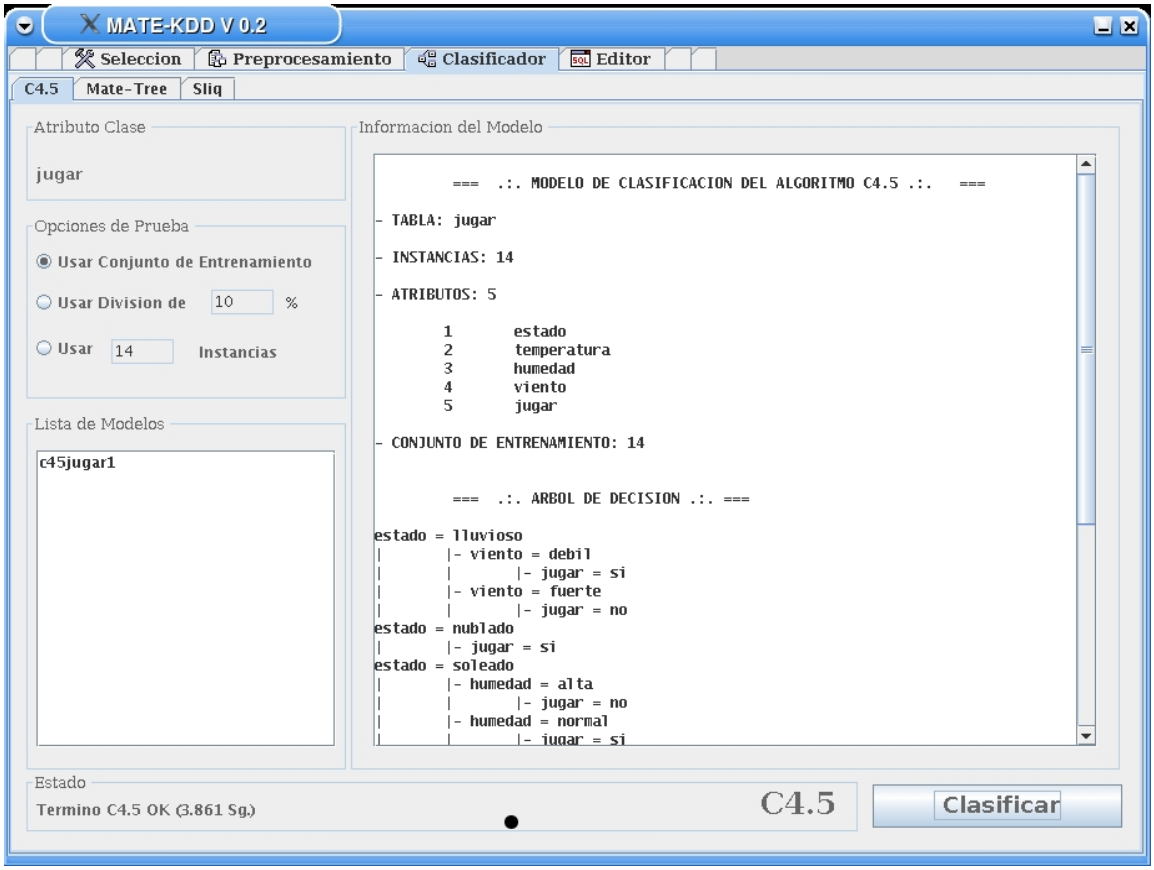
- *Usar conjunto de entrenamiento:* De esta forma se utiliza todo el conjunto de datos para la construcción y para la prueba del modelo.

- *Usar Division de X %:* De esta forma se divide el conjunto de datos para entrenamiento y prueba.

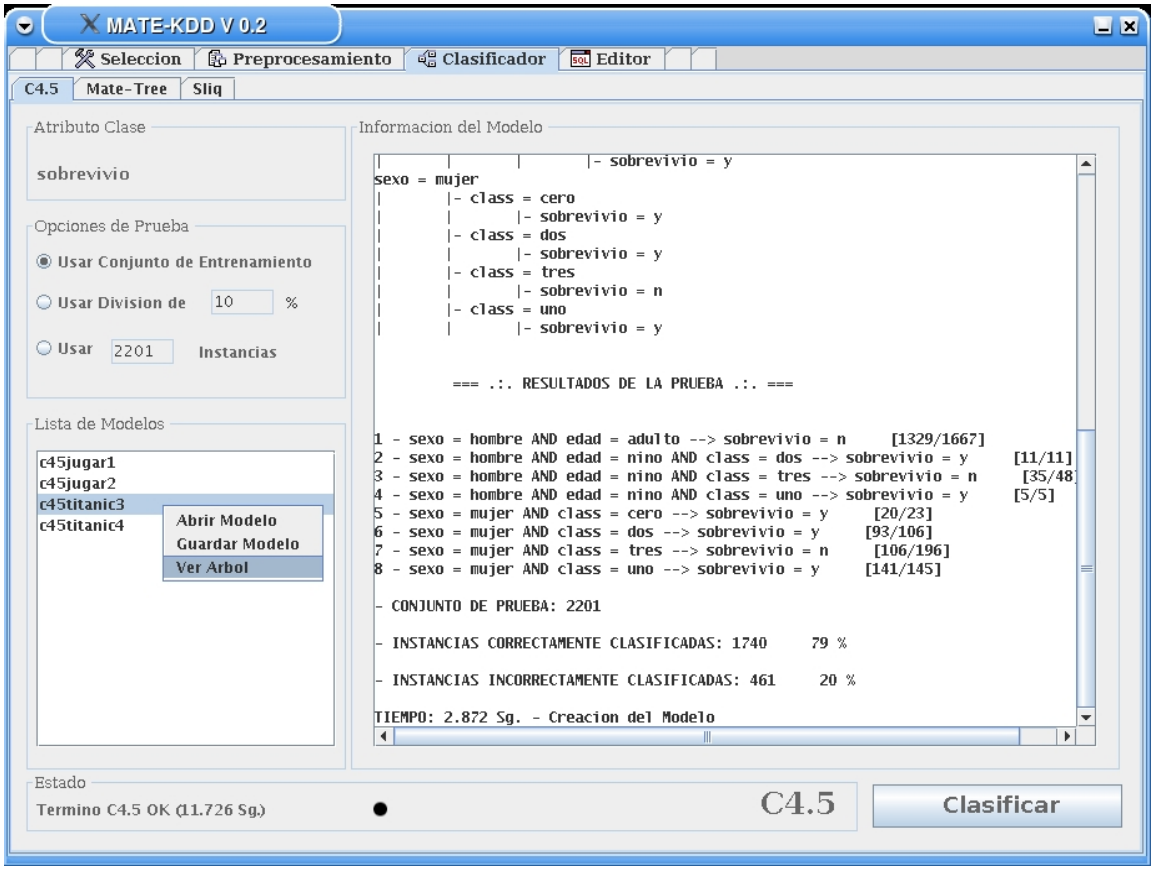
- *Usar X Instancias:* De esta forma se define cuantos registros van a componer el conjunto de datos para la prueba del modelo.

Una vez se haya configurado la opción de prueba, se inicia el modelaje con clic sobre el botón “Clasificar”.

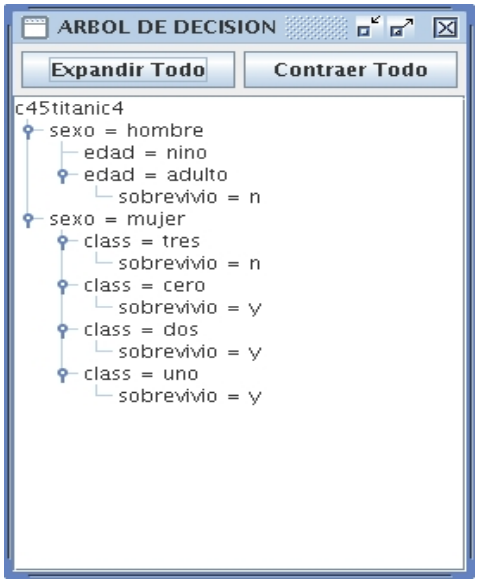
El modelo creado se visualiza en el recuadro blanco al lado derecho de la pantalla, y en el recuadro pequeño al lado izquierdo de la ventana se muestra en detalle el modelo.



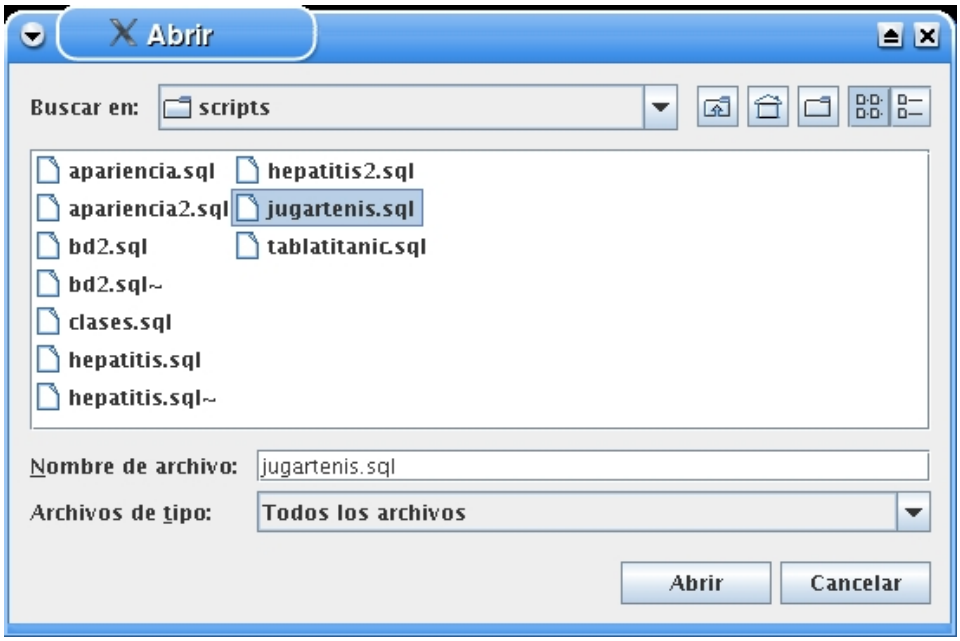
Podemos guardar el modelo dando clic derecho sobre el nombre que aparece por defecto en el recuadro “Lista de modelos”.



Otra opción es ver árbol, cuando la seleccionamos podemos visualizar en la pantalla el árbol del modelo de clasificación.

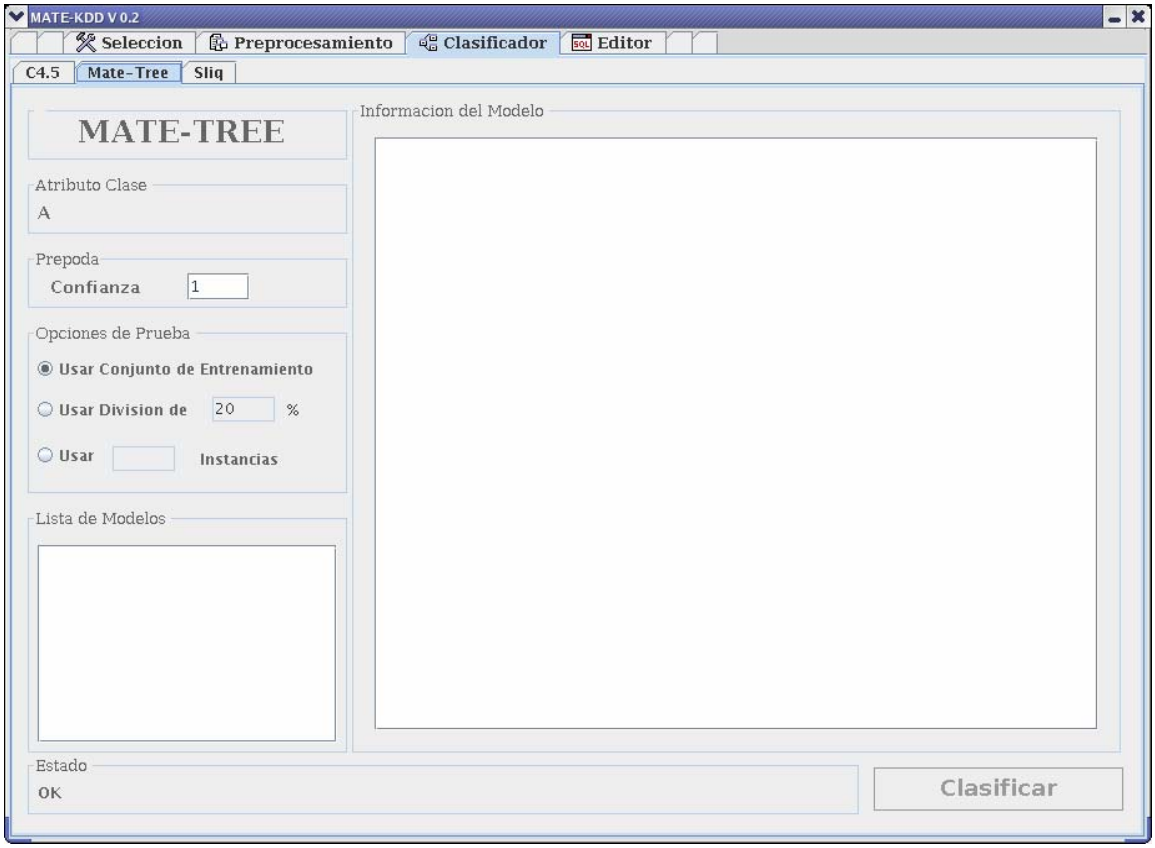


La opción “Abrir Modelo” permite escoger y visualizar en la pantalla un modelo que ya se haya creado y guardado anteriormente.



3.1.3.2 Mate-Tree

En la pestaña Mate-Tree tenemos que configurar la tabla para clasificar los datos para entrenamiento y prueba del modelo. Para aplicar el modelo los datos deben estar discretizados en caso de que hallan atributos numéricos, con el botón *“clasificar”* aplicamos el algoritmo.

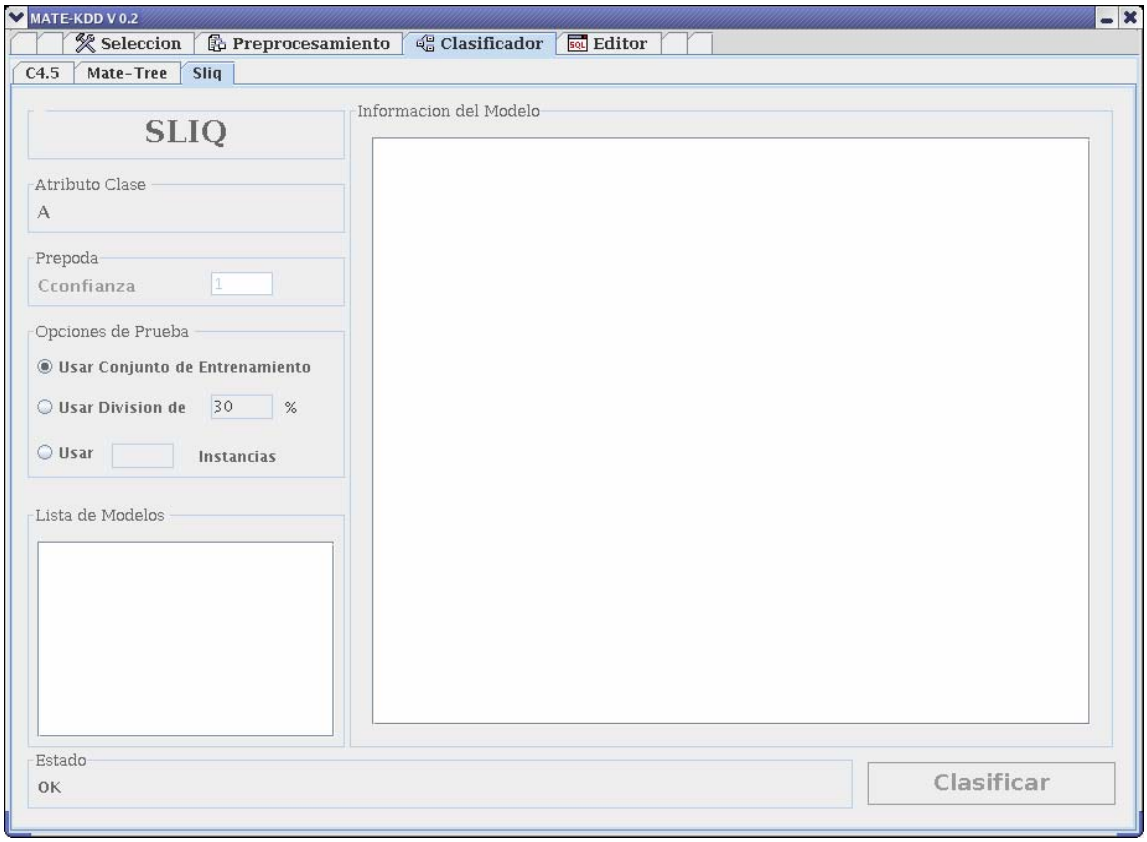


Podemos guardar el modelo, visualizar el árbol y si ya tenemos modelos guardados anteriormente podemos abrirlos y revisar.

3.1.3.3 **Sliq**

Al seleccionar la pestaña Sliq nos encontramos con la pantalla en donde se presentan los diferentes procesos que permiten el proceso clasificación utilizando el algoritmo Sliq.

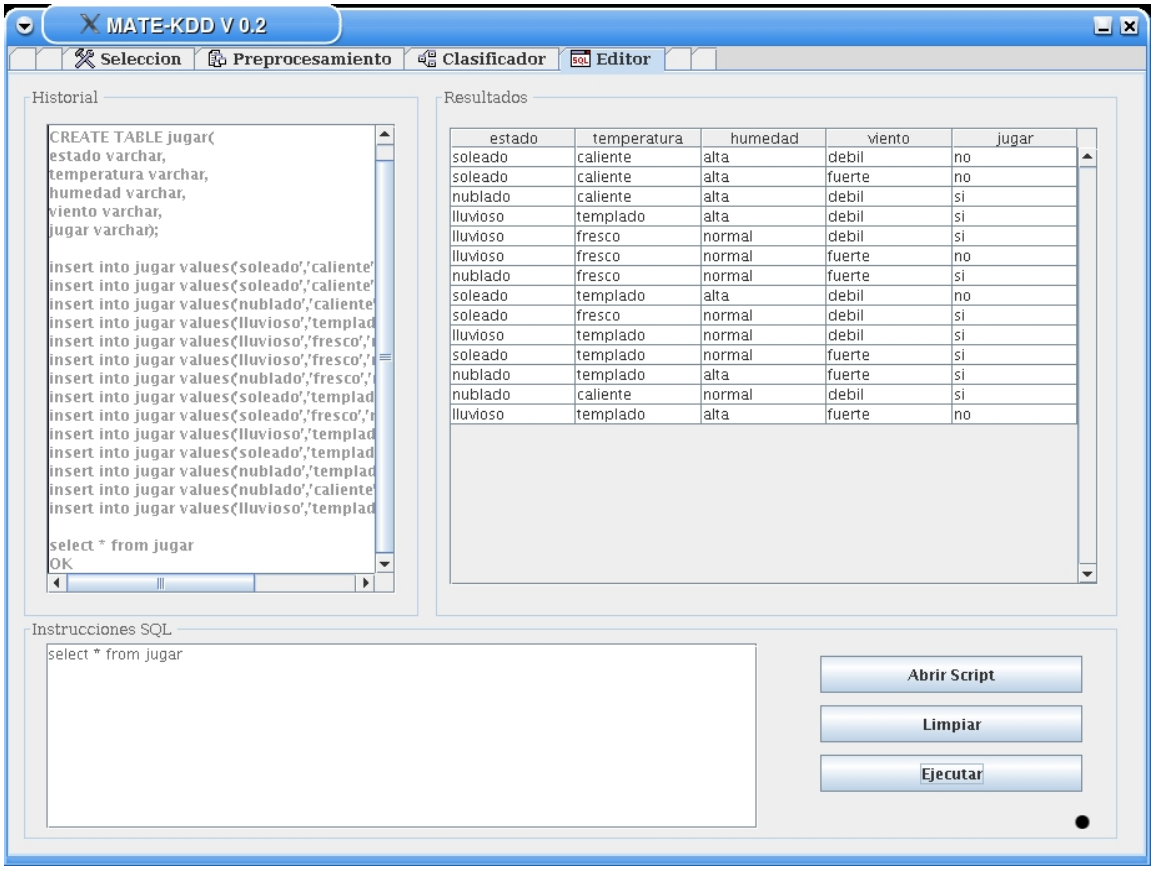
Para este algoritmo no es necesario realizar una discretización previa de los atributos numéricos a menos que el analista lo considere necesario, ya que Sliq trabaja con atributos numéricos, Y se da clic en el botón “Clasificar” para inicial el modelaje.



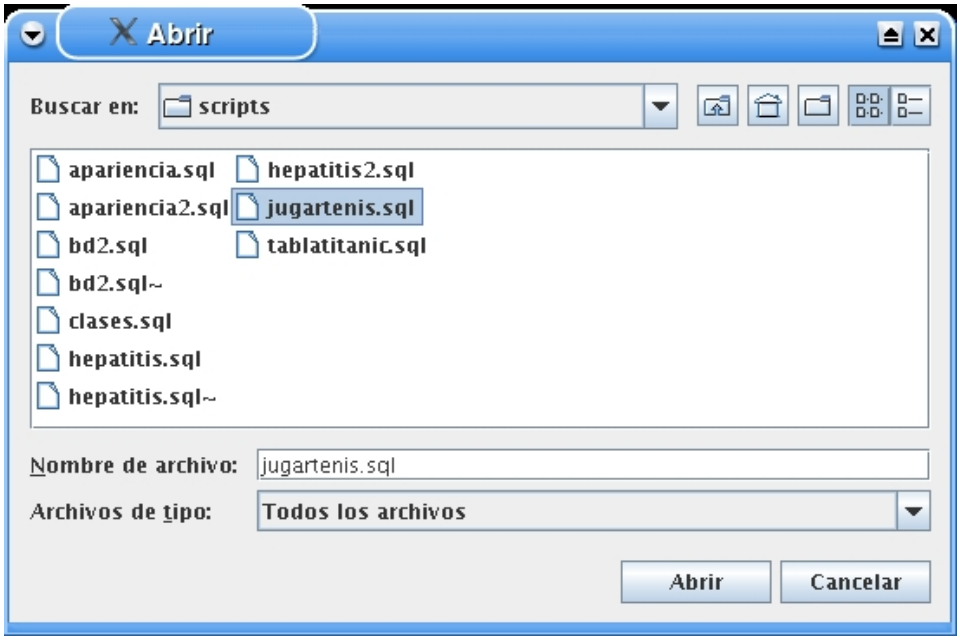
Los modelos creados los podemos guardar, abrir y mirar el árbol de cada modelo, igual que en los anteriores algoritmos, esto con el fin de visualizar los diferentes modelos resultado de aplicar cada algoritmo a una tabla en particular.

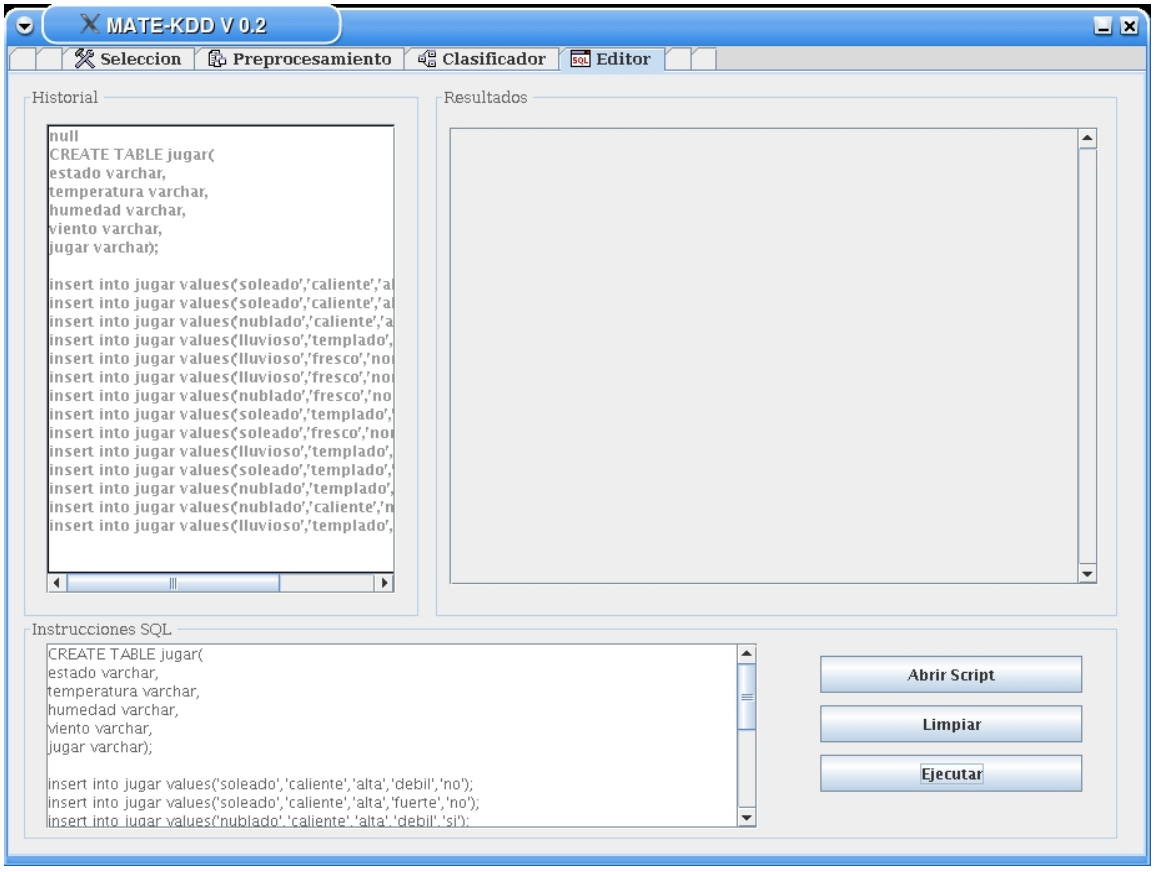
3.1.4 Editor

El editor es una interfaz que nos permite ejecutar sentencias SQL en el SGBD Postgresql, y tiene un àrea en la que visualizamos el resultado de la instrucción si es una consulta.



Otra opción muy útil en el editor es la de abrir y ejecutar Scripts, así que no se tiene que ir a la Base de datos Postgres pues desde la aplicación podemos crear tablas y cargar datos para comenzar a utilizar la aplicaron.


Abrir
Buscar en:
scripts
apariencia.sql
apariencia2.sql
bd2.sql
bd2.sql~
clases.sql
hepatitis.sql
hepatitis.sql~
hepatitis2.sql
jugartenis.sql
tablatitanic.sql
Nombre de archivo:
jugartenis.sql
Archivos de tipo:
Todos los archivos
Abrir
Cancelar


MATE-KDD V 0.2
Seleccion
Preprocesamiento
Clasificador
Editor
Historial
null
CREATE TABLE jugar(
estado varchar,
temperatura varchar,
humedad varchar,
viento varchar,
jugar varchar);
Resultados
Instrucciones SQL
CREATE TABLE jugar(
estado varchar,
temperatura varchar,
humedad varchar,
viento varchar,
jugar varchar);
insert into jugar values('soleado','caliente','alta','debil','no');
insert into jugar values('soleado','caliente','alta','fuerte','no');
Abrir Script
Limpiar
Ejecutar

Como miramos a lo largo de este manual, la aplicación es fácil de utilizar debido a su interfaz amigable, podemos seguir cada paso del proceso de Minería en detalle y hacer una seguimiento hasta el final, además que con el editor podemos interactuar con el SGBD desde la aplicación sin tener que ir a Postgres si tenemos los scripts de las tablas que queremos minar.

ANEXO

INSTRUCCIONES DE INSTALACION DE PostgreSQL Y LAS FDU

Este anexo describe la instalación del código fuente de PostgreSQL, las funciones FDU y la forma correcta de usarlas.

PROCESO DE INSTALACIÓN DE POSTGRESQL

- Se ingresa como usuario root y se desplaza al directorio </usr/local/src/> donde se ubicaran las fuentes de postgresql

  $ su
  # cd /usr/local/src/
  # cp /cdrom/postgresql-7.4.3.tar.gz.

- Se desempaqueta el archivo y se ingresa en el directorio generado.

  # tar -xvfz postgresql-7.4.3.tar.gz
  # cd postgresql-7.3.4/

- Antes de iniciar la instalación se configura el código fuente de la siguiente forma:

  # ./configure

- Para compilar e instalar, se digitan las siguientes secuencias de comandos:

  # gmake
  # gmake install

- Se crea el súper-usuario postgres.

  # adduser postgres

- Se crea el directorio de datos y de generación de archivos de seguimiento, y se le asigna permisos de propietario al usuario postgres.
  # mkdir /usr/local/pgsql/data

  # chown postgres:postgres /usr/local/pgsql/data

- Se inician las bases de datos básicas para el correcto funcionamiento de PostgreSQL y se ejecuta el servidor de PostgreSQL. Para esto, es necesario identificarse como usuario postgres.

  # su – postgres

  # /usr/local/pgsql/bin/initdb -D /usr/local/pgsql/data

  # /usr/local/pgsql/bin/postmaster –s –i -D /usr/local/pgsql/data

- Se crea la base de datos a utilizar, se crea el lenguaje plpgsql en la base de datos y se inicia la terminal interactiva de postgresql.

  # /usr/local/pgsql/bin/createdb <nombre_base_datos>

  # /usr/local/pgsql/bin/createlang plpgsql <nombre_base_datos>

  # /usr/local/pgsql/bin/psql <nombre_base_datos>

CREACIÓN DE LAS FUNCIONES FDUs

- Como usuario postgres se inicia la terminal interactiva de postgresql y se crean las funciones dentro de la base de datos

  # /usr/local/pgsql/bin/psql <nombre_base_datos>

  # \i /usr/local/src/postgresql-7.3.4/contrib/kdd/Functions/matekdd.sql

OBTENER Y VALIDAR EL MODELO C4.5

select * from trans_c45_mate('<nombre_tabla>', '<nombre_att_clase>', <rango_ discretizacion>,<num_instacias_prueba>);

select * from c45('<nombre_att_clase>');

select * from test_rules('c45');

select * from translate_rules(<rango_discretizacion>, 'c45', '<nombre_att_clase>');

select * from arm_rules('<nombre_att_clase>', 'c45');

OBTENER Y VALIDAR EL MODELO MATE-TREE

select * from trans_c45_mate('<nombre_tabla>', '<nombre_att_clase>', <rango_ discretizacion>,<num_instacias_prueba>);

select * from matetree('<nombre_att_clase>');

select * from test_rules('mate');

select * from translate_rules(<rango_discretizacion>, 'mate', '<nombre_att_clase>');

select * from arm_rules('<nombre_att_clase>', 'mate');

OBTENER Y VALIDAR EL MODELO SLIQ

select * from trans_sliq(‘<nombre_tabla>’, ’<nombre_att_clase>’, <num_instacias_prueba>);

select * from sliq(‘<nombre_att_clase>’);

select * from test_sliq(‘<nombre_att_clase>’);

select * from arm_rules(‘<nombre_att_clase>’,’sliq’);