YACHAY

DATA MINING TOOL

MANUAL DE USUARIO

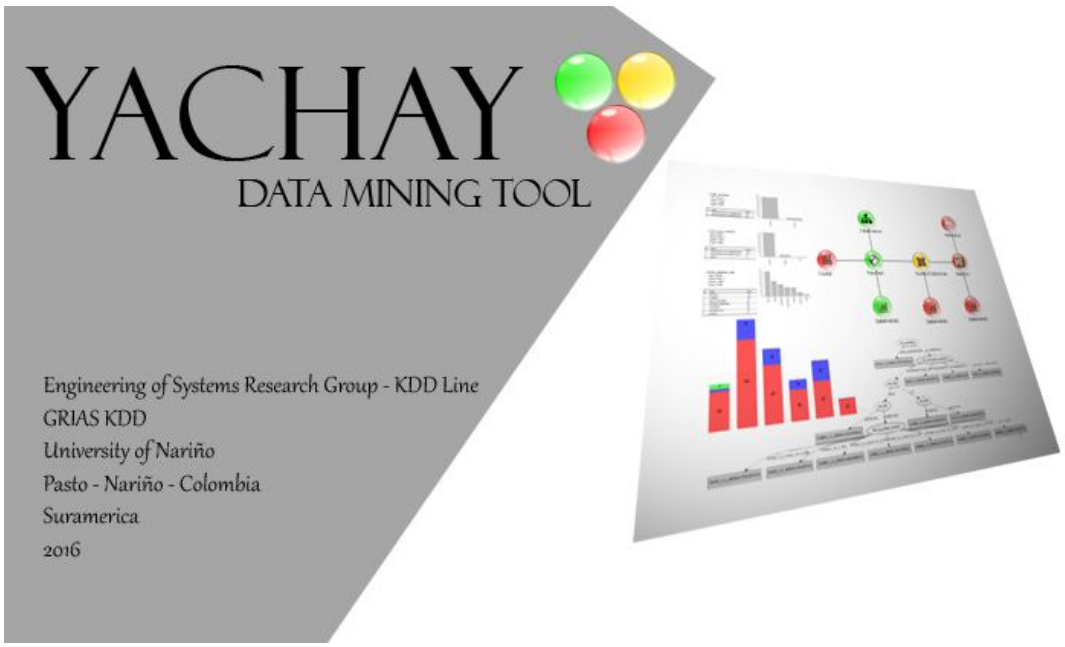


TABLA DE CONTENIDO

1. GENERALIDADES

1.1. INTRODUCCION

En este documento se presenta la primera versión de YachayDTM, una herramienta web para el descubrimiento de conocimiento en bases de datos.

Una herramienta para el descubrimiento de conocimiento en base de datos integra diferentes componentes (limpieza, transformación, análisis, visualización) que permiten extraer patrones interesantes y útiles de los datos suministrados.

Los usos son variados como pueden ser: relaciones entre síntomas y enfermedades, perfiles de estudiantes según características socio económicas, patrones de compra de los clientes, entre muchas otras.

YachayDTM es una herramienta desarrollada en el laboratorio de KDD del departamento de Ingeniería de Sistemas de la Universidad de Nariño (Colombia), está compuesta por cinco módulos:

- Data Source: Permite la recuperación de datos desde archivos planos y bases de datos relacionales.
- Data Saver: Permite la exportación y descarga de archivos en los formatos csv y arff.
- Filters: utilidades que permiten realizar los procesos de selección limpieza y transformación de datos.
- Data Mining: Algoritmos de minería de datos para las tareas de Asociación, Clasificación y Clusters
- View: Tareas de Análisis de datos y Ranking de atributos

El uso de YachayDTM se realiza de manera visual e intuitiva mediante la creación de en un grafico jerárquico que lo arma el usuario dependiendo de las necesidades mediante la conexión de diferentes componentes.

1.2. A QUIÉN VA DIRIGIDO EL MANUAL

Este manual va dirigido a toda persona con deseo de conocer una herramienta que le permita de manera compresible e intuitiva el descubrimiento de conocimiento en bases de datos; Proceso que se realiza mediante la creación de gráficos compuestos por nodos conectados y que siguen una jerarquía determinada que permiten orden y entendimiento, desde la carga de datos hasta la visualización de resultados.

1.3. GLOSARIO DE TERMINOS

- **Árbol**: Es un gráfico que imita la forma de un árbol (conjunto de nodos conectados) para nuestro caso cada uno de los nodos corresponde a cualquiera de los componentes del menú principal, estos pueden ser guardados y cargados mediante su almacenamiento en un archivo plano.
- **Componente o Nodo**: Corresponde a cada una de los componentes del menú principal, y que pueden ser arrastrados al área de trabajo generando un círculo que lo representa y permite su configuración, ejecución o conexión con otros componentes o nodos del árbol.
- **Área de trabajo**: Es la sección de la pantalla que nos permite agregar y conectar componentes del menú principal formando un árbol que nos permite realizar las operaciones del sistema.

2. PROCESO DE INSTALACION

2.1. REQUISITOS DEL SISTEMA

La aplicación YachayDTM es una aplicación web la cual se instala en un equipo servidor y se accede esta aplicación desde uno o más equipos clientes a través de un explorador web conectados en una red (intranet ó internet).

Características mínimas y software servidor:

Glassfish Server 4.1
JAVA JDK 1.7
Explorador Google Chrome
Mínimo: Memoria de 8 GB
Mínimo: Procesador Intel core I3 (Similar ó superior)

Características mínimas Clientes:

Explorador Google Chrome

Paso adicional en Windows

Algo adicional que se debe realizar si la instalación se realiza en Windows es copiar la carpeta "graphvizWin" contenida en el CD de instalación en la ruta c:/
En Linux no es requerido ya que por defecto viene instalada la aplicación DOT, que se utiliza para mostrar Grafos

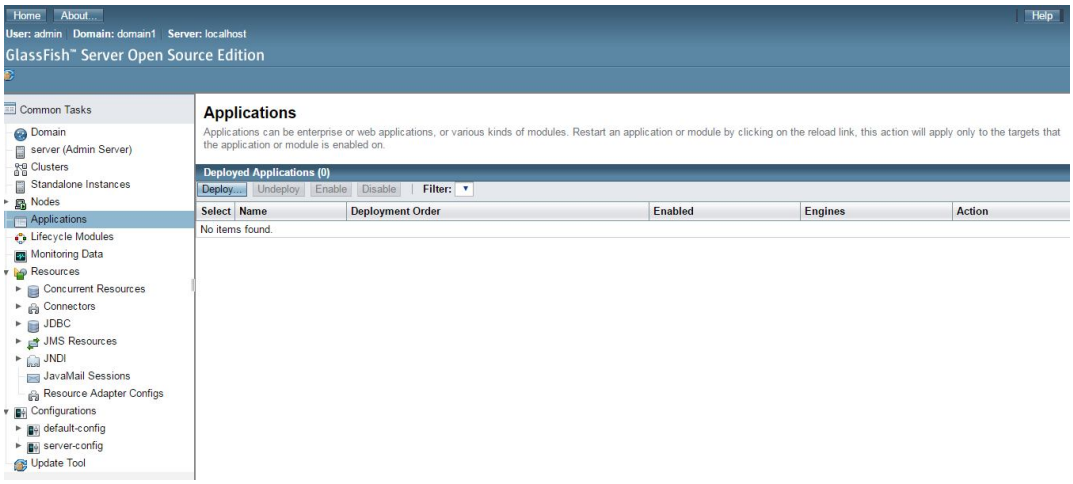
2.2. Instalación de JAVA

Inicialmente se instala Java ejecutando el instalador entregado ***jdk-7u65-windows-x64.exe***

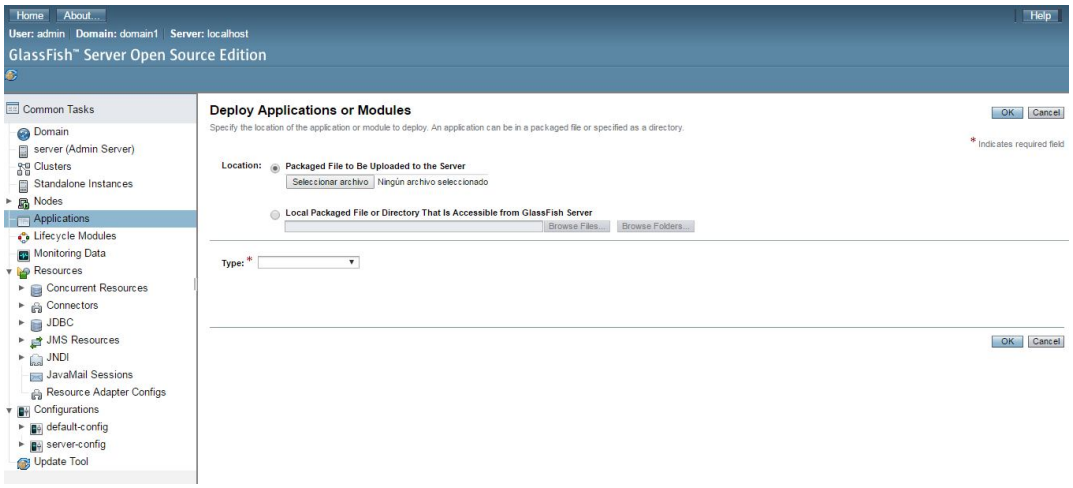
Se aceptan los términos y se da continuar hasta terminar la instalación

2.3. Instalación de Servidor GlasFish

1. Se descomprime la carpeta ***glassfish-4.1.zip*** en la ruta C:\
2. Se ingresa la consola de comandos nos dirigimos a la ruta c:\glasfish\bin
   Se ejecuta el comando: *asadmin start-domain domain1*
   Este comando instala el servidor como servicio de Windows y hace que arranque automáticamente cuando se encienda el computador
3. Abrimos google Chrome e ingresamos a: http://localhost:4848 que es la pagina de administración de glasfish
4. Dirigirse a la sección Nos dirigimos a la sección de aplicaciones



5. Damos click en Deploy y luego en seleccionar archivo, se carga el archivo YachayDTM.war y para finalizar se presiona el botón “Ok”



6. Para verificar que la instalación se realizo correctamente accedemos en el explorador web a la dirección: http://localhost:8080/YachayDTM

3. ARQUITECTURA YACHAY DTM

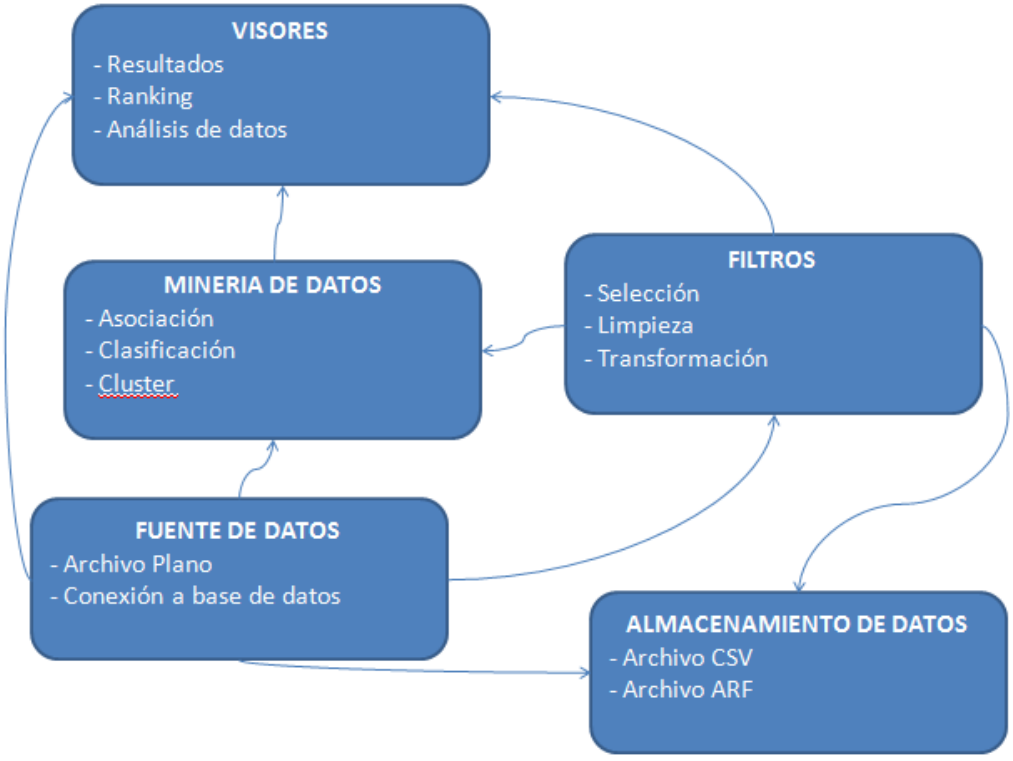


Ilustración 1 – Pantalla de aterrizaje YachayDTM

La arquitectura de YACHAY DTM la componen 5 secciones que corresponden a las 5 secciones en que se divide el menú principal, aquí se explican brevemente ya que serán detallados más adelante

Fuente de datos: Permite la carga de datos al sistema mediante un archivo plano o la conexión a una base de datos

Almacenamiento de datos: permite exportar un archivo en los formatos:

- CSV: Archivo de texto separado por un delimitador que generalmente es coma
- ARF: (Archivo relacional de atributos) el cual contiene metadatos y hace mas ágil la carga de información

Filtros: los filtros nos permiten transformar los datos según nuestros requerimientos agrupados en tres subgrupos: selección, limpieza y transformación

Minería de datos: Esta sección contiene los algoritmos que permiten aplicar las técnicas específicas de minería de datos agrupadas según las tareas de: Asociación, Clasificación y Clúster.

Según el algoritmo que se esté aplicando se permite la visualización de un grafo que lo representa así como la exportación de resultados en un archivo de texto

Visores: Nos brinda la opción de visualizar un análisis de datos, la exportación de este en un archivo PDF y también permite realizar un Rankin de datos según diferentes algoritmos

4. INGRESO E INTERFAZ GENERAL

El ingreso general a la aplicación se realiza abriendo un explorador web (recomendado google crome) y accediendo a la url:

http://localhost:8080/YachayDTM

Cuando se ingresa a la página inicial se nos muestra la pantalla de presentación:

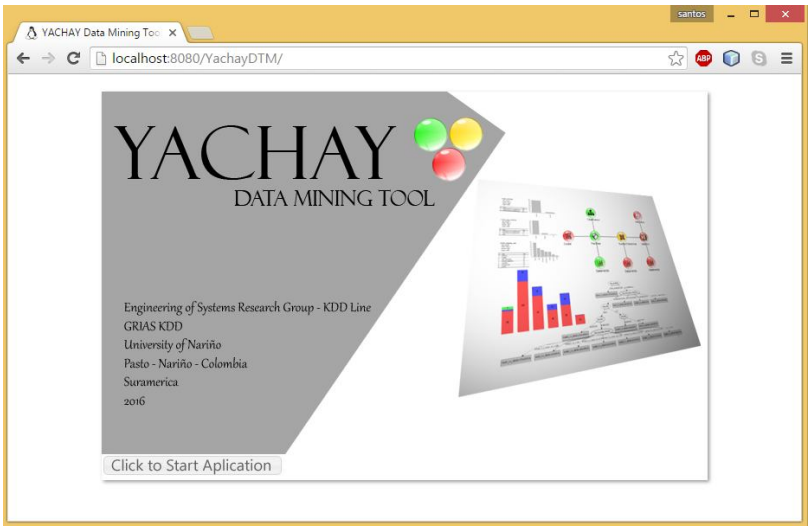


Ilustración 2 – Pantalla de aterrizaje YachayDTM

Para iniciar el uso de la aplicación se da clic sobre el botón “Clic to Start Aplication”.

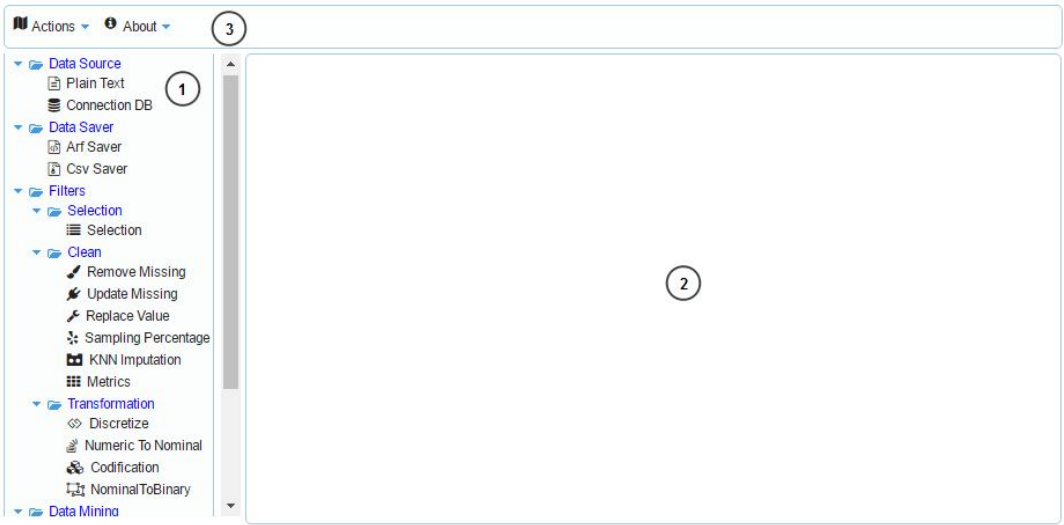


Ilustración 3 – Pantalla de la aplicación YachayDTM

Una vez accedemos a la aplicación se nos muestra la interfaz y menús de la aplicación que se divide en tres secciones: 1 Menú de Principal, 2 Área de trabajo, 3 Menú de acciones y acerca de.

4.1. Menú principal

El menú principal lo encontramos a la izquierda de la pantalla, es un árbol que contiene todas las actividades que se pueden realizar en el sistema. Las opciones que se encuentran en color azul simplemente agrupan a los diferentes componentes; y las opciones de color negro son los diferentes componentes del sistema que pueden ser arrastrados hacia el área de trabajo.

4.2. Área de trabajo

El área de trabajo es la sección de la pantalla en donde se agregan y conectan componentes del menú principal generando uno o varios árboles según se requiera. El área de trabajo permite operaciones creación, configuración, conexión, diferentes acciones que estados, conexiones, configuraciones

4.2.1. Nodos

Cada uno de los componentes que se encuentran en el menú principal (los de color negro) pueden ser arrastrados hacia el área de trabajo. Cuando se suelta un componente en el área de trabajo genera un 'Nodo' que es un círculo que lo representa y permite su configuración, ejecución o conexión con otros nodos.

4.2.2. Estado de Nodos

Cada uno de los nodos pueden estar en tres estados:
Rojo (Desconfigurarado): No están ni configurado ni ejecutándose
Amarillo (Configurado): Se encuentra configurado pero no ejecutándose.
Verde (Ejecutado): Se encuentra configurado y ejecutándose.

Ilustración 4 – Estado de nodos y Menú Emergente

4.2.3. Menú emergente

En la ilustración podemos observar un menú que se despliega al hacer clic derecho, cada nodo tiene más o menos opciones según sea el componente que represente.

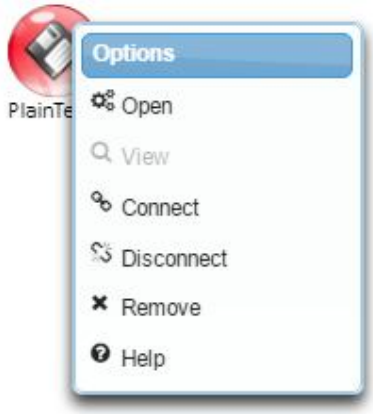


Ilustración 5 – Menú Emergente de un nodo

Las opciones que pueden desplegarse en el menú pueden ser:

Configure: Cuando aparece esta opción, se requiere una configuración previa para ejecutar el nodo.

Run: Esta opción ejecuta el proceso correspondiente al nodo, si el nodo tiene la opción de 'Configure' se debe ejecutar primero.

View: Esta opción permite la visualización de los resultados de proceso, para usar esta opción el nodo debe estar en estado ejecutándose (verde)

Connect: Opción que permite la conexión de dos nodos, en el momento que se usa la opción 'connect' aparece una línea, el que se dirige desde el nodo origen, para finalizar la conexión se da clic encima del nodo destino. El nodo puede estar en cualquier estado para realizar conexiones, simplemente seguir las conexiones permitidas (Ver 3.2.3 Conexiones permitidas)

Disconect: Elimina todas las conexiones que tenga este nodo. Al realizar la desconexión de un nodo, los nodos que se encuentren conectados a él quedan en estado desconfigurado (Rojo)

Remove: Elimina el nodo del área de trabajo. Al realizar la desconexión de un nodo, los nodos que se encuentren conectados a él quedan en estado desconfigurado (Rojo)

4.2.4. Conexiones permitidas

Los nodos deben seguir un orden en la forma que se conectan, a modo de ejemplo para nodo tipo "Plain Text" con un nodo "Data Análisis" debemos:

1) Clic derecho sobre el nodo "Plain text" (**nodo origen**)
2) Seleccionar la opción "Connect" del menú emergente
3) dar clic en el nodo "Plain text" (**nodo destino**)

Se debe seguir el anterior orden por cuanto es "Plain text" quien suministra los datos a "Data análisis".

Para saber que conexiones son permitidas nos basamos en la grafica de la arquitectura de YACHAY DTM en la cual observamos cómo están conectados los diferentes módulos

Si no ejecutamos este orden el sistema nos mostrar un mensaje indicando que la conexión que se intenta realizar no está soportada, por lo cual se debe dirigir al manual para ver las conexiones soportadas.

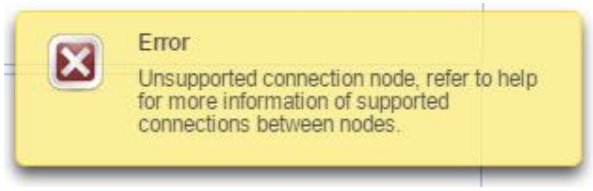


Ilustración 6 – Menú Emergente de un nodo

4.3. Menú de proyectos

Le denominamos proyecto a uno o diferentes componentes del menú principal agregados al área de trabajo así como las conexiones que se hayan realizado para su almacenamiento y posterior apertura mediante un archivo plano.

- New: Nos permite limpiar el área de trabajo
- Save: Carga un cuadro de dialogo en donde se solicita un nombre para descargar un archivo de extensión ydtm que representa los componentes y conexiones del árbol de trabajo
- Open: Permite buscar y cargar archivos de extensión ydtm
- Exit: Mediante esta opción salimos del sistema y se re direcciona a la página de aterrizaje de la aplicación.

5. FUENTE DE DATOS

Este módulo de la aplicación es el encargado de la conexión a una fuente de datos la cual puede ser un archivo plano o una base de datos:

5.1. Plain text

Se dio soporte para los archivos de archivos CSV (archivo de texto separado por un delimitador) así como los archivos ARFF (utilizados por Weka)

Cuando se hace uso de la opción '*Open'* del menú emergente para este nodo nos carga las diferentes opciones:

Date Attributes: Permite especificar que columnas del archivo deben ser interpretadas como fecha, se pueden usar rangos 'first-last', '1,4,7-14, 50-last'.
Date Format: Se especifica cuál es el formato que siguen las fechas
Enclosure Characters: Se ingresa el carácter que encierra los textos
Field Separator: Carácter que sirve como separador de columnas o campos, normalmente es coma
Missing value: valor que se ingresa cuando un campo es nulo
No Header row Present: determina si el archivo contiene o no cabecera

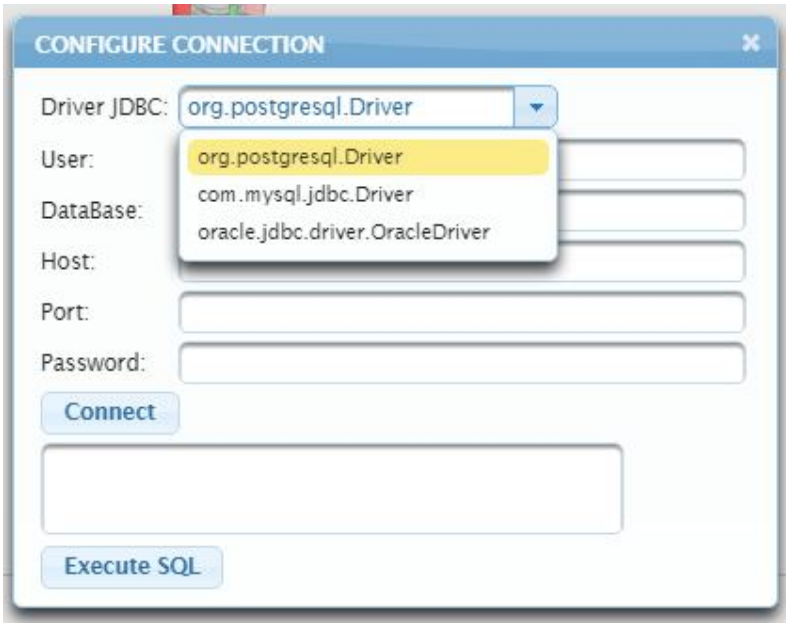
Nominal atributes: se ingresa las columnas que son nominales , se pueden usar rangos 'first-last', '1,4,7-14, 50-last'.
String atributes: se definen las columnas(atributos) que son de tipo texto, se pueden usar rangos 'first-last', '1,4,7-14, 50-last'.
Current file: muestra el nombre del archivo que se haya cargado
File: se presiona este botón para cargar el archivo, una vez seleccionado se activa el botón "Process" que carga el archivo al sistema

Los campos no son obligatorios así que se configuran los que se considere necesario, una vez cargado el archivo, si todo esta correcto el nodo cambia a color verde

5.2. ConnectioDB

Permite la conexión a una base de datos mediante la configuración de sus propiedades

Cuando se hace uso de la opción '*Configure'* del menú emergente para este nodo nos carga las diferentes opciones:



Driver JDBC: Se escoge cual será el gestor de base de datos de entre los tres permitidos: postgreSql, MySql ó Oracle
User: usuario de la base de datos
Password: clave asignada al usuario de la base de datos

Port: numero de puerto de la base de datos

Host: Hace referencia a la dirección donde se encentra la base de datos, se puede colocar una dirección IP o si es local se coloca localhost

Connect: se utiliza este botón para verificar si se realizo la conexión y el sistema nos muestra que la conexión fue satisfactoria o nos muestra un mensaje diciéndonos la causa del fallo en caso de no conexión

Execute SQL: se coloca en esta sección la consulta SQL de los datos que queremos sacar de la base de datos

6. ALMACENAR DATOS

Este módulo permite descargar un archivo en formato CSV o ARF a partir de un nodo perteneciente al modulo de “filtros” o “fuente de datos”

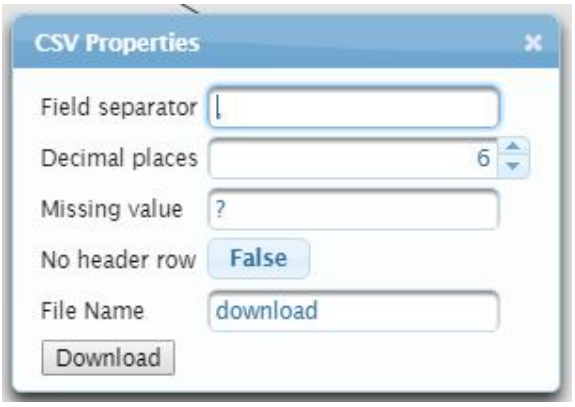
6.1. ARF Saver

Permite la descarga de un archivo en formato arff, se usa la opción “configure” del menú emergente para este componente, mostrando las opciones

Compres Output: permite que el archivo sea o no comprimido para reducir su tamaño
Decimal places: numero de cifras decimales para datos numéricos
File Name: Nombre del archivo que se descargara
Download: inicia la descarga del archivo

6.2. CSV Saver

Permite la descarga de un archivo en formato csv, se usa la opción “configure” del menú emergente para este componente, mostrando las opciones



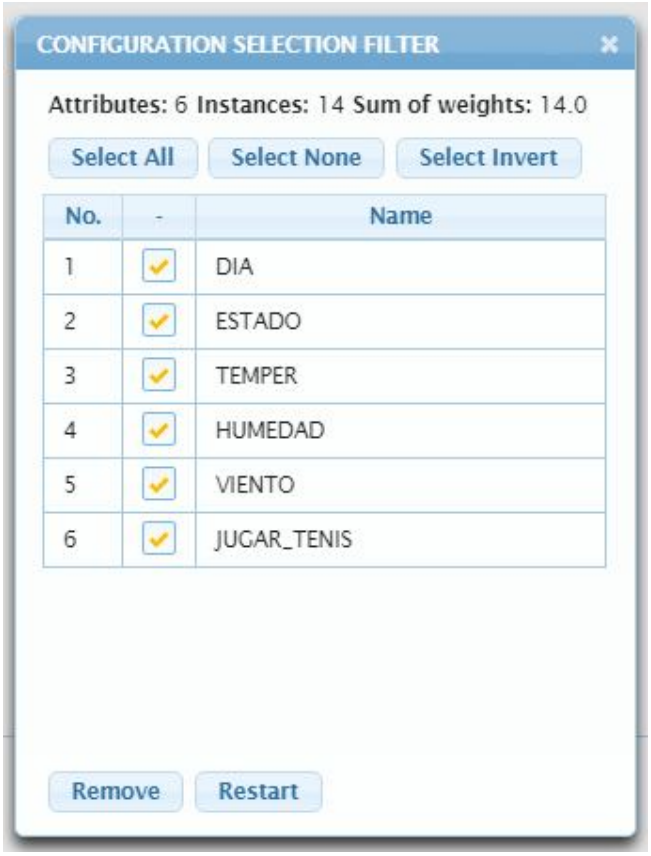
Field separator: Caracter que se usará como separador de las columnas o campos, para usar tabulación se usa '\t'
maxDecimalPlaces: Máximo número de dígitos para valores numéricos
missingValue: caracter que se coloca cuando existan valores nulos en los datos
noHeaderRow: almacenar o no cabecera con los títulos de las columnas o atributos

7. FILTROS

Los filtros realizan modificaciones sobre los datos de entrada permitiendo diferentes operaciones. Se dividen en tres grupos según su función pueden ser de Selección, Limpieza o transformación

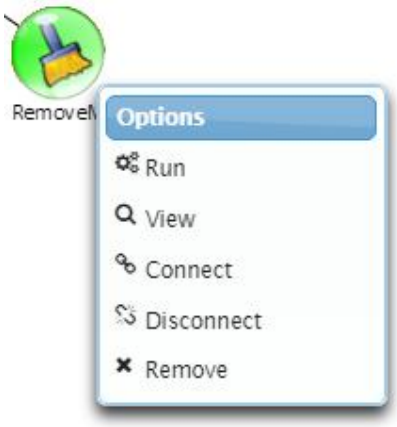
7.1. Selection

Permite seleccionar que atributos se desean eliminar y cuales se dejan para el posterior uso



Select All: permite la selección de todos los atributos listados en la tabla
Select none: quita la selección de todos los atributos listados en la tabla
Select invert: Invierte la selección de los atributos seleccionados en la tabla
Remove: remueve los atributos seleccionados del conjunto de datos
Restart: permite volver a listar todos los atributos como estaban originalmente

7.2. Remove Missing

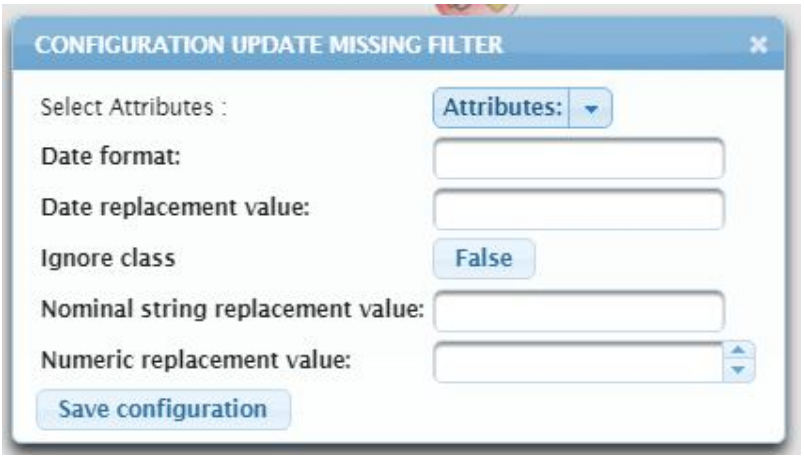


Elimina todos los registros que contengan valores nulos en cualquiera de sus atributos

Este componente no tiene configuración por lo cual para su uso simplemente se utiliza la opción "Run" del menú emergente

7.3. Update Missing

Permite actualizar los valores nulos con los valores que el usuario especifique



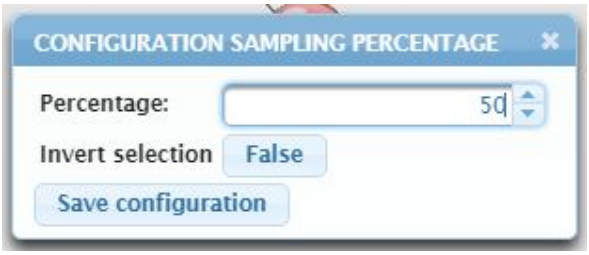
Select Attributes: en este combo se seleccionan los atributos sobre los cuales se realizará la aplicación del filtro
DateFormat: se especifica el formato de fecha que se utilizara
Date replacement value: Fecha que se usara cuando el atributo sea de tipo fecha
IgnoreClass: no efectuar el reemplazo si se trata del atributo clase
Nominal String replacement: Texto que se usara como reemplazo de nulos cuando el atributo sea de tipo texto
Numeric replacement: Número que se usara como reemplazo de nulos cuando el atributo sea de tipo numérico

7.4. Replace Value

Permite reemplazar valores, se debe seleccionar el atributo y así lista cuales son los diferentes valores que contiene ese atributo, en la columna replace se ingresa cual es el nuevo valor que se desea, para finalizar y aplicar el filtro se hace uso del botón "Save configuration"

7.5. Sampling Percentage

Permite la reducción del tamaño de los registros indicando a que porcentaje se desea reducción los datos

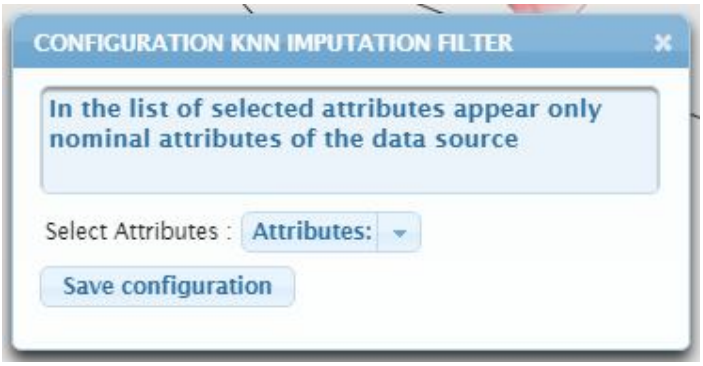


Percentage: Indica el porcentaje de registros que al final de la operación quedaran

Invert selection: Toma el restante porcentaje del indicado, por ejemplo si en "percentaje" se usa 30 y "Invert Selection"=true entonces el porcentaje de datos al final será de 70%
Save configuration: se usa este botón para finalizar y aplicar el filtro

7.6. KNN Imputation

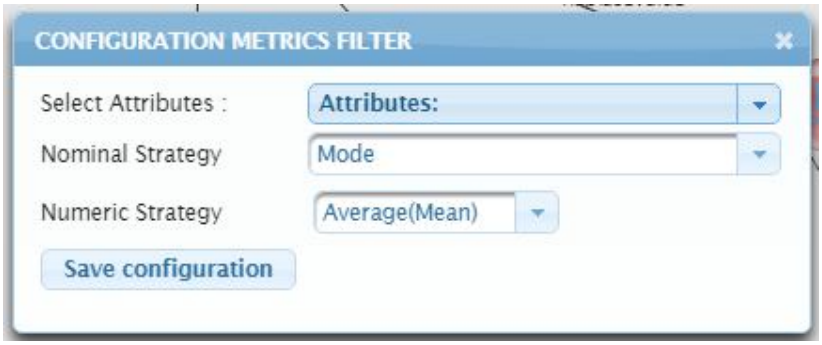
El proceso de imputación consiste en reemplazar los valores nulos mediante el uso del algoritmo de Vecino más cercano, este algoritmo solo funciona solo sobre atributos de tipo nominal, no se permite numérico ni fecha, por lo cual aparece un mensaje indicando esta restricción



Select Atributes: se deben seleccionar los atributos que el usuario desee para la aplicación del algoritmo
Save Configuration: se hace uso de este botón para finalizar y aplicar el algoritmo a los atributos indicados

7.7. Metrics

Permite el reemplazo de valores nulos mediante el cálculo de diferentes métricas



Select attributes: En este combo se realiza la selección de los atributos sobre los cuales se aplicara el filtro
Nominal strategy: en este combo se selecciona el tipo de cálculo (moda o mínimo) que se realizara cuando se trate de atributos nominales
Numeric Strategy: Cuando se trata de datos numéricos el cálculo que se realizara para reemplazar los valores nulos puede ser: Promedio, Mediana, Moda, Máximo, Mínimo y Preservación de la desviación

7.8. Discretize

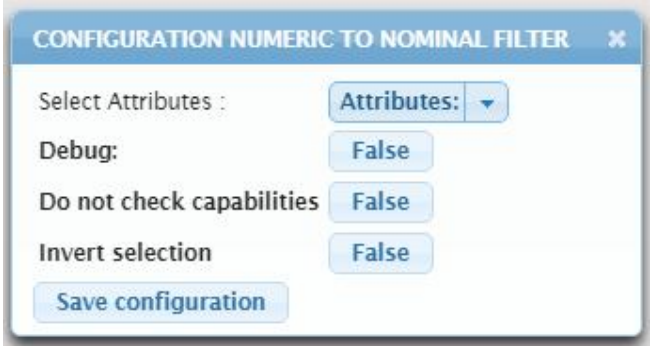
Este filtro permite convertir un atributo de tipo numérico en un atributo de tipo nominal mediante la creación de intervalos

Select Attributes: en este combo se selecciona los atributos a los cuales se les aplicara el filtro
IgnoreClass: el índice de la clase se desactivará temporalmente antes de que se aplique el filtro.
UseEqualFrequency: Si se establece en true, se utilizará binning de igual frecuencia en lugar de binning de igual anchura.
UseBinNumbers: Se nombrara a los valores con números en lugar de intervalos para los atributos discretizados
DesiredWeightOfInstancesPerInterval: Establece el peso deseado de instancias por intervalo para binning de igual frecuencia.
MakeBinary: Hacer los atributos resultantes binarios.

FindNumBins: Optimizar el número de compartimientos de anchura igual usando leave-one-out. No funciona para binning de igual frecuencia
InvertSelection: Establece el modo de selección de atributos. Si es falso, sólo los atributos seleccionados (numéricos) en el rango serán discretizados; Si es cierto, sólo se discretizarán los atributos no seleccionados.

7.9. Numeric to Nominal

Convierte los atributos numéricos en nombres. A diferencia de la discretización, sólo toma todos los valores numéricos y los agrega a la lista de valores nominales de ese atributo. Útil después de importar CSV, para obligar a ciertos atributos a convertirse en nominal, por ejemplo, el atributo de clase, que contiene valores de 1 a 5.

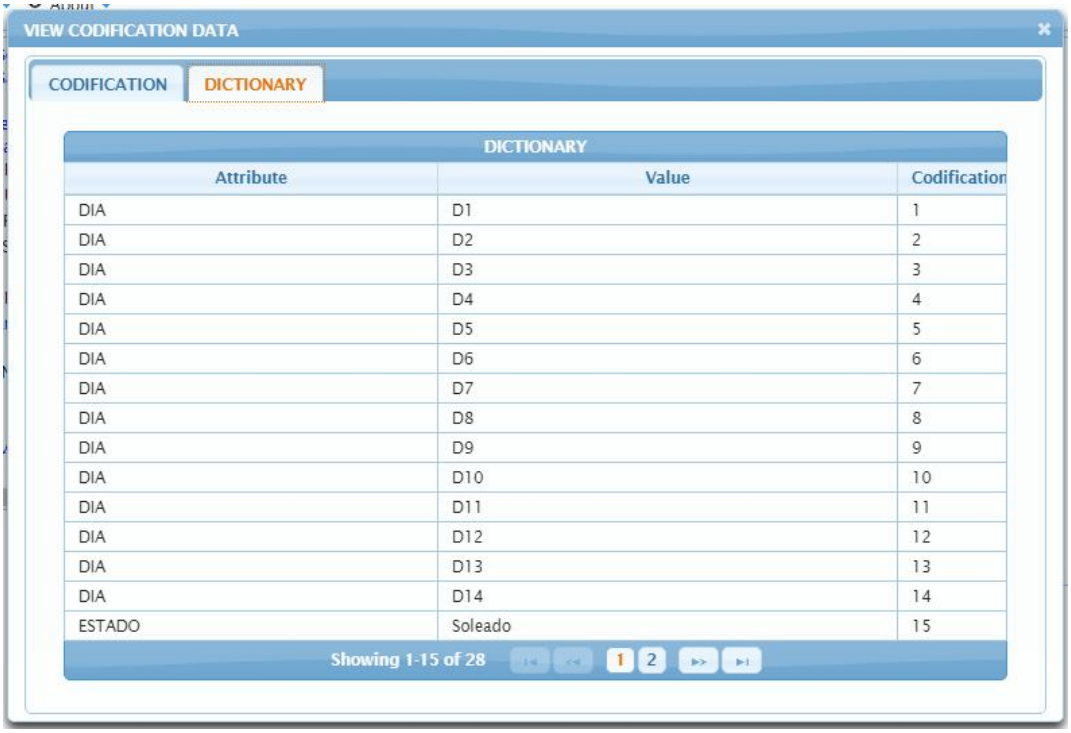


Select Atributes: en este combo se selecciona los atributos a los cuales se les aplicara el filtro
Debug: Si se establece en true, el filtro puede generar información adicional en la consola.
DoNotCheckCapabilities: Si se establece, las capacidades del filtro no se comprueban cuando se establece el formato de entrada (Use con precaución para reducir el tiempo de ejecución).
InvertSelection: Establece el modo de selección de atributos. Si es falso, sólo los atributos (numéricos) seleccionados en el rango serán "nominalizados"; Si es cierto, sólo los atributos no seleccionados serán "nominalizados".

7.10. Codification

Mediante este componente se codifican todos los valores que contengan el conjunto de datos, es útil cuando se desea agilizar los procesos cuando se apliquen los algoritmos de minería de datos

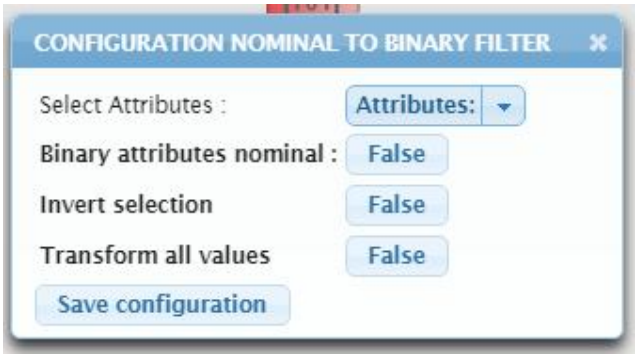
Se hace uso de la opción “run” del menú emergente para ejecutar la codificación, una vez finalizado el se activa la opción “view” del menú emergente, y se podrá visualizar el resultado de la codificación



Codificación: Muestra como quedaron los datos
Diccionario: Muestra cual eran los valores originales y por cuales fueron reemplazados

7.11. Nominal to Binary

Convierte todos los atributos nominales en atributos binarios numéricos. Un atributo con valores k se transforma en k atributos binarios si la clase es nominal (utilizando el enfoque de un atributo por valor). Los atributos binarios se dejan binarios



Select Atributes: en este combo se selecciona los atributos a los cuales se les aplicara el filtro

TransformAllValues: Si todos los valores nominales se convierten en nuevos atributos, no sólo si hay más de 2.

BinaryAttributesNominal: Si los atributos binarios resultantes serán nominales.

InvertSelection: Establece el modo de selección de atributos. Si es falso, sólo los atributos seleccionado, Si es cierto, sólo se discretizarán los atributos no seleccionados.

8. MINERIA DE DATOS

Esta sección contiene los algoritmos que permiten aplicar las técnicas específicas de minería de datos agrupadas según las tareas de: Asociación, Clasificación y Clúster.
Según el algoritmo que se esté aplicando se permite la visualización de un grafo que lo representa así como la exportación de resultados en un archivo de texto

8.1. Association

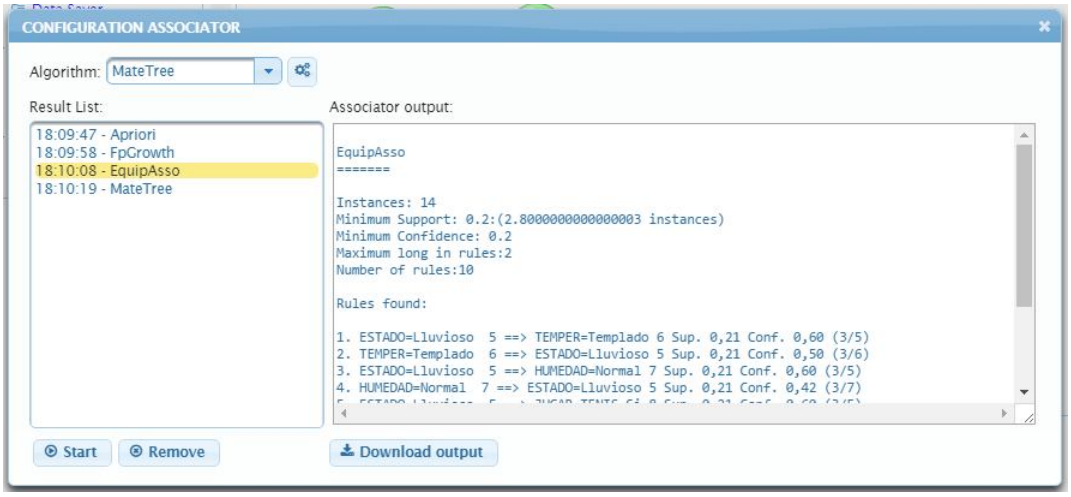
Este componente permite la aplicación de técnicas de minería de datos para las tareas de asociación con los siguientes algoritmos:

Apriori: tiene como objetivo reducir el número de conjuntos considerados, generando un conjunto de itemsets frecuentes a partir de itemsets candidatos.

FpGrowth, utiliza una estructura de datos llamada árbol de patrones frecuentes o FP-tree, la cual es una estructura que almacena información crucial ycuantitativa acerca de los patrones frecuentes

EquipAsso, es un algoritmo, para el cálculo de los itemsets frecuentes basado en dos operadores del álgebra relacional para Asociación: Associator y EquiKeep e implementado en el lenguaje SQL mediante las primitivas SQL Associator Range y EquiKeep On

Mate-tree: Es un algoritmo basado en el operador algebraico relacional Mate [34][35] que conjuntamente con los operadores agregados Entro y Gain, facilitan el cálculo de la Ganancia de Información y con el operador algebraico relacional Describe Classifier, la construcción del árbol de decisión.



Uso de Interfaz grafica:

Algorithm: en este combo seleccionados el algoritmo que deseamos usar, al lado se encuentra un botón que nos permite realizar la configuración del algoritmo seleccionado

Result List: En esta lista se van agregando todos los resultados de los análisis que se vallan realizando

Associator Output: En esta área de texto se muestra el resultado de la aplicación del algoritmo de asociación que se haya realizado

Boton Start: Inicia el proceso de rankeo y una vez terminado muestra el resultado

Boton Remove: Remueve un resultado de rankeo de la lista "Result list"

Boton Download Output: descarga el resultado del rankeo en un archivo de texto

8.2. Classification

Este componente permite la aplicación de técnicas de minería de datos para las tareas de clasificación con los siguientes algoritmos:

J48: es una implementación open source en lenguaje de programación Java del algoritmo C4.5, genera un árbol de decisión C4.5 podado o sin podar

ID3: Construye un árbol de decisión no podado basado en el algoritmo ID3. Sólo puede tratar con atributos nominales. No se permiten valores perdidos. Las hojas vacías pueden resultar en casos no clasificados.

LMT: Clasificador para la construcción de 'árboles de modelos logísticos', que son árboles de clasificación con funciones de regresión logística en las hojas. El algoritmo puede manejar variables binarias y multiclase, atributos numéricos y nominales y valores faltantes.

M5P: Implementa rutinas base para generar Modelos M5 de árboles y reglas El algoritmo original M5 fue inventado por R. Quinlan y Yong Wang le hizo mejoras.

DesicionStump: Construye y usar un pivot de decisión. Generalmente se utiliza en conjunción con un algoritmo de impulso. Realiza la regresión (basada en el error cuadrático medio) o clasificación (basada en la entropía).

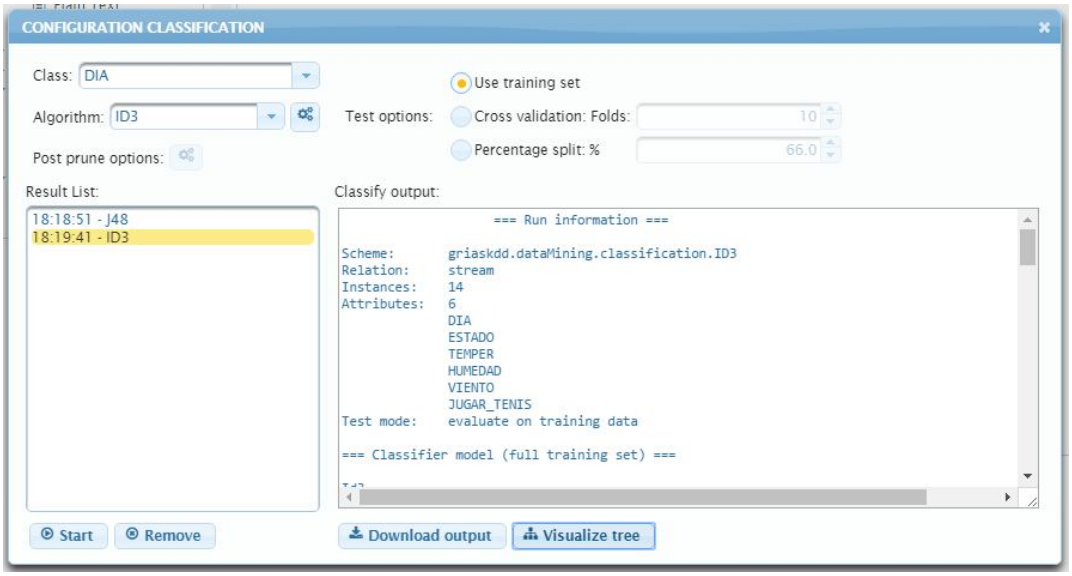
HoeffdingTree: Un árbol Hoeffding (VFDT) es un algoritmo de inducción de árbol de decisión incremental, en cualquier momento que es capaz de aprender de flujos de datos masivos, suponiendo que los ejemplos de generación de distribución no cambian con el tiempo.

RandomForest: Algoritmo que construye un bosque de árboles al azar.

RandomTree: Algoritmo que construye un árbol que considera K atributos elegidos al azar en cada nodo. No realiza podas. También tiene una opción para permitir la estimación de las probabilidades de clase basadas en un conjunto de hold-out

REPTree: Arbol de decisiones rápido. Construye un árbol de decisión / regresión usando la ganancia / varianza de información y la elimina usando poda de reducción de errores (con ajuste posterior). Sólo clasifica valores para atributos numéricos una vez. Los valores faltantes se tratan dividiendo las instancias correspondientes en fragmentos (es decir, como en C4.5).

Uso de Interfaz grafica:



Class: en este combo se listan todos los atributos del cual se debe seleccionar el atributo que se desea como clase

Test Options: brinda las diferentes formas en que se realiza el entrenamiento

Use full trainin set: El valor del subconjunto de atributos se determina Utilizando el conjunto completo de datos de formación.

Cross validation: El valor del subconjunto de atributos está determinado por un Proceso de validación cruzada, el campo Fold establece el número de Pliegues para usar

Percentaje Split: divide el conjunto de datos de acuerdo con el porcentaje dado

Algorithm: en este combo seleccionados el algoritmo que deseamos usar, al lado se encuentra un botón que nos permite realizar la configuración del algoritmo seleccionado

Boton Post prune: este botón solo se activa en el algoritmo J48 y permite realizar una operación de post-poda mediante la especificación del mínimo soporte y/o mínima confianza deseada

Result List: En esta lista se van agregando todos los resultados de los análisis que se vallan realizando

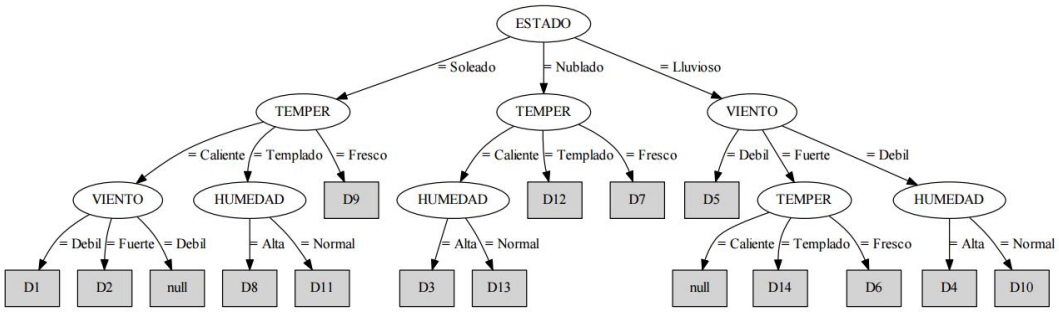
Classify Output: En esta área de texto se muestra el resultado de la aplicación del algoritmo de clasificación que se haya realizado

Boton Start: Inicia el proceso de rankeo y una vez terminado muestra el resultado

Boton Remove: Remueve un resultado de rankeo de la lista "Result list"

Boton Download Output: descarga el resultado del algoritmo de clasificación en un archivo de texto

Botón Visualice tree: este botón se activa si la salida del algoritmo permite la generación de un archivo PDF con el grafico del árbol



8.3. Cluster

Este componente permite la aplicación de técnicas de minería de datos para las tareas de cluster haciendo uso de los algoritmos:

SimpleKMeans: Datos del cluster usando el algoritmo k means. Puede usar la distancia euclidiana (predeterminada) o la distancia de Manhattan. Si se utiliza la distancia de Manhattan, los centroides se calculan como la mediana de los componentes en lugar de la media.

Canopy: Datos del clúster mediante el algoritmo de agrupación de capopy, que requiere un solo paso sobre los datos. Se puede ejecutar en modo bybatch o incremental. Los resultados generalmente no son tan buenos cuando se ejecutan de forma incremental como el min / max para cada atributo numérico no se conoce de antemano. Tiene una heurística (basada en las desviaciones estándar del atributo), que se puede utilizar en modo por lotes, para establecer la distancia T2.

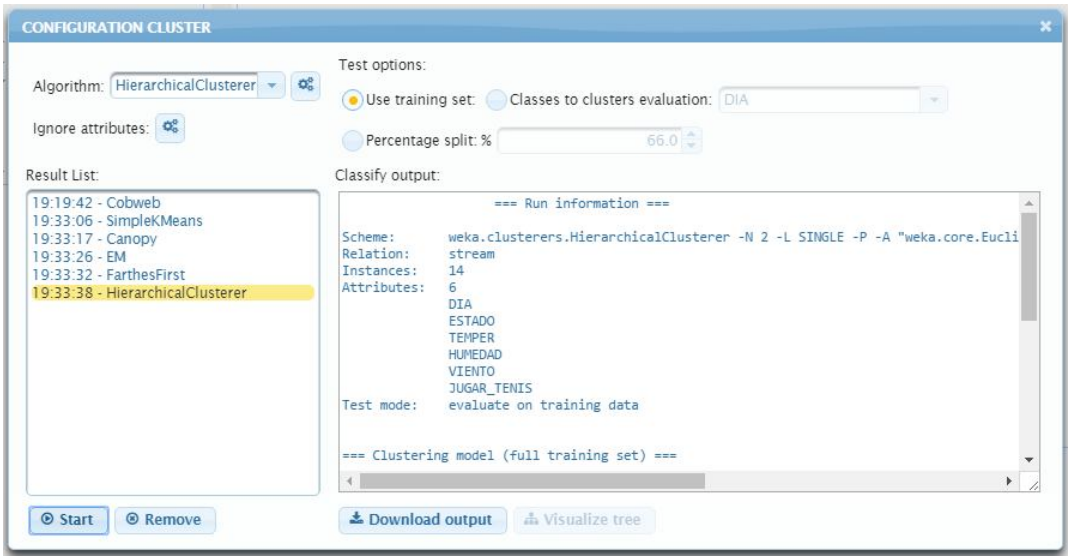
Cobweb: Algoritmo que implementa los algoritmos de agrupación Cobweb y Classit. la aplicación de operadores de nodo (fusión, división, etc.) en términos de ordenación y prioridad difiere (y es algo ambigua) entre los papeles originales de Cobweb y Classit. Este algoritmo compara siempre el mejor anfitrión, agregando una nueva hoja, combinando los dos mejores anfitriones, y dividiendo el mejor anfitrión al considerar donde colocar una nueva instancia.

EM: algoritmo EM simple (maximización de la expectativa). EM asigna una distribución de probabilidad a cada instancia que indica la probabilidad de que pertenezca a cada uno de los conglomerados. EM puede decidir cuántos clústeres crear mediante validación cruzada, o puede especificar apriori cuántos clústeres para generar.

FarthesFirst: Datos del clúster utilizando el algoritmo de Primejo lo mas lejano.

HierarchicalClusterer: Algoritmo de agrupación jerárquica. Implementa un número clásico de aglomeración clásica (es decir, de abajo hacia arriba).

Uso de Interfaz grafica:



Algorithm: en este combo seleccionados el algoritmo que deseamos usar, al lado se encuentra un botón que nos permite realizar la configuración del algoritmo seleccionado

Test Options: brinda las diferentes formas en que se realiza el entrenamiento

Use trainin set: El valor del subconjunto de atributos se determina Utilizando el conjunto completo de datos de formación.

Clases to cluster evaluation: selección del atributo clase

Percentaje Split: divide el conjunto de datos de acuerdo con el porcentaje dado

Result List: En esta lista se van agregando todos los resultados de los análisis que se vallan realizando

Cluster Output: En esta área de texto se muestra el resultado de la aplicación del algoritmo de cluster que se haya realizado

Boton Start: Inicia el proceso de rankeo y una vez terminado muestra el resultado

Boton Remove: Remueve un resultado de rankeo de la lista "Result list"

Boton Download Output: descarga el resultado del algoritmo de Cluster en un archivo de texto

Botón Visualice tree: este botón se activa si la salida del algoritmo permite la generación de un archivo PDF con el grafico del árbol

9. VISORES

El módulo de visores nos permite visualizar el análisis de datos así como realizar un archivo PCF con este análisis, también realizar el ranking de atributos haciendo uso de diferentes algoritmos

9.1. Data Análisis

Este componente brinda un análisis de datos, inicialmente se debe usar la opción de "run" del menú emergente para que genere el análisis y active la opción "view"

La opción view nos muestra la pantalla de análisis que contiene tres secciones:

Attributes:

Lista todos los atributos asi como los datos de, numero de atributos, numero de instancias, y suma de longitudes

Selected Atribute:

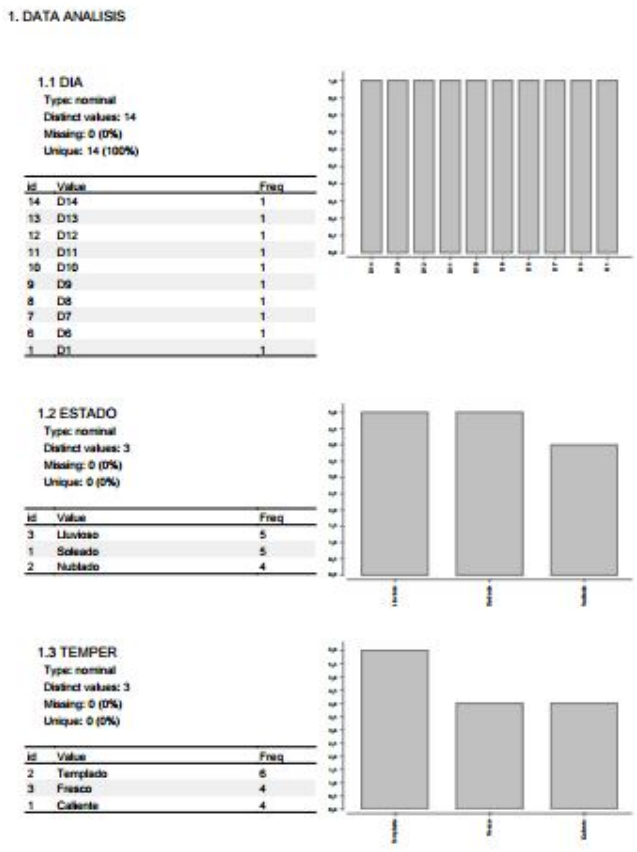
Nos muestra la información detallada del atributo que sea seleccionado en la sección de Attributes

Visualize

Esta sección muestra gráficamente la distribución de los datos en un grafico de barras, también permite seleccionar una variable clase para determinar la relación entre dos atributos

Esta sección contiene dos botones adicionales que permiten ampliar la grafica asi como generar un archivo PDF con los resultados del análisis

El reporte generado contiene toda la información del análisis de datos asi como el resumen de datos nulos y tabla de contenido



1. DATA ANALISIS

1.1 DIA

Type: nominal
Distinct values: 14
Missing: 0 (0%)
Unique: 14 (100%)

id	Value	Freq
14	D14	1
13	D13	1
12	D12	1
11	D11	1
10	D10	1
9	D9	1
8	D8	1
7	D7	1
6	D6	1
1	D1	1

1.2 ESTADO

Type: nominal
Distinct values: 3
Missing: 0 (0%)
Unique: 0 (0%)

id	Value	Freq
3	Lluvioso	5
1	Soleado	5
2	Nublado	4

1.3 TEMPER

Type: nominal
Distinct values: 3
Missing: 0 (0%)
Unique: 0 (0%)

id	Value	Freq
2	Templado	6
3	Fresco	4
1	Caliente	4

9.2. Ranking

El componente de rankin se usa para determinar cuáles son los atributos con más relevancia mediante el uso de los algoritmos:

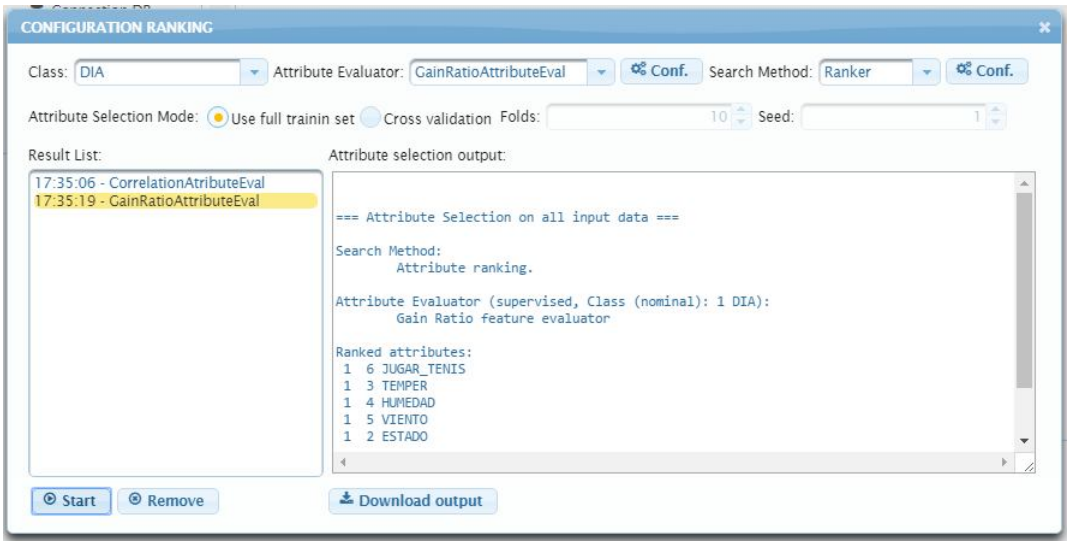
GainRatioAttributeEval: Evalúa el valor de un atributo midiendo la relación de ganancia con respecto a la clase.

CorrelationAttributeEval: Evalúa el valor de un atributo midiendo la correlación (de Pearson) entre él y la clase. Los atributos nominales se consideran valor por valor tratando cada valor como un indicador. Se obtiene una correlación global para un atributo nominal a través de un promedio ponderado.

InfoGainAttributeEval: Evalúa el valor de un atributo midiendo la ganancia de información con respecto a la clase.

OneRAttributeEval: Evalúa el valor de un atributo utilizando el clasificador OneR.

SymetricalUncertAttributeEval: Evalúa el valor de un atributo midiendo la incertidumbre simétrica con respecto a la clase.



Uso de Interfaz grafica:

Class: permite la selección del atributo clase

Attribute Evaluator: en ese combo se selecciona cual es el algoritmo que se desa utilizar para la realización del ranking, al lado de este contro se encuentra el botón que permite realizar la configuración del algoritmo seleccionado

Search Method: se especifica el método de búsqueda, por defecto siempre será Ranker, al lado de este control se encuentra el botón para realizar la configuración de el ranker

Attribute Selection mode: Determina el modo de selección de atributos que puede ser de dos tipos:

Use full trainin set: El valor del subconjunto de atributos se determina Utilizando el conjunto completo de datos de entrenamiento.

Cross validation: El valor del subconjunto de atributos está determinado por un Proceso de validación cruzada, Los campos Fold y Seed establecen el número de Pliegues para usar y la semilla aleatoria utilizada al barajar los datos.

Result List: En esta lista se van agregando todos los resultados de los análisis que se vallan realizando

Atribute Selection Output: En esta área de texto se muestra el resultado del rankeo que se haya realizado

Boton Start: Inicia el proceso de rankeo y una vez terminado muestra el resultado

Boton Remove: Remueve un resultado de rankeo de la lista "Result list"

Boton Download Output: descarga el resultado del rankeo en un archivo de texto