# ADAMIX 1.0
# HERRAMIENTA SOFTWARE PARA LA CLASIFICACIÓN DE DATOS CON VARIABLES CUALITATIVAS Y CUANTITATIVAS

**PAOLA JANNETH VELÁSQUEZ CALDERÓN**
**VICTORIA ISABEL FERNÁNDEZ PIÑA**

**UNIVERSIDAD INDUSTRIAL DE SANTANDER**
**FACULTAD DE INGENIERÍAS FISICOMECÁNICAS**
**ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA**
**BUCARAMANGA**
**2004**

# ADAMIX 1.0
# HERRAMIENTA SOFTWARE PARA LA CLASIFICACIÓN DE DATOS CON VARIABLES CUALITATIVAS Y CUANTITATIVAS

**PAOLA JANNETH VELÁSQUEZ CALDERÓN**
**VICTORIA ISABEL FERNÁNDEZ PIÑA**

**Proyecto de Grado presentado como requisito para optar al título de Ingenieras de Sistemas.**

**Director**
**ING. ADDISSON SALAZAR AFANADOR**
**Ingeniero de Sistemas – Área de la Inteligencia Artificial**

**Codirector**
**ING. FERNANDO RUIZ DÍAZ**
**Ingeniero de sistemas**

**UNIVERSIDAD INDUSTRIAL DE SANTANDER**
**FACULTAD DE INGENIERÍAS FISICOMECÁNICAS**
**ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA**
**BUCARAMANGA**
**2004**

*A Dios, por ser todo en mi vida.*

*A mis padres y hermanos por su amor y apoyo incondicionales.*

*A Roger porque me ha enseñado muchas cosas en la vida..*

*Y a mis amigos por su compañía y afecto*

*Paola Janneth*

*A Dios, por todas las oportunidades que me ha dado en la vida.*

*A mis padres y hermanos por estar siempre con migo a pesar de la distancia.*

*A mi familia por su cobijo.*

*Y mis amigos por su aprecio y amistad incondicional.*

*Victoria Isabel*

## AGRADECIMIENTOS

Las autoras desean expresar su agradecimiento al Ing. Addisson Salazar Afanador, Ingeniero de Sistemas, Master en Informática, DEA en Telecomunicaciones y MBA por su ayuda, asistencia y colaboración durante la realización del presente proyecto de investigación.

## TABLA DE CONTENIDO

## LISTA DE TABLAS

## LISTADO DE FIGURAS

**CAPITULO 5**

## LISTADO DE ANEXOS

# GLOSARIO

**Clasificación:** técnicas que realizan la separación de datos por medio de la asignación de los mismos a clases predefinidas.

**Cluster (grupo):** colección de objetos que guardan una relación de semejanza de acuerdo con alguna medida de similaridad numérica determinada.

**Clustering (agrupamiento):** partición de un conjunto de datos en grupos o clases de objetos semejantes de acuerdo con un criterio determinado de agrupamiento y unas medidas de similaridad entre objetos.

**Clustering aglomerativo:** partición de un conjunto de datos en grupos. El algoritmo crea una jerarquía calculando la similitud entre todos los objetos y agrupando a cada paso la pareja más similar. Esta pareja constituye un nuevo objeto que sustituye los dos primeros objetos.

**Clustering conceptual:** agrupamiento de objetos en clases conceptualmente simples basándose en los valores de sus atributos. El algoritmo tiene en consideración el conocimiento acerca de las relaciones semánticas entre los atributos de los objetos o conceptos globales que puedan ser usados para caracterizar las agrupaciones de objetos.

**Clustering jerárquico:** partición de un conjunto de datos en grupos, buscando el árbol que refleja la estructura jerárquica de los datos. Según el nivel por el que se corte el árbol se obtendrá una partición más o menos precisa del conjunto objeto de estudio.

**Complejo Lógico:** es uno de los esquemas de representación de cluster, en agrupamiento conceptual conjuntivo; es utilizado en cluster/2 junto con la semilla. También conocido como descripción maximal.

**DCBD:** descubrimiento de conocimiento en bases de datos. Una definición completa se incluye en el documento.

**Dendrograma:** representación gráfica en forma de árbol invertido, de los agrupamientos secuenciales sucesivos a partir de $n$ objetos hasta obtener uno.

**Descripción maximal:** ver complejo lógico.

**Dominios poco estructurados:** aquellos en los cuales el consenso entre los expertos es débil o inexistente.

**Estrella:** estructura de datos utilizada en cluster/2, obtenida a partir de la descripción maximal de una semilla y conformada por un conjunto de complejos que cubren una serie de observaciones.

**Índice de discriminación:** medida para la evaluación de la calidad de un agrupamiento, es utilizado en cluster/2 y corresponde al número de atributos que toman valores diferentes en todas las clases.

**KNN:** algoritmo del vecino más cercano. ( K –Nearest Neighbour ), que clasifica datos, basado en distancias euclidianas estáticas.

**LEF:** función de Evaluación Lexicográfica (Lexicographic Evaluation Function). Método que consiste en una secuencia de parejas $criterio_i$ – $tolerancia_i$, donde $criterio_i$ es un criterio de calidad de conglomerados (clusteres) y $tolerancia_i$ es el porcentaje que determina el umbral de los mejores conglomerados evaluados en dicho criterio, los cuales pasan a la siguiente ronda, hasta que se obtiene un solo conglomerado que se convierte en el mejor.

**Medida de similaridad:** se define usualmente por proximidad en un espacio multidimensional. Para datos numéricos, usualmente se pasa primero por un proceso de estandarización.

**Minería de datos:** es el conjunto de técnicas para la exploración y análisis de grandes cantidades de datos, por medios semiautomáticos ó totalmente automáticos, con el fin de descubrir reglas o patrones de datos significativos.

**Selector:** afirmación relacional utilizada en clustering conceptual conjuntivo, que tiene la siguiente forma : [*x* # *Ri*], donde x es una variable, # es uno de estos operadores =,≠,,≤,≥ <, > o el operador rango (..)- y *Ri* es una lista de elementos del dominio de la variable x conectados por el operador disyunción.

**Semilla:** esquema de representación de cluster utilizado en cluster/2 que consiste en un objeto singular distintivo del cluster, a partir del cual se obtiene el mismo.

**TITULO**[*]

ADAMIX 1.0
"HERRAMIENTA SOFTWARE PARA LA CLASIFICACIÓN DE DATOS CON VARIABLES CUALITATIVAS Y CUANTITATIVAS."

**AUTORAS**[**]

PAOLA JANNETH VELÁSQUEZ CALDERÓN
VICTORIA ISABEL FERNÁNDEZ PIÑA

**PALABRAS CLAVES**

CLASIFICACIÓN, AGRUPAMIENTO EN DOMINIOS MIXTOS, MINERÍA DE DATOS, INTELIGENCIA ARTIFICIAL.

## DESCRIPCIÓN O CONTENIDO

ADAMIX 1.O, es una herramienta software para la clasificación mediante agrupamientos de datos con variables cualitativas y cuantitativas, utilizando para ello algoritmos que permiten agrupar datos con variables mixtas. Los algoritmos implementados son: Algoritmo de Agrupamiento Conceptual Conjuntivo basado en Cluster/2, Algoritmo de Agrupamiento Simbólico basado en una Nueva Medida de Similaridad y Algoritmo de Agrupamiento basado en Métricas Mixtas Ponderadas Klass.

El software cuenta con 5 módulos: Módulo de Entrada de Datos, que permite al usuario realizar la entrada de datos ya sea del formato universal ASCII, formatos comerciales como Excel ó del formato predeterminado por herramientas anteriores creadas por el grupo LINCE, así como la creación de nuevos conjunto de datos mediante la inserción de los mismos a través de cuadrículas proporcionadas por la herramienta.

El Módulo de Preparación de Datos, el cual permite al usuario realizar la preparación y depuración de los datos, tales como normalización, llenado de campos vacíos y cálculos estadísticos básicos.

El Módulo de Procesamiento permite al usuario llevar a cabo la clasificación de los datos por medio de los algoritmos de agrupamiento anteriormente expuestos. Estos pueden ser ejecutados en forma interactiva o en procesamiento por lotes.

El Módulo de Presentación de Resultados, muestra al usuario los resultados de los procesos de agrupamiento realizados a los datos, por los diferentes algoritmos implementados, mediante gráficos y tablas.

El Módulo de Ayuda es un módulo externo que contiene la información relativa al manejo de la herramienta, y los conceptos teóricos utilizados en la clasificación de los datos.

---

[*] Trabajo de grado.

[**] Facultad de Ingenierías Físico-Mecánicas. Escuela de Ingeniería de Sistemas e Informática. Director: Ing. Addisson Salazar Afanador; Master en Informática, DEA en Telecomunicaciones y MBA. Codirector: Ing. Fernando Ruiz Díaz; Magíster of Engineering, Profesor de la Escuela de Ingeniería de Sistemas e Informática.

**TITLE**[*]

# ADAMIX 1.0

SOFTWARE TOOL FOR THE CLASSIFICATION OF DATA WITH QUALITATIVE AND QUANTITATIVE VARIABLES.

**AUTHORS**[**]

PAOLA JANNETH VELÁSQUEZ CALDERÓN
VICTORIA ISABEL FERNÁNDEZ PIÑA

**KEYWORDS:**

CLASSIFICATION, CLUSTERING IN MIXED DOMAINS, DATA MINING, ARTIFICIAL INTELLIGENCE.

**DESCRIPTION:**

ADAMIX 1.0, is a software tool for the classification by means of data clustering with qualitative and quantitative variables, using for it algorithms that allow to group data with variables in mixed domains. Implemented algorithms are: Conjunctive Conceptual Clustering Algorithm based on Cluster/2, Symbolic Clustering based on a New Similarity Measure, Clustering Algorithm based in Pondered Mixed Metrics Klass.

The software has 5 modules: Data Entry Module that allows the user to realize the entry of data of the ASCII universal format, formats commercial as Excel or of the format predetermined by previous tools created by the group LINCE, as well as the creation of new data set by means of the insert of the same ones through grids provide by the tool.

The Data Preparation Module, which allows the user to realize preparation and depuration of the data, such as normalization, missing data filling and basic statistics calculation.

The Processing Module allows the user to realize the classification of the data by means of clustering algorithms previously exposed. They can be run in interactive or batch processing mode.

The Results Presentation Module, displays to the user the result from the clustering processes realized to the data, for the different implemented algorithms, by means of graphics and chart.

The Help Module is an external module that contains information regarding to the handling of the tool, and theoretical concepts used in data classification.

---

[*] Final Degree Project.

[**] Faculty of Physical and Mechanical Engineering. School of Computer Science and Informatics. Director: Ing. Addisson Salazar Afanador. Codirector: Ing. Fernando Ruiz Díaz.

# INTRODUCCIÓN

Desde la antigüedad el hombre ha clasificado los conjuntos de objetos por instinto. De hecho la capacidad de clasificar o agrupar es una de las características básicas de la actividad mental humana. El proceso de clasificación fue manual y fundamentalmente artesanal hasta bien entrado el siglo XVIII en el que Adanson[1] establece los principios básicos de la taxonomía numérica y hace el primer esfuerzo para sistematizar este proceso. La primera formulación moderna de las técnicas de clasificación automática debida a Sokal y Sueath, aparece en 1963[2]; precisamente cuando la informática toma cuerpo como herramienta de gran potencia de cálculo.

Desde entonces los métodos de clasificación automática han recuperado la popularidad[3] y son utilizados en todas sus variedades para conocer la estructura de grandes conjuntos de datos, lo cual incide de lleno en los objetivos básicos de los emergentes procesos de Minería de Datos que tan de moda ha puesto la Sociedad de la Información y las Nuevas Tecnologías; sin embargo dichos métodos han encontrado un reto crucial en lo concerniente a la capacidad de manejar objetos representados por la combinación de atributos numéricos y no numéricos (cuantitativos y cualitativos) y todas las derivaciones de estos dos tipos fundamentales de datos.

El proyecto aquí expuesto conjuga el uso de tecnologías informáticas en la implementación de una herramienta software para la clasificación de datos con variables cualitativas y cuantitativas. Dicha clasificación se aborda desde diferentes métodos o algoritmos de agrupamiento, entre los que encontramos:

---

[1] M.Adanson.Histoire naturelle du Senegal, Bauche, Paris. 1757.
[2] R.R.Sokal y P.H.A.Sneath.Principles of Numerical Taxonomy.W.H.Freeman, San Francisco.1963.
[3] K.Gibert y U.Cortés. Una herramienta estadística para la creación de prototipos en dominios poco estructurados. Ed. Grupo Noriega Editores. México. Febrero,1992.

El algoritmo de agrupamiento basado en métricas mixtas ponderadas para variables cualitativas y cuantitativas KLASS de Karina Gibert, orientado a la clasificación de dominios poco estructurados.

El algoritmo de agrupamiento simbólico, jerárquico, aglomerativo y no paramétrico de Gowda y Diday, basado en una medida de similaridad constituida por tres componentes: la posición, la extensión, y el contenido; orientado a la adquisición de conocimiento a través de la descripción de objetos simbólicos compuestos.

El algoritmo de agrupamiento conceptual conjuntivo jerárquico de Michalski y Stepp basado en la metodología de estrella, el cual optimiza un criterio de calidad de agrupamiento predefinido, y está orientado a la formación de grupos que posean una fácil interpretación conceptual y a su vez tengan en cuenta los conceptos emergentes en la descripción de una colección de objetos.

Los resultados obtenidos a lo largo del proyecto se presentan en este documento de la siguiente manera: En primer lugar se encuentra la descripción del proyecto donde se plantea el problema y los objetivos generales y específicos del mismo; a continuación el marco teórico en donde se hace una descripción detallada de los algoritmos y métodos implementados en la herramienta; posteriormente se presentan las etapas seguidas durante el diseño y elaboración de la herramienta software; seguido de un análisis comparativo entre los algoritmos implementados y por último se incluye un capítulo sobre la evaluación de la herramienta software, consistentes en análisis sobre datos de prueba, datos reales y pruebas alfa realizadas con usuarios.

# 1 DESCRIPCIÓN DEL PROYECTO

## 1.1 PLANTEAMIENTO DEL PROBLEMA

Uno de los problemas que se presenta en muchas disciplinas de carácter aplicado lo constituye el análisis de datos; y aumenta proporcionalmente con el volumen de los mismos, de igual manera el problema se hace aún más complicado en la medida en que los datos se hacen más complejos, por ejemplo, -si éstos dejan de ser exclusivamente numéricos para presentarse mezclados con datos de naturaleza cualitativa, con subjetividad, imprecisión y otros elementos de esta índole, - mayor es la dificultad de extraer información útil de ellos. Es en este punto donde entran a jugar un papel importante las técnicas de minería de datos que soportan dominios mixtos.

## 1.2 JUSTIFICACIÓN

El gran volumen de datos, generado actualmente, ha enfocado el interés de muchos investigadores alrededor del mundo, hacia el aprovechamiento de esta circunstancia para la obtención de información útil. Parte de los esfuerzos se han dirigido al desarrollo de técnicas de extracción de datos y patrones, que conforman lo que se ha denominado *Minería de Datos (MD);* dichas técnicas entre otras, dan soporte a los procesos de *Descubrimiento de Conocimiento (DCBD)* en bases de datos u otras fuentes de información.

Entre la extensa variedad de técnicas que existen dentro del proceso de descubrimiento de conocimiento, diferenciadas unas de otras principalmente por sus estructuras de datos y por la arquitectura de cómputo que se formula para su ejecución, encontramos los métodos de *agrupamiento*; una amplia gama de algoritmos basados en la idea de formar grupos automáticamente de tal manera que

los objetos pertenecientes a un grupo tengan la máxima similaridad entre ellos y la máxima disimilaridad con los pertenecientes a otros grupos.

Los métodos de clasificación o separación de datos en grupos son interesantes desde el punto de vista de la Inteligencia Artificial porque abren una puerta a la generación automatizada de reglas, enormemente útil en los ambientes basados en el conocimiento, en particular en aquellos orientados al diagnóstico. Por tal motivo los trabajos realizados alrededor de las técnicas de agrupamiento han sido bastante profusos, encontrando entre sus principales aplicaciones: la caracterización de clientes, formación de taxonomías y clasificación de documentos entre otros; sin embargo han encontrado un reto crucial en lo concerniente a la capacidad de manejar objetos representados en dominios mixtos.

Este reto de clasificación en dominios mixtos se ha convertido en primordial área de estudio para muchos investigadores, entre ellos autoridades en el tema, tales como Chidananda Gowda, E. Diday[4], R. Michalski y R. E. Stepp[5] entre otros; quienes han producido gran parte de la fundamentación teórica existente en este campo, a través de múltiples publicaciones.

El presente proyecto pretende abordar dicho problema de la clasificación de datos en dominios mixtos -que son los más comunes en las bases de datos actuales-, desde el punto de vista de diferentes algoritmos, propuestos por algunos de los autores anteriormente citados, implementándolos y comparándolos en su desempeño a fin de profundizar en el conocimiento sobre cómo operan estas técnicas; y partiendo de una experiencia propia a través de su aplicación a un caso real, formarnos un criterio válido acerca de qué tan efectivos son realmente cada uno de estos métodos al servicio del proceso de descubrimiento de conocimiento.

[4] K. Chidananda. Gowda y E. Diday. "Symbolic clustering using a new similarity measure", IEEE Trans On Systems, man, and cib, 22(2). 1992. pp 368-378.
[5] R. Michalski y R.E. Stepp. "Automated construction of classifications: Conceptual clustering versus numerical taxonomy", IEEE Trans. Pattern Anal Machine Intell; vol.PAMY-5, pp 396-410. 1983.

## 1.3 OBJETIVOS

### 1.3.1 Objetivo General

Desarrollar una herramienta software para la clasificación mediante agrupamientos de datos con variables cualitativas y cuantitativas.

### 1.3.2 Objetivos Específicos

1.3.2.1 Realizar la implementación de varios algoritmos que aborden la clasificación de datos con atributos cualitativos y cuantitativos, desde diferentes enfoques como lo son:

- El algoritmo de agrupamiento basado en métricas mixtas ponderadas para variables cualitativas y cuantitativas KLASS de Karina Gibert, orientado a la clasificación de dominios poco estructurados.

- El algoritmo de agrupamiento simbólico, jerárquico, aglomerativo y no paramétrico de Gowda y Diday, basado en una medida de similaridad constituida por tres componentes: la posición, la extensión, y el contenido; orientado a la adquisición de conocimiento a través de la descripción de objetos simbólicos compuestos.

- El algoritmo de agrupamiento conceptual conjuntivo jerárquico de Michalski y Stepp basado en la metodología de estrella; orientado a la formación de grupos que posean una fácil interpretación conceptual y a su vez tengan en cuenta los conceptos emergentes en la descripción de una colección de objetos.

1.3.2.2 Desarrollar los componentes de software que permitan las siguientes funciones generales: inserción de datos, clasificación de registros con atributos

cualitativos y cuantitativos, y visualización de resultados. La herramienta estará constituida por los siguientes módulos:

a) Módulo de entrada de datos

   A través de este módulo la herramienta leerá los datos de entrada provenientes ya sea del formato universal ASCII, formatos comerciales como Excel ó del formato predeterminado por herramientas anteriores creadas por el grupo LINCE; con el fin de mantener la coherencia y complementariedad entre los proyectos realizados al interior del grupo.

   Se permitirá la creación de nuevos conjuntos de datos, mediante la inserción de los mismos a través de cuadrículas proporcionadas por la herramienta y se crearán archivos de salida con el formato especificado anteriormente.

b) Módulo de preparación de datos

   En este módulo se proporcionarán utilidades que permitan la preparación y depuración de los datos tanto a nivel interno como a nivel de usuario para facilitar los procesos posteriores de agrupamiento.

c) Módulo de procesamiento

   Es el módulo principal de la herramienta, el cual llevará a cabo la clasificación de los datos con variables cualitativas y cuantitativas por medio de los diferentes algoritmos de agrupamiento anteriormente citados.

d) Módulo de presentación de resultados

   Este módulo presentará al usuario los resultados de los procesos de agrupamiento realizados a los datos, por los diferentes algoritmos, mediante gráficos y tablas.

e) Módulo de ayuda

   Este es un módulo externo a través del cual la herramienta presentará las ayudas correspondientes a su manejo, al igual que los conceptos básicos utilizados en la clasificación de los datos.

1.3.2.3 Realizar un estudio comparativo entre los algoritmos de agrupamiento implementados según los siguientes criterios:

- Desempeño: En cuanto al rendimiento, es decir velocidad de procesamiento, consumo de memoria, y al cumplimiento de los objetivos del algoritmo.
- Complejidad: En cuanto a la facilidad de implementación, a la densidad y reutilización de código, y a la convergencia más rápida.

- Conceptualización: En cuanto a la fundamentación teórica de los algoritmos, cuál es más fuerte conceptualmente.

- Precisión: En cuanto a la aproximación de los resultados obtenidos por cada algoritmo a resultados estándares.

1.3.2.4 Validar la herramienta software mediante la realización de pruebas que evalúen los siguientes criterios:

- Calidad: En cuanto a la implementación de los algoritmos, mediante pruebas alfa y beta, y otras pruebas de ingeniería del software tales como carga máxima, y caja negra.

- Funcionalidad: En cuanto a la aplicación a casos de prueba y un caso real. En los casos de prueba se cotejarán los resultados obtenidos con resultados estándares y en el caso real mediante la interpretación de los resultados obtenidos con la herramienta por parte de un experto.

# 2 MARCO TEÓRICO

Este capítulo recopila parte de la fundamentación teórica en la cual se basa el desarrollo del proyecto.

## 2.1 MARCO TEÓRICO GENERAL

### 2.1.1 Descubrimiento de Conocimiento en Bases de Datos (DCBD)

El proceso de descubrimiento de conocimiento en bases de datos DCBD consiste en el análisis exploratorio de bases de datos o grandes cantidades de los mismos. Utilizando para ello una amplia variedad de técnicas, como la minería de datos, la estadística inferencial, y la estadística predictiva entre otras.

#### 2.1.1.1 Ventajas del DCBD

Históricamente, el desarrollo de la estadística ha proporcionado métodos para analizar datos y encontrar correlaciones y dependencias entre ellos, por otra parte, dada la creciente expansión de las bases de datos en los negocios, en los gobiernos y en la ciencia, se destina una gran cantidad de recursos para la adquisición, almacenamiento, procesamiento y análisis de estos inmensos volúmenes de datos; pero esta información cruda es tan voluminosa que resulta inútil, pues no aporta conocimiento o fundamento en gran medida para la toma de decisiones. Muchas veces, el conocimiento más valioso suele estar oculto entre los datos, en forma de patrones o reglas que se relacionen entre sí, este conocimiento podría ser muy útil, por ejemplo, en la toma de decisiones en las organizaciones.

No cabe duda que el valor táctico o estratégico de los grandes almacenes de datos[6] está en proporción directa con la capacidad de analizarlos. Dada la gran gama de patrones (reglas de carácter decisorio, sin plena certeza) que se encuentran en los datos y el inevitable crecimiento de estos datos, se ha creado la necesidad de una nueva generación de herramientas y técnicas para la automatización y el análisis inteligente de bases de datos. Estas herramientas son los objetivos del emergente campo de Descubrimiento de Conocimiento en bases de Datos (DCBD).

#### 2.1.1.2 Proceso DCBD

El proceso DCBD es interactivo e iterativo, involucrando la aplicación de varios algoritmos de minería de datos, además de numerosos pasos en los cuales el usuario toma muchas decisiones. Algunos de dichos pasos son ilustrados a continuación:

a. Entendimiento del dominio de aplicación, el conocimiento relevante a usar y las metas del usuario. Ésta es la tarea que puede llegar a consumir el mayor tiempo.

b. Seleccionar un conjunto o subconjunto de bases de datos, seleccionar y enfocar la búsqueda en subconjuntos de variables, y seleccionar muestras de datos (instancias) en dónde realizar el proceso de descubrimiento.

c. Limpieza y preprocesamiento de datos, diseñando una estrategia adecuada para manejar ruido, valores incompletos, secuencias de tiempo, casos extremos (si es necesario), etc.

d. Selección de la tarea de descubrimiento a realizar, por ejemplo, clasificación, agrupamiento o clustering, regresión, etc.

e. Selección del ó los algoritmos a utilizar.

[6] También conocido como DataWarehouse : término referido a almacenes o bodegas de datos.

f. Transformación de datos al formato requerido por el algoritmo específico de minería de datos

g. Llevar a cabo el proceso de minería de datos.

- Se buscan patrones que pueden expresarse como un modelo o simplemente que expresen dependencias de los datos.

- Se tiene que especificar un criterio de preferencia para seleccionar un modelo dentro de un conjunto posible de modelos.

- Se tiene que especificar la estrategia de búsqueda a utilizar (normalmente está predeterminada en el algoritmo de minería)

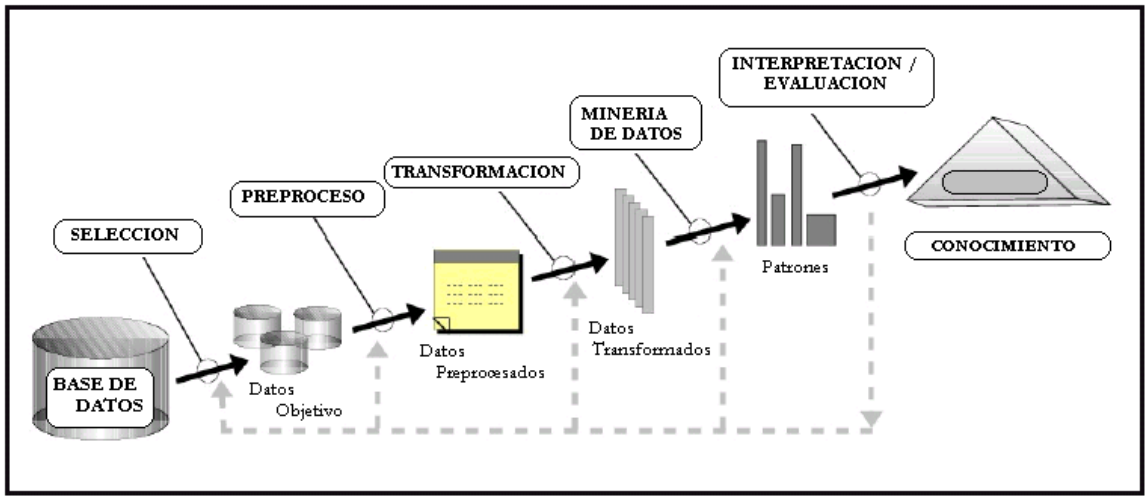
h. Interpretar los resultados y posiblemente regresar a los pasos anteriores. Esto puede involucrar repetir el proceso, quizás con otros datos, otros algoritmos, otras metas y otras estrategias. Este es un paso crucial en donde se requiere tener conocimiento del dominio. La interpretación puede beneficiarse de procesos de visualización, y sirve también para borrar patrones redundantes o irrelevantes.

i. Incorporar el conocimiento descubierto al sistema (normalmente para mejorarlo) lo cual puede incluir resolver conflictos potenciales con el conocimiento existente.

j. El conocimiento se obtiene para realizar acciones, ya sea incorporándolo dentro de un sistema de desempeño o simplemente para almacenarlo y reportarlo a las personas interesadas.

A continuación se presenta una gráfica donde se puede apreciar el proceso DCBD[7] (ver figura 1)

Figura 1. Descripción del proceso de DCBD.




### 2.1.1.3 Limitaciones del DCBD

Se hace necesario presentar algunas limitaciones actuales, y una serie de problemas que deben tenerse en cuenta para la aplicación de investigaciones que deseen aplicar el proceso DCBD.[1]

a. Bases de datos gigantescas: Bases de datos con miles de campos, miles de tablas, millones de registros y muchos gigabytes de tamaño son muy comunes hoy en día, pero se aproxima la época en que bases de datos del orden de los terabytes (10 E+12 bytes) ya han aparecido, y esto representa un reto de grandes proporciones para el DCBD. Se podrían tener soluciones a este problema con la utilización de métodos de minería de datos aproximados y con el procesamiento masivo en paralelo.

[7] Estas etapas fueron descritas originalmente por Fayyad en su libro Advances in Knowledge Discovery and Data Mining. (1996) AAAI Press / The MIT Press.

b. Alta dimensionalidad: No solo hay que tener en cuenta un gran numero de datos sino también un gran numero de campos (atributos, variables) lo cual hace muy grande el problema dimensional para el manejo de los datos. Un conjunto de datos dimensionalmente alto crea problemas en términos del aumento en el tamaño del espacio de búsqueda para modelos de inducción de una manera incrementalmente explosiva.

c. Sobre ajuste: Cuando un algoritmo busca los mejores parámetros para un modelo en particular usando un limitado conjunto de datos, se puede caer en un sobre ajuste en este conjunto de datos, lo cual podría representar un pobre desempeño al aplicar los resultados obtenidos en otro conjunto de datos similares.

d. El cambio dinámico de los datos y su conocimiento: El cambio rápido de los datos puede hacer inválidos los patrones previos. Adicionalmente, las variables en una aplicación de bases de datos pueden ser modificadas, borradas o relacionadas con nuevas variables. Algunas posibles soluciones a estos problemas son la utilización de métodos para actualización de patrones y el tratamiento estos cambios para encontrar nuevo conocimiento.

e. Datos incompletos o perdidos: El manejo de datos incompletos y/o perdidos en una base de datos puede deberse a pérdida de valores de algún atributo (al que se asigna entonces el valor desconocido o NULO), o a la ausencia del mismo. En ambos casos, la incidencia en el resultado dependerá de si el dato incompleto es relevante o no para el objetivo de descubrimiento de conocimiento.

f. Relaciones complejas entre atributos: Los atributos jerárquicamente estructurados, y las relaciones no triviales entre atributos hacen suponer la aplicación de formas mas sofisticadas de representar el conocimiento de las bases de datos, esto sugiere la utilización de algoritmos que puedan aplicar efectivamente tal representación de la información.

g. Patrones entendibles a los investigadores: En muchas aplicaciones se hace imprescindible la utilización de mecanismos que presenten los patrones encontrados mucho mas fáciles de entender para las personas. Algunas posibles soluciones incluyen representaciones gráficas, reglas estructuradas con grafos acíclicos, generación de lenguaje natural y técnicas para la visualización de datos y conocimiento.

h. Integración con otros sistemas: Se hace necesario la integración con sistemas como las DBMS (Data Base Management System), que permitirían el aumento en el potencial para los procesos de DCBD.

### 2.1.2 Minería de Datos

La minería de datos, es la exploración y el análisis de grandes cantidades de datos, por medios semiautomáticos ó totalmente automáticos, con el fin de descubrir reglas o patrones significativos de datos. El proceso de minería involucra ajustar modelos o determinar patrones a partir de datos. Este ajuste normalmente es de tipo estadístico, en el sentido que se permite un cierto ruido o error dentro del modelo.

Los algoritmos de minería de datos realizan en general tareas de *descripción* (de datos y patrones), de *predicción* (de datos desconocidos) y de *segmentación* (de datos). Otras, como *análisis de dependencias* e *identificación de anomalías* se pueden utilizar tanto para descripción como para predicción.

*Descripción:* Normalmente es usada para análisis preliminar de los datos (resumen, características de los datos, casos extremos, etc.). Con esto, el usuario se familiariza con los datos y su estructura.

Busca derivar descripciones concisas de características de los datos tales como medias, desviaciones estándares, etc.

*Predicción:* La Predicción la podemos dividir en dos: Clasificación y Estimación.

- Clasificación: Los datos son objetos caracterizados por atributos que pertenecen a diferentes clases (etiquetas discretas). La meta es inducir un modelo para poder predecir una clase dados los valores de los atributos. Se usan por ejemplo, árboles de decisión, reglas, análisis de discriminantes, etc.

- Estimación o Regresión: La meta es inducir un modelo para poder predecir el valor de la clase dados los valores de los atributos. Se usan por ejemplo, árboles de regresión, regresión lineal, redes neuronales, Algoritmo del vecino más cercano kNN, etc.

*Segmentación:* Separación de los datos en subgrupos o clases interesantes. Las clases pueden ser exhaustivas y mutuamente exclusivas o jerárquicas y con traslapes.

Se puede utilizar con otras técnicas de minería de datos: considerando cada subgrupo de datos por separado, etiquetando cada uno de éstos y utilizando un algoritmo de clasificación. Se usan algoritmos de agrupamiento, SOM (*self-organization maps*), EM (*expectation maximization*), k-means, etc.

Normalmente el usuario tiene una buena capacidad de formar las clases y por esto existen herramientas visuales interactivas para ayudar al usuario en este aspecto.

*Análisis de dependencias:* El valor de un elemento puede usarse para predecir el valor de otro. La dependencia puede ser probabilística, puede definir una red de dependencias o puede ser funcional (leyes físicas).

También se ha enfocado a encontrar si existe una alta proporción de valores de algunos atributos que ocurren con cierta medida de confianza junto con valores de otros atributos. Se pueden utilizar redes bayesianas, redes causales, y reglas de asociación.

*Detección de desviaciones, casos extremos o anomalías:* Detectar los cambios más significativos en los datos con respecto a valores pasados o normales. Sirve para

filtrar grandes volúmenes de datos que son menos probables de ser interesantes. El problema está en determinar cuándo una desviación es significativa para ser de interés.

#### 2.1.2.1 Taxonomía de Técnicas de Minería de Datos

Las técnicas de minería de datos crean modelos que son *predictivos* y/o *descriptivos.*

Un modelo predictivo responde preguntas sobre datos futuros como:

- ¿Qué tipo de seguro es más probable que contrate el cliente X?
- ¿Cuáles serán las ventas del año próximo?.

Un modelo descriptivo proporciona información sobre las relaciones entre los datos y sus características. Genera información del tipo:

- Los clientes que compran pañales suelen comprar cerveza.
- El tabaco y el alcohol son los factores más importantes en la enfermedad.
- Los clientes sin televisión y con bicicleta tienen características muy diferenciadas del resto.

A continuación se presenta la taxonomía de algunos de los algoritmos de minería de datos más conocidos.

**Descriptivos**

*Técnicas de Análisis exploratorio:*

- Estudios correlacionales
- Asociaciones.
- Dependencias.
- Detección datos anómalos.
- Análisis de dispersión.

*Técnicas Segmentación:*

- K-medias (aprendizaje competitivo).

- Redes neuronales de Kohonen.
- EM (Medias estimadas) (Dempster et al. 1977).
- Cobweb (Fisher 1987).
- AUTOCLASS (Cheeseman & Stutz 1996).

**Predictivos**

*Dependiendo de si se estima una función o una correspondencia:*

- clasificación: se estima una función (las clases son disjuntas).
- Basado en la similaridad: Algoritmo del Vecino más cercano.
- Basado en Limitar y completar: k-medias (Aprendizaje competitivo), Aprendizaje del perceptron, Métodos multicapa como Redes Neuronales con propagación hacia atrás, Funciones de base radial, Aprendizaje mediante árboles de decisión, por ejemplo. ID3, C4.5, CART, Clasificadores de bayes, Método de separación de centros, Reglas(CN2).
- categorización: se estima una correspondencia (las clases pueden solapar).

*Dependiendo del número y tipo de clases:*

- Clase *discreta*: se conoce como "clasificación". Ejemplo: determinar el grupo sanguíneo a partir de los grupos sanguíneos de los padres.
- Si sólo tiene dos valores (V y F) se conoce como "Aprendizaje de conceptos". Ejemplo: Determinar si un compuesto químico es cancerígeno.
- Clase *continua* o discreta ordenada: se conoce como "estimación". Ejemplo: estimar el número de hijos de una familia a partir de otros ejemplos de familias.

### 2.1.3 Aprendizaje Automático

La principal cuestión que pretende resolver el aprendizaje automático es la de cómo hacer que un programa de computadora aprenda y mejore a partir de la experiencia. Para esto se tendría que recurrir a diversas técnicas que permitan su programación, cumplir esto sería de un gran impacto para la sociedad, puesto que, se tendrían computadoras que aprenderían de la experiencia de muchos científicos para

prevenir y solucionar una gran cantidad problemas que aquejan a la humanidad. Hay que destacar la importancia del aprendizaje automático, sobre todo en la minería de datos, por sus técnicas para descubrir leyes, que describen relaciones cualitativas y/o cuantitativas existentes entre los datos.

Es necesario tener en cuenta que aún no se sabe como hacer que las computadoras aprendan tan bien como lo hacen las personas, sin embargo, existen algoritmos que han sido desarrollados y que son efectivos en cierto tipo de tareas, una comprensión por lo menos teórica sobre el aprendizaje ha comenzado a surgir. Actualmente existen programas de computadora que realizan cierto tipo de aprendizaje y que han hecho surgir un buen número de aplicaciones comerciales que son utilizadas por diversas organizaciones.

Para poder entender del concepto de aprendizaje automático se hace necesario tener una definición comprensible y que se puede adoptar como referencia, con este objetivo se ha tomado la definición que expresa Tom Mitchell en su libro Machine Learning*: An Artificial Intelligence Approach*. McGraw-Hill. (1997). (Prefacio, Cap. 1,2,3).

*Definición. Un programa de computador se dice que aprende de una experiencia E con respecto a una clase de tarea T y una medida de desempeño P, si el desempeño de esta tarea T medida por P, es mejorado con la experiencia E.*

#### 2.1.3.1 Aprendizaje Inductivo

Se trata del paradigma más estudiado dentro del aprendizaje automático, aunque se suele asociar únicamente a aproximaciones simbólicas, las aproximaciones no simbólicas también son inductivas. Su base es partir de un gran número de ejemplos o muestras que corresponden a uno o varios conceptos y construir una representación de éstos que permita hacer predicciones al ver nuevos ejemplos, sin utilizar otro conocimiento más que los propios ejemplos.

La debilidad de este aprendizaje es que no tiene una base exclusivamente heurística, las reglas que permiten crear descripciones generales a partir de ejemplos no están basadas en mecanismos fundamentados lógicamente. Por lo tanto, es posible que la aparición de nuevos ejemplos invalide el nuevo conocimiento que se ha generado. No obstante, la mayoría del aprendizaje que realiza un ser humano es de naturaleza inductiva, es imposible disponer de todos los ejemplos posibles de un concepto para intentar definirlo.

Se denomina inferencia a todo proceso lógico que a partir de un determinado nivel de conocimiento genera conocimiento de un nivel mayor. Dos procesos de inferencia usuales son la *inducción* y la deducción que representan los dos principales tipos de aprendizaje.

Un sistema de aprendizaje aplica la inducción a los hechos u observaciones suministradas, para obtener nuevo conocimiento. La inferencia inductiva no preserva la verdad del conocimiento, solo su falsedad; si partimos de hechos falsos, el conocimiento adquirido por inducción será falso, pero si los hechos son verdaderos el conocimiento inducido será valido con cierta probabilidad (y no con certeza absoluta, como ocurre con la deducción). La inducción en un sentido amplio consiste en encontrar propiedades comunes a un subconjunto finito de elementos de un cierto dominio y considerar esas propiedades extensibles a cualquier elemento del dominio.

Hipótesis del aprendizaje inductivo. *Cualquier función que aproxime bien a la función objetivo sobre un conjunto de entrenamiento suficientemente grande, también la aproximará bien sobre el resto de los ejemplos del dominio.(*Mitchell*)*

Desde cierto punto de vista se podría llegar a controvertir sobre la palabra "*bien*" de la hipótesis expresada anteriormente, puesto que al aproximarse muy bien al conjunto de entrenamiento se haría a esta función muy sesgada hacia el mismo.

De las distintas disciplinas que convergen en el DCBD es el aprendizaje automático (Machine Learning), el que da el soporte principal en la construcción de herramientas de Minería de Datos.

Existen dos tipos de aprendizaje inductivo: el aprendizaje supervisado y el aprendizaje no supervisado.

*Aprendizaje supervisado:* El nuevo conocimiento es inducido mediante la generalización a partir de una serie de ejemplos y contraejemplos, que son avalados por un experto en el área. Este método también se conoce como adquisición de conceptos o aprendizaje por ejemplos.

En el aprendizaje inductivo supervisado existe un atributo especial, normalmente denominado *clase,* presente en todos los ejemplos que especifica si el ejemplo pertenece o no a un cierto concepto, que será el objetivo del aprendizaje. El atributo clase normalmente toma los valores + y -, que significan la pertenencia o no del ejemplo al concepto que se trata de aprender; es decir, que el ejemplo ejemplifica positivamente al concepto -pertenece al concepto- o bien lo ejemplifica negativamente -que no pertenece al concepto. Mediante una generalización del papel del atributo clase, cualquier atributo puede desempeñar ese papel, convirtiéndose la *clasificación* de los ejemplos según los valores del atributo en cuestión en el objeto del aprendizaje.

*Aprendizaje no supervisado:* El sistema de aprendizaje analiza una serie de entidades y determina características comunes, que pueden ser agrupadas formando un concepto previamente desconocido, esto se realiza sin supervisión especializada. Se conoce como *formación de conceptos* o *aprendizaje por observación y descubrimiento.*

El aprendizaje inductivo no supervisado estudia el aprendizaje sin la ayuda del *maestro*; es decir, se aborda el aprendizaje sin supervisión, que trata de ordenar los ejemplos en una jerarquía según las regularidades en la distribución de los pares atributo-valor sin la *guía* del atributo especial *clase*. Este es el proceder de los

sistemas que realizan *clustering* conceptual y de los que se dice también que adquieren nuevos conceptos. Otra posibilidad contemplada para estos sistemas es la de sintetizar conocimiento cualitativo o cuantitativo, objetivo de los sistemas que llevan a cabo tareas de *descubrimiento.*

#### 2.1.3.2 Aprendizaje Analítico o Deductivo

Los métodos usados en este tipo de aprendizaje intentan reformular el conocimiento que posee un sistema a base de generalizaciones a partir de analizar las resoluciones obtenidas por algún mecanismo de resolución de problemas incluido en el propio sistema.

La base de este aprendizaje no es pues la generación de nuevo conocimiento, sino hacer más eficiente el que ya se posee. En estos términos se ha criticado este tipo de aprendizaje, considerando que no es aprendizaje real dado que no genera nuevo conocimiento. No obstante, si lo es en función de que permite mejorar el rendimiento del sistema que lo utiliza.

En contraste con el aprendizaje inductivo, la base del proceso será el uso de grandes cantidades de conocimiento del dominio, que asociado con ejemplos de resolución de problemas permitirán reformular el conocimiento del sistema. Esta reformulación dará como resultado tanto, nuevas formas de expresar el conocimiento que ya se posee en términos mas sencillos, como el desarrollo de conocimiento que indique la forma más efectiva para resolver un problema, o qué condiciones permiten descartar posibles formas de resolución por ser probablemente erróneas o inútiles [2].

El mayor problema de este método de aprendizaje consiste en decidir cuándo se debe incorporar este nuevo conocimiento al sistema, ya que incluir todo lo que es posible aprender, puede hacer que las nuevas resoluciones sean menos eficientes, al tener que revisar el nuevo conocimiento adquirido.

Otro problema está en que el conocimiento del dominio que se posee no siempre es completo y consistente, por lo que a veces es necesario combinar este aprendizaje con el aprendizaje inductivo para completarlo [3].

## 2.2 MARCO TEÓRICO ESPECÍFICO

### 2.2.1 Agrupamiento en Dominios Mixtos

*Agrupamiento* es el proceso de formar grupos de tal manera que los objetos de un grupo tienen una *similaridad* alta entre ellos, y baja con objetos de otros grupos. En el caso particular de los dominios mixtos el problema que se plantea es cómo obtener una medida de similaridad consistente para comparar dos objetos entre sí, de modo que se consideren todas las variables que los describen, sea cual sea su naturaleza, la cual puede catalogarse dentro de uno de los siguientes tipos:

I. Cuantitativa
   a. Valores de radio continuo. Ejp: velocidad, Longitud.
   b. Valores discretos absolutos. Ejp: personas, casas.
   c. Intervalo de valores. Ejp: intervalo de duración de un evento.

II. Cualitativa
   a. Nominal (no ordenada). Ejp: color, género.
   b. Ordinal (ordenada). Ejp: tamaño, rango militar.

III. Estructurada (ordenada en forma de árbol)

La ventaja de poder tratar con los objetos considerando todas las variables para su descripción, radica en que los conjuntos de datos reales, a menudo describen los individuos con variables de tipo cuantitativo y cualitativo simultáneamente. Ésta es una situación habitual en el caso de los dominios de estructura muy compleja donde

se manejan muchas características mezcladas, relevantes en la representación de los objetos, las cuales por separado no presentan una descripción completa.

El manejo de conjuntos de datos descritos por este tipo de variables mixtas o heterogéneas requiere un tratamiento especial, debido a que los métodos de agrupamiento estándar fueron concebidos originalmente para tratar sólo con variables cuantitativas, lo que algunos autores consideran el campo de la taxonomía numérica, basado en la vecindad entre los elementos de la población; siendo la primera aproximación al denominado agrupamiento cualitativo la presentada Por Michalski en 1983 en lo que llamó clustering (agrupamiento) conceptual, que posteriormente fue retomado por otros autores adquiriendo nuevas concepciones. El agrupamiento en dominios mixtos conjuga estos dos tipos de agrupamiento cualitativo y cuantitativo utilizando para esto funciones que permitan fusionar métricas cuantitativas y cualitativas en métricas mixtas.

La medida de similaridad está basada en los atributos que describen a los objetos. Los grupos pueden ser exclusivos, con traslapes, probabilísticos, jerárquicos. Entre las aplicaciones encontramos: caracterizar clientes, formar taxonomías, clasificar documentos, etc.

Los principales retos a los que se deben enfrentar los algoritmos de agrupamiento en dominios mixtos, y en general los diferentes tipos de algoritmos de agrupamiento, para constituirse en herramientas de gran potencia al servicio del proceso DCBD son:

a. Escalabilidad: normalmente corren con pocos datos.
b. Capacidad de manejar diferentes tipos de atributos: numéricos, binarios, nominales, ordinales, etc.
c. Grupos de formas arbitrarias: los basados en distancias numéricas tienden a encontrar clusteres esféricos.
d. Requerimientos mínimos para especificar parámetros, como el número de grupos a formar.
e. Manejo de ruido: muchos son sensibles a datos erróneos.

f. Independientes del orden de los datos.
g. Poder funcionar eficientemente con alta dimensionalidad.
h. Capacidad de añadir restricciones.
i. Que los grupos formados sean interpretables y utilizables.

***Medidas de similaridad:*** La medida de similaridad se define usualmente por proximidad en un espacio multidimensional. Para datos numéricos, usualmente se pasa primero por un proceso de estandarización. Las medidas más utilizadas son:

1) Para variables numéricas (lineales):

- Distancia Euclideana:

$$d(i,j) = \sqrt{\left|x_{i1} - x_{j1}\right|^2 + \left|x_{i2} - x_{j2}\right|^2 + \cdots + \left|x_{in} - x_{jn}\right|^2}$$

- Distancia Manhattan:

$$d(i,j) = \left|x_{i1} - x_{j1}\right| + \left|x_{i2} - x_{j2}\right| + \cdots + \left|x_{in} - x_{jn}\right|$$

- Distancia Minkowski:

$$d(i,j) = \left(\left|x_{i1} - x_{j1}\right|^q + \left|x_{i2} - x_{j2}\right|^q + \cdots + \left|x_{in} - x_{jn}\right|^q\right)^{1/q}$$

Si $q=1$ es Manhattan y si $q=2$ es Euclideana.

- Distancia Pesada (por ejemplo Euclideana):

$$d(i,j) = \sqrt{w_1\left|x_{i1} - x_{j1}\right|^2 + w_2\left|x_{i2} - x_{j2}\right|^2 + \cdots + w_n\left|x_{in} - x_{jn}\right|^2}$$

propiedades de las distancias: (i) $d(i,j) \geq 0$ , (ii) $d(i,i) = 0$ , (iii) $d(i,j) = d(j,i)$ , (iv) $d(i,j) \leq d(i,h) + d(h,j)$

2) Variables Binarias (0,1):

- Simétricas (ambos valores tienen el mismo peso):

$$d(i,j) = \frac{r+s}{q+r+s+t}$$

donde: $q$ = número de valores que son 1 en las dos, $r$ = número de valores que son 1 en $i$ y 0 en $j$, $s$ = número de valores que son 0 en $i$ y 1 en $j$, y $t$ = número de valores que son 0 en las dos.

- No-simétricas (el más importante y más raro vale 1), conocido como el coeficiente Jaccard:

$$d(i,j) = \frac{r+s}{q+r+s}$$

3) Variables nominales (ejemplo: color):

$$d(i,j) = \frac{p-m}{p}$$

donde: $m$ = número de valores iguales, $p$ = número total de casos.
Se pueden incluir pesos para darle más importancia a $m$.
Se pueden crear nuevas variables binarias asimétricas a partir de las nominales (ejemplo: es amarillo o no).

4) Variables ordinales: nominales con un orden relevante. El orden es importante, pero no la magnitud. Pasos:

a) Cambia el valor de cada variable por un índice $r_{if} \in \{1,...,M_f\}$, donde $M_f$ es el índice del valor más alto de la variable.

b) Mapeo el índice entre 0 y 1 para darle igual peso

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

c) Usa cualquiera de las medidas numéricas anteriores.

5) Variables escalares no lineales, por ejemplo, variables que siguen una escala exponencial. Posibilidades:

a) Tratarlas como numérica normal.

b) Obtener su logaritmo (o alguna otra transformación) antes para convertirlas en lineales.

c) Considerarlas como variables ordinales.

6) Variables mixtas:
Una posibilidad es escalar todas las variables a un intervalo común (entre 0 y

$$d(i,j) = \frac{\sum_{f=1}^{P} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{P} \delta_{ij}^{(f)}}$$

donde: $\delta_{ij}^{(f)} = 0$ si $x_{if}$ o $x_{jf}$ se desconocen o si los dos valores son 0 y la variable es asimétrica binaria. En caso contrario vale 1. $d_{ij}^{(f)}$ depende del tipo:

- Si $f$ es binaria o nominal: $d_{ij}^{(f)} = 0$ sí $x_{if} = x_{jf}$, sino, $d_{ij}^{(f)} = 1$.
- Si $f$ es numérica lineal: $d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{\max_{hx,hf} - \min_{hx,hf}}$
- Si $f$ es ordinal o numérica no lineal: calcula los índices $r_{if}$ y $z_{if} = \frac{r_{if} - 1}{M_f - 1}$ y toma a $z_{if}$ como numérica lineal.

#### 2.2.1.1 Agrupamiento Conceptual

El término de agrupación conceptual se debe a Ryszard S. Michalski [4], él lo define como:

"Agrupar objetos en clases conceptualmente simples basadas en los valores de los atributos tomando en consideración todo conocimiento acerca de las relaciones semánticas entre los atributos de los objetos o cualquier concepto global que pueda ser usado para caracterizar las clases que se forman."[8] [9].

Por lo tanto, el rasgo distintivo de la agrupación conceptual es que intenta introducir la mayor cantidad de conocimiento sobre el contexto en el que se quiere realizar el aprendizaje que pueda ser útil

El origen de estos métodos parte de la constatación de la falta de contexto de las típicas medidas de similaridad. Éstas sólo tienen en cuenta a la hora de contrastar dos objetos los valores de sus atributos, y no consideran los conceptos que pueden ayudar a describirlos. Se comprobó que las características que permiten describir a un grupo de objetos como pertenecientes a una categoría no se encuentra únicamente en la comparación de sus propiedades, sino que hace falta más información proveniente del contexto.

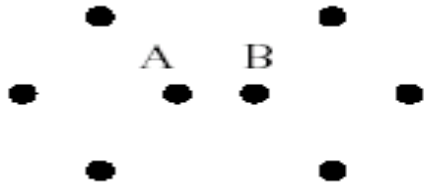
La agrupación conceptual pretende pues asignar los objetos a clases no sobre la base de una distancia entre ellos, sino a su pertenencia a cierto concepto que les da sentido, es lo que se denomina *pertenencia conceptual (concept membership)***.** Debido a esto, las tareas de división y de clasificación de los objetos no son independientes entre sí, una división en clases de un grupo de objetos sólo será buena sí y sólo sí existe una buena interpretación de las clases. Un ejemplo de agrupamiento conceptual se presenta en la Figura 2, una persona considerando esta figura describiría típicamente los puntos observados (A y B) representando dos

[8] R.S. Michalski, Stepp. "Learning from observation: conceptual clustering. En R.S. Michalski, J.G. Carbonell, T.M. Mitchell, editores, Machine Learning: An Artificial Intelligence Approach, capitulo 11, pags 331-363, Springer, Berling Heidelberg, 1984.

[9] R.S. Michalski, Stepp. "Conceptual clustering: Inventing goal oriented classifications of structured". En , J.G. Carbonell, editores, Machine Learning: An Artificial Intelligence Approach II, capitulo 11, pags 331-363. Ed Tioga, Palo Alto, California, 1986.

diamantes, así que los puntos A y B a pesar de estar uno cerca del otro, son colocados en grupos diferentes.

Figura 2.Una ilustración de agrupamiento conceptual




**Objetos a ser agrupados y sus atributos**

Típicamente, los objetos agrupados provienen de un estudio experimental de algún fenómeno y son descritos por un conjunto específico de atributos(variables) seleccionados por un analista de datos. Los atributos pueden ser medidos en diferentes escalas, tal como nominal, ordinal, intervalo, proporción, y absoluto. En un caso simple, uno puede solamente distinguir atributos cualitativos(medidos sobre la escala ordinal o nominal) de atributos cuantitativos(medidos sobre escalas restantes). Los atributos seleccionados por el analista de datos no son siempre todos relevantes al problema de agrupamiento. En aproximaciones convencionales, la selección de atributos relevantes es tratada como un paso preliminar separado. En el método de agrupamiento conceptual conjuntivo, la selección de atributos es realizada simultáneamente con la formación de grupos. El método selecciona aquellos atributos, desde el punto de vista del criterio asumido, que permite caracterizar los grupos individuales en términos de conceptos disponibles.

**Tipo de Estructura del cluster**

Dado una colección de objetos $E$, el objetivo de agrupamiento es dividir la colección dentro de ciertos subconjuntos significativos. Sea $E_1, E_2, ..., E_K$ subconjuntos (grupos) de $E$, y $\alpha_i$ denota una descripción del subconjunto $E_i$. En general, una

descripción $\alpha_i$ es satisfecha no sólo por objetos en $E_i$, sino también por algunos objetos no observados. Basado sobre la relación entre los grupos y las descripciones del grupo, tres tipos diferentes de estructuras de intercluster son comúnmente distinguidas en la literatura.

- *Estructura de partición:* Un conjunto de grupos cuya unión es el conjunto $E$, y cuyas descripciones son todas disjuntas (esto implica que los grupos por sí mismo son disjuntos)
- *Estructura solapada:* Un conjunto de grupos cuyas descripciones se interceptan.
- *Estructura jerárquica:* Una jerarquía de múltiples niveles en las cuales los grupos en el primer nivel representan una partición del conjunto inicial $E$, y los grupos en el nivel más bajo son elementos de particiones de uno de los grupos padres del nivel más alto.

El método de agrupamiento conceptual conjuntivo genera una estructura de partición o jerárquica de grupos.

**Esquema de representación del cluster**

El propósito de un esquema de representación de grupo es caracterizar simple y generalmente los objetos en un grupo. El agrupamiento conceptual conjuntivo emplea dos esquemas de representación de grupo: un objeto representante único seleccionado de un grupo, llamado la *semilla* del grupo, y una afirmación conjuntiva, que describe todo los objetos en el grupo. Esta afirmación conjuntiva, llamada un *complejo lógico*, es una expresión en el sistema de la lógica estimada de la variable $VL_1$.

Suponer que $x_1, x_2, ..., x_n$ son variables seleccionadas para representar objetos. Asumir que cada variable $x_i$, $i \in \{1,2,...,n\}$, tienen un *dominio* asignado, $DOM(x_i)$, que especifica todos los valores posibles que la variable puede tomar para cualquier objeto en la colección a ser agrupada. El número de tales valores es dado por $d_i$.

Los dominios son asumidos finitos, y representados generalmente como $\mathrm{DOM}(x_i) = \{0,1,2,...d_i - 1\}$.

La descripción de la extensión del espacio por las variables $x_1, x_2, ..., x_n$ es llamada el *espacio del evento*. Cada punto (*evento*) en el espacio es un vector de valores específicos de las variables $x_1, x_2, ..., x_n$. Un evento que es una descripción de algún objeto en la colección a ser agrupada es llamado un *evento observado*. Otros eventos son llamados *eventos no observados.*

Una *afirmación relacional*[10] (o un *selector*) es una forma

$$[x_i \# R_i]$$

donde $R_i$, la *referencia*, es una lista de elementos del dominio de la variable $x_i$, conectados por la *disyunción interna*, denotada por "V." # representa el operador relacional = o $\neq$.

El selector $[x_i = R_i]$*(*$[x_i \neq R_i]$*)* es interpretado como "el valor $x_i$ es uno de los elementos de $R_i$" ("el valor de $x_i$ no es un elemento de $R_i$"). En el caso de variables lineales, la notación de un selector puede ser simplificada por el uso de operadores relacionales $\geq$, $>$, $<$, $\leq$, y un operador de rango "**. .**" como se muestra a bajo. Estos son unos pocos ejemplos de un selector, en donde las variables y sus valores son representados por términos lingüísticos:

| | |
|---|---|
| [longitud $>$ 2] | (la longitud es más grande que 2) |
| [color = azul $\vee$ rojo] | (el color es azul o rojo) |
| [tamaño $\neq$ medio] | (el tamaño no es medio) |
| [Peso = 2. .5] | (el peso está entre 2 y 5, inclusivamente). |

Un producto lógico de selectores es denominado un *complejo lógico(l* – complejo). Un conjunto de objetos que satisfacen todos los selectores en un *l* – complejo es

[10] Esta forma fue introducida primero en el sistema lógico Variable-estimado uno($VL_1$)

denominado un *s – complejo* (set – complex). Así, un *l* – complejo es la descripción de un *s – complejo.* Por ejemplo el *l – complejo*:

[altura = alto] [color = azul $\vee$ rojo] [longitud > 2] & [tamaño $\neq$ medio] [peso = 2..5]

(la operación AND es denotada por & o implícito por la concatenación de los selectores) describe aquellos objetos que son altos, azules o rojos, con longitud mayor que 2, tamaño no medio, y peso entre 2 y 5. El conjunto de tales objetos constituye el correspondiente *s – complejo*. La distinción entre *l* – y *s – complejo* es que se usa para permitir la aplicación de operadores lógicos o teoría de conjunto, respectivamente, cualquiera que sea es conveniente. Cuando esta distinción es sin importancia, el termino complejo se usará sin un prefijo.

No toda colección de objeto constituye un *s–complejo*, es decir, no toda colección puede ser precisamente descrita por un *l–complejo*. Esto es, sin embargo, posible para describir cada colección de objetos por un *l–complejo*, sí el *l–complejo* es permitido también para describir algunos objetos adicionales(es decir, sí esto es permitido para ser una descripción generalizada de la colección). Por ejemplo, los eventos

$e_1$ : (azul, grande, circular)

$e_2$ : (rojo, medio, circular)

pueden ser descritos por el complejo

[Color = azul $\vee$ rojo][tamaño $\geq$ medio][forma = circular].

Este complejo, sin embargo, también cubre los eventos no observados

$e_3$ : (rojo, grande, circular)

$e_4$ : (azul, medio, circular)

los cuales son distintos de $e_1$ y $e_2$. El número de eventos no observados cubiertos por un complejo es llamado *dispersión absoluta* del complejo.

Dado un conjunto de complejos, la dispersión absoluta del conjunto es definido como la suma de dispersiones de los complejos en éste. En adición para la

dispersión absoluta, se introduce otro tipo de dispersión, la dispersión *proyectada* de un conjunto de complejos. Este tipo de dispersión es aplicable únicamente para un agrupamiento, el cual es un conjunto de complejos disjuntos de parejas. Dado que los complejos son parejas disjuntas, existe, por algún par de complejos, por lo menos dos selectores con la misma variable y referencias disjuntas. Las variables involucradas en tales selectores son llamadas las *variables discriminantes* de un agrupamiento. La dispersión proyectada de un agrupamiento es la suma de las dispersiones determinadas en la extensión de espacio del evento justo sobre las variables discriminantes. Por ejemplo, si dado dos eventos observados $e_1$ y $e_2$ citados anteriormente, los dos complejos

- [color = azul][tamaño = grande][forma = circular]
- [color = rojo][tamaño $\leq$ medio][forma = circular $\vee$ cuadrada]

tienen dispersiones proyectadas de 0 y 1, respectivamente, en extensión de espacio para las dos variables discriminantes color, y tamaño(asumiendo que tamaño toma solamente los valores pequeño, medio, y grande).

**Criterio de agrupamiento**

El principio tradicional para agrupar objetos dentro de grupos, es utilizar alguna medida numérica de similaridad de objetos, usualmente un recíproco de una medida de distancia. En agrupamiento conceptual, los objetos son reunidos dentro de grupos que representan un concepto singular(términos lingüísticos o funciones de lógica simple definida sobre tales términos). En agrupamiento conceptual conjuntivo, los objetos son agrupados dentro de grupos que son caracterizados por relaciones de productos lógicos sobre los atributos de los objetos seleccionados, es decir, conceptos conjuntivos. Estas relaciones también pueden incluir disyunción de propiedades, pero sólo si la disyunción envuelve valores del mismo atributo. Este tipo de disyunción es llamado *disyunción interna*.

**Criterio de parada**

El criterio de parada del algoritmo *CLUSTER/2* está estrechamente ligado al criterio de calidad de agrupamiento, de tal manera que cuando la calidad no mejora en determinado número de agrupamientos sucesivos el algoritmo se detiene. Lo cual conduce al problema de cómo juzgar la calidad de un agrupamiento, y parece no haber una respuesta universal para esto. Uno puede, sin embargo, indicar dos criterios principales. El primero es que las descripciones formuladas por los grupos (clases) deberían ser "simples," para hacer fácil la asignación de objetos a grupos y diferenciar los grupos. Este criterio solo, sin embargo, podría llevar a clasificaciones triviales y arbitrarias. El segundo criterio es que las descripciones de grupos deben "ajustar bien" los datos reales. No obstante para lograr un "ajuste" muy preciso, una descripción puede llegar a ser compleja, por consiguiente, las demandas por simplicidad y buen ajuste son contradictorias, y la solución es encontrar un balance entre las dos.

Otras medidas pueden ser introducidas para la evaluación de la calidad de un agrupamiento. En *CLUSTER/2* se usa una medida combinada la cual incluye cualquiera de los siguientes criterios elementales:

- El ajuste entre el agrupamiento y los eventos
- La simplicidad
- La Comunalidad
- La separación Intergrupal
- El índice de discriminación.

El *ajuste* entre un agrupamiento y los datos es computado de dos maneras diferentes, denominadas como $T$ y $P$. La medida $T$ es el negativo del total de dispersión (absoluta) del agrupamiento, es decir, el negativo de la suma de las dispersiones absolutas de los complejos en el agrupamiento. La medida $P$ es el negativo de la dispersión proyectada del agrupamiento. La razón por la cual se usan los valores negativos es para incrementar el grado de igualdad de la disminución de la dispersión.

Donde la dispersión absoluta es el número de eventos no observados cubiertos por un complejo, y la dispersión proyectada es un conjunto de complejos disjuntos.

La *simplicidad* de un agrupamiento es definida como el negativo de su complejidad, la cual es la suma del costo atribuido a cada selector presente en los complejos. Una posible función de costo de selector es basada sobre el número de elementos encontrados en su lista de referencia. Los selectores con pocos elementos de referencia son menos complejos que aquellos con muchos elementos, y por lo tanto deben tener un costo más pequeño. Una simple medida de complejidad puede ser computada como el número de selectores contenidos en los complejos, es decir, usan una función de costo de selector constante que da a cada selector un costo de 1.

La *comunalidad* de un agrupamiento es el número total de propiedades compartidas por los eventos en cada uno de los grupos. La comunalidad es medida por el descubrimiento del número total de selectores que aparecen en los complejos. Este criterio es análogo al criterio de agrupamiento tradicional de maximizar la similaridad (Ej. , número de propiedades compartidas) de eventos dentro de un grupo.

La *separación intergrupal* de un agrupamiento es medida por la suma del grado de separación entre cada par de complejos en el agrupamiento. El grado de separación de un par de complejos es el número de selectores en ambos que involucran la misma variable y tienen valores de referencia que no hacen intersección. Por ejemplo, el par de complejos

- [color = rojo][tamaño = pequeño $\vee$ mediano][forma = círculo]
- [color = azul][tamaño = mediano $\vee$ grande]

tienen grado de separación 2, porque 2 de 5 selectores no hacen intersección(los selectores que no se cruzan son subrayados). Este criterio promueve agrupamientos con clases que tienen muchas propiedades diferentes, y es análogo

al criterio de requerimiento de distancia máxima entre grupos, usado en métodos de agrupamiento convencional.

El *índice de discriminación* de un agrupamiento es el número de variables que individualmente se discriminan entre todo los grupos, es decir, las variables tienen diferentes valores en cada descripción de grupo.

Las definiciones de los criterios anteriores son tal que el incremento de cualquier valor de criterio mejore la calidad del agrupamiento. La influencia relativa de cada criterio es especificada usando la *función de evaluación lexicográfica* (LEF: lexicographical evaluation functional). La LEF es definida por una secuencia de parejas de "criterio-tolerancia" $(c_1, \tau_1), (c_2, \tau_2), \ldots,$ donde $c_i$ es un criterio elemental seleccionado de la lista anterior, y $\tau_i$ es un "umbral de tolerancia" $(\tau \in [0 \ldots 100\%])$. En este primer paso todos los agrupamientos son evaluados sobre el primer criterio $c_1$, y aquellos de mejor puntuación o dentro del rango definido por la tolerancia $\tau_1$ son retenidos. Los próximos agrupamientos retenidos son evaluados sobre el criterio $c_2$ con umbral $\tau_2$. Este proceso continúa hasta que cualquier conjunto de los agrupamientos retenidos sea reducido a uno solo (el "mejor" agrupamiento) o la secuencia del par criterio- tolerancia sea exhaustiva. En el último caso, los agrupamientos retenidos tienen calidad equivalente con respecto al LEF dado, y cualquiera puede ser escogido arbitrariamente. La selección del criterio elemental, su ordenación y la especificación de tolerancia es hecha por un analista de datos.

#### 2.2.1.2 Agrupamiento Simbólico

El término de agrupamiento simbólico fue introducido por Gowda y Diday para referirse al agrupamiento realizado sobre datos que representan lo que Diday define como objetos simbólicos [5]. Definiendo como base de dicho agrupamiento una nueva medida de similaridad constituida por tres componentes: la posición, la extensión y el contenido; la cual es aplicable tanto a los atributos cuantitativos como cualitativos de los objetos simbólicos.

**Objetos a ser agrupados y sus atributos**

Los objetos a ser agrupados son más complejos que los datos convencionales, éstos están definidos por una conjunción lógica de eventos que une valores y variables, en donde las variables pueden tomar más de un valor incluso un conjunto infinito de valores (rangos numéricos) y todos los objetos de un conjunto de datos simbólicos pueden no estar definidos en términos de las mismas variables (se permiten valores vacíos). Los atributos pueden ser medidos en las siguientes escalas: nominal, ordinal, intervalo, proporción, absoluta y estructurada. En donde los tipos de variables cuantitativa proporción y absoluta se manejan como casos especiales del tipo intervalo, debido a la forma de composición de los grupos utilizada por el algoritmo (operador cartesiano) [6].

Un *evento* es una pareja (variable-valor) que enlaza las variables y valores de las mismas en los objetos.
Ejemplos de eventos:

e1 = [color = {verde, azul, rojo}]
e2 = [altura = 156]
e3 = [tiempo = [1.2,...,3.1]].

Existen diferentes tipos de objetos simbólicos: aseverativos (assertion), acumulativos (hoards) o de tipo síntesis (synthetic).

Un *objeto aseverativo* es una conjunción de eventos pertenecientes a las descripciones (individuales) de objetos reales.
Ejemplo:

a = e1 & e2 & e3
= [color = {verde, azul, rojo}]&[altura =156]&[tiempo =[1.2,...,3.1]]

Aquí a es un objeto simbólico aseverativo que tiene las siguientes propiedades:

1) el color es verde, azul ó rojo.
2) la altura es igual a 156.
3) el tiempo está entre 1.2 y 3.1.

Un *objeto simbólico acumulativo* es una conjunción de 2 o más objetos simbólicos aseverativos y eventos.
Ejemplo:

h=[VDU(Comp1)=color] & [RAM(Comp1)=64k] & [Keys(Comp1)=[57,..,63]] & [VDU(Comp2)=B&N] & [RAM(Comp2)=48k] & [Keys(Comp2)=[64,..,73]]

Esto significa que el objeto **h** consiste de 2 objetos elementales(aseverativos):

1) Computadora1 con VDU a color, RAM de 64k y Keys entre 57 y 63.
2) Computadora2 con VDU en B&N, RAM de 48k y Keys entre 64 y 73.

Un *objeto simbólico tipo síntesis* es una conjunción de 2 o más objetos simbólicos acumulativos y eventos.
Ejemplo:

S = h1 & h2 =[Tipo(r1)=autopista] & [Vehículos(r1)=2] & [Tipo(r2)=carretera] & [Vehículos(r2)=1] & [Tipo(v1)=carro] & [Color(v1)=azul] & [Movimiento(v1)=r1] & [Tipo(v2)=trailer] & [Color(v2)=rojo] & [Movimiento(v2)=r1] & [Tipo(v3)=autobús] & [Color(v3)=verde] & [Movimiento(v3)=r2].

**Tipo de Estructura del cluster**

El método de agrupamiento simbólico aquí expuesto particiona el conjunto de datos inicial en conjuntos de objetos disjuntos (estructura de partición), por medio de agrupaciones sucesivas, es decir es de carácter *aglomerativo*, el cual consiste en formar un objeto compuesto a partir de la pareja de objetos que tengan la más alta similaridad mutua, reduciendo en 1 el número de grupos en cada iteración hasta que sea igual a 1.

Por medio de los valores máximos y mínimos de similaridad obtenidos en todos los niveles de combinación de los objetos se obtiene el número óptimo de grupos, es decir que este método de agrupamiento no requiere conocimiento previo del número de clusteres porque lo calcula por sí mismo; dándole la característica deseable de ser un "método de agrupamiento *no paramétrico"*.

**Esquema de representación del cluster**

Dada una colección de objetos simbólicos a ser agrupados, el esquema de representación de dicha colección consiste en la descripción de un *objeto simbólico compuesto*, creado por medio de la aplicación de un operador de unión cartesiana sobre el conjunto de objetos inicial.

Dos objetos simbólicos *A* y *B* son escritos como el Producto cartesiano de características $A_k$ y $B_k$ como:

$$A = A_1 \mathrm{X}\ A_2 \mathrm{X} \ldots \mathrm{X} A_d$$
$$B = B_1 \mathrm{X}\ B_2 \mathrm{X} \ldots \mathrm{X} B_d$$

$U_k$ denota el dominio de la k-exima característica. Entonces el espacio característico puede ser escrito como el Producto cartesiano:

$$U^{(d)} = U_1 \mathrm{X}\, U_2 \mathrm{X} \ldots \mathrm{X} U_d .$$

La combinación de dos objetos puede ser vista como la formación de una "conexión" entre ellos, en un procedimiento de agrupamiento, si dos objetos pertenecientes a dos grupos diferentes son unidos, esto equivale a la combinación de los dos grupos; si los dos objetos que son combinados son para ser representados por un objeto único, uno de los métodos usados frecuentemente en agrupamiento convencional es el uso de la media de los dos como un indicativo único. La metodología de agrupamiento aquí propuesta forma un objeto compuesto cuando dos objetos

seleccionados son combinados. Este objeto compuesto, junto con el resto de los objetos del conjunto es usado en análisis de similaridades posteriores.

Un objeto simbólico compuesto es un nuevo objeto resultante de la combinación de dos objetos simbólicos utilizando el Operador de Unión Cartesiano [6].
Sea $A = A_1 X A_2 X \ldots A_d$ y $B = B_1 X B_2 X \ldots B_d$ dos objetos en $U^{(d)}$. Entonces el objeto compuesto $C$ resultante de la combinación de $A$ y $B$ es

$$C = A ++ B = (A_1 ++ B_1) X (A_2 ++ B_2) X \ldots X (A_d ++ B_d)$$

Donde $++$ es un operador de unión cartesiana. Cuando la *k*-ésima característica es cuantitativa o cualitativa ordinal, $A_k ++ B_k$ es definida como el mínimo intervalo que incluye a ambos $A_k$ y $B_k$. Esto es,

$$A_k ++ B_k = [\, min(A_{kl}, B_{kl}), max(A_{ku}, B_{ku})]$$

Donde $A_{kl}$ y $A_{ku}$ representan el limite inferior y superior respectivamente de $A_k$, cuando la *k*-ésima característica es cualitativa nominal, $A_k ++ B_k$ es la unión de $A_k$ y $B_k$.

**Criterio de agrupamiento**

Los objetos son agrupados mediante el criterio del vecino mutuo más cercano (*Mutual Nearest Neighborhood)* expuesto a continuación. En un conjunto de datos, si un objeto $X_i$ es el primer vecino más cercano (*Nearest Neighbor)* de un objeto $X_j$ , y $X_j$ es el primer vecino más cercano de $X_i$, entonces $X_i$ y $X_j$ constituyen un "par mutuo". Se ha observado que aproximadamente el 62% del numero total de individuos en poblaciones generadas al azar y naturales están en pares mutuos[11]. El par mutuo posee el valor de similaridad más alto correspondiente a los dos objetos del conjunto de datos que poseen la más alta similaridad.

[11] R.Hamming and E.Gilbert, "Probability of occurrence of a constant number of isolated pairs in a random population," *univ. Wisconsin Computing News,* no. 32,1974.

El método de agrupamiento simbólico asocia el par mutuo más similar en cada paso, para seleccionar éste se procede de la misma manera que en el agrupamiento jerárquico stepwise-optimizado haciendo el menor incremento posible de la suma del error cuadrado en cada etapa. Éste criterio, que tiene la mínima varianza, toma en consideración el número de muestras en cada grupo así como la similaridad entre los grupos. Él tiende a preferir la combinación de grupos unitarios o la combinación de grupos pequeños con grandes, en vez de combinación de grupos de tamaño medio.

La medida de similaridad que utiliza el método de agrupamiento simbólico entre dos objetos $A$ y $B$ es una asignación de $U(d)$ en $R+$ que tiene las siguientes propiedades:

1) $S(\mathrm{A},\mathrm{A}) = S(\mathrm{B},\mathrm{B}) > S(\mathrm{A},\mathrm{B})$
2) $S(\mathrm{A},\mathrm{B}) = S(\mathrm{B},\mathrm{A})$

La similaridad entre A y B es escrita como

$$S(A,B) = S(A_1,B_1) + S(A_2,B_2) + \ldots + S(A_k,B_k). \qquad (1)$$

Para la k-ésima característica $S(A_k,B_k)$ es definida usando los siguientes tres componentes:

1. $S_p(A_k,B_k)$ debido a la posición ***p***.
2. $S_s(A_k,B_k)$ debido a la extensión ***s***
3. $S_c(A_k,B_k)$ debido al contenido ***c***.

El componente de similaridad debido a la “posición” surge sólo cuando el tipo de característica es cuantitativa. El componente debido a la “extensión” indica el tamaño relativo de los valores de las características sin referirse a las partes comunes entre ellas. El componente debido al “contenido” es una medida de las

partes comunes entre los valores de las características. Los componentes $S_p$, $S_s$ y $S_c$ son definidos tales que sus valores son normalizados entre 0 y 1.

*Tipo de intervalo cuantitativo de* $A_k$ *y* $B_k$*:* La definición de similaridad entre dos intervalos cuantitativos es importante ya que los tipos cuantitativos proporción y absoluto son casos especiales del anterior:

Sea:

*al* = limite inferior del intervalo $A_k$

*au* = limite superior del intervalo $A_k$

*bl* = limite inferior del intervalo $B_k$

*bu* = limite superior del intervalo $B_k$

*inters* = longitud de intersección de $A_k$ y $B_k$

$l_s$ = longitud de extensión de $A_k$ y $B_k$

$= |\max(au, bu) - \min(al, bl)|$

donde max(.) y min(.) representan los valores máximos y el mínimos respectivamente.

Los tres componentes de similaridad son definidos como sigue:

El componente de similaridad debido a la posición es

$$S_p(A_k, B_k) = 1\text{-}al\text{-}b / |U_k| \qquad (2)$$

donde $U_k$ denota la longitud del máximo intervalo de la k-ésima característica.

El componente de similaridad debido a la extensión es

$$S_s(A_k, B_k) = (l_a + l_b)/2.l_s \qquad (3)$$

donde $l_a = |au\text{-}al|$, y $l_b = |bu\text{-}bl|$.

El componente de similaridad debido al contenido es

$$S_c(A_k, B_k) = inters/l_s. \quad (4)$$

El neto de la similaridad entre $A_k$ y $B_k$ es

$$S(A_k, B_k) = S_p(A_k, B_k) + S_s(A_k, B_k) + S_c(A_k, B_k) \quad (5)$$

*Tipo proporción / absoluto cuantitativo de* $A_k$ *y* $B_k$*:* La proporción y el tipo absoluto de características cuantitativas son casos especiales del tipo intervalo que tienen las siguientes propiedades:

$$al = au \; ; \; bl = bu \; ; \; l_a = l_b = inters = 0 \quad (6)$$

*Tipo cualitativo de* $A_k$ *y* $B_k$*:* Para tipo cualitativo de características, el componente de similaridad debido a la posición es ausente. Los dos componentes que contribuyen a la similaridad son

1) extensión *s* y
2) contenido *c.*

Sea $la$ = *longitud de* $A_k$ o número de elementos en $A_k$, $lb$ = longitud de $B_k$ o número de elementos $B_k$, $inters$ = número de elementos comunes a $A_k$ y $B_k$, $l_s$ = longitud de extensión de $A_k$ *y* $B_k$ combinado = $la + lb$ - $inters$.

El componente de similaridad debido a la extensión es

$$S_s(A_k, B_k) = (la+lb) / 2.l_s. \quad (7)$$

El componente de similaridad debido al contenido es

$$S_c(A_k, B_k) = inters / l_s. \quad (8)$$

El neto de similaridad entre $A_k$ y $B_k$ es

$$S(A_k, B_k) = S_s(A_k, B_k) + S_c(A_k, B_k) \quad (9)$$

**Criterio de parada**

Dado que este método de agrupamiento es de carácter aglomerativo, el algoritmo se detiene cuando el número de grupos alcanza su valor óptimo, pero para conocer éste, primero debe calcularlo por medio de un parámetro denominado *Indicador de Cluster* $CI$. Explicado a continuación.

En cada paso, la máxima similaridad mutua indica los objetos individuales del conjunto a ser combinados para formar un objeto simbólico compuesto. Los valores de similaridad mínima en cada paso indican la similaridad entre los grupos más distantes. Al disminuir este valor, mejor es la separación entre los grupos. El valor indicador de grupo en el *p*-ésimo paso es:

$$CI = \left|\max_{p+1} - \max_{p}\right| \Big/ \min_{p} .$$

Donde $\max_{p+1}$, $\max_{p}$ son las máximas similaridades en el paso *(p+1)*-ésimo y *p*-ésimo respectivamente, $\min_{p}$ es la mínima similaridad en el paso *p*-ésimo.

El paso en el cual $CI$ es máximo da el número de grupos.

### 2.2.1.3 Agrupamiento con Métricas Mixtas Ponderadas

La estrategia que sigue el método de agrupamiento empleado en Klass consiste en utilizar un conjunto de medidas que sean compatibles entre ellas y cubran cualquier combinación de tipos de variables. Se trata de realizar un tratamiento homogéneo de todas las variables definiendo una distancia (o eventualmente una medida de disimilitud) entre individuos que utilice expresiones diferentes para cada tipo de variable y que tenga sentido.

**Objetos a ser agrupados y sus atributos**

El método trabaja con lo que denomina dominios poco estructurados, y los define como aquellos que presentan las siguientes características:

- *Matrices con datos heterogéneos*. Las variables que describen los objetos pueden ser cuantitativas o cualitativas. Las cualitativas suelen tener muchas modalidades, tanto mayor cuanto mayor sea la experiencia del usuario.

- *Existencia de información adicional sobre la estructura del dominio*. Es común que se cuente con conocimiento declarativo sobre la estructura del dominio de estudio (relaciones entre variables, objetivos de agrupación, etc.).

- *Conocimiento parcial y no homogéneo*. Los expertos suelen disponer de grandes cantidades de conocimiento implícito, además de manejar diversos grados de especifidad, lo que lo hace no homogéneo.

Los tipos de variable que maneja el algoritmo son variables de radio continuo (cuantitativas) y variables nominales (cualitativas).

**Tipo de Estructura del cluster**

El objetivo de agrupamiento en dominios mixtos es realizar una partición, donde cada objeto es colocado de acuerdo a sus características en un grupo que lo caracterice, evitando así que los grupos formados se traslapen, es decir sean disjuntos (estructura de partición). Las variables numéricas son representadas por la media aritmética de los componentes del grupo, lo que coincide con el centro de gravedad de la clase en esa variable. Para las variables cualitativas la propuesta es el *objeto extendido*, donde dada una clase $C = \{i_1 ... i_{nC}\}$ y la variable cualitativa $X_K$ que toma valores en el conjunto $D_K = \{c_1^k ... c_{nk}^k\}$, la k-ésima componente del representante de $C$, $\bar{I}_C$ es:

$$\bar{x}_{ck} = \left(\left(f_c^{k_1} c_1^k\right) .. \left(f_c^{k_{nk}} c_{nk}^k\right)\right), \quad f_c^{kj} = \frac{card\{i \in C : x_{ik} = c_j^k\}}{n_c}$$

No hace falta decir que si $X_k$ toma un mismo valor en toda la clase ($c_s^k$), el representante de clase será también $\bar{x}_{ck} = c_s^k$ (lo que sería equivalente a $\bar{x}_{ck} = \left(\left(0c_1^k\right)..\left(1c_s^k\right)..\left(0c_{nk}^k\right)\right)$.

Esto es en pocas palabras representar al prototipo de una variable categórica detallando la forma cómo los elementos del grupo se distribuyen en las diferentes modalidades de la variable, y hacerlo mediante proporciones por razones técnicas.

**Esquema de representación del cluster**

El esquema de representación que se utiliza en agrupamiento con métricas mixtas ponderadas es el siguiente: Dado un conjunto de n individuos $i$, $\{i_1, i_2, ..., i_n\}$, descritos por un conjunto de variables $X_k$, $k = 1,2,...k$, de las cuales $n_\zeta$ son variables numéricas y $n_Q$ categóricas; de tal manera que un individuo $i$ es descrito por un vector de observaciones $(x_{i1}, x_{i2}, x_{i3}, ..., x_{iK})$, y una matriz $(n, K)$ formada con los valores $x_{i,k}$. Las filas de la matriz de datos contienen información de los individuos a ser agrupados, mientras que cada columna corresponde a una de las variables usadas para describir la muestra.

*KLASS* representa los grupos que se van formando a través de un elemento prototipo. La idea de prototipo $\bar{I}_C$ de un conjunto de individuos $C = \{i_1 ... i_{nC}\} \subseteq I$ lleva a pensar en un individuo, real o no, que sintetice las características de sus representados.

La explicitación de un elemento que represente prototípicamente cada uno de los grupos formados por *Klass* permite:

- Por un lado, tratar de forma homogénea individuos propiamente dichos y los grupos, los cuales se identificaran con su representante.

- Proporcionar una descripción conceptual (o prototípica) de los grupos formados, lo que va a ser fundamental en el proceso de interpretar los resultados finales desde un punto de vista cognitivo y ver si los grupos encontrados tienen *significado* o no.

**Criterio de agrupamiento**

Dado un conjunto de $n$ individuos $I$, $\{i_1, i_2, ..., i_n\}$, descritos por un conjunto de variables $X_k$, $k = 1,2,...k$, de las cuales $n_\zeta$ son variables numérica y $n_Q$ categóricas, en [7] se propone la siguiente familia de distancias:

$$d_{(\alpha,\beta)}{}^2(i,i') = \alpha\, d_\zeta^2(i,i') + \beta\, d_Q^2(i,i') \qquad (1)$$

donde:

$\alpha, \beta$ son los pesos para balancear la influencia de los grupos de atributos numéricos y categóricos respectivamente y son positivos.

$d_\zeta^2(i,i')$ es la distancia euclidiana normalizada calculada con todas las variables numéricas, o subdistancia cuantitativa:

$$d_\zeta^2(i,i') = \sum_{\forall k \in \zeta} \frac{(x_{ik} - x_{i'k})^2}{s_k^2} \qquad (2)$$

$\zeta$ es el conjunto de índices de todas las variables numéricas y $s_k^2 = Var(X_k)$.

$d_Q^2(i,i')$ es la distancia de $\chi^2$ [12] calculada sobre toda la matriz de variables cualitativas, o subdistancia cualitativa:

[12] Benzécri, J.P. L'analyse des données. Tome 1:Lataxinomie, Tome 2:L'analyse des correspondences.Paris:Dunod.1980.

$$d_Q^2(i,i') = \frac{1}{n_Q^2} \sum_{\forall k \in Q} d_k^2(i,i') \tag{3}$$

de tal forma que se permite que $i, i'$ tengan componentes categóricas representadas con símbolos o como valores extendidos y Q es el conjunto de índices de las variables cualitativas.

$$d_k^2(i,i') = \begin{cases} 0, & \text{si } x_{ik} = x_{i'k} \\ \frac{1}{I_{k^i}} + \frac{1}{I_{k^{i'}}}, & \text{de otra manera} \\ \frac{\left(f_i^{k_s} - 1\right)^2}{I^{k_{sss}}} + \sum_{j \neq s}^{nk} \frac{\left(f_i^{k_j}\right)^2}{I^{k_j}}, & \text{si } x_{ik} = c_{ss}^k \text{ y } i' \text{ es una subclase} \\ \sum_{j=1}^{nk} \frac{\left(f_i^{k_j} - f_{i'}^{k_j}\right)^2}{I^{k_j}}, & \text{en caso general} \end{cases} \tag{4}$$

donde :

$I^{k_j}$ es el numero de individuos de la muestra de valor $c_j^k$ para la variable $X_k$ .

$I_{k^i} = card(i): x_{\hat{i}k} = x_{ik}$ , nombre de individuos de la muestra que, para la k-ésima variable, son de la misma modalidad que $i$ .

$\left(f_i^{k_1}, f_i^{k_2}, ..., f_i^{k_{n_k}}\right)$ donde $f_i^{k_j} = \frac{I^{k_j}}{n_i}$ , $j = 1,2...,n_k$ , es la proporción de objetos elementales de la clase representada por el objeto cualquiera $i$ sobre la categórica $c_j^k$ de la variable categórica $X_k$

En principio $\alpha, \beta$ pueden tomar cualquier valor entre 0 y 1. En [13]se proponen valores para $\alpha, \beta$ iguales a :

$$\alpha = \frac{n_\zeta}{d^2_{\zeta\,\max^*}} \qquad \& \qquad \beta = \frac{n_Q}{d^2_{Q\,\max^*}} \tag{5}$$

donde $n_\zeta$ = número de variables numéricas, $n_Q$ = número variables categóricas, con $d^2_{\zeta\,\max^*} = \max^*_{i,i'}\left\{d^2_\zeta(i,i')\right\}$, $d^2_{Q\,\max^*} = \max^*_{i,i'}\left\{d^2_Q(i,i')\right\}$, y finalmente $\max^*$ máximo truncado al 95% (lo que garantiza estabilidad numérica y robustez a outliers).

**Criterio de parada**

El algoritmo Klass es de carácter aglomerativo, es decir que en cada paso disminuye en uno el total de elementos del conjunto hasta llegar a uno, y para obtener un número de grupos deseado se corta el dendrograma en el nivel adecuado (corte alfa). También se le puede indicar que agrupe hasta un número de grupos dado a priori.

### 2.2.2 Algoritmos de Agrupamiento en Dominios Mixtos

En esta sección se presentan las partes fundamentales de los algoritmos implementados en la herramienta software desarrollada.

#### 2.2.2.1 Algoritmo de Agrupamiento Conceptual Conjuntivo basado en CLUSTER/2

El algoritmo de agrupamiento conceptual conjuntivo es un algoritmo que obtiene *grupos* jerarquizados y definidos por formas normales conjuntivas; esto es,

[13] Karina, Gibert y Ulises, Cortés. Weighing quantitative and qualitative variables in clustering methods.Journal Mathware and Sof Computing, número especial, pags. 251-266. Mayo de 1997.

clasificación de objetos o sucesos consistentes en formas normales conjuntivas de expresiones que *envuelven* relaciones o atributos.

El algoritmo de agrupamiento conceptual procede de la siguiente manera:

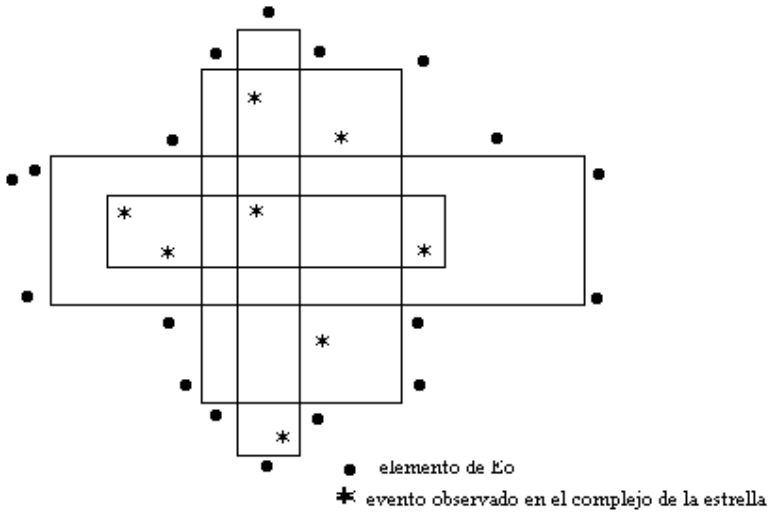
*Paso 1: Se Determinan las Semillas Iniciales*

Selección de *k* semillas entre los sucesos que serán clasificados (el *k* puede ser optimizado por el propio algoritmo probando iterativamente con sucesivos valores de *k*).

*Paso 2: Se Construyen las Estrellas*

Para cada semilla $e$, se construye una estrella reducida, con el conjunto de semillas restantes como los sucesos que deberán ser excluidos de la estrella en construcción. Esto es, generaliza la descripción de cada semilla frente a todas las demás. Una estrella estará constituida por todas las descripciones maximales o *complejas* que generalizan al objeto $e$ frente a las demás semillas. Aplica operadores de generalización tales como encerrar valores en intervalos, eliminar una condición, y ascender en el árbol de generalización.

Figura 3. Ilustración de la estrella G($e \mid E_0$)



*Paso 3: Se Optimiza el Agrupamiento*

En este punto las estrellas tendrán intersecciones invariables. Se construye un *agrupamiento* optimizado de acuerdo a una cierta función, eligiendo un *complejo* de cada estrella y modificándolo. Se asegura que los *complejos* elegidos son disjuntos para lo que se aplicará un procedimiento definido denominado NID.

El procedimiento *NID* transforma un conjunto de complejos No Disjuntos dentro de un conjunto de complejos Disjuntos (es decir, un agrupamiento disjunto), si la entrada de complejos al *NID* son disjuntas, el procedimiento no la altera.

Los pasos que sigue el procedimiento *NID* son:

1. *Se determinan los "centros" de los complejos*: Los eventos observados cubiertos por más de un complejo del conjunto dado son colocados en la *lista de multiplicar - evento cubierto* (*m-list*), si la *m-list* es vacía, entonces los complejos son sólo intersectados débilmente, es decir el área de intersección contiene únicamente eventos no observados; en este caso, el procedimiento termina indicando que la combinación de complejos es un *agrupamiento intersectado débilmente*. Por otro lado, cada complejo es reemplazado por la REFUNION (operación que transforma un conjunto de eventos y/o complejos en un complejo único que cubre los eventos y/o complejos) de los eventos observados contenidos en el complejo que no está en la *m-list* (es decir, que son individualmente cubiertos), las refuniones obtenidas son llamadas "centros" de los complejos.
2. *Se determina un mejor complejo de "patrón" para cada evento en la m-List*: Un evento es seleccionado de la *m-list* y es "añadido" a cada uno de los *k* centro de los complejos por generalización cada complejo a la extensión necesaria para cubrir el evento. Tal generalización es realizada por la aplicación del operador REFUNION al evento y el complejo; como un resultado, se obtienen los *k* complejos modificados. Por reemplazo uno de los centros de los complejos en el conjunto inicial con el correspondiente complejo modificado, en *k* maneras diferente, se obtiene una colección de agrupamientos; éstos son evaluados de acuerdo al criterio de calidad de clustering asumido. El complejo en el mejor agrupamiento que cubre el

evento dado de la *m-list* es considerado el mejor "patrón" para este evento. El mejor agrupamiento es retenido y los restantes son eliminados. Por repetición la operación citada anteriormente para cada evento en la *m-list*, se obtiene un conjunto de *k* complejos disjuntos cuya unión cubre los mismos eventos observados como el conjunto original de complejos no disjuntos.

Si un evento no puede ser añadido a cualquier complejo sin causar el efecto para intersectar otros complejos, entonces el evento es colocado en la *lista de excepciones.*

*Paso 4: ¿Termina?*

1. Si ésta es la primera iteración, se almacenan las descripciones del agrupamiento
2. En otro caso, se almacena únicamente la descripción del agrupamiento si mejora el rendimiento del anterior de acuerdo a LEF (lexicographical evaluation function).
3. Si después de un cierto número de iteraciones no mejora el agrupamiento entonces termina el algoritmo.

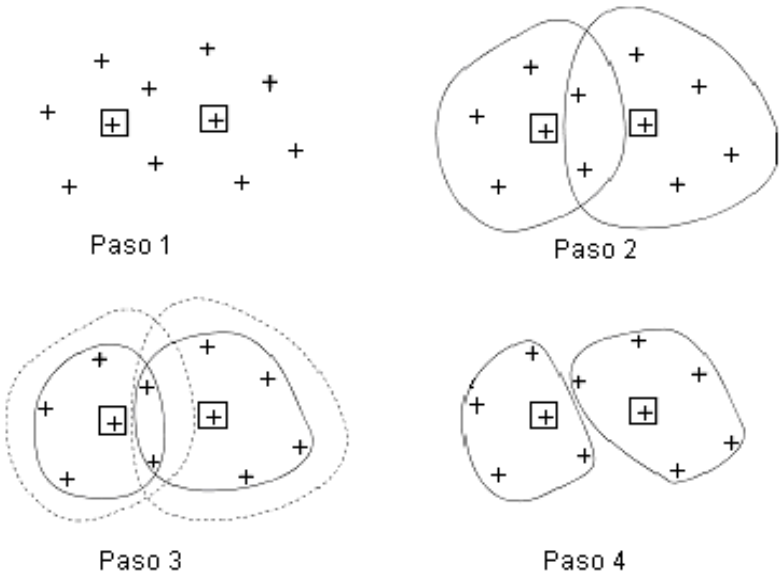
*Paso 5: Selección de Nuevas Semillas*

1. Se seleccionan *k* nuevas semillas, una de cada complejo.
2. Si el agrupamiento mejora, se seleccionarán sucesos *centrales*, esto es, sucesos que estén más cerca de los centros geométricos de los complejos.
3. Si el agrupamiento no mejora, se seleccionarán sucesos de los *bordes*, esto es, sucesos que estén alejados de los centros.

*Paso 6: Repetir*

Repetir desde el paso 2.

Figura 4. Formación de grupos en agrupamiento conceptual.




#### 2.2.2.2 Algoritmo de Agrupamiento Simbólico basado en una Nueva Medida de Similaridad

El algoritmo de agrupamiento simbólico aquí propuesto es aglomerativo, jerárquico y no paramétrico, basado en una nueva medida de similaridad, constituida por tres componentes: la posición, la extensión y el contenido de los objetos simbólicos.
El algoritmo de agrupamiento simbólico procede de la siguiente manera:

1) Sea $\{X_1, X_2, \ldots, X_N\}$ un conjunto de $N$ objetos simbólicos. Sea $N$ el número inicial de grupos, con cada grupo considerando un peso de grupo(numero de objetos) de 1.
2) Computar los pesos de las similaridades entre todos los pares de objetos simbólicos en el conjunto de datos como

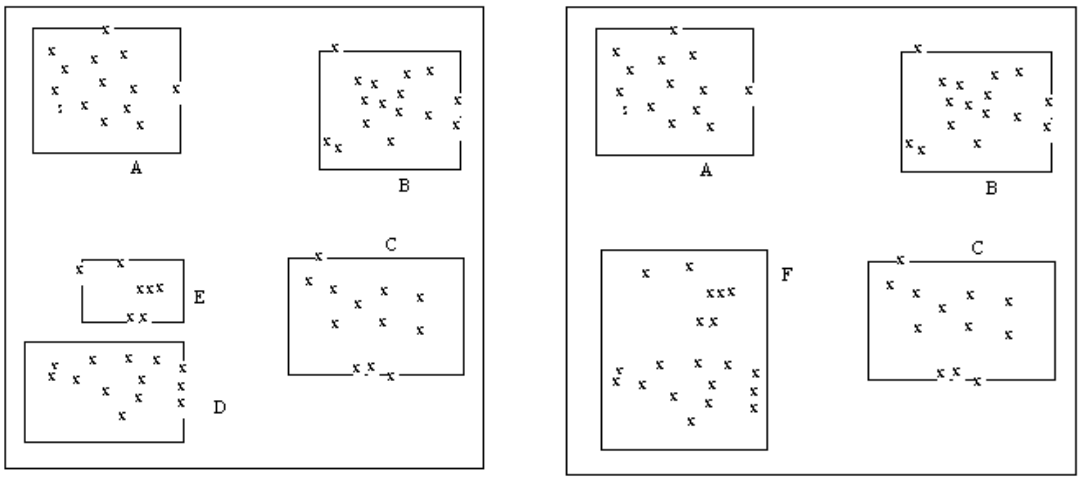
$$S_w(X_i, X_j) = S(X_i, X_j) / sqrt(n_i * n_j / (n_i + n_j)).$$

Donde $n_i$, $n_j$ son pesos de Grupos(numero de objetos) de $X_i$ y $X_j$ respectivamente, y *sqrt* representa la raíz cuadrada. $S(X_i, X_j)$ es el valor de similaridad dado por $S(A,B) = S(A_1,B_1) + S(A_2,B_2) + ... + S(A_k,B_k)$. Determinar el valor máximo y mínimo de similaridad, asimismo el par mutuo considerando la similaridad más alta y formar un objeto compuesto asociando los individuos de este par; Reducir el número de grupos a 1.

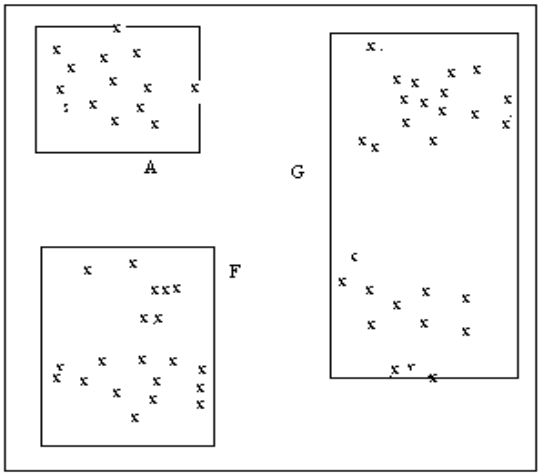
Repetir el paso 2 hasta que el número de grupos sea igual a 1.

Calcular el valor del indicador de cluster ***CI*** utilizando $CI = /max_{p+1} - max_p // min_p$ donde $max_{p+1}$, $max_p$ son las máximas similaridades en el paso (*p +1*)-ésimo y *p*-ésimo respectivamente, $min_p$ es la mínima similaridad en el *p*-ésimo paso. La etapa donde ***CI*** es máximo, indica el número de grupos en los datos. Los objetos compuestos de esta etapa describen los objetos simbólicos representando los grupos.

Figura 5. Ejemplo de objetos simbólicos compuestos representando grupos




(a) (b)




(c)

#### 2.2.2.3 Algoritmo de Agrupamiento basado en Métricas Mixtas Ponderadas KLASS

El algoritmo de agrupamiento de Karina Gibert *Klass* es un sistema de agrupamiento orientado a la clasificación de dominios poco estructurados; como los métodos

estadísticos clásicos tienen pobre desempeño en estos dominios, surge la idea de trabajar en una nueva dirección: usar restricciones declarativas para resolver las deficiencias detectadas en los métodos tradicionales con el fin de enriquecer el agrupamiento. Este algoritmo de agrupamiento implementa una versión adaptada del algoritmo de los vecinos recíprocos; además toma ventaja de cualquier información adicional que un experto pueda suministrar sobre los conceptos designados. *Klass* permite trabajar con las siguientes métricas:

- Euclidiana
- Euclidiana estandarizada
- $X^2$ (Chi-cuadrado).
- Métricas mixtas.

Considérese la descripción de $\bar{I}_C$, $\bar{x}_C = (\bar{x}_{C1}...\bar{x}_{CK})$. El cálculo de $\bar{x}_C$ se hace componente a componente, y hace falta estudiar por separado el caso de las componentes numéricas y el de las categóricas.

Para variables numéricas, en [14] justifica que este representante sea la media aritmética de los componentes de la clase, lo que coincide con el centro de gravedad de la clase en esa variable.

Para variables categóricas, la propuesta es el *objeto extendido* que se define a continuación.

*Objetos compactos y objetos extendidos*

Dada una clase $C = \{i_1...i_{nC}\}$ y la variable cualitativa $X_K$ que toma valores en el conjunto $D_K = \{c_1^k...c_{nk}^k\}$, la k-ésima componente del representante de ***C***, $\bar{I}_C$ es:

[14] Gibert, K.. L'ús de la informació simbòlica en la automatizaci'o del tractament estadístic de dominis poc estructurats. Ph D. Thesis, UPC, Barcelona, España, 1994.

$$\bar{x}_{ck} = \left(\left(f_c^{k_1} c_1^k\right)..\left(f_c^{k_{nk}} c_{nk}^k\right)\right), \quad f_c^{kj} = \frac{card\left\{i \in C : x_{ik} = c_j^k\right\}}{n_c}$$

No hace falta decir que si $X_k$ toma un mismo valor en toda la clase ($c_s^k$), el representante de clase será también $\bar{x}_{ck} = c_s^k$ (lo que sería equivalente a $\bar{x}_{ck} = \left(\left(0c_1^k\right)..\left(1c_s^k\right)..\left(0c_{nk}^k\right)\right)$.

La propuesta Gibert es representar al prototipo de una variable categórica detallando la forma cómo los elementos de la clase se distribuyen en las diferentes modalidades de la variable, y hacerlo mediante proporciones por razones técnicas. En la referencia original aparece una descripción más detallada de por qué se hace de esta manera.

Nombrando $I_c^{k_j}$ al número de individuos de la clase $C$ que toman valor $c_j^k$ para la variable $X_k$, se pueden denotar las componentes del vector $\bar{x}_{ck}$ como

$$f_c^{k_j} = \frac{I_c^{k_j}}{\sum_{j=1}^{n_k} I_c^{k_j}} = \frac{I_c^{k_j}}{n_c}$$

A partir de ahora, para simplificar la notación, se representaran las componentes cualitativas del centro de gravedad de una forma equivalente a la anterior:

$$\bar{x}_{ck} = \left(\left(f_c^{k_1} c_1^k\right)..\left(f_c^{k_{nk}} c_{nk}^k\right)\right) \equiv \left(f_c^{k_j} ... f_c^{k_{nk}}\right)$$

Con esta definición de prototipo de una clase para las variables cualitativas, utilizada por primera vez en [8], aparece un nuevo tipo de objeto que puede tener componentes vectoriales en algunas variables categóricas.

Un objeto cualquiera *i* puede tener, para la variable categórica $X_k$, un valor del tipo $f_i^{k_j},...,f_i^{k_{nk}}$ donde $f_i^{k_j}(j = 1 : n_k)$ es la proporción de objetos elementales de la clase representada por *i* sobre la categoría $c_j^k$. Bajo esta representación, los

valores posibles de las variables categóricas serán un subconjunto de $D_k \cup [0,1]^{n_k}$, dado que las $f_i^{k_j}$ son proporciones, y siempre satisfacen

1. $f_i^{k_j} \geq 0, \forall i$.
2. $\sum_{j=1}^{n_k} f_i^{k_j} = 1$

Por tanto, en [8] se define como valor en forma extendida a todo valor vectorial de una variable cualitativa, y valor en forma compacta al que es representable mediante un símbolo.

Se considera objeto compacto a todo aquél que tiene valores compactos en todas sus variables categóricas, mientras que los objetos extendidos presentan valores extendidos al menos en una de las variables categóricas.

El algoritmo funciona de la siguiente forma:

*Paso 1*

Para las variables categóricas se hallan las cardinalidades por categoría $f_c^{k_j} = \frac{I_c^{k_j}}{\sum_{j=1}^{n_k} I_c^{k_j}} = \frac{I_c^{k_j}}{n_c}$ y se asignan por elemento $\bar{x}_{ck} = \left( \left(0c_1^k\right)..\left(1c_s^k\right)..\left(0c_{nk}^k\right) \right)$; igualmente para las variables numéricas se hallan la media y la varianza, medidas necesarias para hallar las distancias entre cada par de registros.

*Paso 2*

Hallar todas las distancias cualitativas y cuantitativas, y a partir de estas hallar alfa y beta por medio de las ecuaciones dadas en el capítulo anterior:

$$\alpha = \frac{n_\zeta}{d^2_{\zeta \max^*}} \qquad \& \qquad \beta = \frac{n_Q}{d^2_{Q \max^*}}$$

*Paso 3*

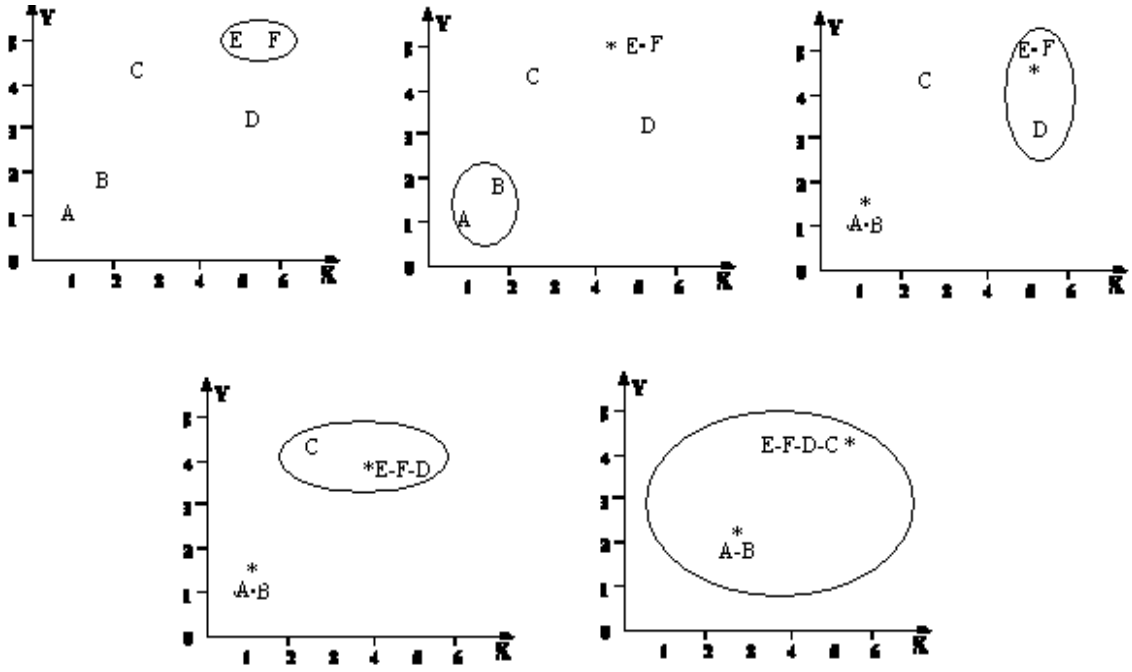
Se halla la distancia para cada par de elementos del conjunto sumando las distancias cualitativas y cuantitativas halladas con su respectivo factor de ponderación:

$$d_{(\alpha,\beta)}{}^{2}(i,i') = \alpha\, d_{\zeta}^{2}(i,i') + \beta\, d_{Q}^{2}(i,i')$$

y se agrupan dos elementos por medio del criterio del vecino más cercano recíproco.

Repetir desde el paso 1 hasta que el número de grupos sea igual a 1.

Figura 6.Formación de grupos en Agrupamiento con métricas mixtas



# 3 DISEÑO Y ELABORACIÓN DEL SOFTWARE

En este capítulo se presenta el diseño y elaboración de la herramienta software para la clasificación de datos con variables cualitativas y cuantitativas denominada ADAMIX 1.0. El capítulo está dividido en las siguientes partes:

- Plataforma de desarrollo.
- Metodología utilizada.
- Proceso de elaboración o plan de trabajo del software.
- Características de la herramienta software.

## 3.1 PLATAFORMA DE DESARROLLO

ADAMIX 1.0 fue desarrollada bajo la plataforma de Windows de 32 bits y es funcional para sistemas operativos WIN98/2000/NT/XP.

Las interfaces y el entorno visual fueron desarrollados con la herramienta de programación Borland Delphi 7, una herramienta de desarrollo que permite crear aplicaciones rápidas y robustas, gracias a que el código de Delphi se apoya en un lenguaje de programación orientado a objetos: *Object Pascal* que tiene su origen en el compilador: *Turbo Pascal* [9].

Para la elaboración de los diagramas de clases correspondientes a los diferentes modelos creados durante el desarrollo de la herramienta incluyendo el prototipo final entregado junto con este documento se utilizó el lenguaje de modelado UML que es un lenguaje gráfico para visualizar, especificar y documentar la estructura de un software o sistema por medio de conceptos orientados a objetos - específicamente se utilizó la herramienta UML estudio 6 – [10].

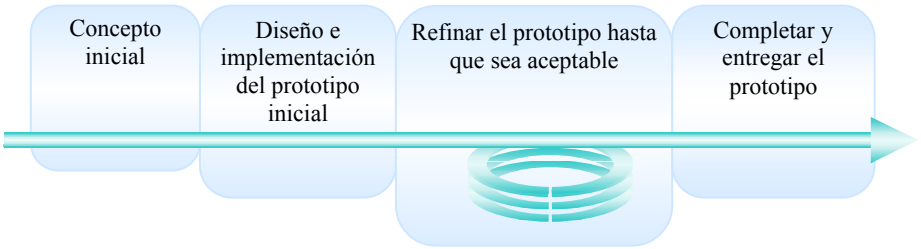
## 3.2 METODOLOGÍA DE DESARROLLO UTILIZADA

Para el proceso de desarrollo de ADAMIX 1.0 se utilizó la metodología de prototipado evolutivo, la cual se explica brevemente a continuación, al igual que su uso a lo largo del proyecto.

### 3.2.1 Prototipado Evolutivo

El modelo de prototipado evolutivo toma sus bases del prototipado simple, pero a diferencia de éste, que es de carácter exploratorio ó para identificar requisitos, el evolutivo posee mayores controles sobre la calidad y desarrolla primero las áreas de mayor riesgo del sistema, de tal forma que el prototipo pueda ser tomado como producto final una vez se llegue a su fin. Es decir, en este modelo se desarrolla el concepto del sistema a medida que avanza el proyecto. El prototipo evolutivo es un enfoque donde se desarrolla primero las partes seleccionadas del sistema y luego el resto a partir de estas partes. A diferencia de otros tipos de prototipado, en el evolutivo no se descarta el código del prototipo; sino se transforma en el código entregado finalmente. El desarrollo de prototipos continúa hasta que se decide que el prototipo es lo suficientemente bueno y se puede entregar como producto final.

Figura 7. Prototipado evolutivo




Durante el proyecto se fueron desarrollando prototipos exploratorios de los algoritmos de agrupamientos implementados, con el fin de identificar las partes críticas del desarrollo; y guiando dicha implementación en un proceso de reducción

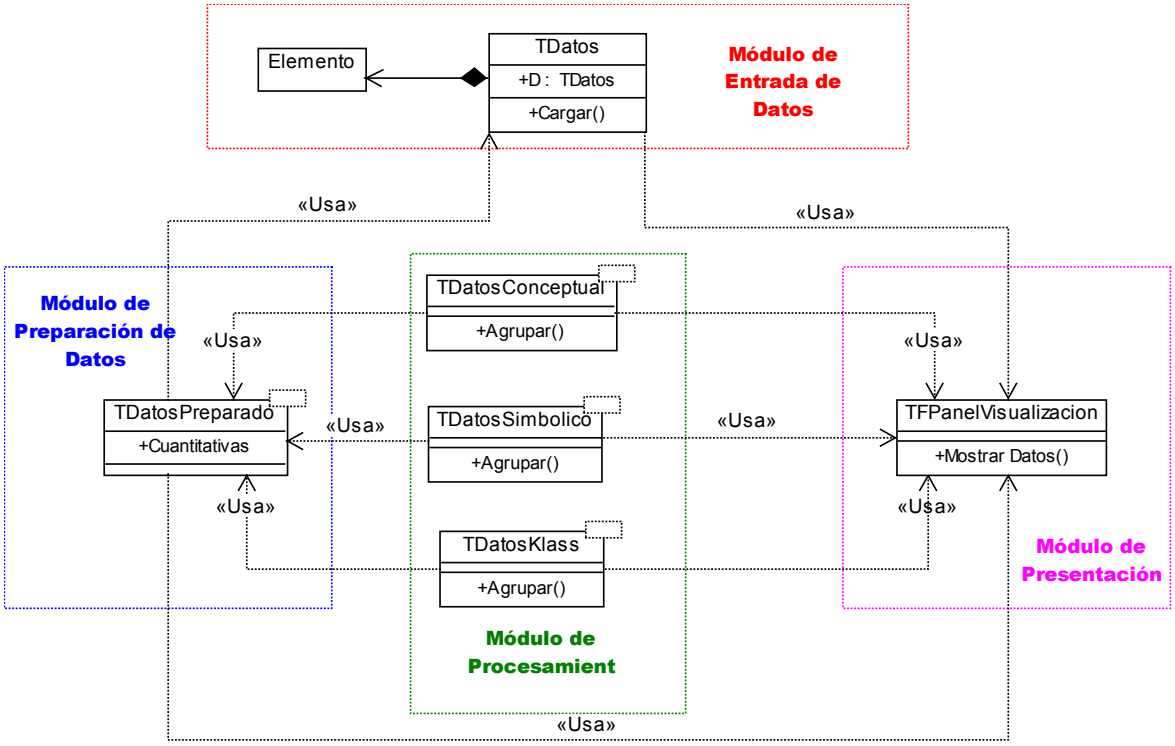
de riesgos, mediante los prototipos evolutivos que se requerían en la elaboración de esta herramienta, en cada uno de los cuales se fue incorporando funcionalidad, complejidad y optimización hasta llegar al punto deseado.

A continuación se presentan los diagramas de clases y casos de uso más importantes de los prototipos elaborados durante el proceso de desarrollo de la herramienta.

#### 3.2.1.1 Prototipo Inicial

En este prototipo se implementaron las funciones básicas necesarias para el manejo de datos, como son la lectura de archivos en formados de texto y excel, estas funciones son manejadas a través de la clase *TDatos* dentro del módulo de entrada de datos, igualmente dentro del módulo de preparación se creó una clase *TDatosPreparado* para contener todos los procedimientos y funciones necesarias para la preparación de los datos para los algoritmos, como son convertir las variables cuantitativas a formatos numéricos y obtener todas las categorías para una variable cualitativa, entre otras. Finalmente el prototipo contiene las clases relativas a cada algoritmo las cuales constituyen el módulo de procesamiento y ofrecen las funciones necesarias para llevar a cabo el agrupamiento en cada caso, las cuales son *TDatosConceptual*, *TDatosSimbolico* y *TDatosKlass* respectivamente. En la Figura 8 se presenta el diagrama de clases del prototipo inicial dividido en los diferentes módulos.

Figura 8. Diagrama de clases del Prototipo Inicial




### 3.2.1.2 Prototipo Intermedio

En este prototipo se agregaron nuevas clases y se implementaron nuevos procedimientos en las clases ya existentes con el fin de agregar funcionalidad en algunos casos y de optimizar la programación en otros, quedando mejor definidos los diferentes módulos de la herramienta, los cuales se describen a grandes rasgos a continuación:

*Módulo de Entrada de datos:*

Constituido principalmente por la clase *TDatos* en la cual se implementan funciones de E/S al disco duro, al igual que los procedimientos necesarios para presentar, y actualizar los datos a través de componentes visuales proporcionados por el Módulo de Presentación.

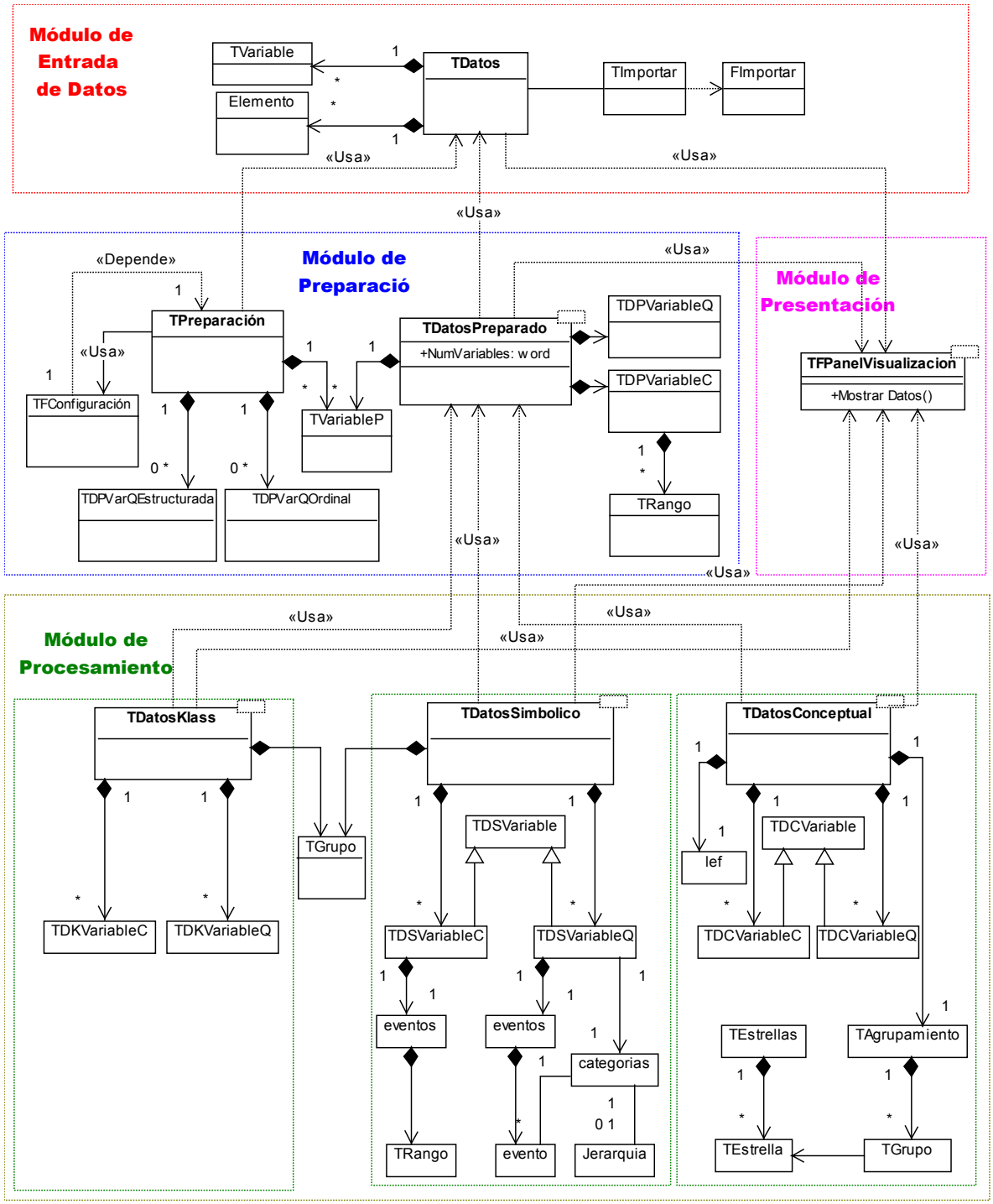
*Módulo de Preparación de datos*:

Constituido principalmente por la clase *TDatosPreparado.* Al igual que *TDatos*, ésta clase también ofrece los procedimientos necesarios para presentar, (pero no para actualizar) los datos preparados a través de componentes visuales del Módulo de Presentación. En este prototipo se agrega la clase *TPreparacion* la cual mantiene algunos parámetros necesarios para la creación de los datos preparados, éstos parámetros son actualizados a través del formulario de configuración, componente visual por medio del cual el usuario accede a las funciones del Módulo de Preparación.

*Módulo de Procesamiento*:

Constituido por tres unidades en cada una de las cuales se implementa una clase fundamental y algunas clases adicionales que dependen de ésta. Las clases fundamentales son *TDatosConceptual*, *TDatosSimbolico* y *TDatosKlass* y corresponden a los tres algoritmos de agrupamientos implementados Conceptual, simbólico y Klass respectivamente. Entres las clases adicionales de cada unidad encontramos los tipos de variable cualitativa y cuantitativa que son diferentes para cada algoritmo. En la Figura 9 se presenta el diagrama de clases para el prototipo intermedio.

Figura 9. Diagrama de clases del Prototipo Intermedio



#### 3.2.1.3 Prototipo Final

Este prototipo bosqueja la constitución final de la herramienta o producto final, representando los diferentes módulos que la integran, sus relaciones, y sus principales componentes y casos de uso. En la Figura 10 se presenta el diagrama de clases y casos de usos del prototipo final

**Relaciones entre los módulos**

A continuación se mencionan las principales relaciones existentes entre los módulos de la herramienta, y la funcionalidad que ofrecen cada uno de éstos al usuario.

En primer lugar se encuentra la *interfaz con el usuario*, un pseudo-módulo por medio del cual se tiene acceso a todos los módulos; cabe destacar que el *usuario* tiene el control integral sobre el proceso realizado a los datos, es decir, el proceso puede ser detenido y retomado en cualquier momento, gracias al manejo de proyectos; ó de igual forma si no se desea abrir un proyecto se pueden abrir archivos de datos o imágenes en forma independiente, ó asociarlos a un proyecto existente.

En seguida se halla el *Módulo de Entrada de datos* el cual se relaciona con el *usuario* permitiéndole crear un conjunto de datos ya sea cargándolos de un archivo en disco, o ingresándolos directamente a través de la interfaz de la herramienta, igualmente le permite guardar datos en un archivo de cualquiera de los formatos especificados para ello.

El *Módulo de Preparación de datos* tiene una relación de dependencia con el *Módulo de Entrada de datos* a través de un flujo de datos (valga la redundancia) procedente de la clase TDatos hacia la clase TDatosPreparado, durante la creación de esta última, cuando el usuario ejecuta la acción *preparación básica de datos* -a través de los diferentes modos que ofrece la interfaz para esto-; Otras de las funciones o servicios que ofrece el *Módulo de Preparació*n al *usuario* son: tratamiento de campos vacíos, normalizaciones, filtros, eliminación de variables, eliminación de registros, etc. Estas acciones son opcionales y algunas de ellas solo estarán disponibles una vez se halla efectuado la preparación básica, acción sin la

cual a su vez el usuario no puede acceder a los algoritmos de procesamiento de los datos.

Lo anterior hace manifiesta la relación de dependencia del *Módulo de Procesamiento* con el *Módulo de Preparación*, ésta se da a través de un flujo de datos desde la clase TDatosPreparado hacia una de las clases TDatosConceptual, TDatosSimbólico o TDatosKlass respectivamente. El *Módulo de Procesamiento* también interactúa con el *usuario* durante el ingreso de los parámetros propios de cada algoritmo, algunos de ellos opcionales, tal como el número de clases deseado entre otros

Concluyendo los módulos internos de la herramienta, se encuentra el *Módulo de Presentación de resultados*, que se relaciona con los demás módulos ofreciendo funciones para la presentación de datos; esto a través de componentes visuales como paneles, TreeViews, y Grids (cuadrículas), además del componente para hacer gráficos Tchart que solo está disponible para los datos procedentes del *Módulo de Procesamiento*.

Por otra parte se halla el *Módulo de ayuda de la herramienta*, que es un módulo externo, y es accesado a través de la *interfaz de usuario*, para facilitar la navegabilidad del usuario a través de la misma.

**Composición de los módulos**

A continuación se describen brevemente las principales clases que constituyen cada uno de los módulos.

*Módulo de Entrada de datos:*

*TDatos:* Esta clase inicia el proceso sobre los datos, debido a que obtiene y almacena los datos suministrados por el usuario, a través de cualquiera de los medios proporcionados por la herramienta para ello (cargar, importar, digitar, etc).

*TImportar:* Clase utilizada por TDatos solo cuando se importan datos, para guardar información relativa a la forma de importar, como delimitadores, Linea inicial, etc.

*FImportar:* Asistente para importar datos.

*Módulo de Preparación de datos:*

*TDatosPreparado:* clase que transforma los datos suministrados por el usuario en datos listos para ser procesados por los diferentes algoritmos.

*TPreparacion:* Clase utilizada para crear TDatosPreparado con la configuración elegida por el usuario.

*TDPVariableQ:* Tipo de variable cualitativa utilizada por TDatosPreparado. Utilizada en la creación de las variables cualitativas de las estructuras de datos inherentes a los algoritmos de agrupamiento.

*TDPVariableC:* Tipo de variable cuantitativa utilizada por TDatosPreparado. Utilizada en la creación de las variables cuantitativas de las estructuras de datos inherentes a los algoritmos de agrupamiento.

*FConfiguracion:* Formulario de varias páginas donde se presentan y se guardan los diferentes parámetros de preparación tanto básica como avanzada de los datos.

*Módulo de Procesamiento:*

*TDatosConceptual*: Estructura de datos básica correspondiente al algoritmo de agrupamiento conceptual conjuntivo. Esta estructura ofrece las funciones necesarias para realizar el agrupamiento propuesto por este algortimo, sin modificar el conjunto inicial de datos.

*TDCVariableQ:* Tipo de variable Cualitativa para el algoritmo de agrupamiento conceptual conjuntivo, utilizada por *TDatosConceptual*. En ella se almacenan los valores que toma la variable para cada registro en una lista de lo que el algoritmo denomina eventos. Los cuales a su vez pertenecen a una lista de categorías que pueden ser de tipo nominal o estructurado.

*TDCVariableC:* Tipo de variable Cuantitativa para el algoritmo de agrupamiento conceptual conjuntivo, utilizada por *TDatosConceptual*. En ella se almacenan los valores que toma la variable para cada registro en una lista de lo que el algoritmo denomina eventos. Los cuales son de tipo *TRango*.

*TAgrupamiento:* Utilizada por *TDatosConceptual,* en esta estructura se almacena toda la información correspondiente a los grupos, los eventos observados y su descripción en forma de complejos.

*TDatosSimbolico: Estructura* de datos básica correspondiente al algoritmo de agrupamiento simbólico aglomerativo. Esta estructura ofrece las funciones necesarias para realizar el agrupamiento propuesto por este algortimo; transformando un conjunto de n registros en un conjunto de n-1 registros a cada paso.

*TDSVariableQ:* Tipo de variable Cualitativa para el algoritmo de agrupamiento simbólico aglomerativo, utilizada por *TDatosSimbolico*. En ella se almacenan los valores que toma la variable para cada registro en una lista de lo que el algoritmo denomina eventos. Los cuales a su vez pertenecen a una lista de categorías que pueden ser de tipo nominal, ordinal o estructurado.

*TDSVariableC:* Tipo de variable Cuantitativa para el algoritmo de agrupamiento simbólico aglomerativo, utilizada por *TDatosSimbolico*. En ella se almacenan los valores que toma la variable para cada registro en una lista de lo que el algoritmo denomina eventos. Los cuales son de tipo *TRango*.

*TDatosKlass: Estructura* de datos básica correspondiente al algoritmo de agrupamiento Klass, denominada así porque representa los datos en forma de matriz extendida sobre las categorías de las variables cualitativas. Esta estructura ofrece las funciones necesarias para realizar el agrupamiento propuesto por este algortimo; transformando un conjunto de n registros en un conjunto de n-1 registros a cada paso.

*TDKVariableQ:* Tipo de variable Cualitativa para el algoritmo de agrupamiento Klass, utilizada por *TDatosKlass.*

*TDKVariableC:* Tipo de variable Cuantitativa para el algoritmo de agrupamiento Klass, utilizada por *TDatosKlass*.

*Módulo de Presentación de Resultados o Visualización:*
*TFPaneVisualizacion:* Clase utilizada por *TDatos*, *TDatosPreparado*, *TDatosKlass*, *TDatosConceptual*, *TDatosSimbolico*, y *TDatosProcesado* para mostrar datos.

*TFPaneHerramientas:* Formulario que permite la navegación a través de los objetos del proyecto, y presenta las opciones disponibles para cada uno de ellos.

*TFDatos:* Componente visual utilizado por TDatos para la visualización de datos.

*TFPreparados:* Componente visual utilizado por TDatosPreparado para la visualización de datos.

*TFProcesados:* Componente visual utilizado por TDatosProcesado para la visualización de datos.

*TFKlass:* Componente visual utilizado por TDatosKlass para la visualización de datos.

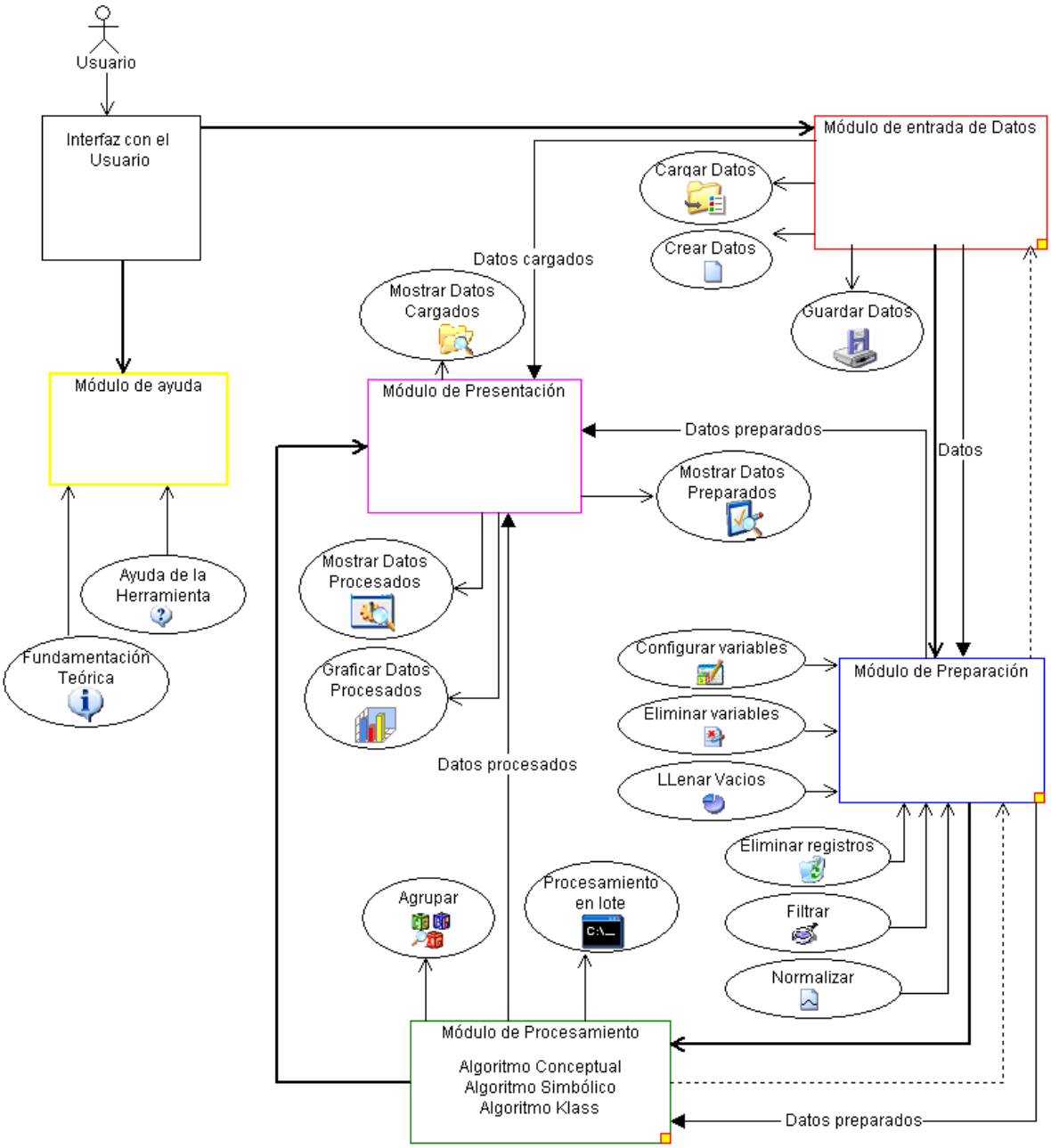
*TFSimbolico:* Componente visual utilizado por TDatosSimbolico para la visualización de datos.

*TFConceptual:* Componente visual utilizado por TDatosConceptual para la visualización de datos.

*TFGraficar:* Componente visual utilizado por TDatosProcesado para graficar datos agrupados.

*TFGrafico:* Componente visual utilizado por TGrafico para visualización de gráficos.

Figura 10. Diagrama de clases y casos de uso del Prototipo Final



## 3.3 PROCESO DE ELABORACIÓN DEL SOFTWARE

Dentro del proceso de elaboración de ADAMIX 1.0 se llevaron a cabo las siguientes fases:

- Fase de iniciación.
- Fase de elaboración.
- Fase de construcción.
- Fase de transición.

### 3.3.1 Fase de Iniciación

Durante esta fase, se estableció la planificación del proyecto, se delimitó su alcance y se estimaron los recursos para su desarrollo; se seleccionó la metodología de desarrollo a utilizar; asimismo se realizó la documentación y revisión bibliográfica de los tópicos de minería de datos, agrupamiento conceptual, agrupamiento simbólico, objetos simbólicos, métricas mixtas, y un estudio detallado de los algoritmos de agrupamiento implementados. También se creo un prototipo ejecutable para determinar la viabilidad del proyecto.

### 3.3.2 Fase de Elaboración

En la fase de elaboración se analizó el dominio del problema, se hizo un análisis y diseño de alto nivel de los diferentes módulos que integraron la herramienta software estableciéndose así una base arquitectónica sólida para la construcción de la misma. Asimismo se elaboró el plan de proyecto y se llevó a cabo el estudio de la herramienta de programación Borland Delphi 7 para la codificación del software desarrollado.

### 3.3.3 Fase de Construcción

Durante la fase de construcción, se desarrolló de forma iterativa e incremental un producto completo; empezando por la construcción de prototipos, que permitieron el desarrollo de versiones cada vez más complejas hasta conseguir el producto final preparado para los usuarios.

- **Construcción del Modulo de entrada de datos**

En la construcción de este módulo se tuvieron en cuenta las funciones necesarias para la entrada / salida de datos, partiendo con la implementación de las funciones básicas, como lo son: la lectura de archivos con los formatos especificados (texto, excel, csv), luego las funciones y procedimientos necesarios para la manipulación de los datos de entrada ya sean cargados o ingresados manualmente por el usuario mediante cuadros de diálogo y cuadriculas(guardar, digitar, importar etc) y finalmente las opciones de guardar los diferentes archivos. Todas las funciones y procedimiento de este módulo se encuentran en las clases *TDatos*, *TImportar*, *FImportar*.

- **Construcción del Modulo de preparación de datos**

En este modulo se implementaron las utilidades que permiten la preparación y depuración de los datos por parte del usuario, con soporte de la herramienta, para facilitar el posterior proceso de agrupamiento de los mismos. Partiendo con la validación de datos numéricos, el llenado de campos vacíos, algunas operaciones opcionales sobre los datos tales como normalizaciones, al igual que otras rutinas de tipo matemático y estadístico que aligeren los cálculos posteriores de procesamiento. Luego las funciones y procedimientos que permiten la creación y actualización de los datos preparados. Las principales clases que se implementaron en la construcción de este módulo son: *TDatosPreparado, TPreparacion, FConfiguracion*.

El módulo se divide en dos partes, una que es el núcleo de la preparación misma de los datos, y otra que es la interfaz visual para la interacción con el usuario, por medio de la cual éste tiene acceso a los servicios de la primera.

- **Construcción del Modulo de procesamiento**

Este es el modulo principal de la herramienta donde se implementaron los algoritmos de agrupamiento, y se consideraron los conceptos que manejan cada uno de ellos para realizar la clasificación de los datos, tales como son: la metodología de estrella ,la formación de grupos que posean una fácil interpretación conceptual y a su vez tengan en cuenta los conceptos emergentes en la descripción de una colección de objetos, en agrupamiento conceptual conjuntivo jerárquico de Michalski y Stepp; agrupamiento simbólico, jerárquico y aglomerativo, y medidas de similaridad para objetos simbólicos, en agrupamiento simbólico de Gowda y Diday; métricas y distancias mixtas, pesaje de la influencia de variables cualitativas y cuantitativas, en Klass.

El Módulo de procesamiento se dividió en tres partes que corresponden a los algoritmos implementados, cada algoritmo tiene un objeto principal asociado y otras estructuras de datos ligadas a éste. Dicho objeto está conformado a su vez por atributos de corresponden a la estructura de datos propuesta por el algoritmo, y por métodos que llevan a cabo las funciones necesarias para realizar el agrupamiento propuesto por el algoritmo, al igual que ofrece rutinas de acceso a los datos necesarias para la representación de éstos en forma tabular y gráfica, para que la interfaz visual de presentación de resultados haga uso de éstas. Las clases principales que se implementaron en la construcción de este módulo fueron: TDatosConceptual, *TDatosSimbolico, TDatosKlass*.

- **Construcción del Módulo de presentación de resultados**

Este módulo es básicamente visual y consta de formularios, paneles, cuadrículas, tcharts, treeviews, entre otros componentes cuyo objetivo es presentar al usuario los diferentes tipos de datos a través del proceso de agrupamiento y navegar a través

de ellos. Los datos agrupados se muestran en la manera de representación de cada algoritmo, y se transforman en lo que denominamos datos procesados para poder ser presentados en forma gráfica a través de un tchart, que ofrece diferentes diagramas para representar las series de datos. La mayor utilidad de este módulo se aprecia en la capacidad de presentación de los resultados de los agrupamientos, y de aquí toma su nombre el módulo.

- **Construcción del Módulo de ayuda del sistema**

En este módulo se implementó un sistema de ayudas correspondiente al manejo de la herramienta, al igual que los conceptos básicos utilizados en la clasificación de los datos. Este sistema de ayuda se diseñó con la herramienta RoboHELP Office Tools.

- **Integración de los Módulos**

Durante ésta etapa se hizo el encadenamiento de los módulos anteriormente descritos, los cuales se fueron relacionando en la unidad principal del sistema, empezando por el enlace del módulo de entrada, seguido por el de preparación y posteriormente el de agrupamiento, y a lo largo del proceso de construcción el módulo de visualización evolucionó, el cual estuvo presente durante todo el proceso. Finalmente se enlazó el sistema de ayuda.

- **Validación de la herramienta**

Durante esta etapa se realizaron diferentes tipos de pruebas, con el fin de buscar posibles errores de ejecución en la herramienta, algunas de ellas fueron: pruebas alfa realizadas con usuarios, al igual que otras pruebas como caja negra, y carga máxima, y algunas pruebas con datos reales.

Las pruebas alfa se llevaron a cabo en un entorno controlado en presencia de los desarrolladores como observadores del usuario, con el fin de registrar los errores

del software que se iban presentando, corregirlos, y repetir el proceso, hasta depurar completamente la herramienta. La ampliación sobre la aplicación de estas pruebas se encuentra en el capitulo 5.

- **Depuración de la herramienta**

Durante esta etapa se realizaron los cambios, arreglos y mejoras necesarias para el correcto funcionamiento de la herramienta, de acuerdo con las observaciones obtenidas en las pruebas alfa.

  - Cambios en la forma

    Se llevaron a cabo mejoras en cuanto a la presentación de las opciones para cada tipo de datos, esto conllevó a la creación del navegador del proyecto, y un panel de opciones que mostrara solo las opciones disponibles para cada objeto. Esto con el fin de evitar confusiones con opciones no disponibles. Estos cambios fueron realizados dentro de la unidad Panel de Herramientas del módulo de visualización.

  - Cambios de fondo

    Dentro de las mejoras de fondo se encuentran todas las validaciones realizadas en todos los puntos del programa donde surgían errores por falta de una restricción de tipo numérico o de otro tipo. También se incluyeron aquí las validaciones en la carga de archivos del proyecto, especialmente los archivos de datos cuando presentaban errores en sus cabeceras.

### 3.3.4 Fase de Transición

En esta fase ADAMIX 1.0 se puso a disposición de los usuarios realizándose así las pruebas *beta*.

Las pruebas beta se realizaron con estudiantes de ingeniería de sistemas, previamente orientados en el funcionamiento del software. Para esto se les suministró una copia de la herramienta, para que la utilizaran libremente.

A diferencia de las pruebas alfa, la prueba beta fue una aplicación "en vivo" del software en un entorno que no pudo ser controlado. Los estudiantes registraron todos los problemas (reales o imaginarios) que encontraron durante la prueba e informaron a intervalos regulares. Como resultado de los problemas informados en las pruebas beta, se realizaron modificaciones y así se preparó una versión del software para todos los usuarios.

## 3.4 CARACTERÍSTICAS DE LA HERRAMIENTA SOFTWARE

ADAMIX 1.0 es una herramienta Software de Minería de Datos para la clasificación de datos con variables cualitativas y cuantitativas que utiliza varios algoritmos de agrupamiento, que abordan la clasificación de datos con variables en dominios mixtos desde diferentes enfoques. Entre las principales características de la herramienta tenemos:

- Está estructurada en forma modular; los módulos que la forman son: Módulo de entrada de datos, preparación, procesamiento, presentación de resultados y de ayuda del sistema.

- Su construcción está basada en la programación orientada a objetos haciendo uso de conceptos tales como *Clases*, *Objetos* y *Herencia*, que permiten la facilidad del mantenimiento y actualización del código para ampliaciones o versiones posteriores.

- Posee aproximadamente 21.000 líneas de código y un total de 80 clases, las principales son: TDatos, TDatosPreparado, TPreparacion, TFConfiguracion,

TDatosConceptual, TDatosSimbolico, TDatosKlass, TDatosProcesado, TDatosProyecto, TFPrincipal, TFPanelVisualizacion, TFPanelHerramientas.

- Tiene un rendimiento alto en cuanto a la velocidad de procesamiento y el tiempo de respuesta. Teniendo en cuenta el tipo de algoritmos que se trabajaron, algoritmos de búsqueda exhaustiva y extendidos sobre categorías cualitativas.

- La interfaz con el usuario comprende de una serie de componentes visuales tales como formularios, paneles, cuadrículas, y componentes para crear gráficos. Fue construida de forma tal que resultara amigable e intuitiva para el usuario, permitiendo su fácil comprensión, aprendizaje y operatividad. Para esto se tuvieron en cuenta criterios tales como que el usuario no tuviera que manejar una cantidad de formularios muy extensa; esto se hizo mediante el uso de páginas dentro de los formularios de visualización y configuración entre otros. Igualmente se buscó usar una distribución estándar.

A continuación se presentará un vistazo general de la constitución de la herramienta, después de los cambios y ajustes realizados a ésta de acuerdo con las observaciones y recomendaciones obtenidas en las pruebas alfa y beta, a través de algunos casos de uso que permiten observar la funcionalidad de los diferentes módulos.

### 3.4.1 Módulo de Entrada de Datos

Este módulo permite que los usuarios realicen los siguientes casos de uso:

- **Caso de uso cargar o abrir archivo de datos**
  **Propósito:** Permite al usuario abrir un archivo de datos previamente creado y guardado o ya existente.

**Resumen:** El usuario selecciona la opción *abrir* del menú “Archivo” o de la barra de herramientas respectiva, el sistema le pide la ubicación y puede seleccionar el tipo de archivo que desea abrir, una vez seleccionado el archivo el sistema le asigna un alias y procede a cargarlo en memoria, visualizarlo y agregarlo al proyecto.

- **Caso de uso crear nuevo conjunto de datos**

  **Propósito:** Permite al usuario crear un nuevo conjunto de datos para posteriormente ser procesado.

  **Resumen:** El usuario selecciona la opción *nuevo* del menú “Archivo” o de la barra de herramientas respectiva, el sistema le solicita una información básica y procede agregarlo al proyecto posteriormente presenta una malla que permite la edición del nuevo conjunto de datos.

- **Caso de uso guardar datos**

  **Propósito:** Permite al usuario guardar un conjunto de datos creados o modificado en un dispositivo de almacenamiento magnético.

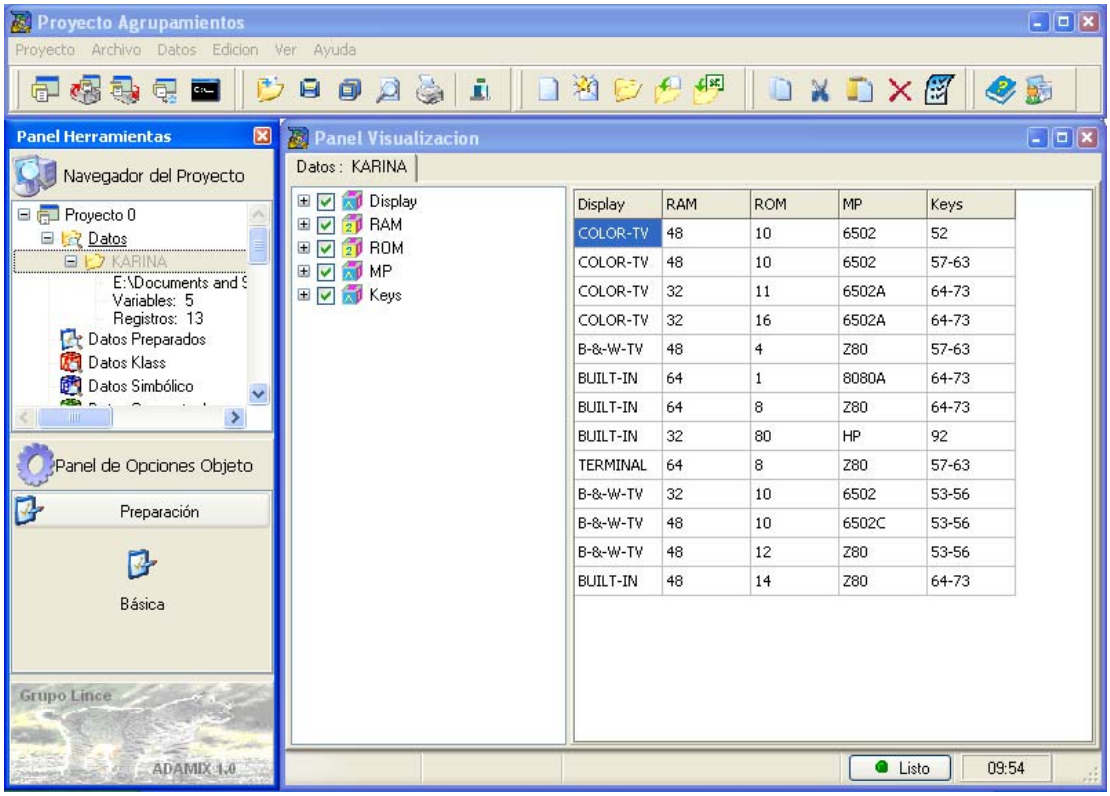
  **Resumen:** El usuario selecciona la opción *guardar* del menú “Archivo” o de la barra de herramientas respectiva, el sistema verifica que el conjunto de datos a guardar sea nuevo o uno existente y lo almacena en la ubicación establecida por el usuario.

- **Caso de uso importar**

  **Propósito:** Permite al usuario importar datos.

  **Resumen:** el usuario selecciona la opción *importar* del menú “Archivo”, el sistema le pide la ubicación y el archivo que desea importar, una vez seleccionado el archivo el sistema procede a cargarlo en memoria y presentar una vista previa del mismo, para permitirle al usuario elegir una configuración adecuada, que permita cargar los datos en forma correcta.

Figura 11. Entrada de datos




### 3.4.2 Módulo de Preparación de Datos

Este módulo permite que los usuarios realicen los siguientes casos de uso sobre los datos cargados en memoria:

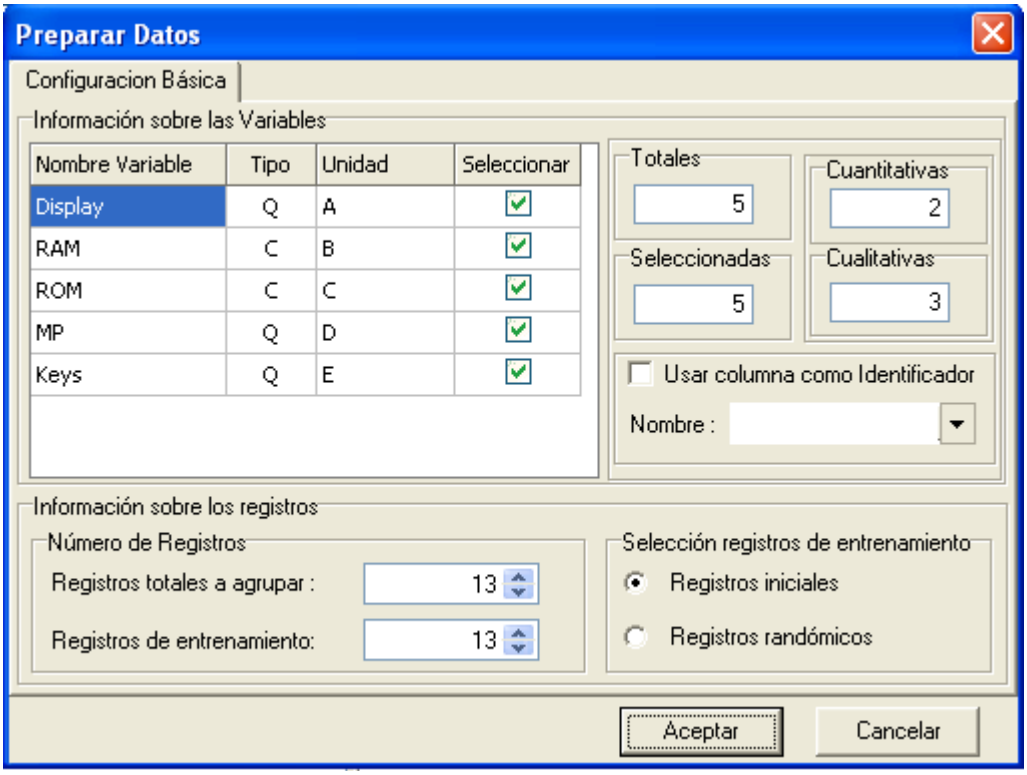
- **Caso de uso preparación básica**

  **Propósito:** Permite al usuario realizar una preparación básica a los datos. Que consiste en convertir los valores de las variables cualitativas a formatos numéricos y los valores de las variables cuantitativas en categorías de la misma.

  **Resumen:** El usuario selecciona la opción *Básica* de la sección 'Preparación' del Panel de herramienta; el sistema muestra la página

correspondiente a la preparación básica del formulario de configuración, el usuario puede determinar con cuales variables desea trabajar así como el número de registros. Una vez aceptada la configuración, el sistema procede a crear un conjunto de datos preparados y procede a cargarlo en memoria, visualizarlo y agregarlo al proyecto.

Figura 12. Preparación básica




- **Caso de uso llenar vacíos**

  **Propósito:** Permite al usuario llenar campos vacíos o eliminar los registros correspondientes a los mismos.

  **Resumen:** Esta opción solo está disponible cuando los datos preparados presentan campos vacíos. El usuario selecciona la opción *vacíos* de la sección "Preparación", del Panel de herramienta; el sistema muestra la página correspondiente al manejo de vacíos del formulario de configuración,

el usuario realiza el llenado de vacíos de acuerdo al método de su agrado, el sistema realiza los cambios sobre la variable datos preparado y los visualiza.

- **Caso de uso filtrar**

  **Propósito:** Permite al usuario filtrar los datos de acuerdo con determinadas restricciones de valores.

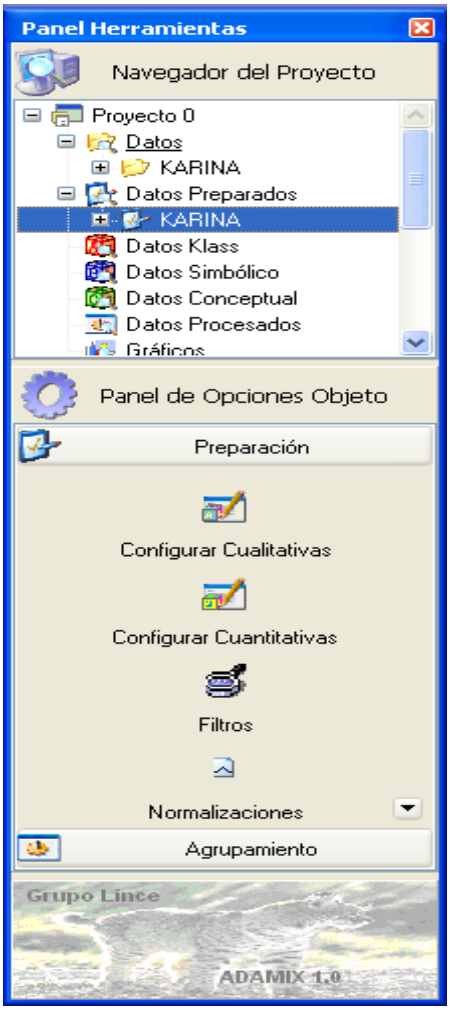
  **Resumen:** El usuario selecciona la opción *filtros* de la sección “Preparación”, del Panel de herramienta; el sistema muestra la página correspondiente al manejo de filtros del formulario de configuración, el usuario realiza el filtrado a través de las opciones que allí se presentan, y posteriormente el sistema realiza los cambios sobre la variable datos preparados y los visualiza.

- **Caso de uso normalizar**

  **Propósito:** Permite al usuario normalizar datos.

  **Resumen:** El usuario selecciona la opción *normalizaciones* de la sección “Preparación”, del Panel de herramienta; el sistema muestra la página correspondiente a normalizaciones del formulario de configuración, el usuario realiza la normalización a través de las opciones y métodos que allí se presentan, y posteriormente el sistema realiza los cambios sobre la variable datos preparados y los visualiza.

Figura 13. Preparación avanzada de datos




### 3.4.3 Módulo de Procesamiento de Datos

Este módulo permite que los usuarios realicen los siguientes casos de uso sobre los datos preparados.

- **Caso de uso agrupar conceptual**

  **Propósito:** Permite al usuario formar grupos a partir de un conjunto de datos preparados a través del algoritmo de agrupamiento Conceptual conjuntivo de Michalski y Stepp.

**Resumen:** El usuario selecciona la opción *conceptual* de la sección 'Agrupamiento' del Panel de herramienta; el sistema muestra el formulario de configuración inicial conceptual, donde el usuario puede determinar con cuales criterios de calidad de agrupamiento desea trabajar, y la preponderancia que da a cada uno de ellos, así como el número de clases o grupos que desea obtener. Una vez aceptada la configuración, el sistema procede a crear un conjunto de datos conceptual y procede a cargarlo en memoria, visualizarlo y agregarlo al proyecto. Este tipo de datos tiene una página asociada que permite el agrupamiento de los datos pulsando el botón 'Agrupar' de la misma.

- **Caso de uso agrupar simbólico**

  **Propósito:** Permite al usuario formar grupos a partir de un conjunto de datos preparados a través del algoritmo de agrupamiento simbólico aglomerativo de Gowda y Diday.

  **Resumen:** El usuario selecciona la opción *simbólico* de la sección 'Agrupamiento' del Panel de herramienta, el sistema procede a crear un conjunto de datos simbólico y procede a cargarlo en memoria, visualizarlo y agregarlo al proyecto. Este tipo de datos tiene una página asociada que permite el agrupamiento de los datos pulsando el botón 'Agrupar' de la misma.

- **Caso de uso agrupar klass**

  **Propósito:** Permite al usuario formar grupos a partir de un conjunto de datos preparados a través del algoritmo de agrupamiento basado en métricas mixtas ponderadas de Karina Gibert.

  **Resumen:** El usuario selecciona la opción *klass* de la sección 'Agrupamiento' del Panel de herramienta, el sistema procede a crear un conjunto de datos Klass y procede a cargarlo en memoria, visualizarlo y agregarlo al proyecto. Este tipo de datos tiene una página asociada que permite el agrupamiento de los datos pulsando el botón 'Agrupar' de la misma.

Figura 14. Procesamiento de datos


### 3.4.4 Módulo de Presentación de Resultados

Este módulo permite que los usuarios realicen los siguientes casos de uso

- **Caso de uso mostrar datos cargados**

  **Propósito:** Permite visualizar los datos que se han cargado o digitado por el usuario.

**Resumen:** Este caso de uso se da cuando se crea un nuevo conjunto de datos, ya sean cargados o digitados. El panel de visualización asigna una página de Datos a la variable de tipo Datos correspondiente permitiéndole visualizarlos a través del procedimiento mostrar datos de este último, y ofreciéndole la opción de realizar preparación básica sobre esos datos.

- **Caso de uso mostrar datos preparados**

  **Propósito:** Visualiza los datos una vez realizada la preparación.

  **Resumen:** Este caso de uso se da cuando se crea un nuevo conjunto de datos preparado, el panel de visualización asigna una página de datos Preparados a la variable de tipo Datos Preparado correspondiente permitiéndole visualizarlos a través del procedimiento mostrar datos de este último, y ofreciéndole opciones para agrupar o realizar otras preparaciones sobre esos datos.

- **Caso de uso mostrar datos procesados**

  **Propósito:** Visualiza los datos una vez fueron agrupados por medio de alguno de los tres algoritmos implementados.

  **Resumen:** Este caso de uso se da cuando se crea un nuevo conjunto de datos procesado, a partir de uno de los tres tipos de datos Conceptual, Simbólico, o Klass, el panel de visualización asigna una página datos Procesados a la variable de tipo Datos Procesado correspondiente permitiéndole visualizarlos a través del procedimiento mostrar datos de este último, y ofreciéndole opciones para graficar esos datos.

- **Caso de uso Graficar datos procesados**

  **Propósito:** Permite al usuario representar los datos procesados por medio de gráficas tales como barras, líneas, puntos, tortas, etc..

  **Resumen:** El usuario selecciona una de las opciones de graficar de la sección “Análisis gráfico”, el sistema procede a mostrar la pagina graficar de la variable de tipo Datos Procesado correspondiente, la cual presenta las variables a seleccionar disponibles para ese tipo gráfica, junto con un tchart,

y un menú de comandos asociado a este que permite todas las opciones de edición, visualización, guardado, e incluso impresión.

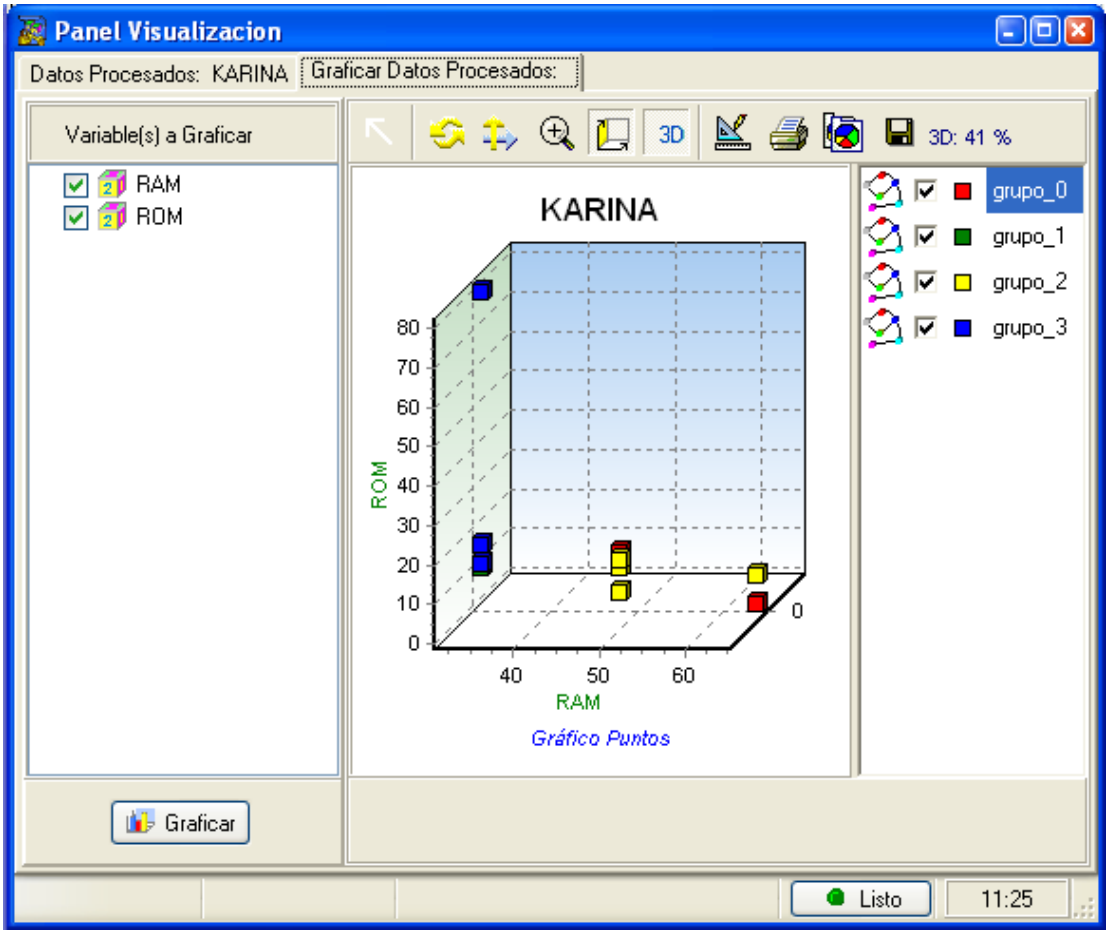
En cuanto se seleccionan las variables se pulsa el botón graficar y el tchart presenta los grupos de datos en la forma del diagrama elegido, y estos a su vez pueden ser seleccionados si no se desean ver en la gráfica. Finalmente cuando la gráfica es aceptada, el sistema procede a crear una variable de tipo gráfico, visualizarla y agregarla al proyecto.

- **Caso de uso Mostrar Gráfica**

  **Propósito:** Permite al usuario ver los gráficos creados por medio de los distintos datos procesados del proyecto.

  **Resumen:** Este caso de uso se da cuando a partir de una variable tipo Datos Procesado, se crea una variable de tipo Gráfico, el panel de visualización asigna una página gráfica a esta variable. Esta gráfica solo está sujeta a las modificaciones propias de las opciones de edición del tchart.

Figura 15. Visualización gráfica de los datos




### 3.4.5 Módulo de Ayuda del Sistema

Este módulo esta formado por la ayuda del sistema referente al manejo de la herramienta así como los conceptos básicos utilizados en la clasificación de los datos. Contiene los siguientes casos de uso:

- **Caso de uso ayuda de la herramienta**

  **Propósito:** Permite al usuario navegar por los diferentes ítem para el manejo de la herramienta.

  **Resumen:** El usuario selecciona la opción *contenido* del menú “Ayuda”, el sistema le presenta el contenido, el usuario abre el libro de ayuda del manejo

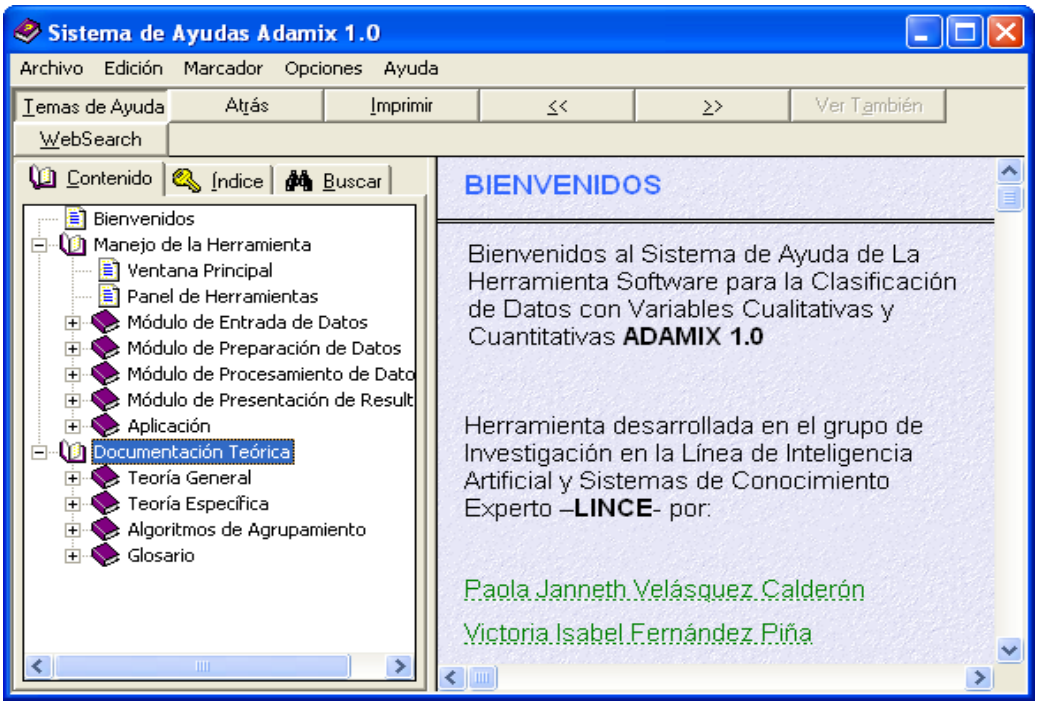
de la herramienta y el sistema le muestra dos tópicos llamados ventana principal y panel de herramienta y los libros módulo de entrada, preparación, y procesamiento de datos, módulo de presentación de resultados y por ultimo el libro de Aplicación, cada uno con sus diferentes tópicos.

- **Caso de uso fundamentación teórica**

  **Propósito:** Permite al usuario navegar por los diferentes ítem de la fundamentación teórica relacionada a conceptos básicos de minería de datos, descubrimiento de conocimiento en base de datos, agrupamiento en dominios mixtos y los algoritmos implementados.

  **Resumen:** El usuario selecciona la opción *contenido* del menú "Ayuda", el sistema presenta el contenido, el usuario abre el libro de ayuda teórica y el sistema muestra los libros de teoría general, específica, algoritmos de agrupamiento y glosario, al abrir cualquiera de estos el sistema muestra sus diferentes tópicos.

Figura 16. Sistema de ayuda



# 4 ANÁLISIS COMPARATIVO DE LOS ALGORITMOS DE AGRUPAMIENTO EN DOMINIOS MIXTOS

En este capítulo se realiza un estudio comparativo entre los algoritmos de agrupamiento implementados según los criterios de desempeño, complejidad, conceptualización y precisión.

Todas las pruebas realizadas tanto en la comparación de algoritmos como en la evaluación de la herramienta se han llevado a cabo utilizando un equipo computacional con las siguientes características:

- Un procesador Intel (R) Pentium (R) 4.
- 256 MB de RAM.
- Sistema operativo Microsoft Windows 2000.

## 4.1 CRITERIO DE DESEMPEÑO

En este criterio se compararon los algoritmos en cuanto al rendimiento, el cual ha sido medido por la velocidad de procesamiento, el tiempo de respuesta, consumo de recursos, y eficacia.

A continuación se presentan los resultados obtenidos con datos de pruebas tomados de la UCI Machine Learning Repository [11], formando 5 clases con cada algoritmo de agrupamiento, estos datos poseen un total de 2000 registros y 10 variables. Primero se realizaron pruebas con sólo datos cuantitativos, luego con datos cualitativos y finalmente con datos mixtos (cualitativos y cuantitativos).

El número de datos se obtiene de multiplicar el número de variables por el número de registros, para los tres casos de datos se utilizaron un total de 5 variables, y se

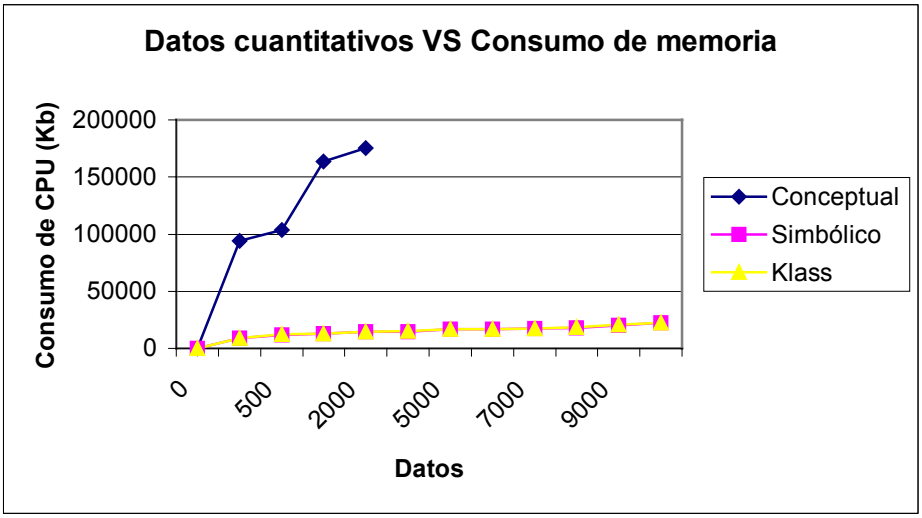
trabajó con 100, 200, 400, 600, 1000 y 2000 registros. Para datos mixtos se escogieron 2 variables cualitativas y 3 cuantitativas.

### 4.1.1 Consumo de Memoria

Tabla 1. Consumo de memoria de los algoritmos con variables cuantitativas.

| Consumo de memoria / Numero de datos | Algoritmo Conceptual | Algoritmo Simbólico | Algoritmo Klass |
|---|---|---|---|
| 500 | 103.476 KB | 11.720 KB | 12.104 KB |
| 1000 | 163.524 KB | 12.826 KB | 12.884 KB |
| 2000 | 175.476 KB | 14.468 KB | 14.828 KB |
| 3000 | Insuficiencia de memoria | 14.556 KB | 14.896 KB |
| 5000 | Insuficiencia de memoria | 16.836 KB | 16.804 KB |
| 10000 | Insuficiencia de memoria | 22.584 KB | 22.452 KB |

Figura 17. Consumo de memoria de los algoritmos con variables cuantitativas.

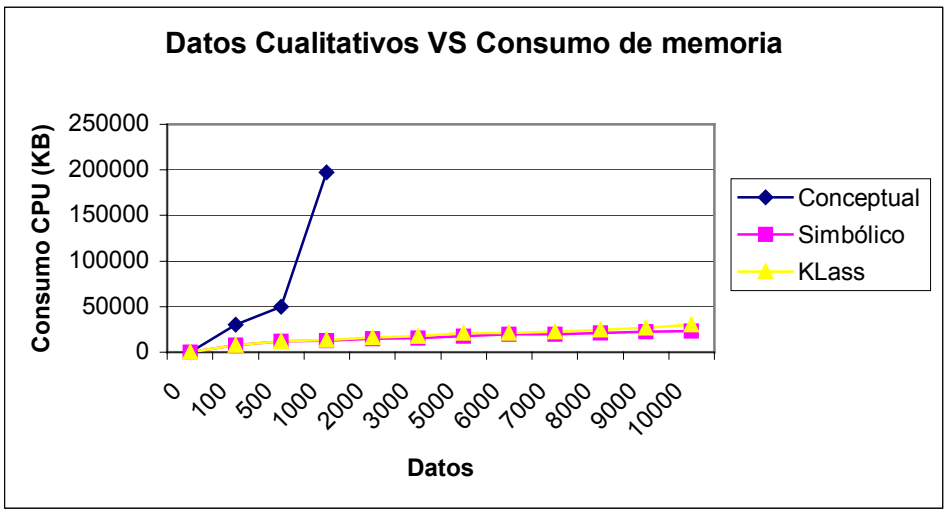


En la Tabla 1 y Figura 17 puede observarse que el algoritmo que obtuvo menor consumo de memoria para datos con variables cuantitativas fue el algoritmo de agrupamiento Simbólico y el de mayor consumo de memoria fue el algoritmo de agrupamiento conceptual basado en CLUSTER/2, el cual con una cantidad de datos superior a los 2000 desbordó el consumo de memoria.

Tabla 2. Consumo de memoria de los algoritmos con variables cualitativas.

| Consumo de memoria / Numero de datos | Algoritmo Conceptual | Algoritmo Simbólico | Algoritmo Klass |
|---|---|---|---|
| 500 | 50.048 KB | 12.054 KB | 12.072 KB |
| 1000 | 197.340 KB | 12.596 KB | 13.360 KB |
| 2000 | Insuficiencia de memoria | 14.612 KB | 16.336 KB |
| 3000 | Insuficiencia de memoria | 15.160 KB | 17.328 KB |
| 5000 | Insuficiencia de memoria | 17.260 KB | 21.312 KB |
| 10000 | Insuficiencia de memoria | 23.472 KB | 30.484 KB |

Figura 18. Consumo de memoria de los algoritmos con variables cualitativas

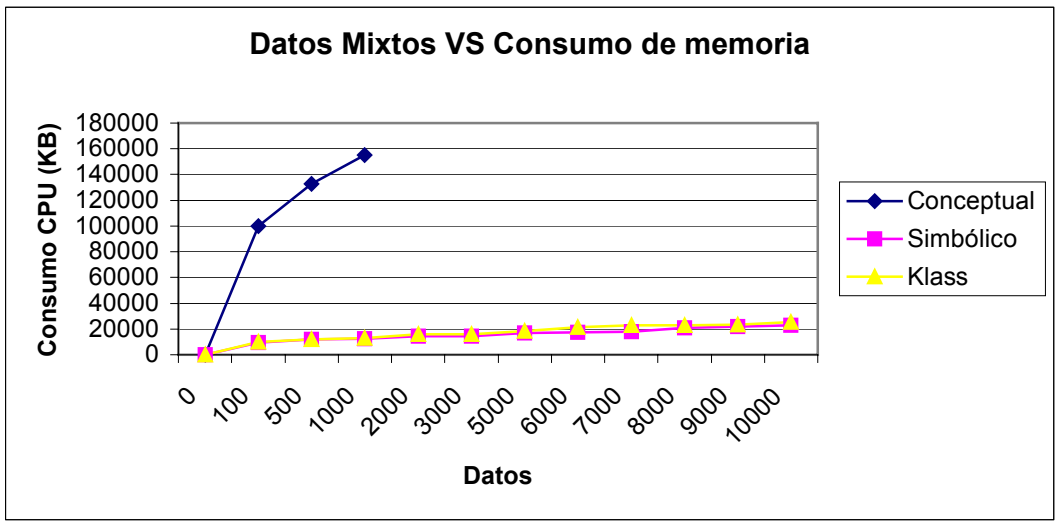


En la Tabla 2 y Figura 18 puede observarse que el algoritmo que obtuvo menor consumo de memoria para datos con variables cualitativas fue el algoritmo de agrupamiento Simbólico y el de mayor consumo de memoria fue el algoritmo de agrupamiento conceptual basado en CLUSTER/2, el cual con una cantidad de datos superior a los 1000 desbordó el consumo de memoria.

Tabla 3. Consumo de memoria de los algoritmos con variables mixtas.

| Consumo de memoria / Numero de datos | Algoritmo Conceptual | Algoritmo Simbólico | Algoritmo Klass |
|---|---|---|---|
| 500 | 132.620 KB | 11.712 KB | 11.832 KB |
| 1000 | 155.212 KB | 12.512 KB | 12.888 KB |
| 2000 | Insuficiencia de memoria | 14.524 KB | 15.832 KB |
| 3000 | Insuficiencia de memoria | 14.580 KB | 15.956 KB |
| 5000 | Insuficiencia de memoria | 16.928 KB | 18.412 KB |
| 10000 | Insuficiencia de memoria | 22.768 KB | 25.232 KB |

Figura 19. Consumo de memoria de los algoritmos con variables mixtas.

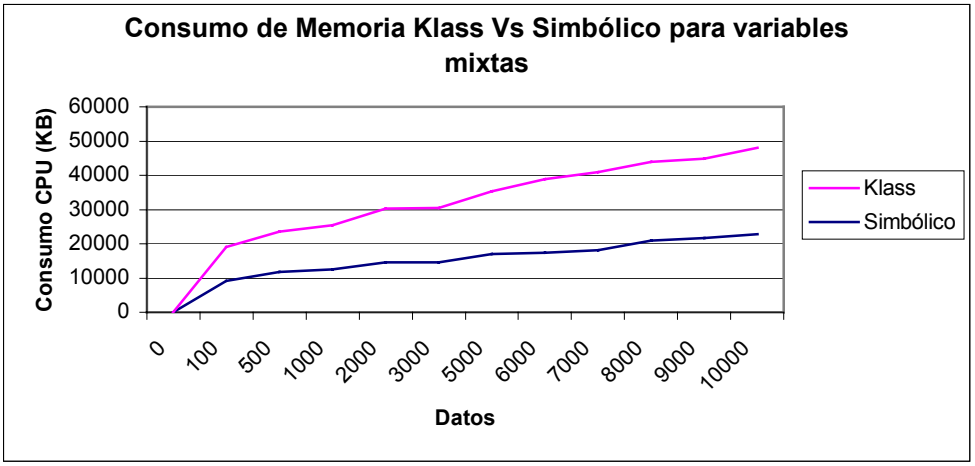


En la Tabla 3 y Figura 19 puede observarse que el algoritmo que obtuvo menor consumo de memoria para datos con variables mixtas fue el algoritmo de agrupamiento Simbólico y el de mayor consumo de memoria fue el algoritmo de agrupamiento conceptual basado en CLUSTER/2, el cual con una cantidad de datos superior a los 1000 desbordó el consumo de memoria.

En la Figura 19 puede apreciarse que la dinámica que describen las curvas de los algoritmos simbólico y klass tiene un comportamiento similar. Estas curvas no presentan puntos de inflexión marcados. La curva correspondiente al algoritmo Conceptual presenta una pendiente muy pronunciada, el consumo de memoria se incrementa rápidamente, alcanzando el nivel de saturación de uso de la memoria disponible sobre los 2000 registros. En el caso de los algoritmos klass y simbólico, todavía no se llega al nivel de saturación de memoria con el procesamiento de 10000 registros.

El pico de consumo de memoria presentado por el algoritmo conceptual enmascara el análisis de las curvas de los algoritmos Klass y simbólico. Para analizar mejor las diferencias entre estos dos últimos algoritmos se presenta la Figura 20, donde puede observarse que a medida que va aumentando la cantidad de datos la distancia entre las dos curvas tiende a incrementarse al igual que el consumo de memoria de los dos algoritmos, siendo el más alto para klass.

Figura 20. Consumo de memoria klass Vs simbólico para variables mixtas

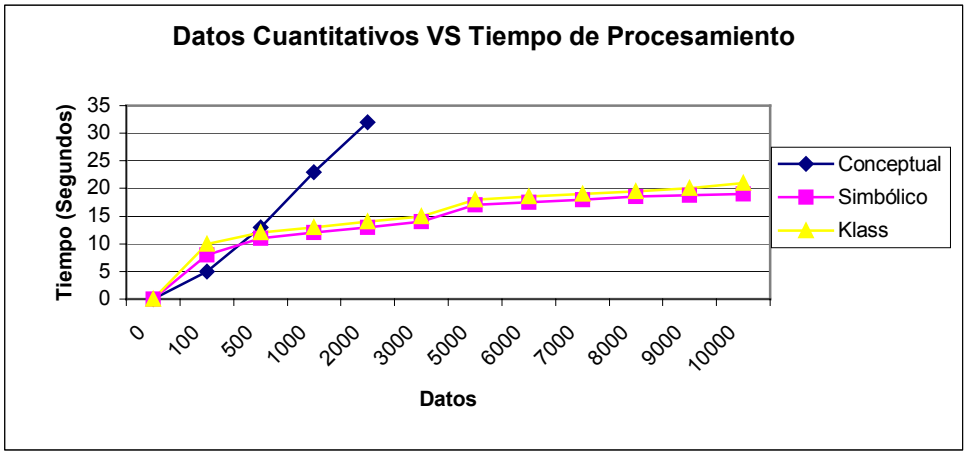


### 4.1.2 Tiempo de Procesamiento

Tabla 4. Tiempo de procesamiento de los algoritmos con variables Cuantitativas.

| Tiempo de procesamiento / Numero de datos | Algoritmo Conceptual | Algoritmo Simbólico | Algoritmo Klass |
|---|---|---|---|
| 500 | 13 Segundos | 11 Segundos | 12 Segundos |
| 1000 | 23 Segundos | 12 Segundos | 13 Segundos |
| 2000 | 32 Segundos | 13 Segundos | 14 Segundos |
| 3000 | Insuficiencia de memoria | 14 Segundos | 15 Segundos |
| 5000 | Insuficiencia de memoria | 17 Segundos | 18 Segundos |
| 10000 | Insuficiencia de memoria | 19 Segundos | 21 Segundos |

Figura 21. Tiempo de procesamiento de los algoritmos con variables Cuantitativas.

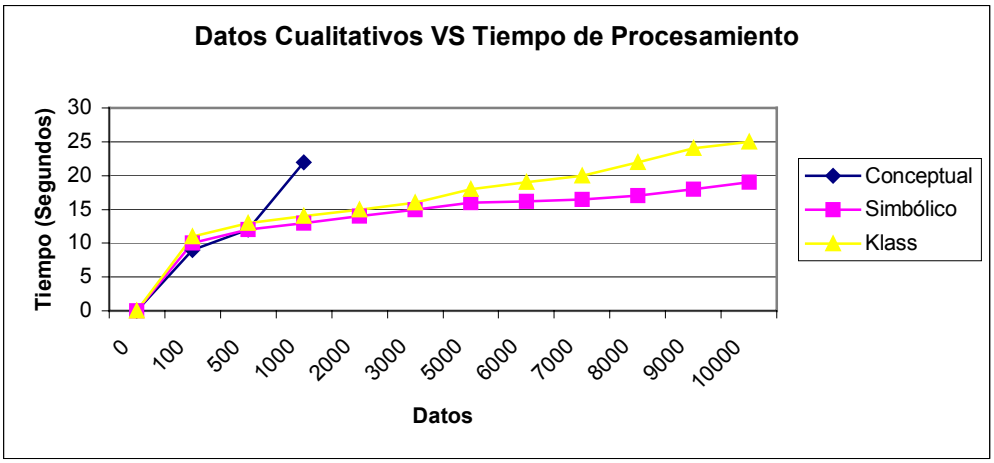



En la Tabla 4 y Figura 21 puede observarse que el algoritmo que obtuvo menor tiempo de procesamiento para datos con variables cuantitativas fue el algoritmo de agrupamiento Simbólico y el de mayor tiempo de procesamiento fue el algoritmo de agrupamiento conceptual basado en CLUSTER/2, el cual con una cantidad de datos superior a los 2000 saturó la capacidad del procesador.

Tabla 5. Tiempo de procesamiento de los algoritmos con variables Cualitativas.

| Tiempo de procesamiento / Numero de datos | Algoritmo Conceptual | Algoritmo Simbólico | Algoritmo Klass |
|---|---|---|---|
| 500 | 12 Segundos | 12 Segundos | 13 Segundos |
| 1000 | 22 Segundos | 13 Segundos | 14Segundos |
| 2000 | Insuficiencia de memoria | 14 Segundos | 15 Segundos |
| 3000 | Insuficiencia de memoria | 15 Segundos | 16 Segundos |
| 5000 | Insuficiencia de memoria | 16 Segundos | 18 Segundos |
| 10000 | Insuficiencia de memoria | 19 Segundos | 25 Segundos |

Figura 22. Tiempo en procesamiento de los algoritmos con variables Cualitativas.

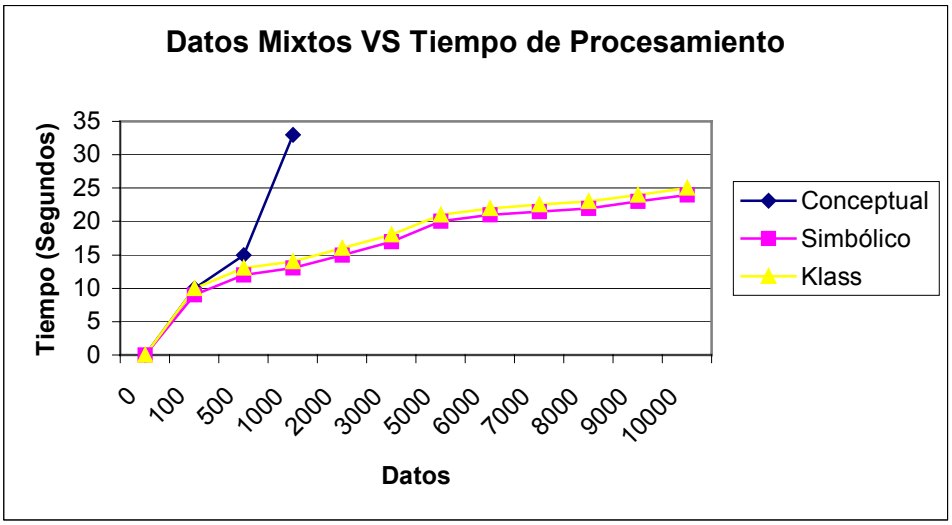



En la Tabla 5 y Figura 22 puede observarse que el algoritmo que obtuvo menor tiempo de procesamiento para datos con variables cualitativas fue el algoritmo de agrupamiento Simbólico y el de mayor tiempo de procesamiento fue el algoritmo de agrupamiento conceptual basado en CLUSTER/2 el cual con una cantidad de datos superior a los 1000 saturó la capacidad del procesador.

Tabla 6. Tiempo de procesamiento de los algoritmos con variables mixtas.

| Consumo de memoria / Numero de datos | Algoritmo Conceptual | Algoritmo Simbólico | Algoritmo Klass |
|---|---|---|---|
| 500 | 15 | 12 Segundos | 13 Segundos |
| 1000 | 33 | 13 Segundos | 14 Segundos |
| 2000 | Insuficiencia de memoria | 16 Segundos | 15 Segundos |
| 3000 | Insuficiencia de memoria | 17 Segundos | 18 Segundos |
| 5000 | Insuficiencia de memoria | 20 Segundos | 21 Segundos |
| 10000 | Insuficiencia de memoria | 24 Segundos | 25 Segundos |

Figura 23. Tiempo de procesamiento de los algoritmos con variables mixtas.




En la Tabla 6 y Figura 23 puede observarse que el algoritmo que obtuvo menor tiempo de procesamiento para datos con variables mixtas fue el algoritmo de agrupamiento Simbólico y el de mayor tiempo de procesamiento fue el algoritmo de agrupamiento conceptual basado en CLUSTER/2, el cual con una cantidad de datos superior a los 1000 saturó la capacidad del procesador.

Como puede observarse en los resultados de las tablas y figuras anteriores el algoritmo de agrupamiento Simbólico basado en una nueva medida de similaridad tiene un tiempo de procesamiento y consumo de memoria menor comparado con los algoritmos de agrupamiento Conceptual conjuntivo basado en CLUSTER/2 y el algoritmo de agrupamiento basado en métricas mixtas ponderadas Klass. Además se pudo establecer que los 3 algoritmos de agrupamiento tienen un tiempo de procesamiento y consumo de memoria mayores con variables cualitativas que con variables cuantitativas y que a medida que aumenta el número de datos aumenta el consumo de memoria así como el tiempo de procesamiento. Igualmente se concluyó que el algoritmo de agrupamiento Conceptual requiere mucha memoria para procesar una pequeña cantidad de datos.

### 4.1.3 Número de Iteraciones

El número de iteraciones para los algoritmos Klass y simbólico está dado por la siguiente formula:

*Numero de iteraciones* = *N – K.*

Donde *N* representa el número de registros y *K* el número de grupos o clases a formar. Cuando este número no se da como parámetro, sino que se pide calcular el número óptimo de clases, La fórmula anterior será reemplazada por:

*Numero de iteraciones* = *2N – K'.*

Donde *K*' es el número óptimo de clases calculado por el algoritmo.

Para el algoritmo de agrupamiento conceptual, el número de iteraciones depende de dos parámetros dados por el usuario, que son:

- Número de iteraciones base, y
- Número de iteraciones de prueba, después de la última mejora.

De tal manera que cuando el algoritmo completa el número de iteraciones base inicia el contador de iteraciones de prueba, y éste incrementa mientras no ocurra una mejora en el agrupamiento obtenido, si este contador supera el parámetro dado el algoritmo termina. Por consiguiente el número de Iteraciones del algoritmo Conceptual, estaría comprendido entre:

*Iteraciones Base + Iteraciones Prueba* $\leq$ *Número de Iteraciones* $\leq$ *n! / k!(n-k)!*

## 4.2 CRITERIO DE COMPLEJIDAD Y CONCEPTUALIZACIÓN

En este criterio se compararon los algoritmos en cuanto a la facilidad de implementación, a la densidad y reutilización de código, a la convergencia más rápida y a la fundamentación teórica, es decir cuál es más fuerte conceptualmente.

### 4.2.1 Facilidad de Implementación

El algoritmo de agrupamiento simbólico fue el menos complejo de implementar, a pesar de que maneja objetos muy complejos que están definidos por una conjunción lógica de eventos que une valores y variables, sólo se limita a hallar similaridades y agrupar sin demasiados cálculos internos, además no requiere conocimiento del número de grupos a priori ya que lo calcula por sí mismo, dándole la característica de ser un método de agrupamiento no paramétrico. La implementación del algoritmo de agrupamiento basado en métricas mixtas ponderadas Klass presentó cierta complejidad en el tratamiento de las variables cualitativas por la extensión de las variables sobre sus categorías (matriz extendida), y el cálculo permanente de cardinalidades por categorías y por registros. Y finalmente el más complejo para su implementación fue el algoritmo de agrupamiento conceptual basado en Cluster/2 por la cantidad de objetos diferentes que intervienen en una iteración del mismo, y más aún por la complejidad de operaciones al interior de dichos objetos.

En la Tabla 7 se presenta la calificación de los algoritmos en cuanto a la facilidad de implementación, donde la posición 1 representa la calificación más alta y 3 la más baja respectivamente.

Tabla 7. Facilidad de implementación.

| Facilidad de Implementación / Algoritmo | Alto | Medio | Bajo | **Posición** |
|---|---|---|---|---|
| Conceptual | X | | | **3** |
| Simbólico | | | X | **1** |
| Klass | | X | | **2** |

### 4.2.2 Densidad de Código

El algoritmo de agrupamiento conceptual conjuntivo tiene una densidad de código alta comparado con los demás algoritmos, con un total de 2021 líneas, 11 clases, 67 procedimientos y funciones, seguido por Klass con un total de 1618 líneas, 6 clases, 49 procedimientos y funciones y por último el algoritmo de agrupamiento Simbólico con 1334 líneas, 5 clases y 41 procedimientos y funciones.

En la Tabla 8 se presenta la calificación de los algoritmos en cuanto a la menor densidad de código donde la posición 1 representa la calificación más alta y 3 la más baja respectivamente.

Tabla 8. Densidad de código.

| Densidad de código / Algoritmo | Alto | Medio | Bajo | **Posición** |
|---|---|---|---|---|
| Conceptual | X | | | **3** |
| Simbólico | | | X | **1** |
| Klass | | X | | **2** |

### 4.2.3 Reutilización de Código

El algoritmo que permitió la mayor reutilización de código fue el algoritmo de agrupamiento simbólico por la forma uniforme en que maneja las variables cualitativas y cuantitativas, y las medidas de similaridad que aplica sobre ellas, a pesar de manejar una mayor variedad de variables, como se muestra en la Tabla 9.

Tabla 9. Tipos de variables soportadas por los algoritmos de agrupamiento.

| Variables / Algoritmos | CUALITATIVAS | | | CUANTITATIVAS | | |
|---|---|---|---|---|---|---|
| | Nominales | Ordinales | Estructuradas | Discretas | Valores de Radio continuo | Intervalos |
| Conceptual | X | | | X | X | X |
| Simbólico | X | X | X | X | X | X |
| Klass | X | | | X | X | |

En la Tabla 10 se presenta la calificación de los algoritmos en cuanto a la reutilización de código donde la posición 1 representa la calificación más alta y 3 la más baja respectivamente.

Tabla 10. Reutilización de código.

| Reutilización de código / Algoritmo | Alto | Medio | Bajo | **Posición** |
|---|---|---|---|---|
| Conceptual | | | X | **3** |
| Simbólico | X | | | **1** |
| Klass | | X | | **2** |

### 4.2.4 Convergencia más Rápida

Los algoritmos de agrupamiento Klass y simbólico convergen a la misma velocidad en cuanto a número de iteraciones se refiere dado su forma aglomerativa de agrupamiento, que agrupa un elemento en cada iteración. Pero en términos reales de tiempo, el algoritmo de agrupamiento simbólico converge más rápido, debido al menor tiempo empleado por iteración.

En cuanto al algoritmo de agrupamiento conceptual se puede decir que converge más rápido a una solución, mas no necesariamente a una solución óptima, puesto que obtiene un agrupamiento en k clases desde la primera iteración. El tiempo empleado por este algoritmo depende del número de iteraciones especificadas por el usuario, y del número de datos a agrupar. Pero en términos generales es más lento que los dos anteriores.

En la Tabla 11 se presenta la calificación de los algoritmos en cuanto a la convergencia más rápida donde la posición 1 representa la calificación más alta y 3 la más baja respectivamente.

Tabla 11. Convergencia más rápida.

| Convergencia / Algoritmos | Alto | Medio | Bajo | **Posición** |
|---|---|---|---|---|
| Conceptual | | | X | **3** |
| Simbólico | X | | | **1** |
| Klass | X | | | **2** |

### 4.2.5 Fundamentación Teórica

En cuanto a la fundamentación teórica, el algoritmo que presenta mayor soporte es el conceptual conjuntivo, el cual ha sido objeto de estudio desde los 80 hasta la actualidad, siendo cluster/2 el primer trabajo conocido sobre agrupamiento con

variables cualitativas, y surgiendo a partir de él muchos trabajos en lo que se denominó la rama del agrupamiento conceptual (Conceptual Clustering). Después se ubica el algoritmo de agrupamiento simbólico que introduce una medida de similaridad basada en tres componentes (la posición, la extensión y el contenido), el cual ofrece la posibilidad de tratar con datos cuantitativos de radio continuo, discretos e intervalos al igual que con datos cualitativos nominales, ordinales e incluso con datos estructurados en forma de árbol. Finalmente el algoritmo de agrupamiento Klass, el más reciente (1994), el cual si bien es cierto no tiene la densa fundamentación teórica del algoritmo conceptual, en la que radica su complejidad, ni la amplitud en el manejo de variables del algoritmo simbólico, maneja una representación de los objetos que le proporciona bastante precisión en el momento de agrupar.

En la Tabla 12 se muestra la calificación de los algoritmos en cuanto a la fundamentación teórica donde la posición 1 representa la calificación más alta y 3 la más baja. respectivamente

Tabla 12. Fundamentación teórica

| Fundamentación teórica / Algoritmos | Alto | Medio | Bajo | **Posición** |
|---|---|---|---|---|
| Conceptual | X | | | **1** |
| Simbólico | | X | | **2** |
| Klass | | | X | **3** |

A continuación se hace un resumen (ver Tabla 13), donde se presenta la posición de cada algoritmo en los anteriores criterios, y su posición final en el criterio de complejidad y conceptualización. La posición 1 representa calificación más alta y 3 la más baja respectivamente.

Tabla 13. Resumen criterio de complejidad y conceptualización.

| Complejidad Conceptualización / Algoritmos | Facilidad de implementación | Densidad de código | Reutilización de código | Convergencia más rápida | Fundamentación teórica |
|---|---|---|---|---|---|
| Conceptual | 3 | 3 | 3 | 3 | 1 |
| Simbólico | 1 | 1 | 1 | 1 | 2 |
| Klass | 2 | 2 | 2 | 2 | 3 |

| Complejidad y Conceptualización / Algoritmos | **Totales** | **Posición** |
|---|---|---|
| Conceptual | **13** | **3** |
| Simbólico | **6** | **1** |
| Klass | **11** | **2** |

### 4.3 CRITERIO DE PRECISIÓN

En este criterio se compararon los algoritmos en cuanto a la aproximación de los resultados obtenidos por cada uno a resultados estándares. Para ello se ha utilizado un ejemplo de clasificación de Microcomputadores documentado en [7]. En la Tabla 14 se presenta el conjunto de datos para los microcomputadores, el cual describe 13 microcomputadores americanos en términos de las variables 'Display', 'RAM', 'ROM', 'MP' y 'Keys', de las cuales, Display, MP y Keys son variables cualitativas.

Tabla 14. Conjunto de datos para los microcomputadores.

| Objetos Microcomputadores | Id. | Display | RAM | ROM | MP | Keys |
|---|---|---|---|---|---|---|
| APPLE-II | AP | COLOR-TV | 48 | 10 | 6502 | 52 |
| ATARI-800 | AT | COLOR-TV | 48 | 10 | 6502 | 57-63 |
| COMMODORE-VIC-20-A | CoA | COLOR-TV | 32 | 11 | 6502A | 64-73 |
| COMMODORE-VIC-20-B | CoB | COLOR-TV | 32 | 16 | 6502A | 64-73 |
| EXIDI-SORCERER | ES | B-&-W-TV | 48 | 4 | Z80 | 57-63 |
| ZENITH-H8 | ZH8 | BUILT-IN | 64 | 1 | 8080A | 64-73 |
| ZENITH-H89 | ZH89 | BUILT-IN | 64 | 8 | Z80 | 64-73 |
| HP-85 | HP | BUILT-IN | 32 | 80 | HP | 92 |
| HORIZON | Ho | TERMINAL | 64 | 8 | Z80 | 57-63 |
| OHIO-SC.-CHALLENGER | OCh | B-&-W-TV | 32 | 10 | 6502 | 53-56 |
| OHIO-SC.-II-SERIES | OS | B-&-W-TV | 48 | 10 | 6502C | 53-56 |
| TRS-80-I | TRI | B-&-W-TV | 48 | 12 | Z80 | 53-56 |
| TRS-80 | TRIII | BUILT-IN | 48 | 14 | Z80 | 64-73 |

El experto en el tema de los microcomputadores consideró que las variables *ROM* y *Keys* son poco importantes en la caracterización de los mismos y estructuró las variables categóricas de la siguiente manera basado en sus conocimientos y experiencia:

$$Display \left| \begin{array}{l} TV \left| \begin{array}{l} Black\ \&\ White \\ Color \end{array} \right. \\ Built\ in \\ Ter\ \min\ al \end{array} \right. \qquad Microprocesador \left| \begin{array}{l} Motorola \left| \begin{array}{l} 6502 \\ 6502A \\ 6502C \end{array} \right. \\ Intel \left| \begin{array}{l} Z80 \\ 8080A \end{array} \right. \\ Hewlett\ Packard \end{array} \right.$$

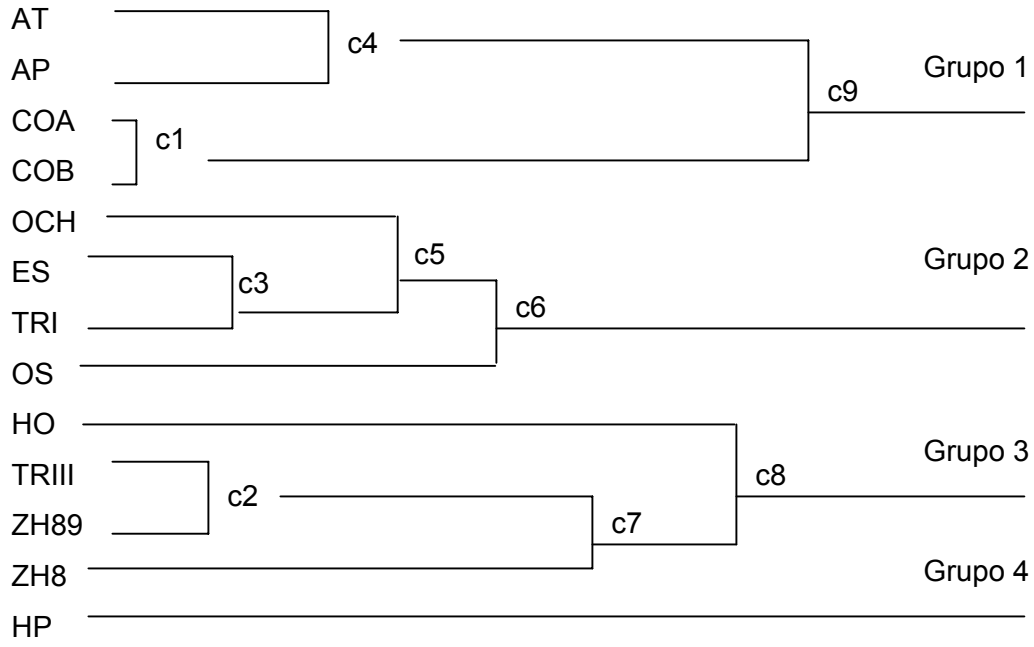
después de esto el experto propuso su clasificación de acuerdo a los valores tomados por las variables *Display*, *MP* y *RAM*. La Tabla 15 muestra la clasificación propuesta por el experto Vs la realizada por los algoritmos de agrupamiento implementados.

Tabla 15. Clasificación de los microcomputadores.

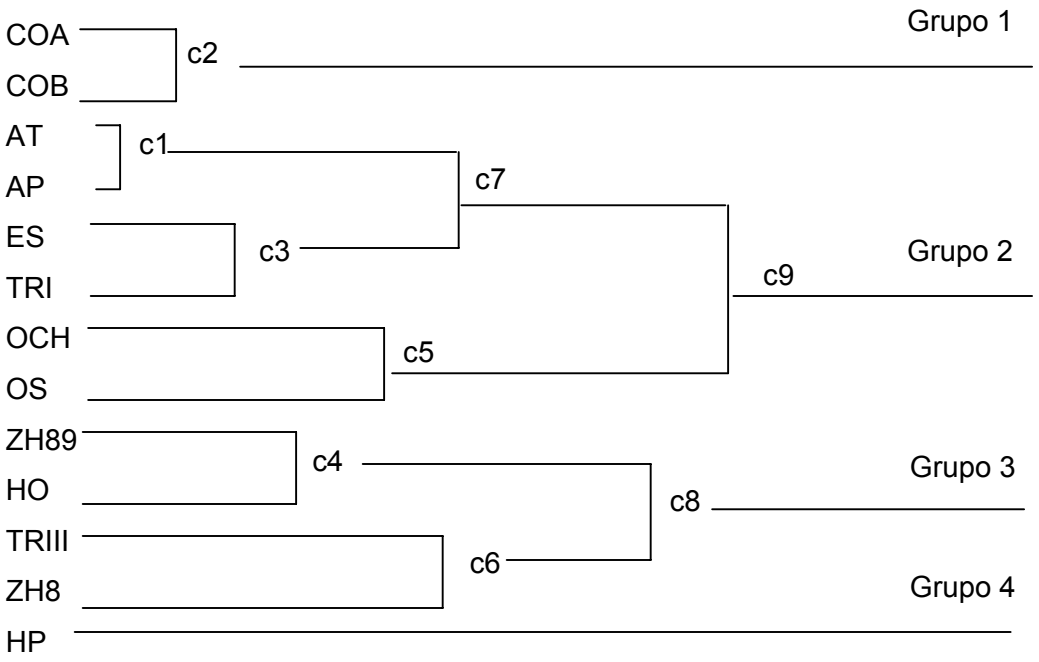
| Id. | | Conceptual | Simbólico | Klass | **Experto** |
|---|---|---|---|---|---|
| 0 | AP | 1 | 2 | 1 | **1** |
| 1 | AT | 1 | 2 | 1 | **1** |
| 2 | CoA | 1 | 1 | 1 | **1** |
| 3 | CoB | 1 | 1 | 1 | **1** |
| 4 | ES | 1 | 2 | 2 | **2** |
| 5 | ZH8 | 3 | 3 | 3 | **3** |
| 6 | ZH89 | 3 | 3 | 3 | **3** |
| 7 | HP | 2 | 4 | 4 | **3** |
| 8 | Ho | 1 | 3 | 3 | **4** |
| 9 | OCh | 1 | 2 | 2 | **2** |
| 10 | OS | 4 | 2 | 2 | **2** |
| 11 | TRI | 1 | 2 | 2 | **2** |
| 12 | TRIII | 1 | 3 | 3 | **3** |

A continuación se muestran los dendrogramas obtenidos para los algoritmos de agrupamiento Klass y Simbólico.

*Dendrograma Klass*



*Dendrograma Simbólico*




Realizada la comparación de los resultados se concluyó que de los tres algoritmos el que más se aproximó a los resultados del experto fue el *Algoritmo de Agrupamiento basado en Métricas Mixtas Ponderadas Klass*, con un porcentaje de error de un 15 % que equivale a 2 objetos ubicados en grupos diferentes, seguido por el *Algoritmo de Agrupamiento Simbólico basado en una Nueva Medida de Similaridad* con un porcentaje de error del 31 % equivalente a 4 objetos ubicados en grupos diferentes, y por último el *Algoritmo de Agrupamiento Conceptual Conjuntivo basado en CLUSTER/2* con un porcentaje de error del 54 % que equivale a 7 objetos ubicados en grupos diferentes.

En la Tabla 16 se muestra el porcentaje de precisión de cada algoritmo en cuanto a su aproximación a la clasificación del experto y además su posición con respecto a los otros.

Tabla 16. Precisión de los algoritmos de agrupamiento.

| Precisión / Algoritmos | Precisión % | **Posición** |
|---|---|---|
| Algoritmo de agrupamiento conceptual conjuntivo basado en CLUSTER/2 | 46 % | **3** |
| Algoritmo de agrupamiento simbólico basado en una nueva medida de similaridad | 69 % | **2** |
| Algoritmo de agrupamiento basado en métricas mixtas ponderadas Klass | 85 % | **1** |

## 4.4 CONCLUSIONES

A continuación se presenta una tabla resumen (ver Tabla 17) donde se incluyen todos los criterios evaluados en los algoritmos, calificándolos de 1 a 3 de acuerdo a los resultados obtenidos, donde 1 representa la más alta calificación y 3 la menor.

Tabla 17. Calificación Final de los algoritmos en el análisis comparativo.

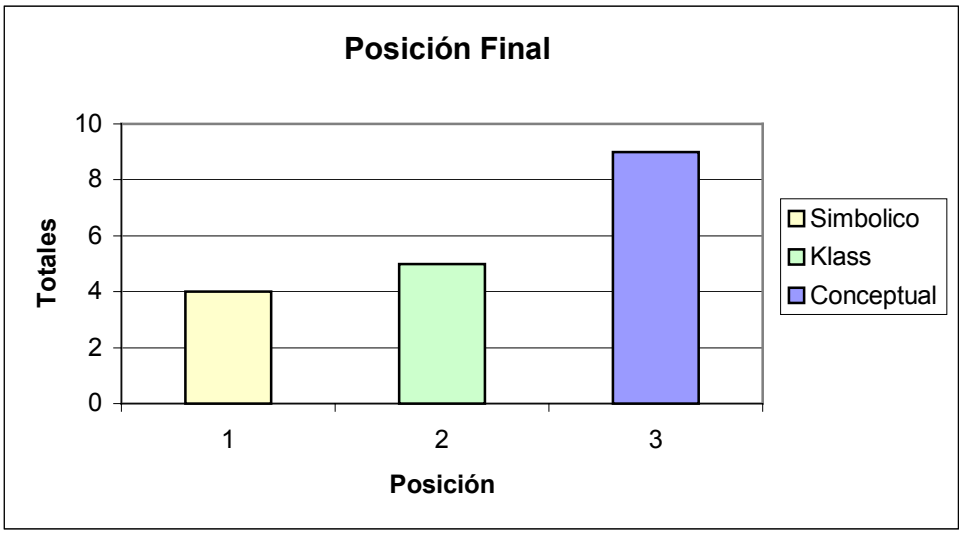
| **Criterios** / **Algoritmos** | Desempeño | Complejida y Conceptualización | Precisión | **Totales** | **Posición** |
|---|---|---|---|---|---|
| Algoritmo de agrupamiento conceptual conjuntivo basado en CLUSTER/2 | 3 | 3 | 3 | **9** | **3** |
| Algoritmo de agrupamiento simbólico basado en una nueva medida de similaridad | 1 | 1 | 2 | **4** | **1** |
| Algoritmo de agrupamiento basado en métricas mixtas ponderadas Klass | 2 | 2 | 1 | **5** | **2** |

Como puede observarse en la Tabla 17 el algoritmo que obtuvo mejor desempeño fue el algoritmo de agrupamiento simbólico, por su poco uso de memoria y tiempo de procesamiento, seguido por el algoritmo agrupamiento klass y finalmente el algoritmo de agrupamiento conceptual. En cuanto a la complejidad y conceptualización el algoritmo simbólico es el primero por su facilidad de implementación, su poca densidad de código. Finalmente en el criterio de precisión el primero es el algoritmo klass, por su cercanía a los resultados comparados con el experto y el menos preciso es el algoritmo conceptual.

Para determinar cual algoritmo es mejor, evaluando todos los criterios se totalizaron las filas, donde el valor más alto corresponde a la posición más baja (Posición 3) y el valor más bajo a la posición más alta (Posición 1), concluyéndose así que el algoritmo de agrupamiento simbólico basado en una nueva medida de similaridad es el mejor con un total de 4, seguido por el algoritmo de agrupamiento klass con un total de 5 y finalmente el algoritmo de agrupamiento conceptual con un total de 9.

A continuación se presentan los totales y posición de los algoritmos después de evaluar todos los criterios (ver Figura 24).

Figura 24. Posición final de los algoritmos en el análisis comparativo.



# 5 EVALUACIÓN DE LA HERRAMIENTA SOFTWARE

En este capitulo se presenta la evaluación realizada a la herramienta software en cuanto a su calidad y funcionalidad. Para ello se han llevado a cabo diversas pruebas como caja negra, pruebas alfa y pruebas beta, cuya teoría se ha explicado anteriormente (ver capitulo 3 secciones 3.3.3 y 3.3.4). Por otra parte, la calidad de la herramienta fue validada por medio de una aplicación a un caso real y su interpretación y constatación de sus resultados por parte de un experto.

La prueba del software es un elemento de un tema más amplio que, a menudo, es conocido como verificación y validación. La *verificación* se refiere al conjunto de actividades que aseguran que el software implementa correctamente una función específica. La *validación* se refiere a un conjunto diferente de actividades que aseguran que el software construido se ajusta a los requisitos del cliente. Las pruebas constituyen el último bastión desde el que se puede evaluar la calidad y, de forma más pragmática, descubrir los errores [12].

La validación de la herramienta se consigue mediante una serie de pruebas de caja negra, que se llevan a cabo sobre la interfaz del software. Es decir, los casos de prueba pretenden demostrar que las funciones del software son operativas, que la entrada se acepta de forma adecuada y que se produce un resultado correcto, de igual forma, que la integridad de la información externa se mantiene [12].

## 5.1 EVALUACIÓN EN APLICACIÓN DE LA HERRAMIENTA SOBRE DATOS DE PRUEBA

Para la evaluación de la herramienta con datos de prueba, se tomó el conjunto de datos de Mushroom de la UCI Machine Learning Repository [11].

### 5.1.1 Análisis de los datos

Cada registro del conjunto de datos Mushroom contiene información que describe las características físicas ( Ej., color, olor, tamaño, forma) de un solo hongo, además presenta etiquetas de comestibles o venenoso.

La base de datos de Mushroom tiene un total de 8124 registros, todas las variables son cualitativas, 22 atributos nominales y 2 sobre información de los atributos (edible (comestible) = e, poisonous (venenoso) = p). A continuación se presentan las variables con sus respectivas categorías:

1. ***Cap-shape (Forma del capuchón):***

   Bell = b, conical = c, convex = x, flat = f, knobbed = k, sunken = s.

2. ***Cap-surface (Superficie del capuchón):***

   Fibrous = f, grooves = g, scaly = y, smooth = s.

3. ***Cap-color (Color del capuchón):***

   Brown = n, buff = b, cinnamon = c, gray = g, green = r, pink = p, purple = u, red = e, white = w, yellow = y.

4. ***Bruises? (Ronchado):***

   Bruises = t, no = f.

5. ***Odor (Olor):***

   Almond = a, anise = l, creosote = c, fishy = y, foul = f, musty = m, none = n, pungent = p, spicy = s.

6. ***Gill-attachment (Adherencia):***

   Attached = a, descending = d, free = f, notched = n.

7. ***Gill-spacing (Espaciamiento):***

   Close = c, crowded = w, distant = d.

8. ***Gill-size (Tamaño):***

   Broad = b, narrow = n.

9. ***Gill-color (Color):***

Black = k, brown = n, buff = b, chocolate = h, gray = g, green = r, orange = o, pink = p, purple = u, red = e, white = w, yellow = y.

***10. Stalk-shape (Forma del tallo):***

Enlarging = e, tapering = t.

***11. Stalk-root (Tallo de la raíz):***

Bulbous = b, club = c, cup = u, equal = e, rhizomorphs = z, rooted = r, missing = ?.

***12. Stalk-surface-above-ring (Superficie del tallo encima del anillo):***

Fibrous = f, scaly = y, silky = k, smooth = s.

***13. Stalk-surface-below-ring (Superficie del tallo debajo del anillo):***

Fibrous = f, scaly = y, silky = k, smooth = s.

***14. Stalk-color-above-ring (Color del tallo encima del anillo):***

Brown = n, buff = b, cinnamon = c, gray = g, orange = o, pink = p, red = e, white = w, yellow = y.

***15. Stalk-color-below-ring (Color del tallo debajo del anillo):***

Brown = n, buff = b, cinnamon = c, gray = g, orange = o, pink = p, red = e, white = w, yellow = y.

***16. Veil-type (Tipo de velo):***

Partial = p, universal = u.

***17. Veil-color (Color del velo):***

Brown = n, orange = o, white = w, yellow = y.

***18. Ring-number (Número de anillos):***

None = n, one =o, two = t.

***19. Ring-type (Tipo de anillo):***

Cobwebby = c, evanescent = e, flaring = f, large = l, none = n, pendant = p, sheathing = s, zone = z.

***20. Spore-print-color(Color de la espora):***

Black = k, brown = n, buff = b, chocolate = h, green = r, orange = o, purple = u, white = w, yellow = y.

***21. Population (Población):***

Abundant = a, clustered = c, numerous = n, scattered = s, several = v, solitary = y.

***22. Habitat (Habitat):***

Grasses = g, leaves = l, meadows = m, paths = p, urban =u, waste = w, woods = d.

### 5.1.2 Preparación de los datos

**Selección de las variables de estudio**

Las variables que son representativas para el estudio de las características de los hongos se presentan en la Tabla 18, donde se muestra el nombre de las variables, la descripción, el tipo y las categorías que presenta cada una.

Tabla 18. Variables seleccionadas de la base de datos Mushroom.

| Nombre Variables | Descripción | Tipo | Valores |
|---|---|---|---|
| Cap-color | Color del capuchon | Cualitativa | n, b, c, g, r, p, u, e, w, y |
| Odor | Olor | Cualitativa | a, l, c, y, f, m, n, p, s |
| Stalk-surface-below-ring | Superficie del tallo debajo del anillo | Cualitativa | f, y, k, s |
| Stalk-color-above-ring | Color del tallo encima del anillo | Cualitativa | n, b, c, g, o, p, e, w, y |
| Spore-print-color | Color de la espora | Cualitativa | k, n, b, h, r, o, u, w, y |
| Population | Población | Cualitativa | a, c, n, s, v, y |
| Habitat | Habita de los hongos | Cualitativa | g, l, m, p, u, w, d |

En la Tabla 19 se muestran los resultados obtenidos después de la preparación básica, las categorías más frecuentes y menos frecuente en cada variable.

Tabla 19. Categoría más frecuente y categoría menos frecuente de la base de datos Mushroom.

| Medidas Estadísticas / variables | Categoría más frecuente | Categoría menos frecuente |
|---|---|---|
| Cap-color | n (café) frecuencia de 2284 | u (púrpura) frecuencia de 16 |
| Odor | n (ninguno) frecuencia de 3528 | m (rancio) frecuencia de 36 |
| Stalk-surface-below-ring | s (suave) frecuencia de 4936 | y (escamoso) frecuencia de 284 |
| Stalk-color-above-ring | w (blanco) frecuencia de 4464 | y (amarillo) frecuencia de 8 |
| Spore-print-color | w (blanco) frecuencia de 2388 | u (púrpura) frecuencia de 48 |
| Population | v (varios) frecuencia de 4040 | c (en racimo) frecuencia de 340 |
| Habitat | d (bosques) frecuencia de 3148 | w (desechos) frecuencia de 192 |

### 5.1.3 Procesamiento de los datos

Se procesaron un total de 8.124 registros y 7 variables, en la Tabla 20 se presentan los resultados obtenidos, donde pueden apreciarse las categorías de las variables cualitativas presentes en cada grupo.

Tabla 20. Grupos formados para la base de datos Mushroom.

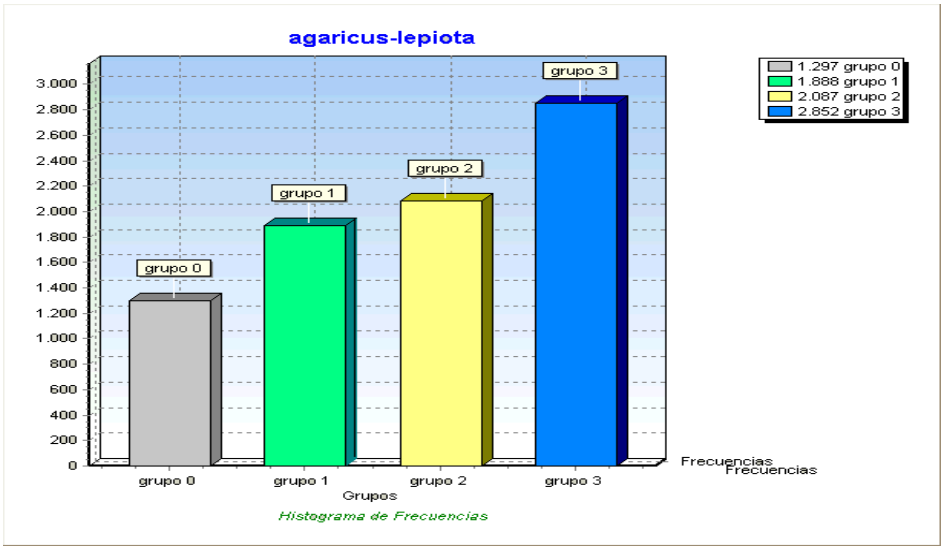
| Grupos / Variables | Grupo 0 | Grupo 1 | Grupo 2 | Grupo 3 |
|---|---|---|---|---|
| Cap-color | {y,g} | {e,g,n } | {e,n,g,c,p,y,w } | {n,w,g,e,b,p,c,u,r,y } |
| Odor | {f} | {n} | {f,y,s,n,m} | {n,p,a,f,l,c, } |
| Stalk-surface-below-ring | {k,s} | {s} | {s,k,y,f} | {s,f,y,k} |
| Stalk-color-above-ring | {p,n,b,w} | {p,g,w,o } | {p,w,c,n.y,e} | {w} |
| Spore-print-color | {h} | {n,k,y,o,b} | {w} | {n,k,w,r,h,u} |
| Population | {v,y} | {y,v,c,a} | {v,y,c} | {a,s,n,c,v,y} |
| Habitat | {d,p,g} | {d,l,u,g} | {d,l,p,w} | {g,w,p,m,d,u} |

Tabla 21. Distancias Intercluster de los grupos obtenidos de los datos Mushroom.

| Grupos | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| **0** | 0 | 0.309 | 0.279 | 0.3 |
| **1** | 0.309 | 0 | 0.297 | 0.89 |
| **2** | 0.279 | 0.297 | 0 | 0.237 |
| **3** | 0.3 | 0.289 | 0.237 | 0 |

En la Figura 25 se muestra el número de registros en cada grupo, el grupo 0 tiene 1297 registros, el grupo 1 tiene 1888 registros, el grupo 2 tiene 2087 registros y el grupo 3 tiene 2852 registros.

Figura 25. Número de Registros en cada grupo de los datos Mushroom.




### 5.1.4 Interpretación de los resultados

Del análisis de los grupos obtenidos se han determinado las siguientes características de los hongos en cada grupo:

- El grupo 0, está conformado por hongos venenosos que presentan las siguientes características: color del capuchón amarillo (y) o gris (g), con un

olor fétido (f), con una superficie del tallo debajo el anillo suave (s) o sedoso (k), con un color de esporas chocolate (h), su población puede ser de varios (v) o solitarios (y), y su hábitat es en bosques (d), caminos (p) o céspedes (g).

- El grupo 1, está compuesto por hongos comestibles que presentan las siguientes características: color del capuchón rojo (e), café (n) o gris (g), sin ningún (n) olor característico, con una superficie del tallo debajo del anillo suave (s), con colores de esporas café (n), negro (k), amarillo (y), naranja (o), o amarillo claro (b), su población puede ser de varios (v), solitarios (y), en racimos (c) o abundantes (a),y su hábitat es en bosques (d), hojas (l), en zonas urbanas (u), o céspedes (g).

- El grupo 2, está constituido en un 87% por hongos venenosos, y en un 13% por hongos comestibles, que en conjunto presentan las siguientes características: color del capuchón rojo (e), café (n), gris (g), canela (c), rosa (p), amarillo (y), o blanco (w), con olor fétido (f), a pescado (y), a condimento (s), a rancio (m) o sin ningún (n) olor característico, con una superficie del tallo debajo del anillo suave (s), sedoso (k), escamoso (y), o fibroso (f), con colores de esporas blanco (w), su población puede ser de varios (v), solitarios (y) o en racimos (c), y su hábitat es en bosques (d), hojas (l), caminos (p) o desechos (w). Por lo tanto a los hongos con estas características, para mayor seguridad es mejor considerarlos como venenosos, y por ende no consumirlos.

- El grupo 3, está conformado en un 72% por hongos comestibles, y en un 28% por hongos venenosos, que en conjunto presentan las siguientes características: color del capuchón café (n), blanco (w), rojo (e), amarillo claro (b), rosa (p), canela (c), púrpura (u), verde (r), o amarillo (y), con olor característico a picante (p), almendra (a), fétido (f), anís (l), creosota (c) o sin ningún (n) olor característico, con una superficie del tallo debajo del anillo suave (s), sedoso (k), escamoso (y), o fibroso (f), con colores de esporas

café (n), negro (k), blanco (w), verde (r), chocolate (h) o púrpura (u), su población puede ser abundante (a), dispersa (s), numerosa (n), en racimos (c), de varios (v) o solitarios (y), y su hábitat es en céspedes (g), desechos (w), caminos (p), prados (m), bosques (d) o en zona urbana. Por lo tanto los hongos con estas características, pueden ser consumidos con precaución, haciendo un análisis más riguroso, de ellos para mayor seguridad; pues a pesar de ser en su mayoría comestibles, el 28% es un riesgo que no se puede descartar. Una opción aquí podría ser aplicar nuevamente la herramienta de clasificación solo a este grupo de datos, e incluir en el análisis las variables, no tenidas en cuenta en la anterior clasificación.

La clasificación arrojada por la herramienta es consistentes con la información manejada por los expertos, ya que tiende a ubicar los hongos comestibles, y los venenosos en grupos separados, brindado para cada grupo, descripciones que permitan discriminar su comestibilidad mediante un número reducido de atributos, y los grupos encontrados constituyen una clasificación lógica de la familia de los hongos agaricus lepiota.

## 5.2 EVALUACIÓN EN APLICACIÓN DE LA HERRAMIENTA SOBRE DATOS REALES

Los métodos de clasificación o separación de datos en grupos son interesantes desde el punto de vista de la Inteligencia Artificial porque abren una puerta a la generación automatizada de reglas, útiles en los ambientes basados en el conocimiento. Por tal motivo los trabajos realizados alrededor de las técnicas de agrupamiento han sido bastante profusos, encontrando entre sus principales aplicaciones: la caracterización de clientes, formación de taxonomías y clasificación de documentos entre otros; sin embargo han encontrado un reto crucial en lo concerniente a la capacidad de manejar objetos representados por la combinación de atributos cuantitativos y cualitativos, y todas las clasificaciones que se derivan de estos dos tipos fundamentales de datos.

### 5.2.1 Selección de los datos

Para la selección de los datos, se tuvo en cuenta los siguientes parámetros:

- La disponibilidad de la información.
- La disponibilidad del experto.
- Seguridad de los datos.

Una vez determinados estos parámetros, se seleccionaron los datos. Los datos recopilados para las pruebas fueron:

- Base de Datos Académica de la UIS.
- Datos de la Red Sismológica de Santander.

### 5.2.2 Base de Datos Académica de la UIS

#### 5.2.2.1 Análisis de los datos

Los datos de la base académica de la UIS fueron suministrados por el proyecto "Descubrimiento de conocimientos en la base de datos académica de la Universidad Industrial de Santander" [13], los cuales se obtuvieron del sistema de información de registro académico y del sistema de información financiero de la UIS, del período comprendido desde 1986 hasta 1999 con un total de 26336 registros. La base de datos contiene información sobre el currículo de estudiantes, currículo de profesores, programas académicos, evaluaciones de estudiantes y evaluaciones de profesores.

Los datos presentan un preprocesamiento, en el cual se eliminaron registros con valores fuera del rango, con códigos inexistentes y con valores nulos o perdidos, obteniéndose así una bodega de datos confiable de una población de 22969 registros.

El propósito de trabajar con estos datos es identificar a los estudiantes con buen y mal desempeño académico, cuáles son las causas de retención y por qué desertan los estudiantes en la universidad.

Las variables que presenta esta base de datos son las siguientes:
Cod_facultad, cod_escuela, cod_carrera, puntos_icfes, prom_acumulado, zona_procede, tipo_colegio, estrato, razcreditos, edad, estancia, desertor, categoría, retención, nota_promedio1, nota_promedio2, nota_promedio3, nota_promedio4, nota_promedio5, nota_promedio6, nota_promedio7, nota_promedio8, nota_promedio9, fecha_grado.

#### 5.2.2.2 Preparación de los datos

**Selección de las variables de estudio**

Para la selección de las variables se tuvieron en cuenta los objetivos de estudio:

- Buen y mal desempeño académico.
- Deserción estudiantil.
- Retención estudiantil.

Tabla 22. Variables seleccionadas de la base de datos de la UIS.

| Nombre Variables | Descripción | Tipo | Valores |
|---|---|---|---|
| Cod_facultad | Código de la facultad | Cualitativas | Las facultades |
| Cod_carrera | Código de la carrera | Cualitativas | Las carreras |
| Prom_acumulado | Promedio acumulado | Cuantitativas | Reales[0-5] |
| Puntos_icfes | Puntos en examen Icfes | Cuantitativas | Enteros[220-418] |
| Razcreditos | créditos aprobados / créditos cursados, | Cuantitativas | Reales[0-1] |
| Edad | Edad del estudiante | Cuantitativas | Reales[15.99-44.70] |
| Estancia | Estancia en la UIS | Cuantitativas | Reales[0-17.32] |
| Sexo | Sexo | Cualitativas | Femenino, Masculino |
| Desertor | Indicador de deserción | Cualitativas | Desertor, No desertor |
| Retención | Indicador de retención | Cualitativas | Retenido, No retenido |

El volumen de datos a manipular fue de 9496 registros y 10 variables, que corresponde a la facultad de físico mecánicas (6500). Las variables eliminadas en el proceso de preparación fueron zona_procede, tipo_colegio, estrato, nota_promedio1, nota_promedio2, nota_promedio3, nota_promedio4, nota_promedio5, nota_promedio6, nota_promedio7, nota_promedio8, nota_promedio9, fecha_grado, ya que no se encuentran actualizadas para toda la población.

La variable tipo_colegio, se ha utilizado solamente para la población cuya zona de procedencia sea Bucaramanga o Santander, la variable estrato sólo se encuentra actualizada para los estudiantes matriculados en el segundo semestre de 1999, que corresponde a la fecha de extracción de los datos, las variables nota_promedio de la 1 a la 9 son variables para un estudio particular sobre áreas de conocimientos y la variable fecha_grado servirá como indicador para seleccionar los registros de estudiantes graduados.

Una vez determinada las variables se procedió a realizar una preparación al conjunto de datos. Se encontraron los siguientes valores estadísticos que se presentan en las Tablas 23, 24 y 25.

Variables Cuantitativas

Tabla 23. Medidas de centralización y valores extremos de la base de datos de la UIS.

| Medidas Estadísticas / variables | Medidas de centralización | | Valores extremos | |
|---|---|---|---|---|
| | Media | Moda | Mínimo | Máximo |
| Prom_acumulado | 3,41 | 3,49 | 0,51 | 4,76 |
| Puntos_icfes | 331,09 | 345 | 225 | 412 |
| Razcreditos | 0,79 | 1 | 0 | 1 |
| Edad | 26,28 | 23,76 | 15,99 | 44,27 |
| Estancia | 4,29 | 0 | 0 | 17,01 |

Tabla 24. Medidas de dispersión y distribución de la base de datos de la UIS.

| Medidas Estadísticas / Variables | Medidas de dispersión | | Medidas de distribución | |
|---|---|---|---|---|
| | Varianza | Desviación | Oblicuidad | Curtosis |
| Prom_acumulado | 0,28 | 0,53 | -1,82 | 8,56 |
| Puntos_icfes | 651,61 | 25,53 | -0,7 | 3,48 |
| Razcreditos | 0,05 | 0,22 | -1,59 | 5,31 |
| Edad | 27,94 | 5,29 | 0,31 | 2,26 |
| Estancia | 10,44 | 3,23 | 0,3 | 2,13 |

Variables Cualitativas

Tabla 25. Categoría más frecuente y categoría menos frecuente de la base de datos de la UIS.

| Medidas Estadísticas / variables | Categoría más frecuente | Categoría menos frecuente |
|---|---|---|
| Cod_Escuela | 6570<br>frecuencia de 1929 | 6520<br>frecuencia de 657 |
| Cod_carrera | 11<br>frecuencia de 1929 | 25<br>frecuencia de 25 |
| Sexo | M<br>frecuencia de 7048 | F<br>frecuencia de 2448 |
| Desertor | No desertor<br>frecuencia de 7472 | Desertor<br>frecuencia de 2024 |
| Retención | Otro<br>frecuencia de 6332 | No retenido<br>frecuencia de 192 |

#### 5.2.2.3 Procesamiento de los Datos

En el procesamiento de los datos se tomaron las variables Cod_carrera, Prom_acumulado, Puntos_icfes, Razcredito, Edad, Estancia para un estudio general de los datos.

Para estudios particulares se procesaron los siguientes grupos de variables:

- Cod_carrera, Prom_acumulado, Estancia, Retención, Desertor.
- Cod_carrera, Prom_acumulado, Estancia, Razcredito, Edad, sexo.

**Algoritmo Klass**

El algoritmo Klass permite determinar el número de grupos a formar o si se desea dejar que éste forme el número de grupos óptimo, también muestra los valores alfa

y beta calculados ya sea para el número de grupos determinados o el óptimo, los cuales se pueden cambiar para el posterior procesamiento.

Una vez procesados los datos el algoritmo suministra para cada grupo 3 tipos de información los cuales son: *Integrantes*, *Variables* y *Distancias*.

En *Integrantes* se presenta el número de registros y la desviación. En *variables* se muestran para las variables cualitativas las categorías que posee, la moda y el menos frecuente, y para variables cuantitativas la media, el rango, la moda y la varianza. En *distancias* se presenta las distancias intergrupo, centro de grupo y intragrupo.

Se procesaron un total de 9496 registros formando 4 grupos, a continuación se presentan los resultados para cada grupo de variables.

*Variables Cod_carrera, Prom_acumulado, Puntos_icfes, Razcreditos, Edad y Estancia:*

En la Tabla 26 se muestran las medias de las variables cuantitativas y las categorías de las variables cualitativas para cada grupo.

Tabla 26. Grupos obtenidos para el estudio general por el algoritmo Klass de la base de datos de la UIS

| Grupos / Variables | Grupo 0 | Grupo 1 | Grupo 2 | Grupo 3 |
|---|---|---|---|---|
| Cod_carrera | {21,25,22, 26,11,24} | {26,27} | {22,24} | {23,11} |
| Puntos_icfes | 330,08 | 336,39 | 328,88 | 331,49 |
| Pro_acumulado | 3,37 | 3,41 | 3,28 | 3,52 |
| Razcredito | 0,79 | 0,79 | 0,74 | 0,83 |
| Edad | 26,66 | 23,57 | 26,98 | 26,46 |
| Estancia | 4,51 | 2,89 | 4,47 | 4,52 |

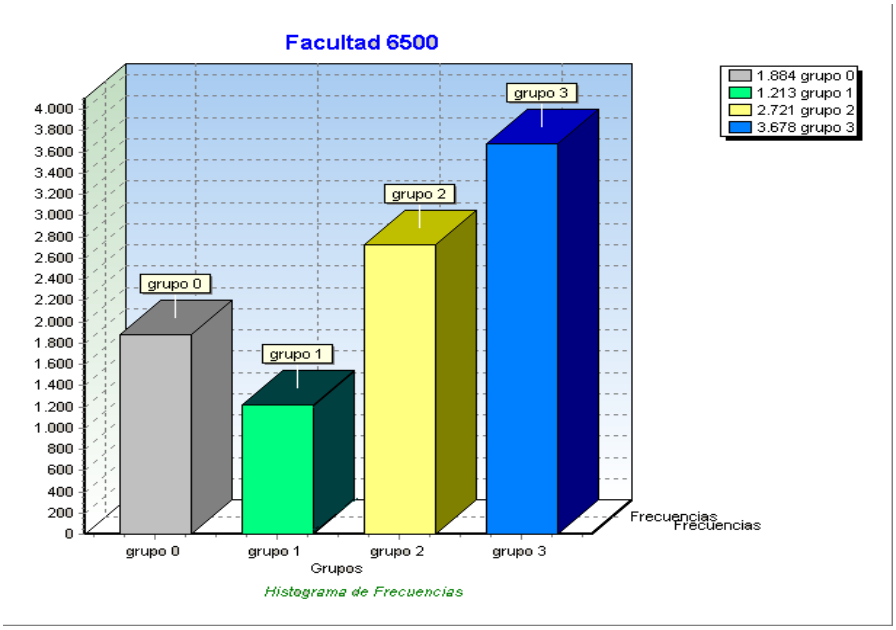
En la Tabla 27 se muestran las distancias que existen entre cada centro de un grupo a otro grupo.

Tabla 27. Distancias Intercluster de los Grupos obtenidos para el estudio general por el algoritmo Klass de la base de datos de la UIS

| Grupos | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| **0** | 0 | 3,999 | 3,896 | 3,948 |
| **1** | 3,999 | 0 | 3,795 | 3,822 |
| **2** | 3,896 | 3,795 | 0 | 3,76 |
| **3** | 3,948 | 3.822 | 3,76 | 0 |

En la Figura 26 se muestra el número de registros en cada grupo, el grupo 0 tiene 1884 registros, el grupo 1 tiene 1213 registros, el grupo 2 tiene 2721 registros y el grupo 3 tiene 3678 registros.

Figura 26. Número de Registros en cada grupo en el estudio general con el algoritmo klass para la base de datos de la UIS.



*Variables Cod_carrera, Prom_acumulado, Estancia, Retención, Desertor:*

En la Tabla 28 se muestran las medias de las variables cuantitativas y las categorías de las variables cualitativas para cada grupo.

Tabla 28. Grupos obtenidos para el estudio de la retención y deserción por el algoritmo Klass de la base de datos de la UIS.

| Grupos / Variables | Grupo 0 | Grupo 1 | Grupo 2 | Grupo 3 |
|---|---|---|---|---|
| Cod_carrera | {21} | {23,25,27, 11,22,21} | {22,11} | {24,26,27} |
| Pro_acumulado | 3,38 | 3,48 | 3,45 | 3,34 |
| Estancia | 4,45 | 4,23 | 5,05 | 3,49 |
| Retención | {No retenido, Retenido} | {No retenido, Retenido} | {No retenido, Retenido} | {No retenido, Retenido} |
| Desertor | {No desertor, Desertor} | {No desertor, Desertor} | {No desertor, Desertor} | {No desertor, Desertor} |

En la Tabla 29 se muestran las distancias que existen entre cada centro de un grupo a otro grupo.

Tabla 29. Distancias Intercluster de los grupos obtenidos para el estudio de la retención y deserción por el algoritmo Klass de la base de datos de la UIS.

| **Grupos** | **0** | **1** | **2** | **3** |
|---|---|---|---|---|
| **0** | 0 | 4,356 | 4,44 | 5,274 |
| **1** | 4,356 | 0 | 5,06 | 5,896 |
| **2** | 4,44 | 5,06 | 0 | 5,98 |
| **3** | 5,274 | 5,896 | 5,98 | 0 |

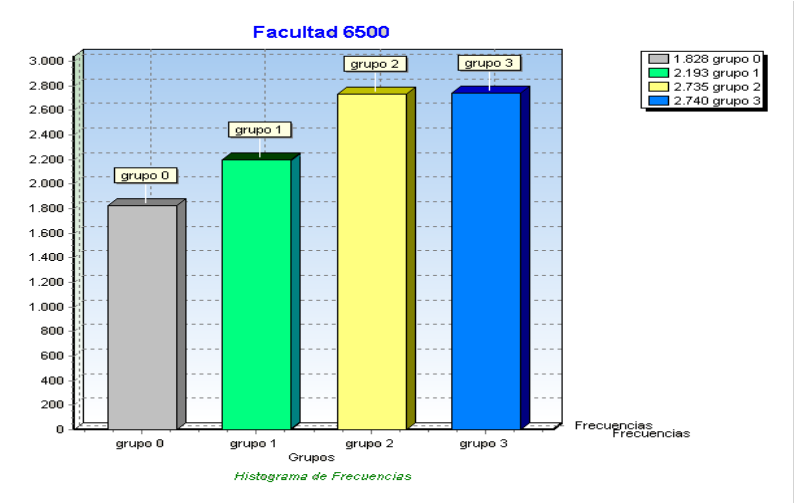
A continuación se presenta en la Tabla 30 el número de muestras (registros) presentes en las categorías de las variables cualitativas Retención y Desertor en los grupos obtenidos.

Tabla 30. Número de muestras en las variables Retención y Desertor por el algoritmo Klass de la base de datos de la UIS.

| Variables Cualitativas / Grupos | Retención | | | Desertor | |
|---|---|---|---|---|---|
| | Retenido | No retenido | Otros | Desertor | No desertor |
| Grupo 0 | 674 | 28 | 1126 | 357 | 1471 |
| Grupo 1 | 757 | 151 | 1285 | 561 | 1632 |
| Grupo 2 | 1124 | 11 | 1600 | 462 | 2273 |
| Grupo 3 | 417 | 2 | 2321 | 644 | 2096 |
| **Subtotal** | 2972 | 192 | 6332 | 2024 | 7472 |
| **Totales** | **9496** | | | **9496** | |

En la Figura 27 se muestra el número de registros en cada grupo, el grupo 0 tiene 1828 registros, el grupo 1 tiene 2193 registros, el grupo 2 tiene 2735 registros y el grupo 3 tiene 2740 registros.

Figura 27. Número de Registros en cada grupo en el estudio de retención y deserción con el algoritmo klass para la base de datos de la UIS.



*Variables Cod_carrera, Prom_acumulado, Estancia, Razcredito, Edad, Sexo:*

En la Tabla 31 se muestran las medias de las variables cuantitativas y las categorías de las variables cualitativas para cada grupo.

Tabla 31. Grupos obtenidos para el estudio del rendimiento por el algoritmo klass de la base de datos de la UIS.

| Grupos / Variables | Grupo 0 | Grupo 1 | Grupo 2 | Grupo 3 |
|---|---|---|---|---|
| Cod_carrera | {11} | {27,23} | {22,24} | {21,26,23, 11,27,25} |
| Pro_acumulado | 3,51 | 3,47 | 3,28 | 3,43 |
| Razcredito | 0,83 | 0,8 | 0,74 | 0,8 |
| Estancia | 4,89 | 3,73 | 4,47 | 4,09 |
| Edad | 26,87 | 25,39 | 26,98 | 25,81 |
| Sexo | {F,M} | {F,M} | {F,M} | {F,M} |

En la Tabla 32 se muestran las distancias que existen entre cada centro de un grupo a otro grupo.

Tabla 32. Distancias Intercluster de los Grupos obtenidos para el estudio del rendimiento por el algoritmo Klass de la base de datos de la UIS.

| **Grupos** | **0** | **1** | **2** | **3** |
|---|---|---|---|---|
| **0** | 0 | 3,74 | 3,942 | 4,197 |
| **1** | 3,74 | 0 | 4,638 | 4,895 |
| **2** | 3,942 | 4,638 | 0 | 5,1 |
| **3** | 4,197 | 4,895 | 5,1 | 0 |

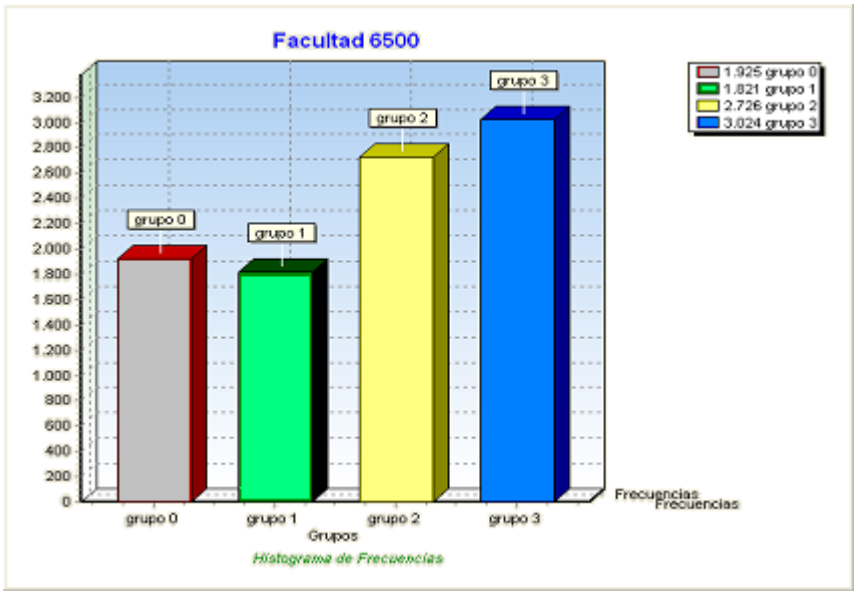
A continuación se presenta en la Tabla 33 el número de muestras presentes en las categorías de la variable Sexo en los grupos obtenidos.

Tabla 33. Número de muestras en la variable sexo por el algoritmo Klass de la base de datos de la UIS.

| Variables Cualitativas / Grupos | Sexo | |
|---|---|---|
| | Femenino F | Masculino M |
| Grupo 0 | 624 | 1301 |
| Grupo 1 | 1116 | 705 |
| Grupo 2 | 176 | 2550 |
| Grupo 3 | 532 | 2492 |
| **Subtotal** | 2448 | 7048 |
| **Totales** | **9496** | |

En la Figura 28 se muestra el número de registros en cada grupo, el grupo 0 tiene 1925 registros, el grupo 1 tiene 1821 registros, el grupo 2 tiene 2726 registros y el grupo 3 tiene 3024 registros.

Figura 28. Número de Registros en cada grupo en el estudio de rendimiento académico con el algoritmo klass para la base de datos de la UIS.



**Algoritmo Simbólico**

El algoritmo simbólico al igual que el algoritmo klass permite determinar el número de grupos a formar o sí se desea dejar que éste forme el número de grupos óptimo.

Una vez procesados los datos el algoritmo suministra para cada grupo 3 tipos de información los cuales son: *integrantes*, *variables* y *Distancias*.

En *Integrantes* se presenta el número de registros y la desviación. En *variables* se muestran para las variables cualitativas las categorías que posee, la moda y el menos frecuente, y para variables cuantitativas la media, el rango, la moda y la varianza. En *distancias* se presenta las distancias intergrupo, centro de grupo y intragrupo.

Se procesaron un total de 9496 registros formando 4 grupos, a continuación se presentan los resultados para cada grupo de variables.

*Variables Cod_carrera, Prom_acumulado, Puntos_icfes, Razcreditos, Edad y Estancia:*

En la Tabla 34 se muestran las medias de las variables cuantitativas y las categorías de las variables cualitativas para cada grupo.

Tabla 34. Grupos obtenidos para el estudio general por el algoritmo simbólico de la base de datos de la UIS.

| Grupos / Variables | Grupo 0 | Grupo 1 | Grupo 2 | Grupo 3 |
|---|---|---|---|---|
| Cod_carrera | {26,23,11,21, 24,22,27} | {23,22,24, 11,27,25} | {11,22,21,23, 24,26,27,25} | {11,21,23,24} |
| Puntos_icfes | 346,37 | 322,53 | 329,56 | 326,67 |
| Pro_acumulado | 3,69 | 3,63 | 2,96 | 3,57 |
| Razcredito | 0,91 | 0,9 | 0,59 | 0,88 |
| Edad | 21,88 | 29,89 | 25,14 | 28,14 |
| Estancia | 2,82 | 7,31 | 1,58 | 6,76 |

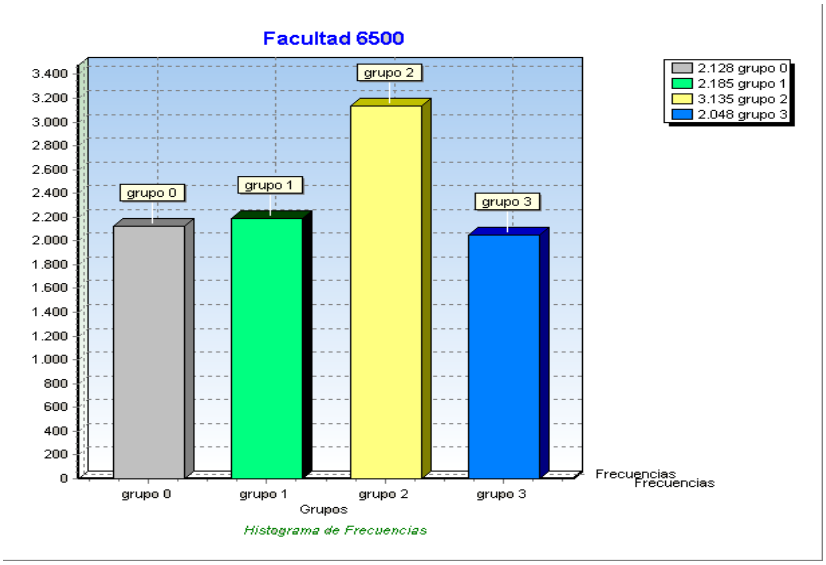
En la Tabla 35 se presentan las distancias que existen entre cada centro de un grupo a otro grupo.

Tabla 35. Distancias Intercluster de los Grupos obtenidos para el estudio general por el algoritmo simbólico de la base de datos de la UIS.

| Grupos | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| **0** | 0 | 6,869 | 7,384 | 6,452 |
| **1** | 6,869 | 0 | 8,052 | 6,213 |
| **2** | 7,384 | 8,052 | 0 | 7,858 |
| **3** | 6,452 | 6,213 | 7,858 | 0 |

En la Figura 29 se muestra el número de registros en cada grupo, el grupo 0 tiene 2128 registros, el grupo 1 tiene 2185 registros, el grupo 2 tiene 3135 registros y el grupo 3 tiene 2048 registros.

Figura 29. Número de Registros en cada grupo en el estudio general con el algoritmo simbólico para la base de datos de la UIS.



*Variables Cod_carrera, Prom_acumulado, Estancia, Retención, Desertor:*

En la Tabla 36 se muestran las medias de las variables cuantitativas y las categorías de las variables cualitativas para cada grupo.

Tabla 36. Grupos obtenidos para el estudio de la retención y deserción por el algoritmo simbólico de la base de datos de la UIS.

| Grupos / Variables | Grupo 0 | Grupo 1 | Grupo 2 | Grupo 3 |
|---|---|---|---|---|
| Cod_carrera | {27,24,21,11, 23,25,22} | {23,11,22,27, 26,25,21,24} | {11,21,27,24, 23,22,25,26} | {11,22,21,26, 24,23,27} |
| Pro_acumulado | 2,84 | 3,2 | 3,64 | 3,51 |
| Estancia | 1,69 | 3,12 | 7,75 | 2,64 |
| Retención | {Otro} | {No retenido, Otro} | {No retenido, Retenido,Otro} | {Otro} |
| Desertor | {Desertor} | {No desertor, Desertor} | {No desertor} | {No desertor} |

En la Tabla 37 se muestran las distancias que existen entre cada centro de un grupo a otro grupo.

Tabla 37. Distancias Intercluster de los grupos obtenidos para el estudio de la retención y deserción por el algoritmo simbólico de la base de datos de la UIS.

| **Grupos** | **0** | **1** | **2** | **3** |
|---|---|---|---|---|
| **0** | 0 | 6,671 | 10,756 | 7,917 |
| **1** | 6,671 | 0 | 9,158 | 8,27 |
| **2** | 10,756 | 9,158 | 0 | 10,914 |
| **3** | 7,917 | 8,27 | 10,914 | 0 |

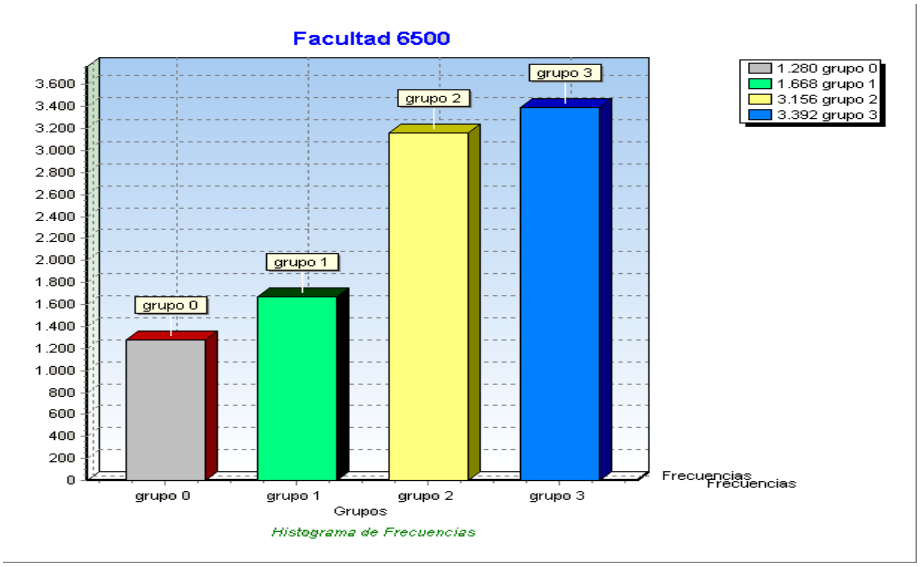
A continuación se presenta en la Tabla 38 el número de muestras o registros presentes en las categorías de las variables cualitativas Retención y Desertor en los grupos obtenidos.

Tabla 38. Número de muestras en las variables Retención y Desertor por el algoritmo simbólico de la base de datos de la UIS.

| Variables Cualitativas / Grupos | Retención | | | Desertor | |
|---|---|---|---|---|---|
| | Retenido | No retenido | Otro | Desertor | No desertor |
| Grupo 0 | 0 | 0 | 1280 | 1280 | 0 |
| Grupo 1 | 0 | 15 | 1653 | 744 | 924 |
| Grupo 2 | 2972 | 177 | 7 | 0 | 3156 |
| Grupo 3 | 0 | 0 | 3392 | 0 | 3392 |
| **Subtotal** | **2972** | **192** | **6332** | **2024** | **7472** |
| **Totales** | **9496** | | | **9496** | |

En la Figura 30 se muestra el número de registros en cada grupo, el grupo 0 tiene 1280 registros, el grupo 1 tiene 1668 registros, el grupo 2 tiene 3156 registros y el grupo 3 tiene 3392 registros.

Figura 30. Número de Registros en cada grupo en el estudio de retención y deserción con el algoritmo Simbólico para la base de datos de la UIS.



*Variables Cod_carrera, Prom_acumulado, Estancia, Razcredito, Edad, Sexo:*

En la Tabla 39 se muestran las medias de las variables cuantitativas y las categorías de las variables cualitativas para cada grupo.

Tabla 39. Grupos obtenidos para el estudio del rendimiento por el algoritmo simbólico de la base de datos de la UIS.

| Grupos / Variables | Grupo 0 | Grupo 1 | Grupo 2 | Grupo 3 |
|---|---|---|---|---|
| Cod_carrera | {11,27,24,22 21,23,25,26} | {23,11,27,21 22,24,25} | {24,21,11,27, 23,26,22} | {22,21,11,23, 26,27,24} |
| Pro_acumulado | 2,8 | 3,63 | 3,38 | 3,65 |
| Razcredito | 0,53 | 0,9 | 0,77 | 0,9 |
| Estancia | 1,16 | 7,44 | 2,6 | 5,29 |
| Edad | 25,17 | 30,16 | 24 | 25,91 |
| Sexo | {M,F} | {M,F} | {M,F} | {M,F} |

En la Tabla 40 se muestran las distancias que existen entre cada centro de un grupo a otro grupo.

Tabla 40. Distancias Intercluster de los grupos obtenidos para el estudio del rendimiento por el algoritmo simbólico de la base de datos de la UIS.

| **Grupos** | **0** | **1** | **2** | **3** |
|---|---|---|---|---|
| **0** | 0 | 7,164 | 6,028 | 7,012 |
| **1** | 7,164 | 0 | 7,36 | 6,732 |
| **2** | 6,028 | 7,36 | 0 | 6,999 |
| **3** | 7,012 | 6,732 | 6,999 | 0 |

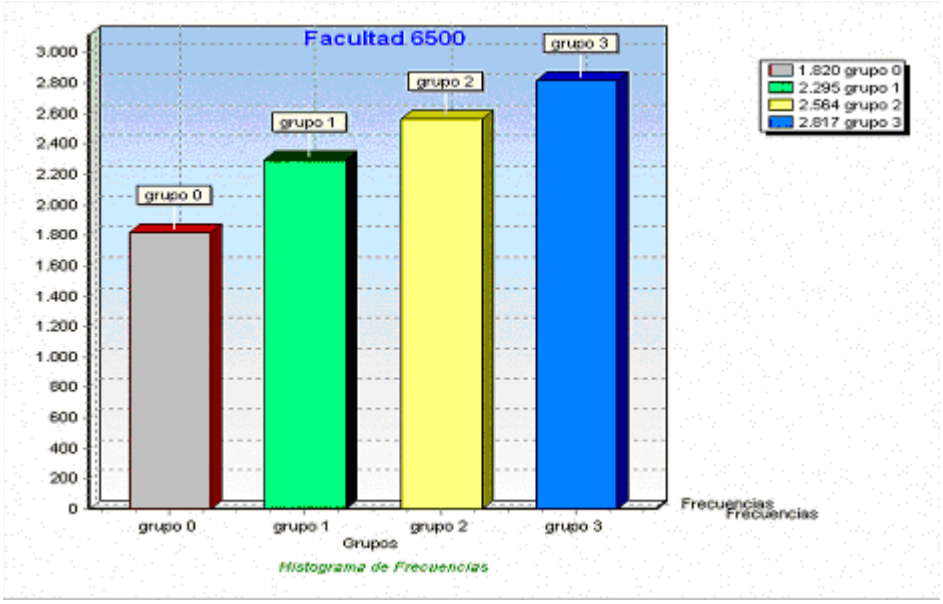
A continuación se presenta en la Tabla 41 el número de muestras presentes en las categorías de la variable Sexo en los grupos obtenidos.

Tabla 41. Número de muestras en la variable sexo por el algoritmo simbólico de la base de datos de la UIS.

| Variables Cualitativas / Grupos | Sexo | |
|---|---|---|
| | Femenino F | Masculino M |
| Grupo 0 | 281 | 1539 |
| Grupo 1 | 363 | 1932 |
| Grupo 2 | 580 | 1984 |
| Grupo 3 | 1224 | 1593 |
| **Subtotal** | **2448** | **7048** |
| **Totales** | **9496** | |

En la Figura 31 se muestra el número de registros en cada grupo, el grupo 0 tiene 1820 registros, el grupo 1 tiene 2295 registros, el grupo 2 tiene 2564 registros y el grupo 3 tiene 2817 registros.

Figura 31. Número de Registros en cada grupo en el estudio de rendimiento académico con el algoritmo simbólico para la base de datos de la UIS.



#### 5.2.2.4 Interpretación de los resultados

A continuación se presenta la interpretación de los resultados obtenidos en los diferentes grupos de variables.

**Variables Cod_carrera, Prom_acumulado, Puntos_icfes, Razcredito, Edad, Estancia.**

Para este grupo de variables se ha podido observar que para los estudiantes de la Facultad de Físico Mecánicas de la Universidad Industrial de Santander, el puntaje obtenido en las pruebas del Icfes está relacionado con el buen desempeño académico y que las carreras con mejor rendimiento académico son Ingeniería de Sistemas e Ingeniería Industrial con un promedio de edad de 26,46 años y una estancia promedio de 4,52 años. Es decir, a mayor puntaje obtenido en las pruebas del Icfes, la expectativa de lograr un mejor promedio académico es mayor y viceversa.

**Variables Cod_carrera, Prom_acumulado, Estancia, Retención, Desertor.**

En este grupo de variables pudo observarse que todos los cuatro conglomerados obtenidos con el algoritmo klass tienen un desempeño académico normal, los grupos 0, 1 y 2 tienen una estancia larga y el grupo 3 una estancia media.

El número total de estudiantes retenidos es de 2972 y no retenidos es de 192, el número de estudiantes desertados es de 2024 y no desertados es de 7472 distribuidos en los grupos como se muestra en la Tabla 42.

También se pudo observar que los estudiantes que tienen un promedio acumulado alto presentan una estancia menor en la universidad.

"Los resultados de la clasificación obtenida en este estudio, trabajando en el dominio mixto que integra variables cuantitativas y cualitativas son consistentes con los resultados obtenidos en [13]. En [13] se realizó un clustering numérico y

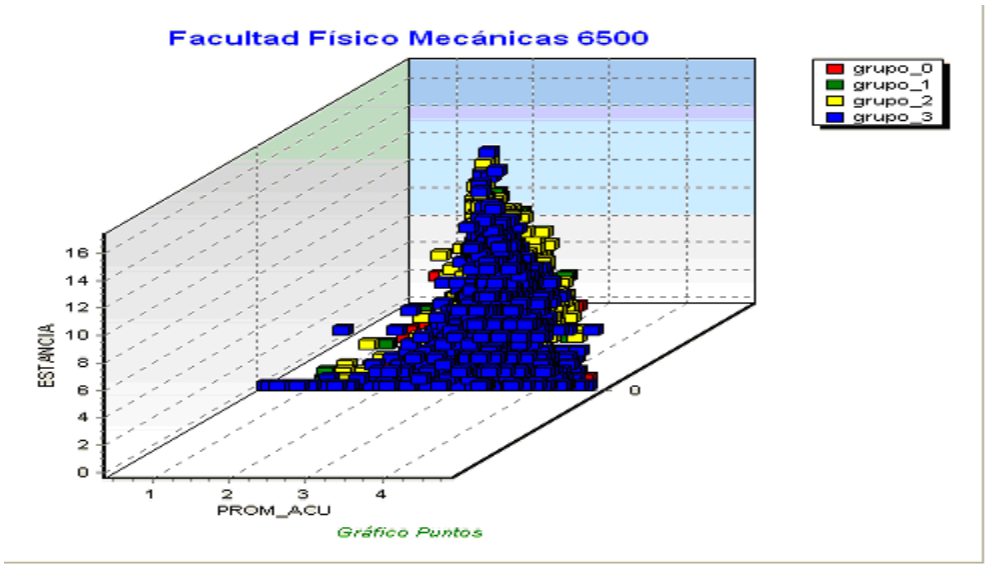
después un proceso de árboles de decisión sobre los conglomerados encontrados involucrando variables mixtas. Realizando el clustering directamente sobre el dominio mixto, se tiene más conocimiento involucrado en la clasificación, por lo que los grupos obtenidos involucran un mayor significado. Lo anterior puede llegar a potenciar los hallazgos de conocimientos en procesos de minería de datos".

Tabla 42. Distribución de las variables de Retención y Desertor de la base de datos de la UIS.

| Categorias / Grupos | Retenido | No retenido | Desertor | No desertor |
|---|---|---|---|---|
| Grupo 0 | 674 | 28 | 357 | 1471 |
| Grupo 1 | 757 | 151 | 561 | 1632 |
| Grupo 2 | 1124 | 11 | 462 | 2273 |
| Grupo 3 | 417 | 2 | 644 | 2096 |

En la Figura 32 se presenta la relación existente entre el promedio acumulado y la estancia de los estudiantes de la Facultad Físico Mecánicas.

Figura 32. Promedio acumulado Versus Estancia.



**Variables Cod_carrera, Prom_acumulado, Estancia, Razcredito, Edad, sexo.**

En los grupos obtenidos con estas variables se observó que todos los grupos están representados por jóvenes maduros con un desempeño académico normal, con una estancia larga en la universidad excepto el grupo 1 que presenta una estancia media, además se observó que en los grupos 0, 1 y 3 los estudiantes repiten ocasionalmente y en el grupo 2 los estudiantes repiten habitualmente con un promedio acumulado bajo.

También se observó que los estudiantes de mayor edad presenta un promedio acumulado bajo, y que la repetición de asignaturas prolonga la permanencia en la universidad.

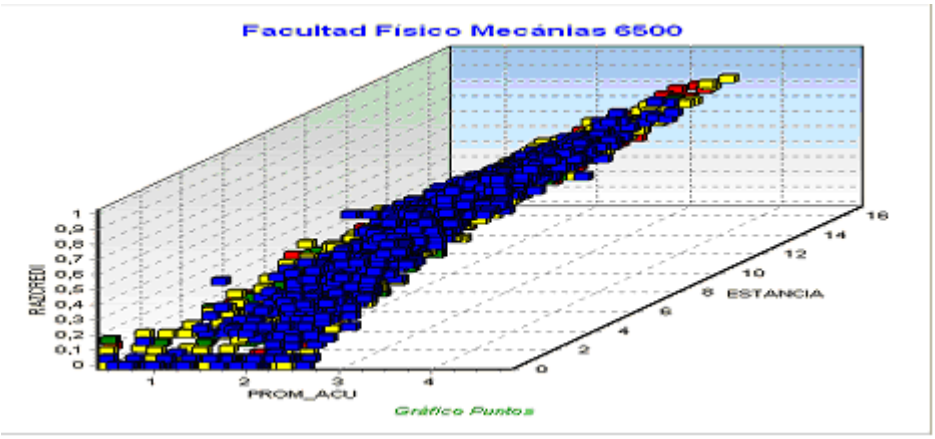
En la Tabla 43 se muestra el número de la población femenina y masculina que hay en cada grupo, donde el sexo masculino presenta la frecuencia más alta cuyo valor es de 7048 y el del sexo femenino es de 2448.

Tabla 43. Distribución de la variable sexo de la base de datos de la UIS.

| Categorias / Grupos | Femenino F | Masculino M |
|---|---|---|
| Grupo 0 | 624 | 1301 |
| Grupo 1 | 1116 | 705 |
| Grupo 2 | 176 | 2550 |
| Grupo 3 | 532 | 2492 |

En la Figura 33 se presenta la relación existente entre el promedio acumulado razcredito y la estancia de los estudiantes de la Facultad Físico Mecánicas.

Figura 33. Promedio acumulado Versus Estancia Versus Razcredito.




### 5.2.3 Datos de la Red Sismológica de Santander

#### 5.2.3.1 Análisis de los datos

Estos datos fueron recolectados a partir del registro de los movimientos sísmicos registrados en el área geológica de Santander. Los datos fueron recopilados a partir del año de 1958 hasta 1998 los cuales contienen información básica sismológica.

El objetivo de trabajar con estos datos es determinar cuales son las características de actividad sismológica de esta región.

Las variables que presenta este archivo de datos son las siguientes:
Día, Mes, Año, Hora, Longitud, Latitud, Profundidad, Magnitud, Intensidad, Fuente y Ubicación.

#### 5.2.3.2 Preparación de los datos

**Selección de las variables de estudio**

Las variables que son representativas para el estudio de las características de actividad sismológica de Santander se muestran en la Tabla 44:

Tabla 44. Variables seleccionadas para Datos de la Red Sismológica de Santander.

| Nombre Variable | Descripción | Tipo | Valores |
|---|---|---|---|
| Año | Año del evento sismológico. | Fecha | Entero[1958-1998] |
| Latitud | Posición georeferenciada de donde ocurrió el epicentro del sismo. | cuantitativa | Grados |
| Longitud | Posición georeferenciada de donde ocurrió el epicentro del sismo. | cuantitativa | Grados |
| Profundidad | Distancia de la superficie al punto donde ocurrió el sismo. | cuantitativa | Kilómetros. |
| Magnitud | Cantidad de energía liberada en el punto del subsuelo donde ocurre el sismo. | cuantitativa | escala de Richter de 1 a 10 |

Las variables eliminadas fueron Fuente, Ubicación e Intensidad, ya que solamente una pequeña porción de los datos posee estos campos.

Una vez determinada las variables se procedió a realizar una preparación al conjunto de datos. Se encontraron los siguientes valores estadísticos que se presentan en las Tablas 45 y 46.

Tabla 45. Medidas de centralización y valores extremos para datos de la Red Sismológica de Santander.

| Medidas Estadísticas / variables | Medida de centralización | | Valores extremos | |
|---|---|---|---|---|
| | Media | Moda | Mínimo | Máximo |
| Año | 1988,7 | 1995 | 1958 | 1988 |
| Latitud | 6,7 | 6,8 | 0 | 8 |
| Longitud | -74,16 | -73,08 | -7295 | -7,03 |
| Profundidad | 123,03 | 150 | 0 | 765,1 |
| Magnitud | 3,4 | 3 | 0 | 6,5 |

Tabla 46. Medidas de dispersión y distribución para datos de la Red Sismológica de Santander.

| Medidas Estadísticas / variables | Medidas de dispersión | | Medidas de distribución | |
|---|---|---|---|---|
| | Varianza | Desviación | Oblicuidad | Curtosis |
| Año | 119,52 | 10,93 | -1,26 | 3,14 |
| Latitud | 0,16 | 0,4 | -1,55 | 14,6 |
| Longitud | 6674,31 | 81,7 | -88,36 | 7810,24 |
| Profundidad | 2942,9 | 54,25 | -0,72 | 7,71 |
| Magnitud | 0,53 | 0,73 | 0,17 | 4,06 |

#### 5.2.3.3 Procesamiento de los datos

En el procesamiento de los datos se tomaron las variables Longitud, Latitud, Profundidad y Magnitud.

**Algoritmo Klass**

Se procesaron un total de 7780 registros formando 6 grupos, a continuación se presentan los resultados obtenidos con el algoritmo de agrupamiento con métricas ponderadas Klass.

En la Tabla 47 se muestran las medias de las variables cuantitativas para cada grupo y en la Tabla 48 las distancias que existen entre cada centro de un grupo a otro grupo.

Tabla 47. Grupos formados para Datos de la Red Sismológica de Santander.

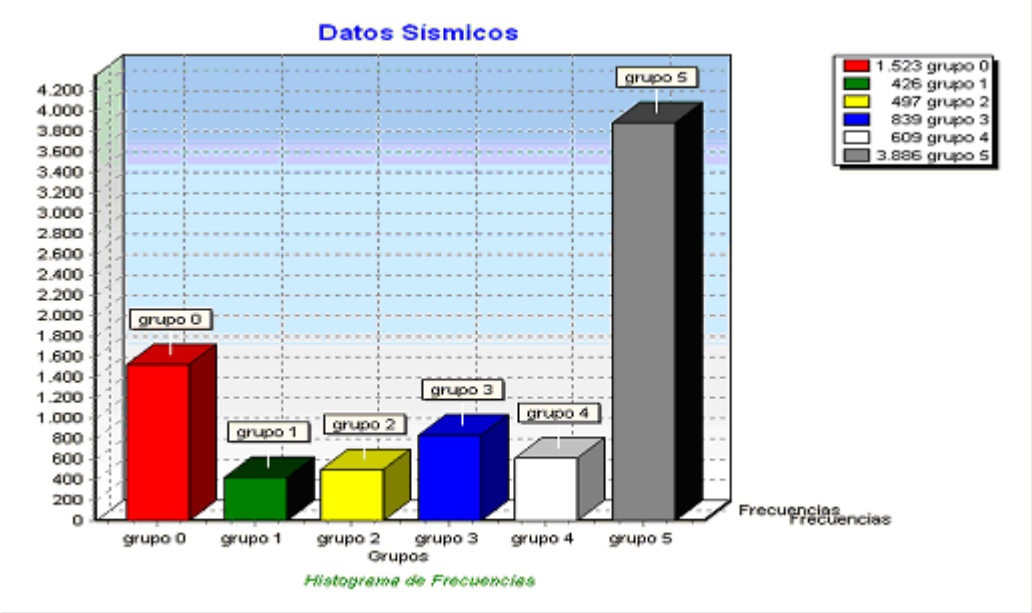
| Grupos / Variables | Grupo 0 | Grupo 1 | Grupo 2 | Grupo 3 | Grupo 4 | Grupo 5 |
|---|---|---|---|---|---|---|
| Latitud | 6,82 | 6,8 | 6,84 | 6,02 | 6,55 | 6,8 |
| Longitud | -73,06 | -73,13 | -73,31 | -73,86 | 73,87 | -73,1 |
| Profundidad | 156,2 | 155,13 | 54,17 | 45,05 | 14,86 | 148,9 |
| Magnitud | 4,03 | 4,84 | 3,89 | 2,76 | 2,64 | 3,2 |

Tabla 48. Distancias Intercluster de los Grupos obtenidos para Datos de la Red Sismológica de Santander.

| **Grupos** | **0** | **1** | **2** | **3** | **4** | **5** |
|---|---|---|---|---|---|---|
| **0** | 0 | 4,55 | 3,87 | 18,68 | 13,06 | 1,14 |
| **1** | 4,55 | 0 | 3,2 | 13,26 | 9,59 | 5,79 |
| **2** | 3,87 | 3,2 | 0 | 11,87 | 4,73 | 4,1 |
| **3** | 18,68 | 13,26 | 11,87 | 0 | 3,04 | 13,67 |
| **4** | 13,06 | 9,59 | 4,73 | 3,04 | 0 | 9,43 |
| **5** | 1,14 | 5,79 | 4,1 | 13,67 | 9,43 | 0 |

En la Figura 34 se muestra el número de registros en cada grupo, el grupo 0 tiene 1523 registros, el grupo 1 tiene 426 registros, el grupo 2 tiene 497 registros, el grupo 3 tiene 839, el grupo 4 tiene 609  y el grupo 5 tiene 3886 registros.

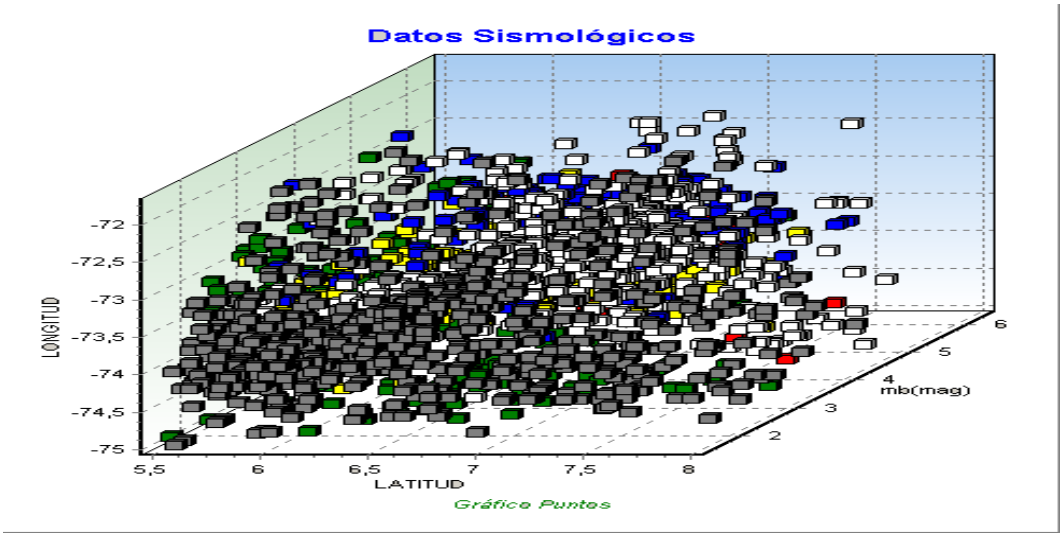
Figura 34. Número de Registros en cada grupo de la Red Sismológica de Santander.




### 5.2.3.4 Interpretación de los resultados

Para una interpretación de los resultados se graficaron las variables LATITUD, LONGITUD, MAGNITUD obteniéndose la siguiente gráfica (ver Figura 35).

Figura 35. Longitud Versus Latitud Versus Magnitud.




Donde el experto observó lo siguiente:

- A lo largo de la falla del Suárez se presenta la mayor activad sísmica.
- Se encontró la actividad en la falla de salinas y aparece un grupo correspondiente a un grupo de fallas superficiales en Yondo (Antioquia), la cual no se Había registrado antes.

## 5.3 EVALUACIÓN DE LA HERRAMIENTA CON USUARIOS

En Ingeniería de Software se definen una serie de *pruebas de aceptación* para permitir que el cliente o usuario valide todos los requisitos, estas pruebas llamadas alfa y beta son realizadas por los usuarios finales principalmente para comprobar la correcta funcionalidad del software, revisar el sistema de ayudas, encontrar errores de lógica en los programas y recoger sugerencias de mejoras [12].

### 5.3.1 Pruebas Alfa

Para la realización de las pruebas alfa con usuarios se desarrolló una guía dividida en 3 secciones, la primera sección presenta una introducción a la herramienta, la segunda sección las actividades a realizar por parte del usuario y la última una encuesta para evaluar la herramienta.

Las pruebas alfa fueron realizadas por 20 estudiantes y 2 profesores (Addisson Salazar Afanador y Fernando Ruiz) de ingeniería de sistemas para lo cual se realizó una pequeña exposición sobre el funcionamiento de la herramienta y se entregó a cada usuario 5 (cinco) archivos de datos con características diferentes en medio magnético y la encuesta.

#### 5.3.1.1 Formulario de la encuesta

El tema de la evaluación de software mediante consulta a usuarios está siendo ampliamente estudiado. Entre otros trabajos consultados para elaborar la presente encuesta se encuentran [14] [15].

A continuación se presenta el formulario realizado para las pruebas alfa.

Formato Pruebas alfa para Herramienta software

ADAMIX 1.0

Nombre del usuario: _____ Clase de usuario: _____

Sitio de la prueba: _____ Fecha: _____

Hora Inicio: _____ Hora Fin: _____

(N) Usuario Normal
(A) Analista de datos
(E) Experto área aplicada

Introducción

ADAMIX 1.0 (Agrupamientos con DAtos MIXtos) es una herramienta software para la clasificación de datos en dominios mixtos* desde diferentes enfoques. La clasificación es automática mediante la técnica de Minería de Datos llamada clustering, ofreciendo para ello tres algoritmos denominados: "*Agrupamiento Conceptual Conjuntivo basado en CLUSTER/2*", "*Agrupamiento simbólico Basado en una Nueva Medida de Similaridad*", y "*Agrupamiento basado en métricas mixtas ponderadas KLASS*".

El núcleo de ADAMIX 1.0 está escrito en *Object Pascal*. Las interfaces y el entorno visual fueron desarrollados con la herramienta de programación *Borland Delphi*, el sistema de ayuda se diseñó con la herramienta RoboHELP Office Tools, este sistema de ayuda suministra información sobre el manejo de la herramienta y la documentación teórica de conceptos básicos utilizados en la clasificación de los datos y temas relacionados a la minería de datos.

A continuación se presenta un bosquejo de la interfaz de la herramienta que permite localizar fácilmente sus componentes visuales, para facilitar las actividades a realizar.


* con variables cualitativas y cuantitativas combinadas.

Actividades a realizar

1. Abrir un archivo de datos de cualquiera de los tipos permitidos: archivos propios del programa, archivos de texto y archivos de Excel básicamente. Recomendación: Junto con la aplicación se entregan diferentes tipos de datos de ejemplo que pueden ser utilizados.
2. Realizar la preparación básica de los datos, una vez aparezcan en la cuadrícula del panel de visualización y hayan sido hechas todas las modificaciones deseadas, dando clic en la opción *Básica* en la sección *Preparación* del panel de opciones, donde puede seleccionar el número de registros con los cuales desea trabajar y las variables que desea incluir en el proceso de agrupamiento.
3. Utilizar alguno de los algoritmos de agrupamiento, desplegados en la sección *Agrupamiento* del panel de opciones, que se activa una vez sean preparados los datos, haciendo click sobre el algoritmo escogido para trabajar.
4. Escoger el número de clases deseada, o de lo contrario el algoritmo calculará un 'número óptimo' de clases. Dar clic en el botón *Agrupar* de la página correspondiente al algoritmo escogido.
5. Una vez realizado el agrupamiento aparecerá un indicador de listo en la barra de estado del panel de visualización de datos, que le indicará que ya puede continuar, y estará disponible el botón *Listo*, que le permitirá aceptar el agrupamiento realizado y crear datos procesados los cuales pueden ser graficados. Haga clic sobre el botón Listo.
6. Ahora puede visualizar los datos procesados y cambiar los nombres de los grupos obtenidos si lo desea, haciendo clic sobre ellos. En el panel de opciones se le presentan las opciones de *Gráficos* que ofrece el sistema para las variables y grupos de datos. Escoja una de ellas haciendo clic sobre la opción respectiva.
7. .Escoja las variables o grupos que desee graficar, y a continuación de clic en el botón Graficar. El sistema grafica el diagrama seleccionado, y si lo desea puede seleccionar los grupos o variables que no desee ver en la gráfica. Si desea conservar la gráfica obtenida, puede pulsar el botón *Listo* que aparece en el panel del gráfico.
8. Haga clic sobre el botón listo y observe como la gráfica es añadida a la lista de gráficas. Ahora ha completado exitosamente un proceso de agrupamiento y puede explorar cualquiera de los objetos del proyecto a través del navegador del proyecto que se encuentra en la parte izquierda-arriba del Panel de herramientas.
9. Navegar a través del proyecto, añadiendo datos, preparándolos, agrupándolos, y graficándolos de tal manera que explore todas las opciones que ofrece el sistema para cada tipo de datos del proyecto.
10. Realizar observaciones y recomendaciones para todo el proceso de ejecución del sistema.

A continuación se presenta una encuesta que tiene como finalidad medir la calidad del software.

| ✓ Facilidad de Manejo | Marque con una X su respuesta. |
|---|---|
| 1. La interfaz de usuario de la herramienta es..<br>a____Muy amigable b____Amigable<br>c____Normal d____Poco amigable<br>Sugerencias:_____<br>_____ | 2. ¿La herramienta es de fácil manejo?<br>a____Muy fácil b____Fácil<br>c____Normal d____Difícil<br>Sugerencias:_____<br>_____ |
| 3. ¿La herramienta permite un manejo bien estructurado de proyectos?<br>a____Sí b ____No<br>Porqué?:_____<br>_____<br><br>5. ¿Considera que la ayuda del software es clara y adecuada?<br>a____Sí b ____No<br>Sugerencias:_____<br>_____ | 4. ¿La herramienta permite la fácil navegabilidad entre los objetos del proyecto y sus funciones?<br>a____Sí b ____No<br>Porqué?: _____<br>_____<br><br>6. ¿Considera que la ayuda del software está completa ybien estructurada?.<br>a____Sí b ____No<br>Sugerencias:_____<br>_____ |
| ✓ Funcionalidad y Cubrimiento | Marque con una X su respuesta |
| 7. ¿Es clara la utilidad de este software?<br>a____Sí b ____No<br>Porqué?:_____<br>_____ | 8. ¿El software cumple con cada una de las funciones ofrecidas? a____Sí b____No<br>Porqué?: _____<br>_____ |

Marque con la letra que corresponda en cada casilla

(M) La Mayoría de las veces (A) Algunas veces (N) Nunca

9. ¿Con que frecuencia realizó las siguientes actividades cuando utilizó el software?

| | |
|---|---|
| a____Cargar datos de una archivo | b____Crear datos manualmente |
| c____Filtrar | d____Normalizar |
| e____Agrupar Conceptual | f____Agrupar Klass |
| g____Agrupar Simbólico | h____Graficar datos procesados |
| i____Agregar gráficos al proyecto | j.____Procesamiento por lotes |

k____Otras. ¿Cuáles?._____

| | |
|---|---|
| 10. ¿Se presentó algún problema con el software durante la realización de las actividades anteriores?.<br>a____No<br>b____Sí ¿con cuál (es)?_____<br>_____ | 11.¿Considera que la forma tabular y la gama de posibilidades de representación gráfica, ofrecidas por la herramienta, permiten una clara y completa interpretación de los resultados obtenidos?.<br>a____Sí<br>b____No Porqué?_____<br>_____ |
| 12. ¿El software cubre las operaciones necesarias para lograr los resultados esperados?<br>a____Sí<br>b____No ¿Porqué?_____<br>_____ | 13.Considera usted que el software cumple con el tópico de agrupamiento en dominios mixtos desde diferentes enfoques, en los algoritmos implementados?<br>a.____Sí<br>b.____No Porqué?:_____<br>_____ |
| ✓ Tiempos | Marque con una X su respuesta |
| 14. En términos generales y considerando el tipo de procesamiento. Considera usted el tiempo de respuesta de la aplicación:<br>a____Bueno b____Razonable<br>c____Excesivo | 15.¿ Son los tiempos de ejecución consistentes con el volumen de datos manejados?<br>a____Sí<br>b____No Porqué?_____<br>_____ |
| 16.Enumere los siguientes procesos de menor a mayor según los tiempos empleados:<br>a____Preparación de datos b____Agrupamiento conceptual<br>c____Agrupamiento simbólico c____Agrupamiento Klass | |
| ✓ Manejo de errores | Marque con una X su respuesta |
| 17.¿Anticipa el software la ocurrencia de errores, alertando al usuario?<br>a____Siempre b____Algunas veces<br>c____Nunca<br>Sugerencias:_____ | 18.¿Realiza el software validaciones al introducir datos erróneos o incompletos?<br>a____Siempre b____Algunas veces<br>c____Nunca<br>Sugerencias:_____ |
| 19.¿El software le permitió retroceder para corregir errores?<br>a____Sí b ____No<br>Sugerencias:_____ | 20.¿Se recupera el sistema fácilmente de errores?<br>a____Sí b ____No<br>Sugerencias:_____<br>_____ |

#### 5.3.1.2 Análisis de los resultados de la encuesta

Los resultados y sugerencias obtenidos en las pruebas alfa en cuanto a las actividades realizadas por los usuarios se resumen de la siguiente forma:

- La herramienta permitió que los usuarios realizaran la apertura de archivos de datos con variables mixtas correctamente así como importar archivos de datos en formato excel, csv (archivo de texto separados por coma), texto y ASCII.
- Los usuarios realizaron la preparación básica de los datos seleccionando las variables que deseaban incluir en el proceso de agrupamiento y el número de registros a agrupar sin ningún inconveniente.
- Los usuarios pudieron llenar campos vacíos, cuando la herramienta los detectó en los datos, al igual que realizar normalizaciones y aplicar filtros a los datos
- Los usuarios realizaron la agrupación de los datos a través de los algoritmos implementados y su representación gráfica.
- Los usuarios recomendaron cambiar el nombre de la sección Presentación de resultados por Análisis gráfico y el nombre de la sección Procesamiento por Agrupamiento.
- Los usuarios sugirieron que al posicionarse el ratón sobre los botones aparezcan leyendas que digan el título de la opción.
- Los usuarios sugirieron colocar las secciones del panel de opciones en el menú principal, además colocar una ventana en la parte inferior de la ventana que se presenta al ejecutar Adamix 1.0 que cubra toda la pantalla de manera que las aplicaciones activas queden ocultas y permita una mejor visualización del entorno de la aplicación.

A continuación se presentan los resultados obtenidos en las encuestas entregadas a cada usuario.

**Facilidad de manejo de la herramienta**

Tabla 49. Resultados en facilidad de manejo de la herramienta.

| Respuesta / No de la Pregunta | Muy amigable | Amigable | Normal | Poco amigable | Muy fácil | Fácil | Difícil | Si | No |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 8 | 9 | 3 | 0 | _ | _ | _ | _ | _ |
| 2 | _ | _ | 6 | _ | 5 | 9 | 0 | _ | _ |
| 3 | _ | _ | _ | _ | _ | _ | _ | 19 | 1 |
| 4 | _ | _ | _ | _ | _ | _ | _ | 20 | 0 |
| 5 | _ | _ | _ | _ | _ | _ | _ | 19 | 0 |
| 6 | _ | _ | _ | _ | _ | _ | _ | 20 | 0 |

**Funcionalidad y cubrimiento**

Tabla 50. Resultados en funcionalidad y cubrimiento.

| Respuestas / No de la pregunta | SI | NO |
|---|---|---|
| 7 | 17 | 3 |
| 8 | 20 | 0 |
| 10 | 0 | 20 |
| 11 | 20 | 0 |
| 12 | 19 | 1 |
| 13 | 20 | 0 |

**Tiempos**

Tabla 51. Resultados en Tiempos

| Respuestas / No de la pregunta | Bueno | Excesivo | Razonable | SI | NO |
|---|---|---|---|---|---|
| 14 | 15 | 0 | 5 | _ | _ |
| 15 | _ | _ | _ | 20 | 0 |

**Manejo de errores**

Tabla 52. Resultados en manejo de errores.

| Respuestas / No de la pregunta | Siempre | Algunas veces | Nunca | SI | NO |
|---|---|---|---|---|---|
| 17 | 20 | 0 | 0 | _ | _ |
| 18 | 20 | 0 | 0 | _ | _ |
| 19 | _ | _ | _ | 20 | 0 |
| 20 | _ | _ | _ | 20 | 0 |

### 5.3.2 Pruebas Beta

Para la realización de las pruebas Beta con estudiantes de ingeniería de sistemas (5 estudiantes) se entregó una versión del software, al igual que la encuesta utilizada en las pruebas Alfa y se les realizó una pequeña charla donde se dio una

explicación del funcionamiento del software, para que lo probaran en sus computadores.

#### 5.3.2.1 Resultados de las pruebas Beta

Los estudiantes de ingeniería de sistemas que realizaron las pruebas Betas lograron efectuar todas las actividades propuestas en la guía sin ningún inconveniente y presentaron las siguientes observaciones y conclusiones:

- Colocar en la barra de herramienta las opciones de importar datos de texto y datos de Excel, así como las opciones de archivo.

- Consideraron que la herramienta es muy amigable, de fácil navegabilidad y muy intuitiva para su comprensión.

- La disponibilidad de utilidades como eliminar y duplicar tipos de datos en el proyecto son muy importantes en la reutilización y abandono de los diferentes tipos de datos presentes en el proyecto.

- Las opciones de preparación de los datos son completas y adecuadas para una buen tratamiento de las variables dependiendo de su tipo.

# CONCLUSIONES

- La clasificación de los datos que soporta dominios mixtos, que son los más comunes en las bases de datos actuales, se hace más compleja que la clasificación tradicional cuantitativa, debido al problema de cómo obtener una medida de similaridad consistente para comparar dos objetos entre sí. Para esto hay que considera simultáneamente todas las variables que lo describen tanto de tipo cuantitativo como cualitativo; ya que en dominios mixtos se manipulan muchas características mezcladas importantes de los objetos, las cuales por separado no presentan una descripción completa. Debido a lo anterior, una herramienta software para la clasificación de datos con variables cualitativas y cuantitativas aporta al campo de la minería de datos un instrumento útil para el analista de datos, que le permite realizar una mejor interpretación de ellos, y así cumplir su objetivo de obtener una información útil de los mismos.

- La dificultad para obtener una medida de similaridad consistente en dominios mixtos, ha devenido en el surgimiento de diversos algoritmos propuestos en un marco carente de estandarización, en donde divergen las soluciones propuestas, algunas de ellas abordadas en este proyecto y descritas brevemente a continuación:

    - El algoritmo de agrupamiento conceptual conjuntivo, elude el problema de las medidas, centrando la clasificación, en la formación de grupos descritos mediante las formas normales conjuntivas, que mejor se ajusten a criterios especificados por el usuario, en este caso el analista de datos Estos criterios pueden variar desde la simplicidad en las descripciones, hasta el mejor ajuste de las mismas a los objetos cubiertos, pasando por la mayor comunalidad entre las representaciones de los elementos de un grupo hasta la mayor

separación entre la representación de los grupos. De tal manera que realiza operaciones excesivamente complejas que conducen a un alto consumo de recursos de computación (memoria, procesador). lo que conlleva a un bajo rendimiento y un campo de acción limitado.

Cabe anotar que este algoritmo, denominado CLUSTER/2 fue uno de los primeros publicados en este campo, teniendo el mérito de ser pionero en el área de investigación que posteriormente se denominó clustering conceptual; y de él han sido realizadas versiones mejoradas, entre las que se encuentra una designada CLUSTER/S, pero estas versiones, no están cubiertas en el marco de investigación de este proyecto.

- El algoritmo de agrupamiento simbólico aborda el problema de las medidas, definiendo una medida de similaridad, constituida por tres componentes: la posición, la extensión y el contenido, la cual aplica indistintamente sobre atributos tanto cuantitativos como cualitativos, con la salvedad que estos últimos no poseen la componente debida a la posición. Dichas componentes están normalizadas y permiten una ponderación intrínseca del peso de la parte cualitativa y cuantitativa de los objetos. Por otra parte el agrupamiento de los objetos se realiza mediante el operador unión cartesiano, lo cual genera un bajo consumo de recursos sacrificando para ello un poco la precisión por el rendimiento.

  A este algoritmo igualmente, se le abona su capacidad en el manejo de diferentes subtipos de variables tanto cualitativas como cuantitativas, encontrándose para las variables cuantitativas los subtipos: valores discretos absolutos, valores de radio continuo, e intervalos; y para las cualitativas: valores nominales, ordinales, y estructurados en forma de árbol. La capacidad de manejo de este último subtipo de atributos le confiere al algoritmo la propiedad de ser jerárquico.

- El algoritmo de agrupamiento Klass hace un manejo separado de las medidas cuantitativas y cualitativas, y posteriormente las combina mediante ponderaciones estimadas de acuerdo a la distribución de los datos. Para las variables cuantitativas utiliza distancias euclidianas sobre la varianza para obtener medidas normalizadas, y en cuanto a las variables cualitativas forma una matriz extendida sobre todas las categorías, para cada una de ellas, realizando una serie de operaciones sobre éstas que permiten igualmente obtener medidas normalizadas. Éstas medidas finalmente son combinadas de acuerdo a los pesos mencionados. El proceso sobre las variables cualitativas es muy costoso en recursos de computación, por tanto el rendimiento de este algoritmo está intrínsecamente relacionado con el número de variables cualitativas que posean los datos, y el número de valores distintos, ó categorías que a su vez posean cada una de éstas variables.

  De este algoritmo se destaca su precisión en la representación de los grupos, utilizando la media para los atributos numéricos, y en el caso de los atributos cualitativos: valores de pertenencia a cada categoría, que el algoritmo denomina cardinalidades, lo cual proporciona criterios más exactos en el momento de agrupar.

- Se ha desarrollado una herramienta software que permite realizar clasificación en dominios mixtos, de una forma amigable e intuitiva. Esta herramienta cuenta, entre otras, con las siguientes características:

  - Fácil navegabilidad a través de su entorno, y rápida accesibilidad a cada objeto y las funciones disponibles para éste, dentro del proyecto. Al igual que una ayuda completa tanto de la herramienta como del marco teórico en el cual está comprendida.

  - Manejabilidad para cargar datos con extensiones ASCII, Excel, y texto con diferentes tipos de separadores como tabulaciones,

espacios, puntos, comas; o delimitadores definidos por el usuario; además de la posibilidad tanto de crear nuevos conjuntos de datos en blanco para insertarlos manualmente, como de generarlos aleatoriamente, interactuando para esto con el usuario con el fin determinar el número de registros y variables a crear así como un nombre o alias para identificar al nuevo conjunto de datos.

- Diversidad en el preprocesamiento de los datos que permite realizar desde una preparación básica, donde se seleccionan las variables a procesar, y el tipo de cada una de ellas, hasta una preparación avanzada que permite eliminar vacíos, al igual que redondear, discretizar y normalizar las variables de tipo cuantitativo, y modificar los nombres de las categorías y el subtipo de las variables de tipo cualitativo, además de permitir la eliminación de registros mediante una amplia gama de condiciones de filtrado.

- La posibilidad de permitir al usuario escoger el algoritmo de agrupamiento, que desee utilizar para el procesamiento, teniendo en cuenta las características de sus datos, y de acuerdo a ellas las ventajas que le puede ofrecer cada uno de los algoritmos.

- La variedad de opciones para el análisis gráfico de los datos procesados, que permite al usuario realizar un análisis pormenorizado de las variables y los grupos. En cuanto a las variables se ofrecen opciones separadas para graficar las variables cualitativas y cuantitativas, de acuerdo a las características de las mismas. En lo referente a los grupos se ofrecen diversas opciones para graficar sus frecuencias, distancias entre ellos, distancias de cada elemento al centro de clase de su grupo, y distancias para cada elemento a todos los elementos de su grupo. En el caso de datos agrupado con el algoritmo de agrupamiento conceptual conjuntivo éstas distancias no están disponibles, puesto que este algoritmo no trabaja con este tipo de medidas.

- Disponibilidad de utilidades adicionales que permiten eliminar y duplicar todo tipo de datos dentro del proyecto; esto con el fin de que el usuario en caso de haber cometido un error o desee abandonar el procesamiento de unos datos en cualquier punto del proceso pueda eliminarlos del proyecto, o por el contrario si no desea perder unos datos pero requiere continuar trabajando sobre ellos pueda duplicarlos y continuar trabajando sobre la copia sin perder los datos originales.

- La metodología utilizada para el diseño y construcción de la herramienta se ha basado en las 4 fases del proceso unificado de desarrollo de software: fase de inicio, elaboración, construcción y transición, las cuales permitieron establecer la razón de ser del proyecto, determinar los requerimientos, hacer un análisis y diseño de alto nivel de los diferentes módulos que integran la herramienta y la construcción de un producto final que se desarrolló de forma iterativa e incremental, empezando por la construcción de prototipos, que fueron evolucionando hasta obtener la herramienta software, preparada para los usuarios. Dicha construcción se llevó a cabo con el lenguaje de programación Delphi, que es un lenguaje de programación orientado a objetos que permitió la creación de un motor eficiente y versátil, que cubrió la alta complejidad del proyecto.

- Mediante el proceso de desarrollo de la herramienta y su posterior utilización con datos de prueba, se realizó un análisis comparativo de los algoritmos implementados teniendo en cuenta criterios tales como el desempeño, la complejidad, y la precisión de los mismos. Los resultados obtenidos se expusieron en el capítulo cuatro de este documento y se describen brevemente a continuación:

    - En el criterio de desempeño, el algoritmo de agrupamiento simbólico basado en una nueva medida de similaridad tiene un tiempo de procesamiento y consumo de memoria menor comparados con los algoritmos de agrupamiento conceptual conjuntivo basado en

CLUSTER/2 y el algoritmo de agrupamiento basado en métricas mixtas ponderadas Klass. Además se observó que los algoritmos de agrupamiento tienen un tiempo de procesamiento y consumo de memoria altos con variables cualitativas y menor con variables cuantitativas y que a medida que aumenta el número de datos aumenta el consumo de memoria así como el tiempo en procesamiento.

  - En el criterio de complejidad, el algoritmo más complejo fue el algoritmo de agrupamiento conceptual conjuntivo basado en cluster/2, seguido por el algoritmo de agrupamiento basado en métricas mixtas ponderadas Klass y finalmente el algoritmo de agrupamiento simbólico basado en una nueva medida de similaridad.

  - En el criterio de precisión, el algoritmo más preciso fue el algoritmo de agrupamiento basado en métricas mixtas ponderadas Klass por su aproximación a los resultados comparados con el experto, seguido por el algoritmo de agrupamiento simbólico y por último el algoritmo de agrupamiento conceptual.

- El proceso de evaluación que permitió valorar la calidad y funcionalidad de la herramienta se llevó a cabo en dos fases, en la primera de éstas se realizó la aplicación de la herramienta tanto a datos de prueba como datos reales, y en la segunda se llevaron a cabo pruebas alfa y beta con diferentes tipos de usuarios. Los resultados obtenidos de éstas pruebas se encuentran en el capítulo cinco de este documento y se describen brevemente a continuación.

  - En la primera fase se valoró la calidad y la funcionalidad de la herramienta con datos de pruebas ampliamente probados tomados de la UCI Machine Learning Repository y datos reales de dos casos de estudio como lo son la Base de Datos Académica de la Universidad Industrial de Santander y la Base de Datos de la Red Sismológica de Santander, obteniéndose resultados coherentes a los

probados y a las apreciaciones de los expertos en los casos de estudio.

- En la segunda fase se valoró con usuarios normales, analistas de datos y expertos en áreas aplicadas. Los usuarios normales los cuales fueron estudiantes de ingeniería de sistemas se enfocaron en la amigabilidad, navegabilidad y facilidad de manejo considerándola buena en estos aspectos, los analistas de datos y expertos se centraron en los resultados obtenidos de los datos reales, catalogándola como una herramienta útil y necesaria para la minería de datos.

# RECOMENDACIONES

A continuación se presentan recomendaciones para futuros trabajos en esta línea de investigación:

- Para un buen desempeño de la herramienta y los algoritmos implementados en ella se recomienda que se ejecute en equipos que cuenten con una buena capacidad de recursos computacionales (procesador Pentium 4 o similar, memoria RAM de 128 MB como mínimo), ya que los algoritmos en dominios mixtos consumen mucha memoria y requieren alta velocidad de procesamiento, especialmente para grandes cantidades de datos.

- Continuar y profundizar en el estudio de agrupamiento en dominios mixtos e implementar otros algoritmos de agrupamientos que aborden este tipo de clasificación desde otras perspectivas con el fin de enriquecer el conocimiento adquirido en este campo, y también fortalecer la herramienta.

- También sería interesante explorar en la creación de nuevos tipos de gráficos para la visualización de grupos, especialmente, en lo que se refiere a la capacidad de hacer gráficas de atributos cuantitativos y cualitativos combinados.

# BIBLIOGRAFÍA

## TEORÍA


- Berry Michael J. A. , Linoff Gordon. Data Mining Techniques, for Marketing, Sales, and Customer Support – Wiley Computer Plublishing, 1997.

- Looney, Carl G. Pattern Recognition Using Neuronal Networks, Theory and Algorithms for Engineers and Scientists - New York Oxford, OXFORD UNIVERSITY PRESS. 1.997 New York.

- Michalski, R.S. Revealing conceptual structure in data by inductive inference. Machine Intelligence 10, eds. Hayes-Michic, D.Michine, and Y.H. Pao, Chichister: Ellis Horwood, New York: Halsted Press (John Wiley), (1981).

- Michalski, R.S; Mitchel, T. y Carbonel, J.(Eds). Machine Learning: An Artificial Intelligence Approach, TIOGA Pub. Co, Palo Alto,1983, cap 11 p 331-363.


## ARTÍCULOS


- Gibert, K. y Cortés, U. Una herramienta estadística para la creación de prototipos en dominios poco estructurados. Ed. Grupo Noriega Editores. México. Febrero,1992.

- Gibert, K y Cortés, U. Weighing quantitative and qualitative variables in clustering methods. Journal Mathware and Sof Computing, número especial, pags. 251-266. Mayo de (1997).

- Gowda, K.C. y Krishna, G. “Agglomerative clustering using the concept of mutual nearest neighborhood”. Pattern Recog; vol. 10, pp. 105-112, 1977.

- Gowda, K. C. y Diday, E. “Symbolic clustering using a new similarity measure”, IEEE Trans On Systems, man, and cib, 22(2). 1992. pp 368-378.

- Michalski, R.S.; Stepp, E. y Diday, E. A recent advanced in data analysis: clustering objects into classes characterized by conjunctive concepts. Progress in Pattern Recognition, Vol. 1, L. Kanal and A Rosenfeld, eds, (1981).

- Michalski, R y Stepp, R.E. “Automated construction of classifications: Conceptual clustering versus numerical taxonomy”, IEEE Trans. Pattern Anal Machine Intell; vol.PAMY-5, pp 396-410. 1983.


**TÉCNICA**


- Moral, Víctor. DELPHI 4.0. Prentice Hall Iberia S.A. 1999.

- Teixeira, Steve Y Pacheco, Xavier. Guía de desarrollo Delphi 5. Pearson educación. Madrid, 2000.


**METODOLOGÍA**


- Larman, UML y Patrones: Introducción al análisis y diseño orientado a objetos, Prentice Hall, México, 1999.

- Pressman, Roger. Ingeniería del software. Un enfoque Practico. Mc Graw Hill, tercera edición, 1993.

## REFERENCIAS BIBLIOGRÁFICAS


- [1] Fayyad, Usama. Advances in Knowledge Discovery and Data Mining. (1996) AAAI Press / The MIT Press. (Cap 1,2,5-7).

- [2] Mitchell. T. M. : "Learning and problem-solving". Proceedings de la 8ª International Joint Conference on Artificial Intelligence. Morgan Kaufmann. (1983).

- [3] Mooney, R.: "Induction over the unexplained : Using overly-general domains theories to aid concept learning". Machine Learning 10(1), (1993).

- [4] Michalski, R.S., Chilausky, R.L. "Learning by being told and Learning from examples: An experimental comparison of the two methods of knowledge acquisition in the context of developing an expert system for soybean disease diagnosus". International Journal of Policy Analysis and Information Systems, 4(2), (1980).

- [5] Diday, E. "The symbolic approach in clustering,"in Classification and Related Methods of Data Analysis , H. H.Bock, Ed Amsterdam: Elsevier Science (North Holland). (1988).

- [6] Ichino, M. "General metrics for mixed features – The Cartesian space theory for pattern recognition." In Proc. IEEE 1988 Int. Conf. Syst, Man. Cybern. (1988).

- [7] Gibert, K y Cortés, U. Weighing quantitative and qualitative variables in clustering methods. Journal Mathware and Sof Computing, número especial, pags. 251-266. Mayo de (1997).

- [8] Gibert, K. Klass: Estudi d'un sistema d'ajuda al tractament estadístic de grans bases de dates. Masters Thesis. UPC, (1991).

- [9] Teixeira, Steve Y Pacheco, Xavier. Guía de desarrollo Delphi 5. Pearson educación. Madrid. (2000).

- [10] Larman, UML y Patrones: Introducción al análisis y diseño orientado a objetos, Pearson.

- [11] http:// w.w.w.uci.edu/mlear/MLRepository.

- [12] Pressman, Roger. Ingeniería del software. Un enfoque Practico. Mc Graw Hill, tercera edición, (1993).

- [13] Salazar, Addisson. Descubrimiento de Conocimientos en la Base de Datos Académica de la Universidad Industrial de Santander. Tesis de maestría en informática. UIS, Bucaramanga, (2003).

- [14] Bonnie, E. Jhon. Using GOMS for User Interface Design and Evaluation: wich technique?. Composicion and contrast. ACM Transactions On Computer – Human Interaction, vol. 3, No. 4, (1996). Páginas 287-319.

- [15] Bonnie, E. Jhon. The GOMS Family of User Interface Analysis Techniques: Composicion and contrast. ACM Transactions On Computer – Human Interaction, vol. 3, No. 4, (1996). Páginas 320-351.

# ANEXO A
# MANUAL DEL USUARIO

## INTRODUCCIÓN

Para tener acceso al sistema de ayuda de ADAMIX 1.0 se puede ir al botón de ayuda en el menú principal de la herramienta. El sistema de ayuda sigue el estándar habitual de los archivos de ayuda de las aplicaciones Windows, ofreciendo dos posibilidades de búsqueda:

- Seleccionando la pestaña de **contenido** para una búsqueda jerárquica navegando por los diferentes libros de la ayuda.
- Seleccionando la pestaña de **buscar** para una búsqueda incremental en una lista con todas las palabras marcadas.

El sistema de ayuda de ADAMIX 1.0 suministra información sobre el manejo de la herramienta y la documentación teórica de conceptos básicos utilizados en la clasificación de los datos, temas relacionados a la minería de datos.

## 1 ¿QUÉ ES ADAMIX 1.0?

ADAMIX 1.0 es una herramienta Software de Minería de Datos para la clasificación de datos con variables cualitativas y cuantitativas, que utiliza varios algoritmos de agrupamiento, que abordan la clasificación de datos con variables mixtas desde diferentes enfoques como:

- El algoritmo de agrupamiento Klass, basado en métricas mixtas ponderadas.
- El algoritmo de agrupamiento simbólico, basado en una nueva medida de similaridad.
- El algoritmo de agrupamiento conceptual conjuntivo, basado en cluster/2.

El núcleo de ADAMIX 1.0 está escrito en *Object Pascal*, un autentico lenguaje orientado a objetos que tiene su origen en un potente compilador: *Turbo Pascal.* Las interfaces y el entorno visual fueron desarrollados con la herramienta de

programación *Borland Delphi 7*, una herramienta de desarrollo cuya velocidad no se restringe al desarrollo y diseño de los programas, sino que las aplicaciones finales son rápidas y robustas.

# 2 REQUERIMIENTO DEL SISTEMA

## 2.1 HARDWARE

**Requerimientos Mínimos:**

- Procesador de 1200 MHz
- Memoria RAM de 128 MB
- Espacio libre en disco 400 MB
- Monitor SVGA, resolución de video: 800*600
- Unidad de CD-ROM
- Mouse

## 2.2 SOFTWARE

- Sistema Operativo Windows 98 (o versión superior)

# 3 INSTALACIÓN DE ADAMIX 1.0

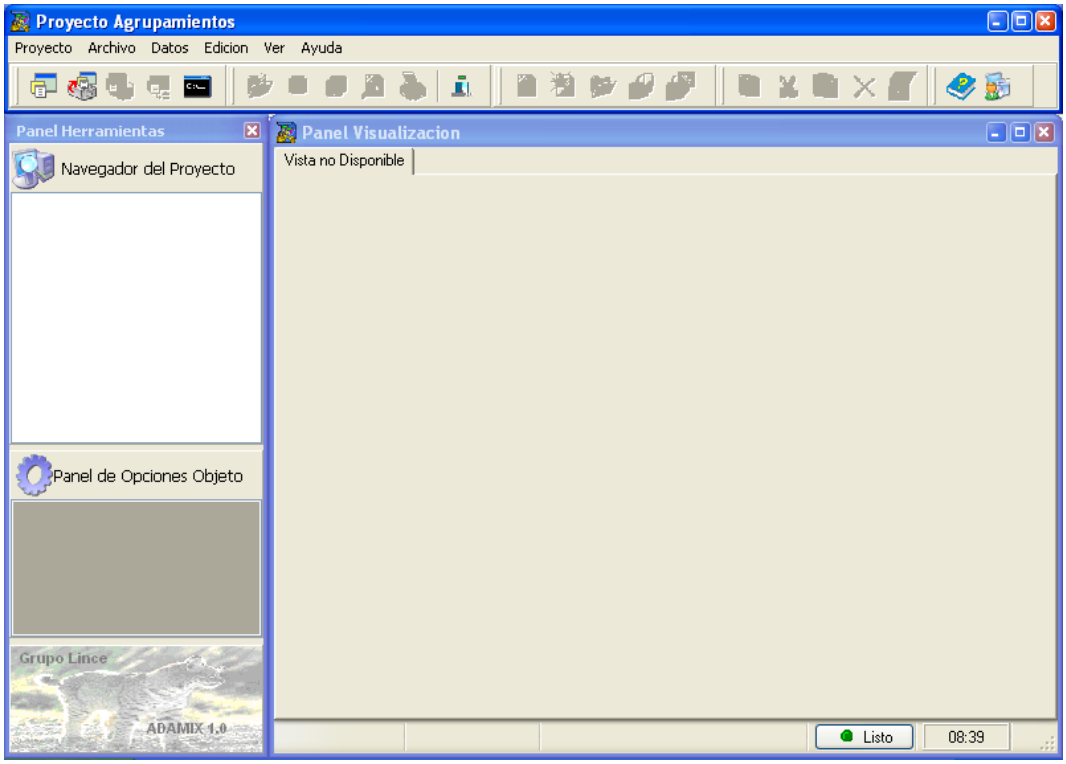
Para instalar ADAMIX 1.0 siga las siguientes instrucciones:

- Inserte el CD de instalación.
- Vaya a Inicio, Ejecutar. Con el botón "Examinar" busque el archivo llamado "Setup.exe" que se encuentra en el CD y haga click en el botón "Aceptar".
- Aparecerá la pantalla de instalación del software que lo guiará de forma fácil y rápida a través del proceso de Instalación.

# 4 INTRODUCCIÓN A LA HERRAMIENTA

Al ejecutar ADAMIX 1.0 aparecerá la siguiente ventana, la cual representa la ventana principal de la herramienta, formada por el ***Menú Principal***, la ***Barra de Herramienta***, el ***Panel de Herramientas*** y el ***Panel de Visualización***.

Figura A.1. Ventana Principal.




La herramienta ADAMIX 1.0 está constituida por 5 módulos los cuales son:

- Módulo de entrada de datos
- Módulo de preparación de datos
- Módulo de procesamiento de datos
- Módulo de presentación
- Módulo de ayuda del sistema

## 4.1 MÓDULO DE ENTRADA DE DATOS

Este módulo permite al usuario realizar la entrada de datos ya sean cargados o digitados manualmente, la creación de proyectos o abrir proyectos ya existentes, guardar, guardar como, guardar proyecto como, entre otras funciones.

## 4.2 MÓDULO DE PREPARACIÓN DE DATOS

Este módulo permite al usuario realizar una preparación básica de los datos una vez cargados o digitados como la selección de variables y registros con los cuales desea trabajar en el proceso de agrupamiento, tratamiento de vacíos, normalización para variables cuantitativas, configurar variables cualitativa y cuantitativas, y filtros para la eliminación de variables que cumplan con ciertas condiciones.

## 4.3 MÓDULO DE PROCESAMIENTO DE DATOS

El módulo de procesamiento de datos presenta al usuario las opciones de los algoritmos de agrupamiento implementados: Conceptual, Simbólico, Klass; los cuales se activan una vez preparados los datos.

## 4.4 MÓDULO DE PRESENTACIÓN DE RESULTADOS

Éste módulo permite mostrar los datos cargado, preparados y procesados, y la visualización de la representación gráfica.

## 4.5 MÓDULO DE AYUDA DEL SISTEMA

Este módulo esta formado por el sistema de ayuda referente al manejo de la herramienta así como los conceptos básicos utilizados en la clasificación de los datos.

# 5 MANEJO DEL MENÚ PRINCIPAL Y LA BARRA DE HERRAMIENTA O DE TRABAJO

El menú principal como su propio nombre lo indica es el menú principal de la herramienta, se puede hacer prácticamente todo desde los diferentes menús, los cuales son: Proyecto, Archivo, Datos, Edición, Ver y Ayuda. La barra de herramienta o de trabajo, es un conjunto de botones de acceso rápido a diferentes opciones del menú principal, lo que permite realizar las acciones más frecuentes sin necesidad de navegar entre los menús.

## 5.1 PROYECTO

Figura A.2. Menú Proyecto.


Este menú contiene las opciones que permiten al usuario crear, abrir, guardar, cerrar proyectos y hacer un proyecto por procesamiento por lotes. Estas opciones son: Nuevo Proyecto Modo Visual o Procesamiento por Lotes, Abrir Proyecto, Guardar Proyecto, Guardar Proyecto Como, y Cerrar Proyecto.

### 5.1.1 Nuevo Proyecto

Al seleccionar el usuario la opción *Nuevo Proyecto* el sistema habilita las opciones de los menús Archivo, Datos, Ver y Ayuda y presenta en el navegador de proyecto todos los tipos de datos que se manejan en el proyecto para que el usuario pueda crear o cargar un conjunto de datos y trabajar sobre ellos. Esta opción se puede efectuar también desde la barra de herramienta en la sección de proyecto dando clic sobre el botón crear nuevo proyecto.

### 5.1.2 Nuevo Proyecto de Procesamiento por Lotes

Para realizar un procesamiento por lotes se selecciona esta opción o desde la barra de herramienta en la sección de proyecto dando clic sobre el botón Nuevo proyecto de procesamiento por lotes, el cual se realiza en dos pasos.

*Paso 1: cargar datos*

Una vez seleccionada la opción Nuevo proyecto de procesamiento por lotes el sistema presenta la ventana de *Asistente para la Configuración de Proyecto de Procesamiento por Lotes*, aquí el usuario determina el conjunto de datos de entrada y un nombre a los datos clasificados de salida, al seleccionar estas opciones el sistema muestra la venta de apertura de Windows donde selecciona el tipo y el nombre del archivo de los datos de entrada y los datos clasificados de salida, después selecciona la configuración de preprocesamiento y la configuración de procesamiento, una vez determinado esto presionar el botón siguiente de la misma ventana, de inmediato aparece la ventana de importar datos en caso que los archivo de datos sean de texto, separados por como, Excel u otra tipo de archivo, una vez cargado o importado los datos se abre la ventana de preparación de datos con la ventana de configuración básica activa donde se seleccionan las variables a procesar, realizado esto en caso de haber escogido alguna opción de preprocesamiento aparecerá la ventana de preparar datos con la página activa correspondiente a la opción seleccionada en configuración de preprocesamiento, después de haber realizado todo lo anterior se presenta la ventana de *Asistente para la Configuración de Proyecto de Procesamiento por Lotes* para continuar con el paso 2.

*Paso 2: configuración de los lotes*

En este paso el usuario realiza la configuración de los lotes, donde determina en la sección de configuración de los lotes el número de lotes a formar. El sistema determina el tamaño de los lotes, el número de registros totales a agrupar, y los registros de entrenamiento dependiendo del número de lotes a formar. Después de esto presionar el botón finalizar, de inmediato el sistema presenta la ventana de

configuración del algoritmo de agrupamiento seleccionado en la configuración de procesamiento, donde el usuario selecciona el número de clases deseado o la forma de determinarlo, una vez determinado esto presionar el botón aceptar, realizándose así el procesamiento por lotes, y finalmente el sistema presenta una información confirmando que la clasificación de datos finalizó satisfactoriamente y que los resultados se encuentran en el archivo dado por el usuario en los datos clasificados de salida.

### 5.1.3 Abrir Proyecto

Para abrir un proyecto ya existente se selecciona esta opción o desde la barra de herramienta en la sección de Proyecto dando clic sobre el botón Abrir Proyecto. De inmediato el sistema presenta la ventana de dialogo estándar de Windows, donde se pide el nombre del proyecto que se desea abrir y la ubicación.

### 5.1.4 Guardar Proyecto

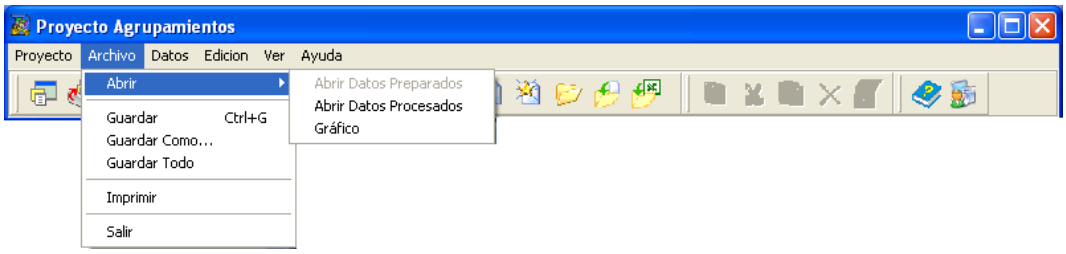
Para guardar un proyecto se presiona sobre esta opción o desde la barra de herramienta en la sección de proyecto dando clic sobre el botón guardar proyecto, si es un proyecto nuevo el sistema presenta la ventana de dialogo guardar estándar de Windows, la cual pide la ubicación, el nombre y tipo de archivo, si es un archivo previamente guardado el sistema guarda los cambios realizados por el usuario en memoria. La opción Guardar Proyecto Como permite guardar un proyecto con un nombre diferente al que posee, para guardar proyecto como seleccionar esta opción en el menú Proyecto del Menú Principal y digitar el nuevo nombre y determinar la ubicación.

### 5.1.5 Cerrar Proyecto

Para cerrar un proyecto se da clic sobre esta opción o desde la barra de herramienta en la sección de proyecto dando clic sobre el botón cerrar proyecto.

## 5.2 ARCHIVO

Figura A.3. Menú archivo.




Este menú contiene las opciones que permiten al usuario Abrir, Guardar, Guardar Como, Guardar Todo, Imprimir, y Salir del sistema

### 5.2.1 Abrir

Al seleccionar la opción *Abrir* el usuario puede cargar datos preparados, procesados o gráfico existente, al escoger alguna de las opciones anteriores, el sistema de inmediato presenta la ventana de diálogo de apertura estándar de Windows, donde se pide la ubicación, el nombre y el tipo del archivo de datos que se desea abrir.

### 5.2.2 Guardar

Para *guardar* un archivo se presiona sobre esta opción o desde la barra de herramienta en el botón *guardar archivo*, si es un archivo de datos nuevos el sistema presenta la ventana de dialogo guardar estándar de Windows, la cual pide la ubicación, el nombre y tipo de archivo, si es un archivo previamente guardado el sistema guarda los cambios realizados por el usuario en memoria.

### 5.2.3 Guardar Como

Esta opción permite al usuario guardar un archivo de datos con un nombre diferente al que posee. Para *guardar como*, se presiona esta opción o desde la barra de herramienta en el botón *guardar archivo como*.

### 5.2.4 Guardar Todo

Esta opción permite al usuario guardar todos los archivos. Para *guardar Todo*, se presiona esta opción o desde la barra de herramienta en el botón *guardar Todos los archivos*.

### 5.2.5 Imprimir

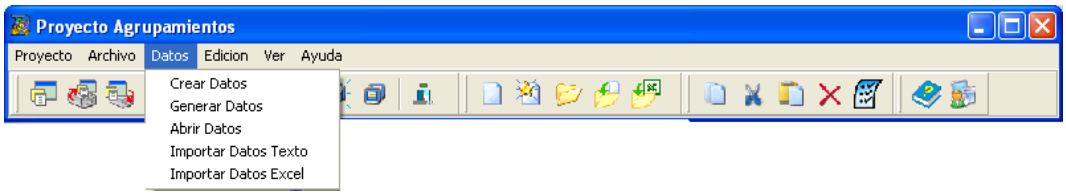
Esta opción permite al usuario imprimir reportes, sobre los resultados obtenidos en el proceso de agrupamiento. Para imprimir escoja esta opción.

### 5.2.6 Salir

Para salir de ADAMIX 1.0 se presiona en esta opción o desde la barra de herramienta en el botón *salir*.

## 5.3 DATOS

Figura A.4. Menú Datos.




Este menú contiene todas las opciones para la creación y apertura de datos. Las opciones del menú datos son: Crear Datos, Generar Datos, Abrir Datos, Importar Datos Texto, Importar Datos Excel.

### 5.3.1 Crear Datos

Para crear un conjunto de datos se selecciona esta opción o desde la barra de herramienta en el botón Crear Datos. Una vez seleccionada el sistema presenta la

ventana de *Asistente para Crear Nuevo Conjunto de Datos*, donde el usuario escribe un nombre o alias para identificar el conjunto de datos, selecciona el número de registros y de variables, después de esto se presiona el botón aceptar, de inmediato el sistema presenta en el panel de visualización la malla con el tamaño correspondiente al número de registros y variables, lista para su posterior llenado.

### 5.3.2 Generar Datos

Para generar un conjunto de datos se selecciona esta opción o desde la barra de herramienta en el botón Generar Datos. Esta opción se realiza en dos pasos:

*Paso 1:*
Una vez seleccionada la opción generar datos el sistema presenta la ventana de *Asistente para Crear Nuevo Conjunto de Datos*, donde el usuario escribe un nombre o alias para identificar el conjunto de datos, selecciona el número de registros y de variables, después de esto presionar el botón siguiente, para realizar el segundo paso.

*Paso 2: Generar datos Aleatoriamente*
En este paso el sistema presenta la ventana de *Asistente para Crear Nuevo Conjunto de Datos*, con la página de generar datos aleatoriamente activa, donde el usuario determina la manera de cómo generar el tipo de valores para variables cuantitativas y cualitativas.

### 5.3.3 Abrir Datos

Esta opción permite abrir archivos de datos con cabecera; al seleccionar esta opción el sistema presenta la ventana de diálogo de apertura estándar de Windows, donde se pide la ubicación, el nombre del archivo de datos que se desea abrir, una vez determinado lo anterior se presiona el botón aceptar, para que el sistema los visualice en el Panel de Visualización para su posterior preparación y procesamiento.

### 5.3.4 Importar Datos Texto

Esta opción permite abrir archivos de datos de texto, separados por coma, o archivos de texto con otra extensión; al seleccionar esta opción el sistema presenta la ventana de diálogo de apertura estándar de Windows, donde se pide la ubicación y el nombre del archivo de datos que se desea abrir.

Una vez seleccionado el archivo se presiona el botón abrir, presentándose la ventana de *Asistente para importar archivos de texto*, con la página de *vista previa del archivo* activa donde se muestra la vista previa de los datos, al presionar el botón siguiente de esta ventana se muestra la página de configuración general activa donde se elige él o los tipos de separadores utilizados en el archivo de datos, la posición del nombre de las variables y desde donde y hasta donde importar los datos, una vez realizado lo anterior presionar el botón siguiente para continuar o atrás si se desea hacer algún cambio.

Al presionar el botón siguiente se muestra la página *vista previa de los datos* presentando la forma como serán importado los datos, presionar el botón finalizar para que el sistema los visualice en el Panel de Visualización para su posterior preparación y procesamiento o atrás si se desea hacer un cambio o cancelar para abandonar la importación de los datos.

### 5.3.5 Importar Datos Excel

Esta opción permite abrir archivos de datos de Excel; para importar datos de Excel seleccione esta opción o desde la barra de herramienta en la sección de datos, una vez seleccionada de alguna de las dos maneras el sistema presenta la ventana de diálogo de apertura estándar de Windows, donde se pide la ubicación y el nombre del archivo de datos que se desea abrir.

Una vez seleccionado el archivo se presiona el botón abrir, presentándose la ventana de *Asistente para importar archivos de Excel*, con la página de *vista previa del archivo* activa donde se muestra la vista previa de archivo y se selecciona la

hoja que se desea importar, al presionar el botón siguiente de esta ventana se muestra la página de configuración general activa donde se seleccionan las columna a importar, la posición del nombre de las variables y desde donde y hasta donde importar los datos, una vez determinado esto presionar el botón siguiente para continuar o atrás si se desea hacer algún cambio.

Al presionar el botón siguiente se muestra la página *vista previa de los datos* presentando la forma como serán importado los datos, presionar el botón finalizar para que el sistema los visualice en el Panel de Visualización para su posterior preparación y procesamiento o atrás si se desea hacer un cambio o cancelar para abandonar la importación de los datos.

## 5.4 EDICIÓN

Figura A.5. Menú edición.


Este menú contiene todas las opciones para la edición de datos y el manejo de la malla. Las opciones del menú edición son: Copiar, Cortar, Pegar, Borrar y Seleccionar Todo.

### 5.4.1 Copiar

Para copiar se selecciona lo que se desee copiar y se presiona esta opción o desde la barra de herramienta en el botón Copiar.

### 5.4.2 Cortar

Para cortar se selecciona lo que se desee cortar y se presiona esta opción o desde la barra de herramienta en el botón Cortar.

### 5.4.3 Pegar

Para pegar una vez de haber copiado o cortado, se ubica en el sitio donde se desee pegar y se presiona esta opción o desde la barra de herramienta en el botón pegar.

### 5.4.4 Borrar

Para borrar se selecciona lo que se desee eliminar y se presiona esta opción o desde la barra de herramienta en el botón borrar.

### 5.4.5 Seleccionar Todo

Para seleccionar todo el contenido de la malla seleccionar esta opción o desde la barra de herramienta en el botón seleccionar todo.

## 5.5 VER

Figura A.6. Menú Ver


En este menú se permite seleccionar las partes del entorno de ADAMIX 1.0 que se deseen visualizar. Las opciones del menú Ver son: Panel de Herramienta, Panel de Visualización y Barra de herramientas.

### 5.5.1 Panel de Herramientas

Para visualizar el Panel de Herramientas seleccionar esta opción, de inmediato el sistema lo visualiza.

### 5.5.2 Panel de Visualización

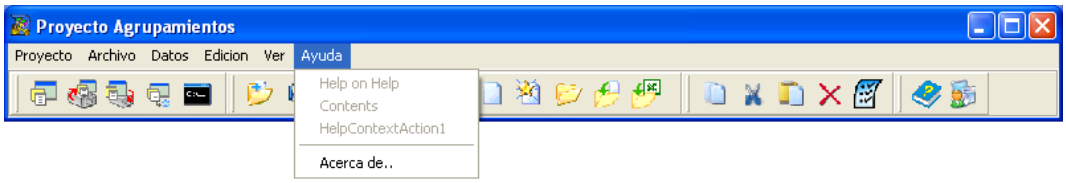
Para visualizar el Panel de Visualización seleccionar esta opción, el panel de visualización presenta al usuario los datos ya sean datos cargados, datos preparados, datos procesados y graficas.

### 5.5.3 Barra de Herramientas

Esta opción permite visualizar las diferentes secciones del panel de herramienta: Proyecto, Archivo, Datos y Ayuda. Para visualizar las secciones anteriores dar clic sobre ellas.

## 5.6 AYUDA

Figura A.7. Menú ayuda.

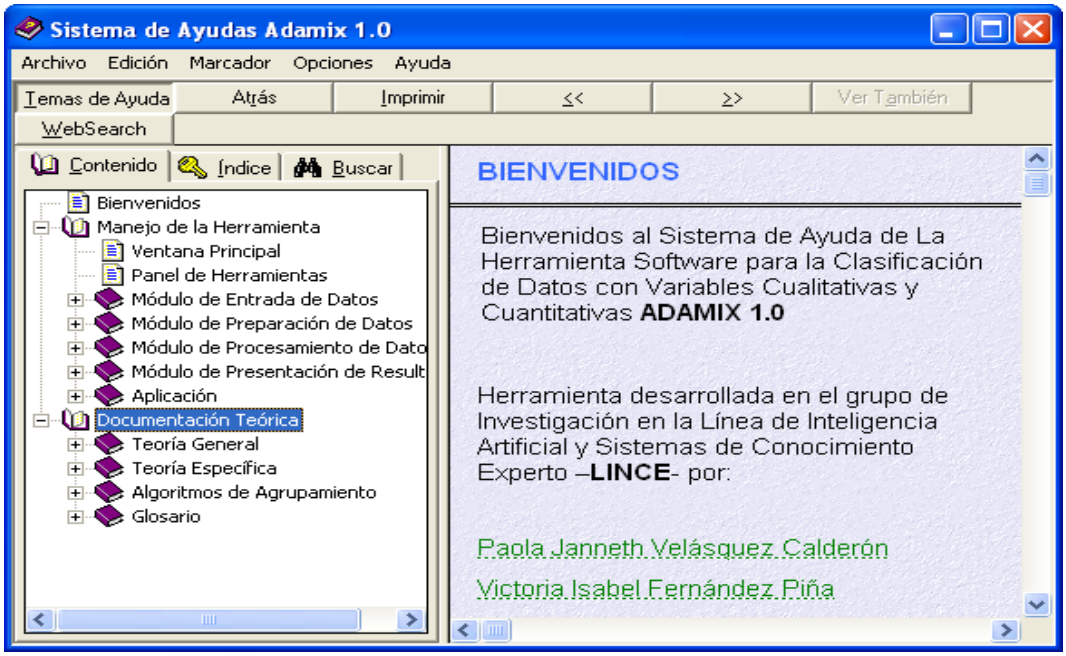



En este menú se encuentran las opciones Contenido y Acerca de

### 5.6.1 Contenido

Al seleccionar la Opción Contenido, el sistema muestra al usuario el sistema de ayuda, donde el usuario puede navegar por los diferentes temas.

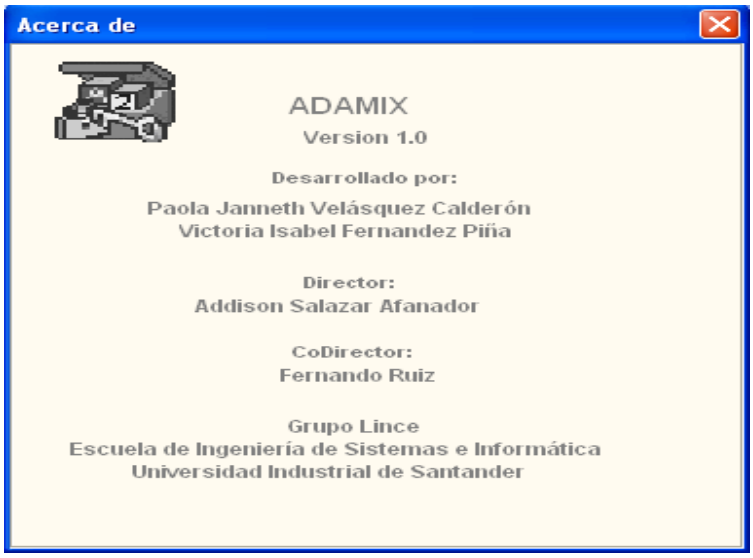
Figura A.8. Ayuda del sistema.




### 5.6.2 Acerca de

Al seleccionar esta opción se presenta la información sobre la herramienta.

Figura A.9. Acerca de.



# 6 MANEJO DEL PANEL DE HERRAMIENTAS: EL NAVEGADOR DEL PROYECTO Y PANEL DE OPCIONES

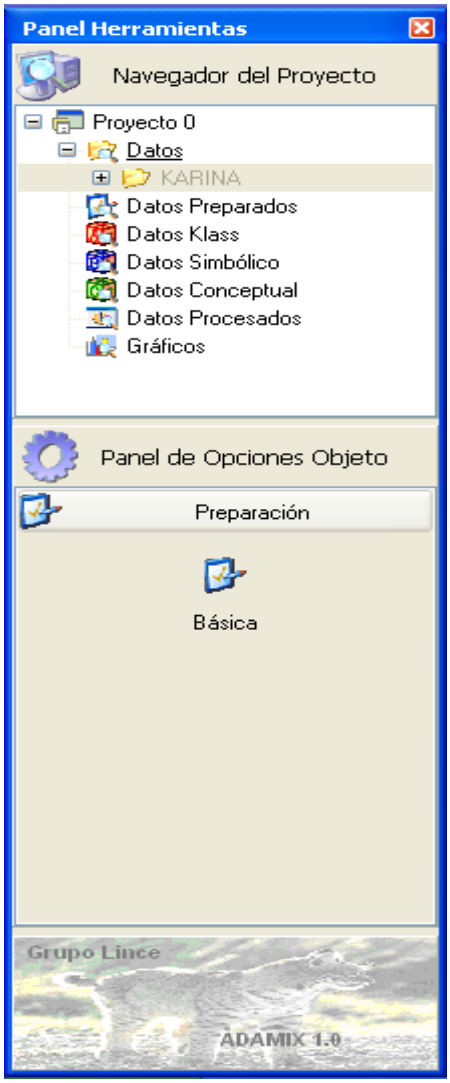
El panel de herramientas esta formado por el ***Navegador del Proyecto*** y el ***Panel de Opciones objeto***.

***El navegador del proyecto*** presenta en forma de árbol la información de los Datos cargados, Datos Preparados, Datos Klass, Datos Simbólico, Datos Conceptual, Datos Procesados y Gráficos; al expandir el + de cada tipo de datos y dar clic sobre su contenido de inmediato el sistema los visualiza en el Panel de Visualización.

La información que se presenta para los datos cargados es la ubicación, el número de variables que posee y el número de registros, para datos preparados, datos klass, datos simbólico y datos conceptual el número de variables y el número de registro, para datos procesados el número de variables, el número de grupos y el número de elementos y para gráficos el origen, el tipo de gráfico, el número de variables y el número de grupos formados.

***El panel de opciones objeto*** esta formado por las secciones que permiten realizar la preparación y procesamiento de los datos, así como el análisis gráfico de los mismos una vez procesados las cuales son: Preparación, Agrupamiento y Análisis Gráfico. *La sección de Preparación* se activa una vez sean cargados los datos, *la sección Agrupamiento* se activa una vez realizada la preparación básica y *la sección Análisis Gráfico* sólo se activa cuando existen datos procesados por alguno de los algoritmos de agrupamiento.

Figura A.10. Panel de Herramientas.



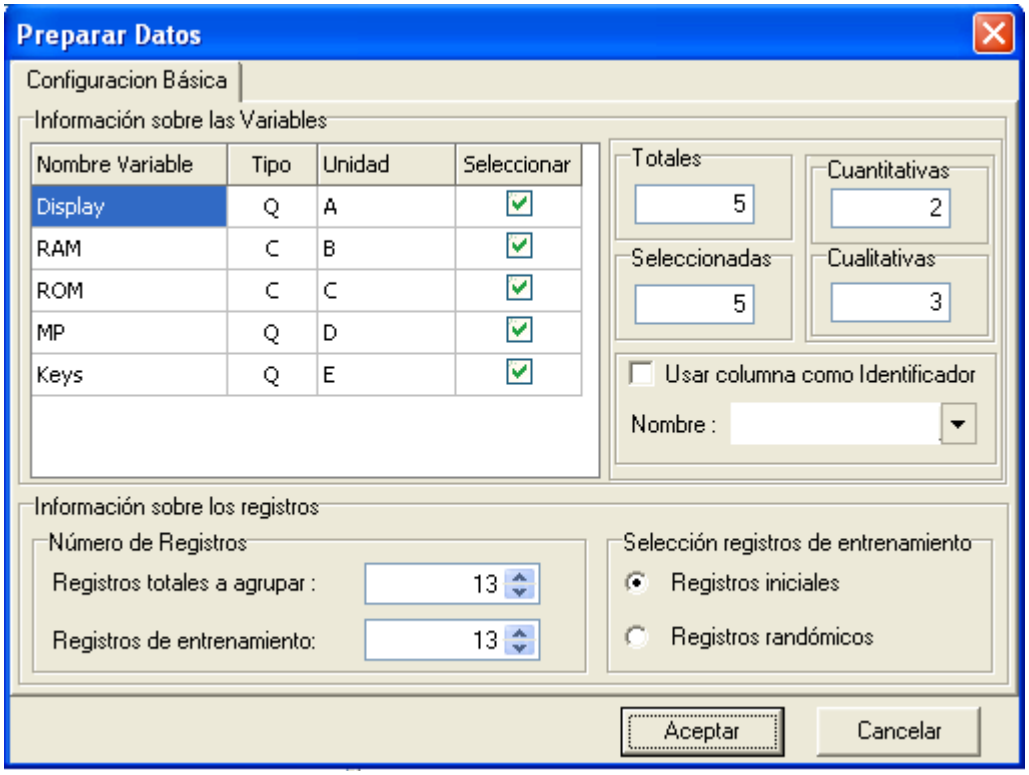

## 6.1 PREPARACIÓN

La sección de preparación contiene las siguientes opciones para que el usuario realice la preparación de los datos ya sean cargados o digitados: Básica, Vacíos, Configuración Cualitativas, Configuración Cuantitativas, Normalizaciones y Filtros. La opción Básica se activa de primero, una vez realizada ésta, se activan las demás opciones.

### 6.1.1 Básica

Esta opción es la primera que se presenta al activarse la sección de Preparación una vez cargado los datos, para realizar la preparación básica seleccionar esta opción, de inmediato el sistema abrirá el formulario de Preparar Datos, activando la página de Configuración Básica (Figura A.11). Donde se le presenta al usuario dos tipos de información de los datos cargados: Información sobre las variables e Información sobre los registros.

Figura A.11. Configuración Básica




*Información sobre las variables*: Como su nombre lo indica se presenta toda la información referente a las variables las cuales son: Nombres, Tipos, Unidades, variables a incluir, el número de variables totales, número de variables seleccionadas, número de variables cuantitativas y cualitativas.

- Nombres: En este campo aparecen los nombres de las variables.
- Tipos: Aquí se presenta el tipo de la variable, si la variable es cualitativa escoja Q y si es cuantitativa escoja C.
- Unidades: En este campo se presenta las unidades de las variables si poseen.
- Incluir: En este campo se seleccionan las variables que se desean procesar, dando clic en la casilla correspondiente a cada variable.
- Totales: En este campo se presenta el número total de variables cargadas o digitadas.
- Seleccionadas: En este campo se presenta el número de variables seleccionadas en el campo incluir.
- Cuantitativas: En este campo se presenta el número de variables cuantitativas seleccionadas.
- Cualitativas: En este campo se presenta el número de variables cualitativas seleccionadas.

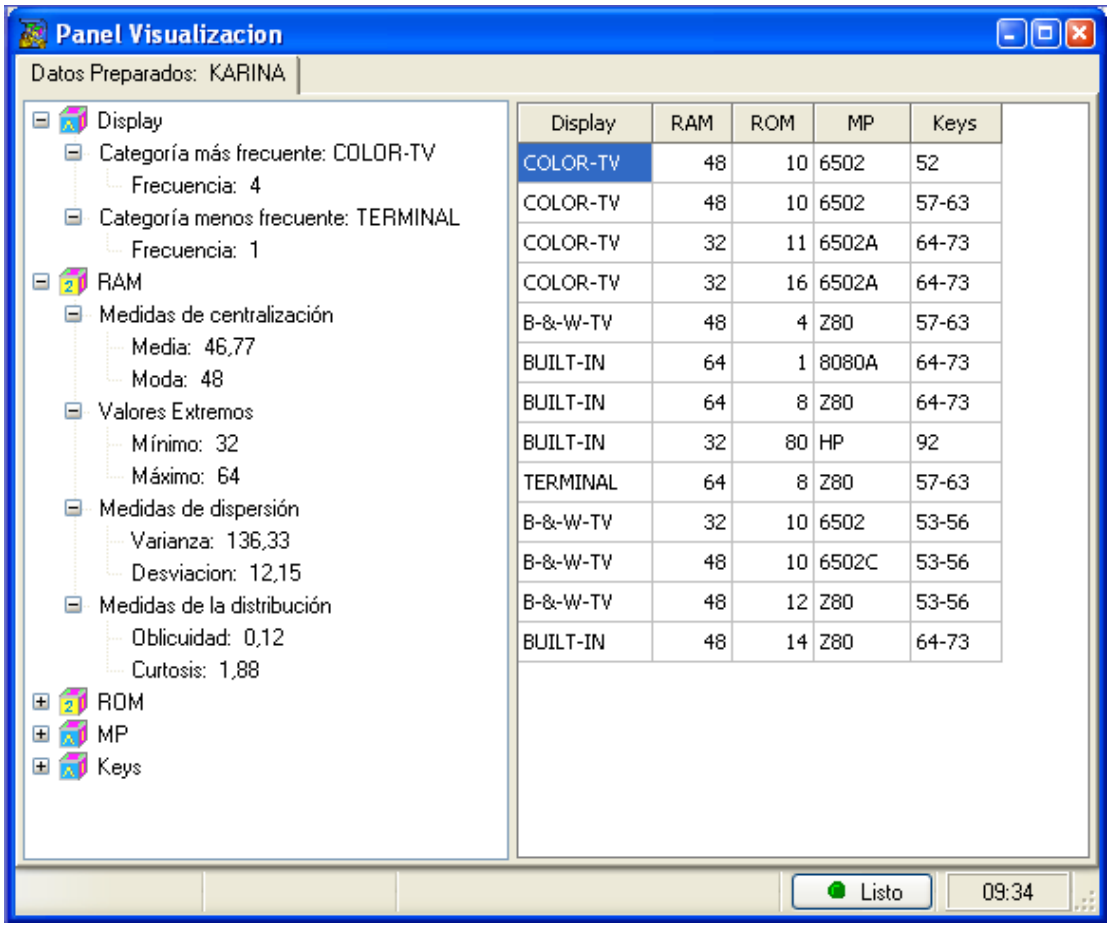
*Información sobre los registros*: En esta parte se presenta la información sobre los registros cargados.

- Número de Registros: Se presenta el número de Registros totales que se van a agrupar, así como el número de Registros Tomados simultáneamente.
- Selección registros simultáneos: Aquí se presenta la forma en que usted desea seleccionar los registros simultáneos, para ser procesados en el Módulo de Procesamiento. Seleccione Registros iniciales si desea trabajar con los registros iniciales o, seleccione Registros randómicos si desea que el sistema determine los registros a trabajar.

Una vez realizado los cambios necesarios presionar la tecla aceptar, de inmediato aparece el formulario del Panel de Visualización con la pagina de Datos Preparados activa, donde se presentan las variables incluidas para el procesamiento. Al expandir el + de cada variable se presenta el Tipo, la Unidad, Medidas Estadísticas: Categoría más y menos frecuentes para variables cualitativas, Medidas de

centralización (Media, Moda), Medidas de dispersión (Varianza, Desviación), Valores Extremos (Mínimo, Máximo) y Medidas de la distribución ( Oblicuidad, Curtosis) para variables cuantitativas.

Figura A.12. Preparación Básica

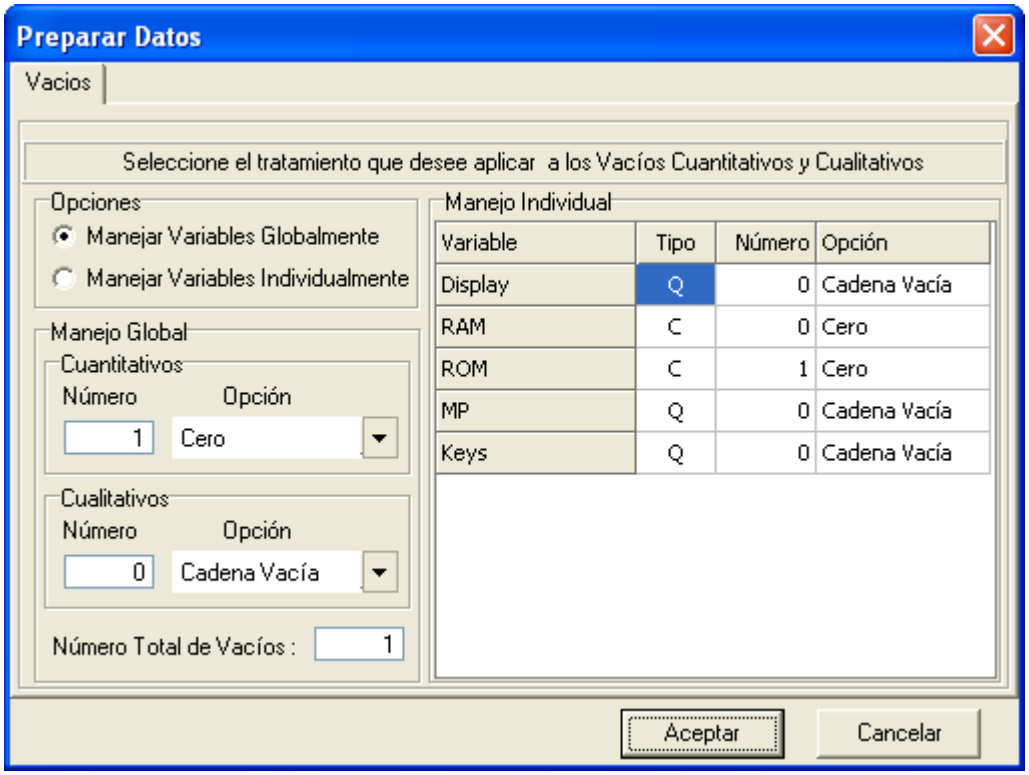



### 6.1.2 Vacíos

Esta opción se activa cuando al presionar la opción básica y al dar aceptar el sistema encuentra algún campo vacío, informando sobre esto y preguntando si desea llenar vacíos, al dar si aparecerá el formulario de Preparar Datos con la

página de Vacíos activa (Figura A.13), constituida por 3 secciones: *Opciones*, *Manejo Global*, *Manejo Individual*.

Figura A.13. Vacíos.




**Opciones**

En esta sección se selecciona el tipo de tratamiento que se desee dar a los vacíos cualitativos y cuantitativos, que puede ser Manejar Variables Globalmente o Manejar Variables Individualmente. Al seleccionar manejar variables globalmente queda habilitada la sección de Manejo Global y deshabilitada la sección de manejo individual, al seleccionar manejar variables individualmente queda habilitada la sección de Manejo Individual y deshabilitada la sección de manejo global.

**Manejo Global**

En esta sección se determina el tipo de tratamiento que se desee dar a campos vacíos para variables cualitativas y cuantitativas.

- Cuantitativos: En el *campo Número* se muestra la cantidad de vacíos numéricos encontrados por el sistema, en el *campo Opción* se presentan las opciones para llenar los campos vacíos al dar clic sobre este campo aparecerán las siguientes opciones: Cero, Media, Moda, Limite Inferior, Limite Superior, Eliminar los registros. Seleccione la opción con que desee llenar los campos vacíos
- Cualitativos: En el *campo Número* se muestra la cantidad de vacíos cualitativos encontrados por el sistema, en el *campo Opción* se presentan las opciones para llenar los campos vacíos al dar clic sobre este campo aparecerán las siguientes opciones: Cadena Vacía, Moda, Eliminar los registros. Seleccione la opción con que desee llenar los campos vacíos.
- Número Total de Vacíos: En este campo se muestra el número total de vacíos que el sistema encontró en el conjunto de datos.

**Manejo Individual**

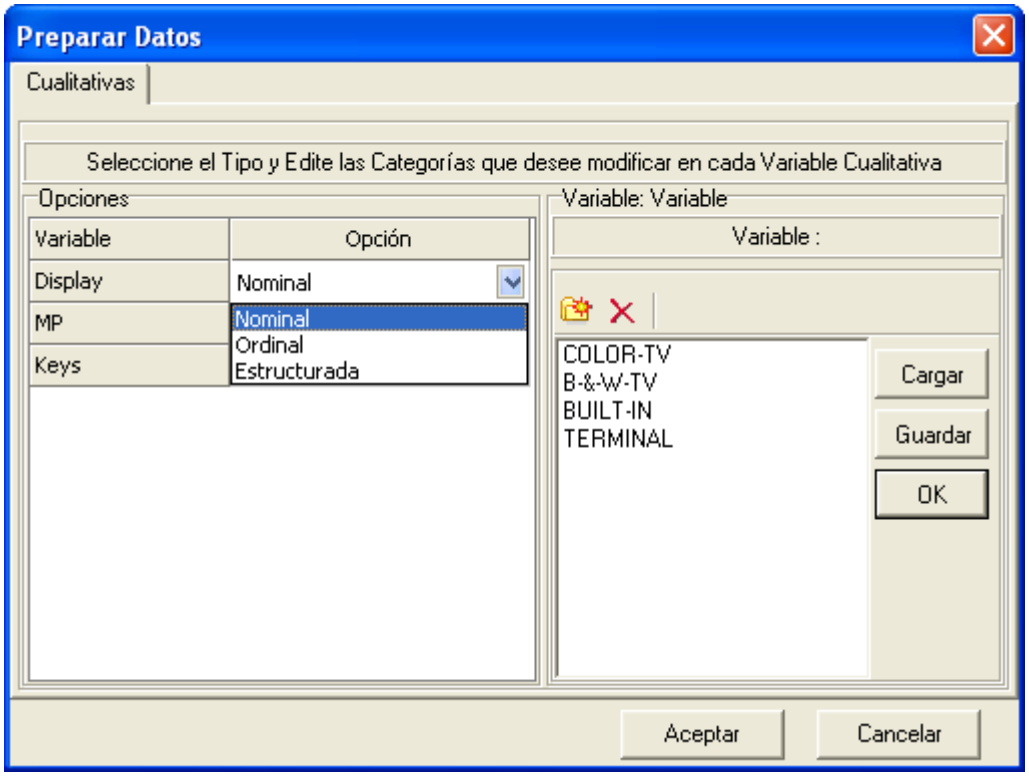
Esta sección está constituida por una cuadricula de cuatro columnas, la primera columna denominada *Variables* donde se presenta el nombre de las variables cargadas, la segunda columna denominada *Tipo* donde se muestra el tipo de cada variable *Q* para variables cualitativas y *C* para variables cuantitativas, la tercera columna denominada *Número* donde se presenta la cantidad de vacíos encontrados por el sistema en cada variable y por ultimo la cuarta denominada *Opción* donde se presentan las opciones para llenar los campos vacíos de las variables, dependiendo si es cuantitativa o cualitativa. Al dar click al frente de cada variable en la columna opción seleccione la opción con que desee llenar los campos vacíos.

Una vez seleccionado el tipo de tratamiento y la opción para llenar vacíos dar clic en aceptar para que el sistema proceda a llenar los campos vacíos.

### 6.1.3 Configurar Cualitativas

Esta opción permite realizar una preparación avanzada de las variables cualitativas. Para realizar una preparación avanzada de las variables cualitativas seleccione esta opción, de inmediato el sistema abrirá el formulario de Preparar Datos con la página de Cualitativas activa dividida en dos secciones: *Opciones* y *variables*.

Figura A.14. Configuración Cualitativa.




**Opciones**

En esta sección al dar clic en la columna de opción al frente de cada variable se despliegan los tipos de variables cualitativas (Nominal, Ordinal y Estructurada) de inmediato se muestra en la sección de variable las categorías de la variable seleccionada.

**Variables**

En esta sección se presentan los nombres de las categorías de la variable seleccionada en la sección de opciones en la estructura de su tipo.

*Tipo nominal*: tipo de variable cualitativa no ordenada, para modificar el nombre de la categoría selecciónela y edite el nuevo nombre.

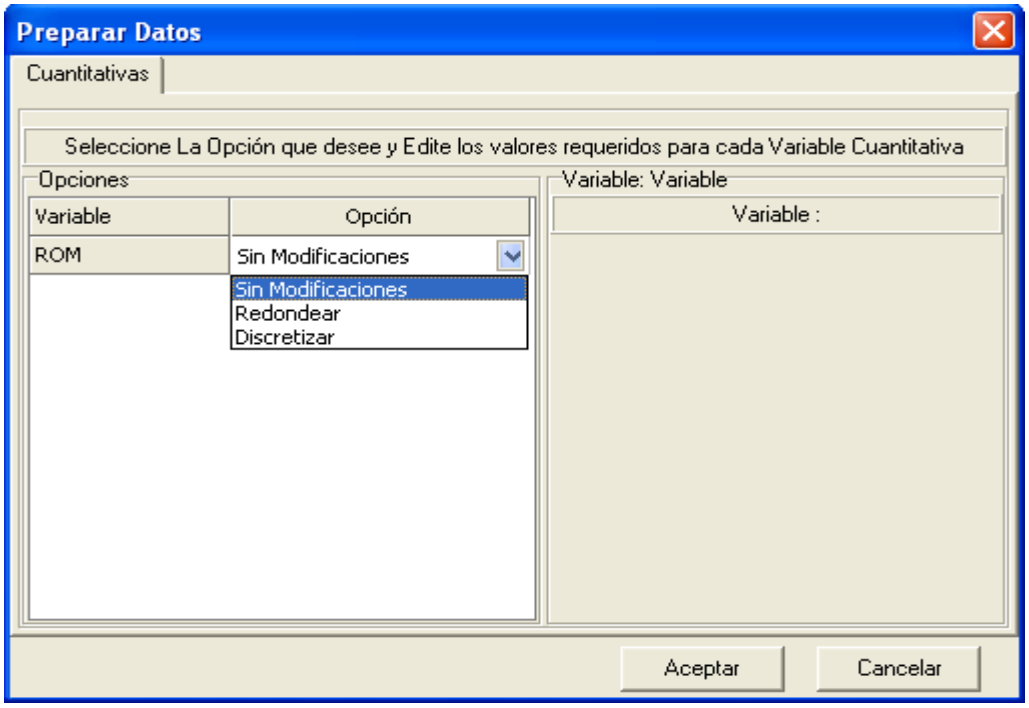
*Tipo ordinal*: tipo de variable cualitativa ordenada, para modificar el nombre de la categoría selecciónela y edite el nuevo nombre y para cambiar el orden dar clic en las flechas de desplazamiento hacia arriba ( ⇑ ) o hacia abajo ( ⇓ ) dependiendo de la ubicación.

*Tipo estructurada*: tipo de variable cualitativa ordenada en forma de árbol, para modificar el nombre de la categoría selecciónela y edite el nuevo nombre, para cambiar el orden dar clic en las flechas de desplazamiento asía arriba ( ⇑ ) o asía abajo ( ⇓ ) dependiendo de la ubicación y para desplazarse en la radicación del árbol dar clic en las flechas de desplazamiento a la derecha ( ⇒ ) y a la izquierda ( ⇐ ).

### 6.1.4 Configurar Cuantitativas

Esta opción permite efectuar una preparación avanzada de las variables cuantitativas. Para realizar una preparación avanzada de las variables cualitativas seleccione esta opción en la sección de preparación, de inmediato el sistema abrirá el formulario de Preparar Datos con la página de Cuantitativas activa dividida en dos secciones: Opciones y variables.

Figura A.15. Configuración Cuantitativa.




**Opciones**

En esta sección al dar clic en la columna de opción al frente de cada variable se despliegan las opciones para modificar las variables cuantitativas (sin modificaciones, discretizar y redondear).

**Variables**

En esta sección se presenta la opción seleccionada en la sección de opciones:

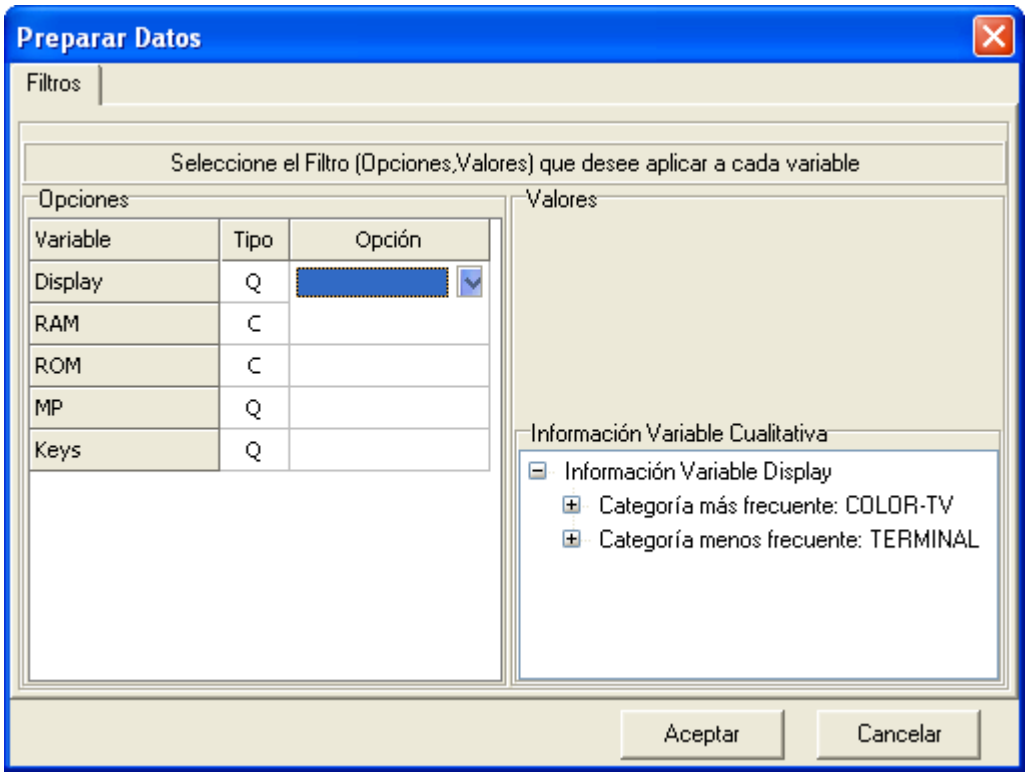
*Redondear:* En esta opción se determina el número de dígitos con la que se desea redondear la variable cuantitativa en el campo Número de Dígitos, una vez determinado el número de dígitos presionar el botón OK para que el sistema acepte el redondeo. Además en la parte inferior se muestra la información estadística básica de la variable seleccionada.

*Discretizar:* En esta opción se determina el número de particiones con la que se desea discretizar la variable cuantitativa en el campo Número de Particones, una vez determinado el número de particiones el sistema muestra en la parte inferior la manera cómo quedan distribuidos los valores en cada partición de la variable seleccionada. Presionar el botón OK para que el sistema acepte la discretización.

### 6.1.5 Filtros

Para filtrar seleccione esta opción, de inmediato el sistema muestra el formulario de Preparar Datos con la página de filtros activada (Figura A.16), la cual está formada por dos secciones (Opciones y Valores.). Esta opción permite al usuario, eliminar los registros que cumplan con las condiciones del filtro seleccionado (Opciones y Valores.).

Figura A.16. Filtros.



**Opciones**

Al dar click en la columna de opciones al frente de cada variable se despliegan las condiciones para eliminar los registros dependiendo del tipo de variable y de inmediato se muestra la sección de valores con la información estadística de la variable.

Para variables cualitativas las condiciones son: Valor igual a, Frecuencia menor a, Frecuencia mayor a, Porcentaje menor a y Porcentaje mayor a.

Para variables cuantitativas las condiciones son: Valor igual a, Dentro del rango, Fuera del rango, Valores extremos e Intervalo de confianza.

**Valores**

En la sección de Valores en la parte superior se presenta la condición seleccionada en la sección de opciones dependiendo del tipo de variable, para que el usuario digite o escoja los valores.

Condiciones para variables cualitativas:

- Valor igual a: escoger el valor con que desee eliminar los registros en el campo Valor Específico.
- Frecuencia menor a: escribir el valor en el campo para eliminar registros cuyos valores tengan frecuencia absoluta inferior al valor digitado.
- Frecuencia mayor a: escribir el valor en el campo para eliminar registros cuyos valores tengan frecuencia absoluta superior al valor digitado.
- Porcentaje menor a: escribir el valor en el campo para eliminar registros cuyos valores tengan frecuencia relativa inferior al valor escrito.
- Porcentaje mayor a: escribir el valor en el campo para eliminar registros cuyos valores tengan frecuencia relativa superior al valor escrito.

Condiciones para variables cuantitativas:

- Valor igual a: escribir el valor con que desee eliminar los registros en el campo Valor Específico.
- Dentro del rango: escribir los valores en los campos Valor Mínimo y Valor Máximo valor para eliminar los registros comprendidos entre estos valores.

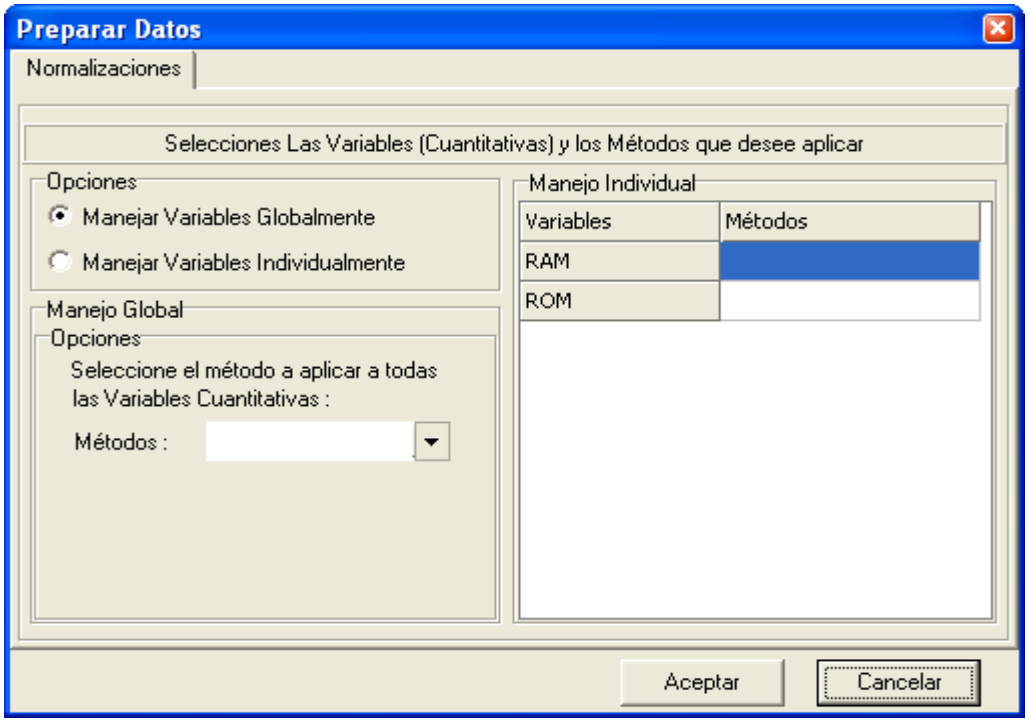
- Fuera del rango: escribir los valores en los campos Valores superiores a y Valores inferiores a dependiendo al seleccionado, para eliminar los registros con valores superiores o inferiores a los digitados o ambos.
- Valores extremos: seleccionar Máximo Valor o Mínimo Valor o ambos para eliminar valores extremos.
- Intervalo de confianza: escribir el valor en el campo de desviaciones típicas para eliminar los registros fuera del intervalo de confianza.

En la parte inferior de esta sección se presenta información estadística sobre las variables, al expandir el + de Información para variables cualitativas se muestra la Categoría más frecuente y la Categoría menos frecuente con sus respectivos valores de frecuencias, para variables cuantitativas se muestran Medidas de centralización (Media, Moda), Medidas de dispersión (Varianza, Desviación), Valores extremos (Mínimo, Máximo.) y Medidas de la distribución (Oblicuidad, Curtosis.).

### 6.1.6 Normalizaciones

Para normalizar seleccione esta opción en la sección de preparación; de inmediato el sistema muestra el formulario de Preparar Datos con la página de normalizaciones activada, esta página está formada por 3 secciones: Opciones, Manejo Global y Manejo Individual.

Figura A.17. Normalizaciones.




**Opciones**

En esta sección se selecciona el tipo de manejo de normalización que se desea dar a los datos, los cuales son: Manejar Variables Globalmente y Manejar Variables Individualmente; al seleccionar Manejar Variables Globalmente queda habilitada la sección de Manejo Global y deshabilitada la sección de Manejo Individual y viceversa.

**Manejo Global**

En esta sección se determina el método que se desee aplicar a las variables cuantitativas, estos métodos son: Estandarización, Normalización, Escala rango y Exponencial.

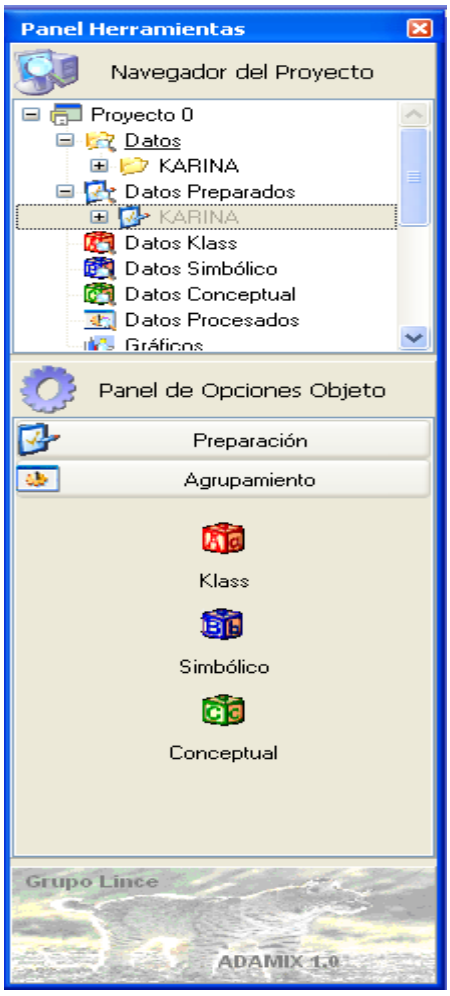
**Manejo Individual**

Esta sección está constituida por una cuadricula de dos columnas, la primera columna denominada Variables donde se presenta el nombre de las variables

cuantitativas cargadas, la segunda columna denominada Métodos donde se muestran los métodos (Estandarización, Normalización, Escala rango y Exponencial) para realizar la normalización. Al dar clic al frente de cada variable en la columna método seleccione la opción con que se desee normalizar.

## 6.2 AGRUPAMIENTO

Figura A.18. Agrupamiento.

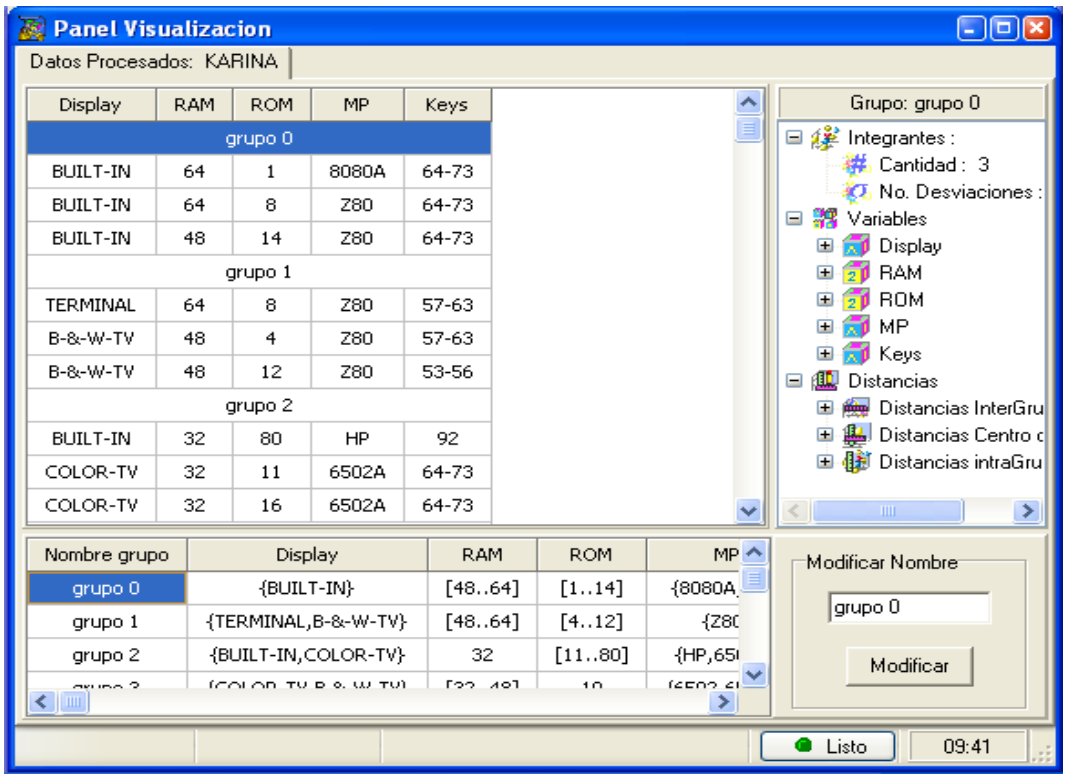



Una vez cargado los datos y preparados se sigue con el procesamiento, el cual se efectúa en la sección de agrupamiento del panel de opciones objeto, esta sección se despliega una vez sean preparados los datos, las opciones de agrupamiento son:

- Klass.
- Simbólico.
- Conceptual.

al seleccionar alguna de las opciones de la sección de agrupamiento se presentará la página correspondiente a cada algoritmo, donde se presiona el botón configurar presentándose el formulario de configurar correspondiente a cada algoritmo, para digitar el número de grupos a formar o si se desea que el sistema calcule el número optimo de clases se selecciona optimo, después de presionar el botón agrupar y listo se muestra la página de datos procesados en el panel de visualización.

Figura A.19. Datos Procesados.




esta página está dividida en cuatro partes, en la parte superior izquierda se muestra una malla con los grupos con sus respectivos integrantes (registros), en la parte superior derecha se presenta un cuadro con la información sobre los integrantes

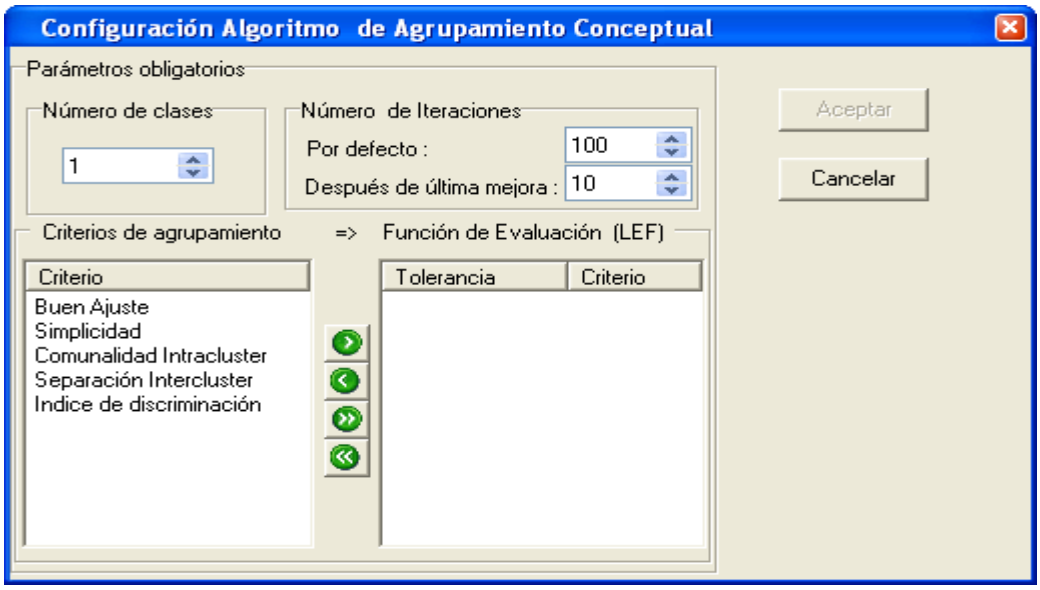
(número, desviación), las variables (valores, moda, valor menos frecuente para variables cualitativas y media, moda, varianza para variables cuantitativas) y las distancias (intergrupo, centro de grupo y intragrupo) del grupo seleccionado en la parte inferior izquierda.

En la parte inferior izquierda se muestra una malla con la distribución de las variables en cada grupo, al dar clic sobre el nombre del grupo de inmediato el sistema lo ubica en la malla de la parte superior izquierda, además muestra la información de éste en la parte superior derecha y en la parte inferior derecha se puede cambiar el nombre del grupo si se desea, digitando el nuevo nombre y dando clic en modificar.

### 6.2.1 Conceptual

Para procesar los datos una vez entrados y preparados mediante el algoritmo de agrupamiento conceptual conjuntivo basado en cluster/2 dar clic en la opción de Conceptual en la sección de Agrupamiento del panel de opciones objeto, una vez presionado ésta aparecerá el formulario de configuración conceptual:

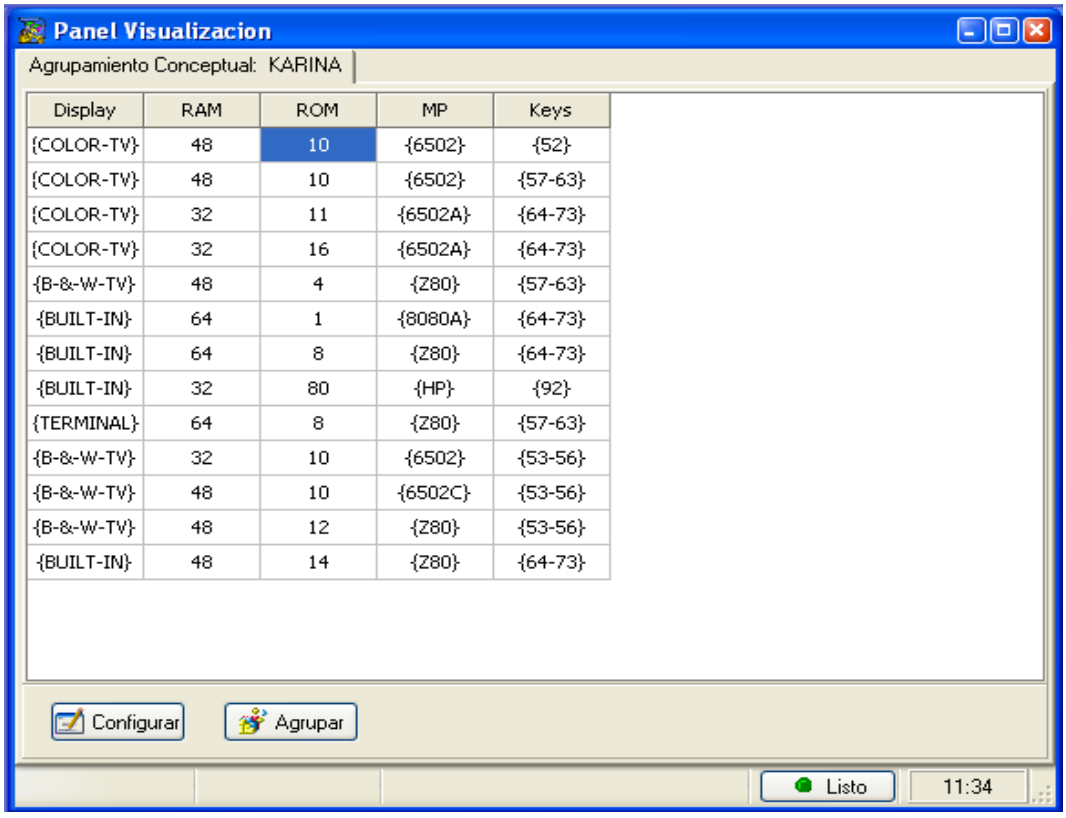
Figura A.20. Configuración Conceptual.



en donde se debe determinar en la sección de Parámetros obligatorios el número de clases, número de iteraciones por defecto y después de la última mejora, y seleccionar los criterios de calidad para el agrupamiento (Buen Ajuste, Simplicidad, Comunalidad Intracluster, Separación Intercluster, Índice de discriminación) para formar la Función de Evaluación Lexicográfica(LEF) que es definida por una secuencia de parejas de "criterio-tolerancia" $(c_1,\tau_1),(c_2,\tau_2),\ldots,$ donde $c_i$ es un criterio elemental seleccionado de la lista anterior, y $\tau_i$ es un "umbral de tolerancia" $(\tau \in [0\ldots100\%])$.

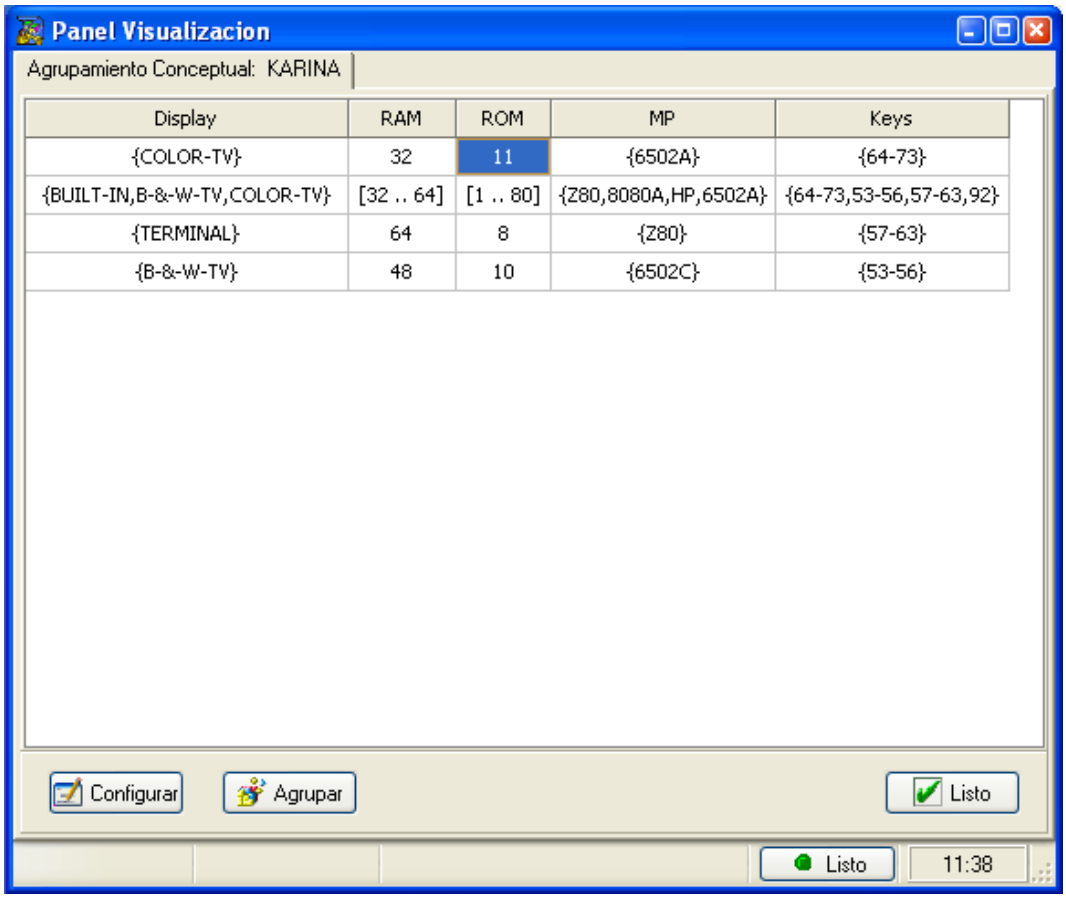
Una vez realizado lo anterior presionar aceptar, de inmediato el sistema muestra el formulario del Panel de Visualización con la página de Agrupamiento Conceptual activa donde se presentan el conjunto de datos conceptual listos para la agrupación, y además son agregados al proyecto.

Figura A.21. Agrupamiento Conceptual.



Presionar el botón agrupar para realizar el agrupamiento; cuando el sistema se demora en cargar los datos aparece un bombillito en la barra de estado del panel de visualización indicando que está ocupado, cuando los datos se encuentran ya listos se cambian de ocupado a listo obteniéndose la agrupación de acuerdo a la configuración realizada.

Figura A.22. Agrupamiento Conceptual Listo.

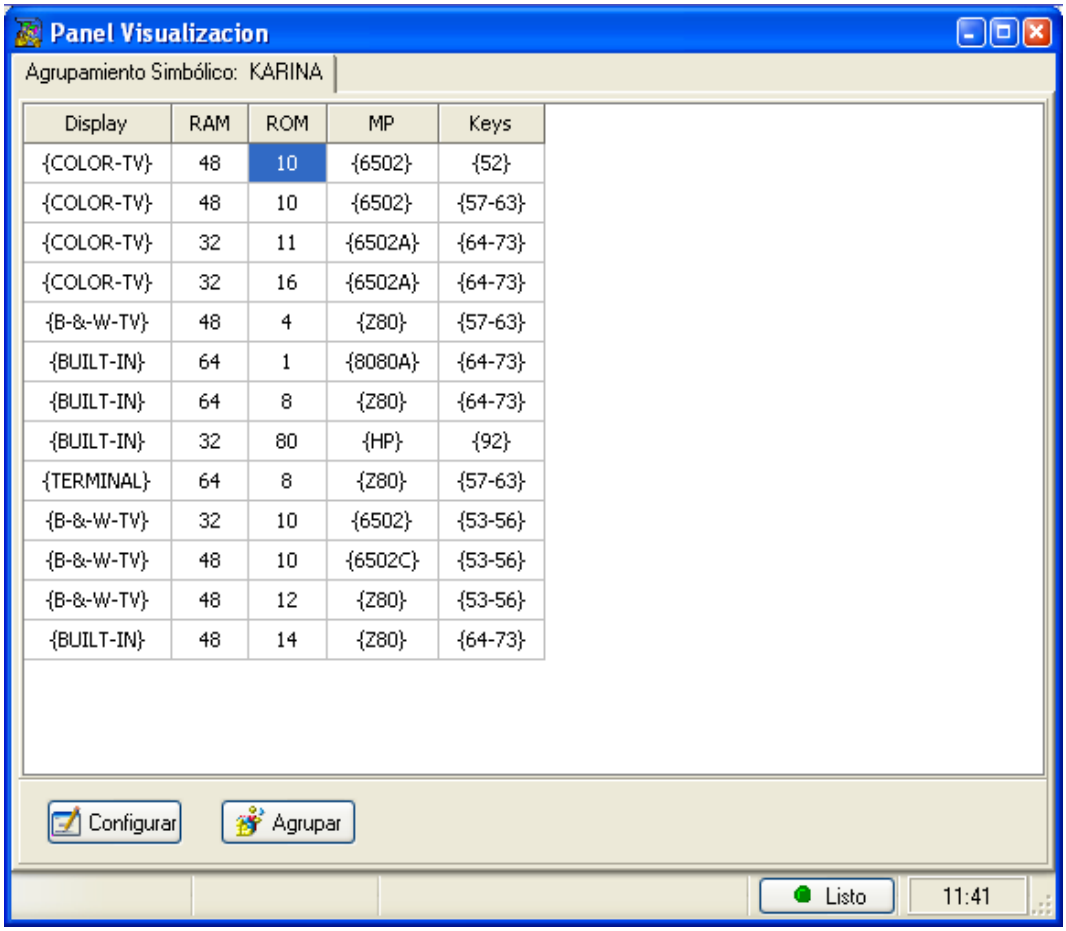



y finalmente presionar el botón listo, para obtener los grupo y seguir con el análisis gráfico.

### 6.2.2 Simbólico

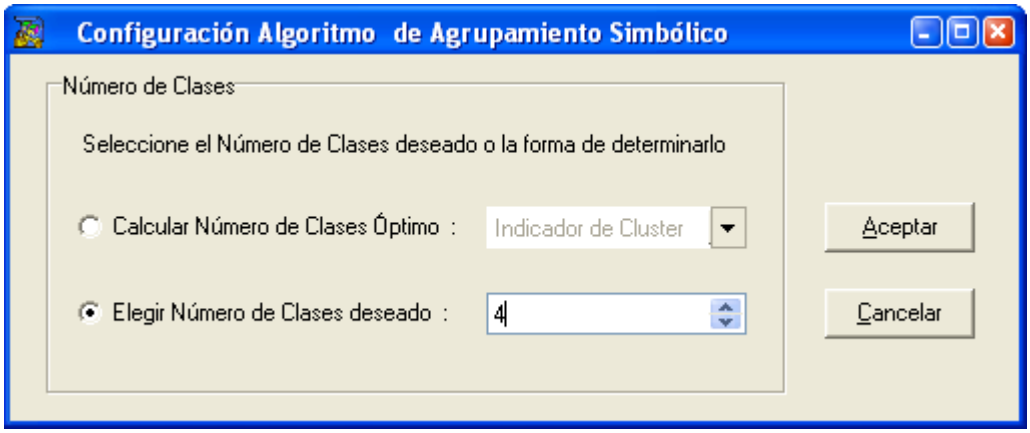
Para procesar los datos una vez entrados y preparados mediante el algoritmo de agrupamiento simbólico basado en una nueva medida de similaridad dar clic en la opción Simbólico en la sección de Agrupamiento del panel de opciones, una vez presionada ésta aparecerá el Panel de Visualización con la página de Agrupamiento Simbólico activa:

Figura A.23. Agrupamiento Simbólico.




donde se muestran los datos listos para la agrupación y son agregados al proyecto, al presionar el botón configurar se presenta el formulario de configuración del algoritmo de agrupamiento simbólico donde se selecciona el número de clases deseado o la forma de determinarlo:

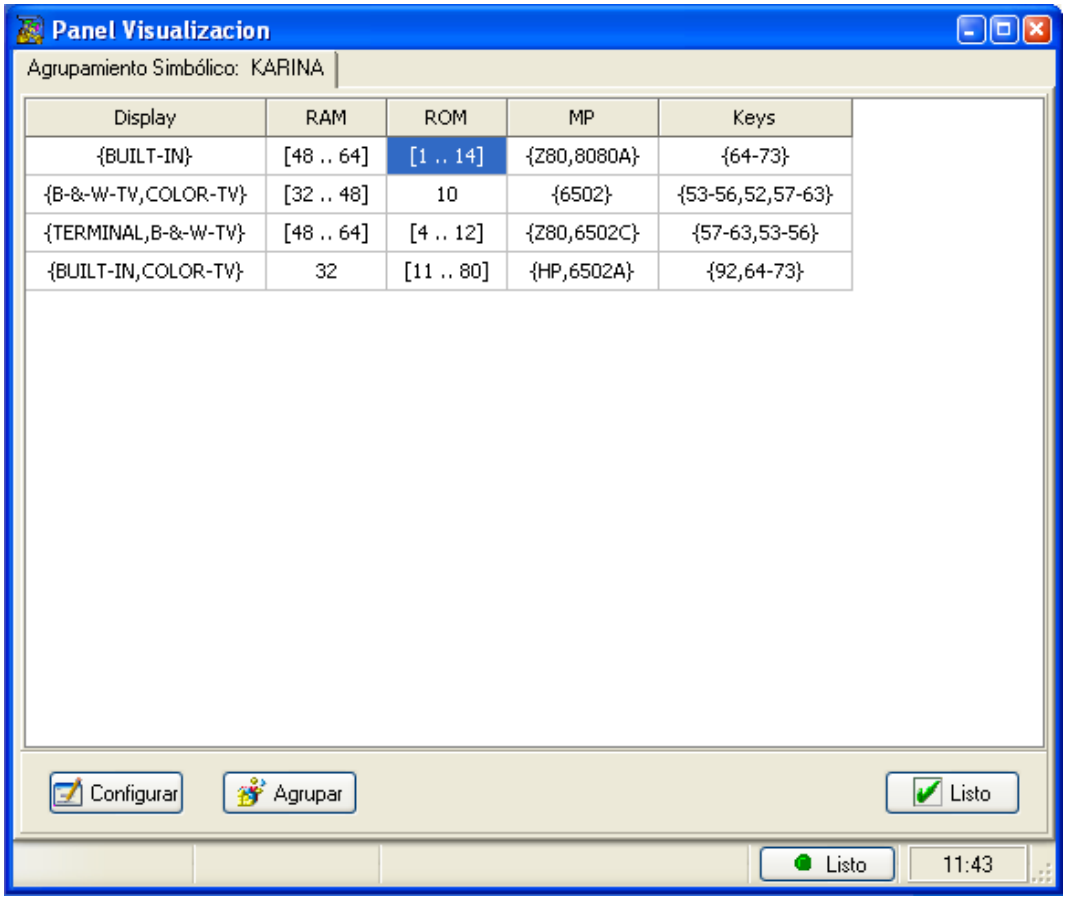
Figura A.24. Configuración Simbólico.




Seleccione Calcular Número De Clases Óptimo si desea que el sistema calcule el número de clases óptimo ya sea por Indicador de cluster o distancia centroides. Seleccione Elegir Número de Clases Deseado si se desea determinar el número de grupos o clases a formar, en este caso digite el número en el campo y finalmente presionar el botón aceptar.

Después en el Panel de Visualización en la página de Agrupamiento Simbólico dar clic en el botón agrupar y finalmente el sistema indicará que puede continuar al pasar de ocupado a listo, y aparecerá en el panel de visualización el botón listo, dar clic sobre este botón, para que el sistema acepte el procesamiento y de esta manera ser agregados al proyecto y pueda cambiar el nombre de los grupos si desea, para ello se posiciona en éste y digita el nombre en la sección de modificar nombre y dar clic en el botón modificar.

Figura A.25. Agrupamiento Simbólico Listo.

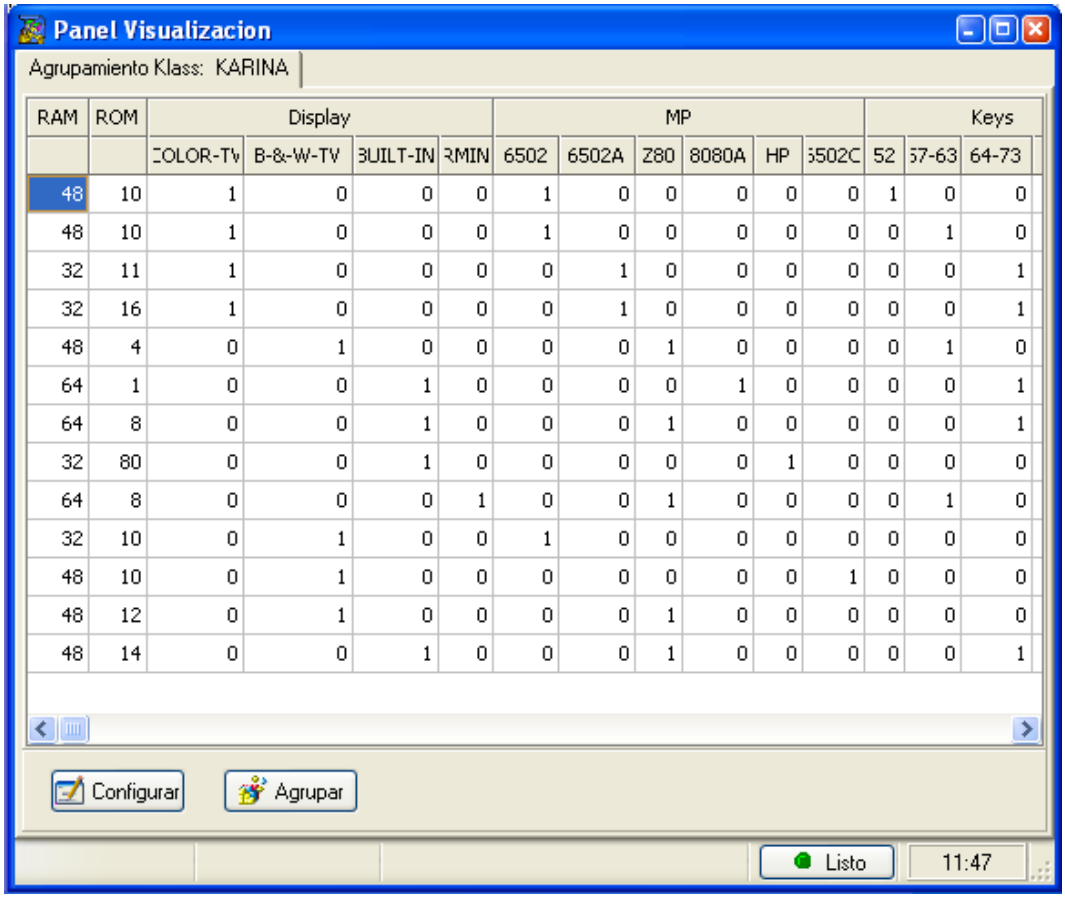



Una vez realizado todo lo anterior queda a disposición la sección de Análisis Gráfico en el Panel de Opciones para realizar la representación gráfica de los grupos obtenidos.

### 6.2.3 Klass

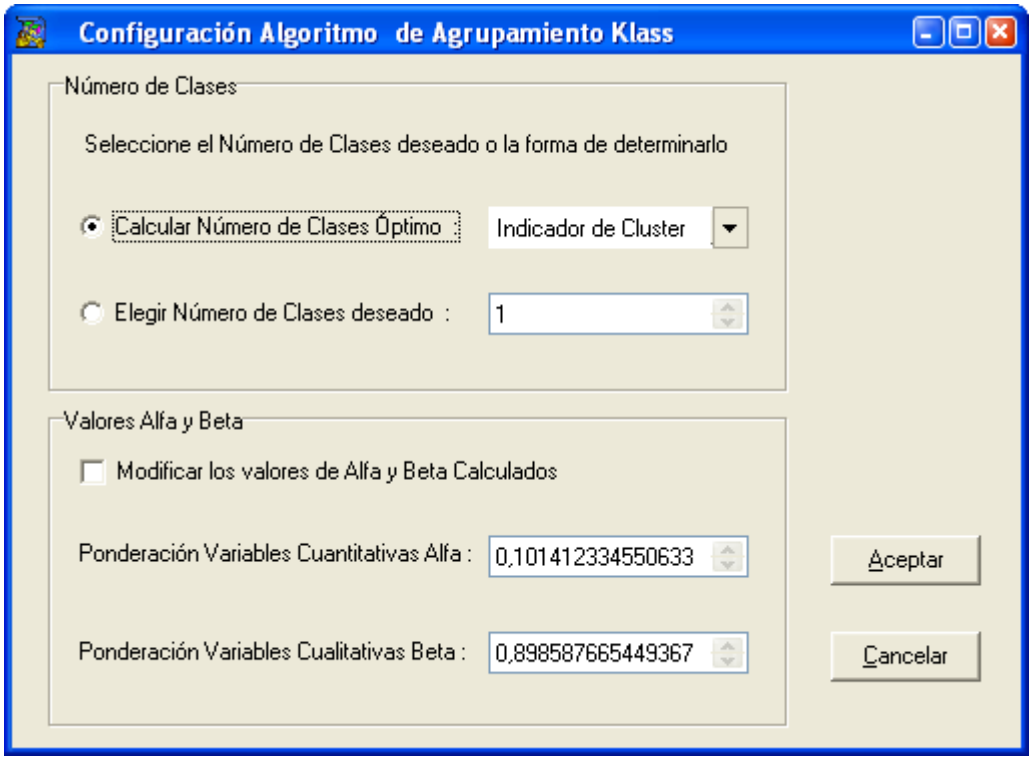
Para procesar los datos una vez entrados y preparados mediante el algoritmo de agrupamiento basado en métricas mixtas ponderadas Klass dar clic en la opción Klass en la sección de Agrupamiento del panel de opciones objeto, una vez presionada ésta aparecerá el Panel de Visualización con la página de Agrupamiento Klass activa:

Figura A.26. Agrupamiento Klass.




donde se muestran los datos listos para la agrupación y son agregados al proyecto, al presionar el botón configurar se presenta el formulario de configuración del algoritmo de agrupamiento Klass donde se selecciona el número de clases deseado o la forma de determinarlo:

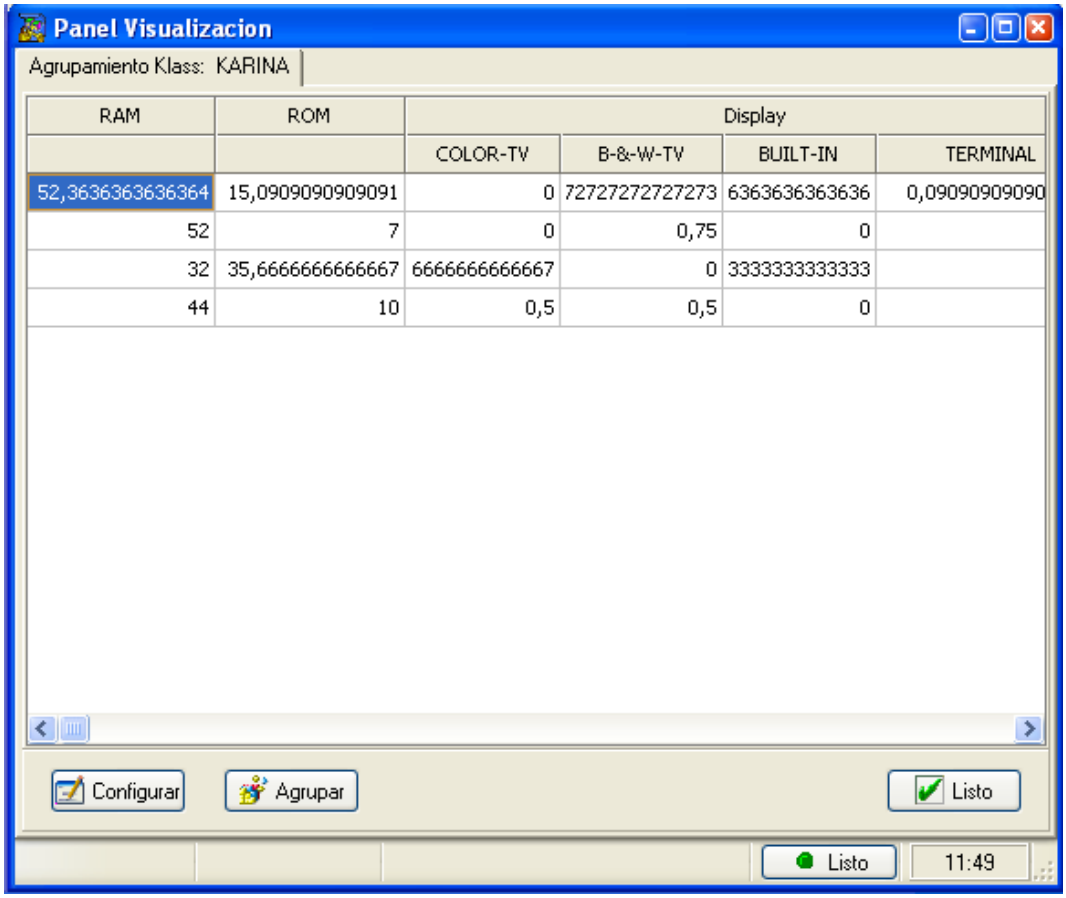
Figura A.27. Configuración Klass.




Seleccione Calcular Número De Clases Óptimo si desea que el sistema calcule el número de clases óptimo ya sea por Indicador de cluster o distancia centroides. Seleccione Elegir Número de Clases Deseado si se desea determinar el número de grupos o clases a formar, en este caso digite el número en el campo, también puede cambiar los valores alfa y beta en la sección valores alfa y beta seleccionando Modificar los valores Alfa y Beta Calculados y finalmente presionar el botón aceptar.

Después en el Panel de Visualización en la página de Agrupamiento Klass dar clic en el botón agrupar y finalmente el sistema indicará que puede continuar al pasar de ocupado a listo, y aparecerá en el panel de visualización el botón listo, dar clic sobre este botón, para que el sistema acepte el procesamiento y de esta manera ser agregados al proyecto y pueda cambiar el nombre de los grupos si desea, para ello se posiciona en éste y digita el nombre en la sección de modificar nombre y dar clic en el botón modificar.

Figura A.28. Agrupamiento Klass Listo.




Una vez realizado todo lo anterior queda a disposición la sección de Análisis Gráfico en el Panel de Opciones para realizar la representación gráfica de los grupos obtenidos.

## 6.3 ANÁLISIS GRÁFICO

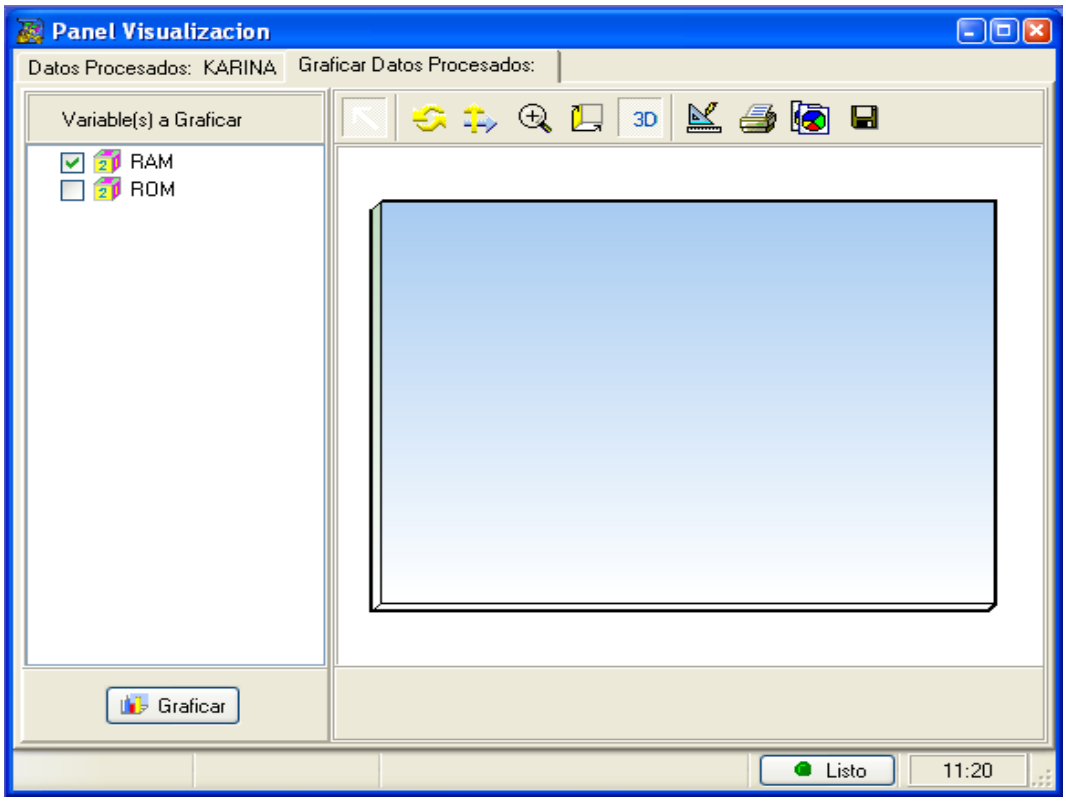
Para graficar los datos procesados existen 10 formas de representación gráfica:

- Puntos.
- Superficies.
- Rangos.
- Barras.

- Barras 3D.
- Histogramas.
- Líneas.
- Torta.
- Máximos y Mínimos.
- Desviación Estándar.

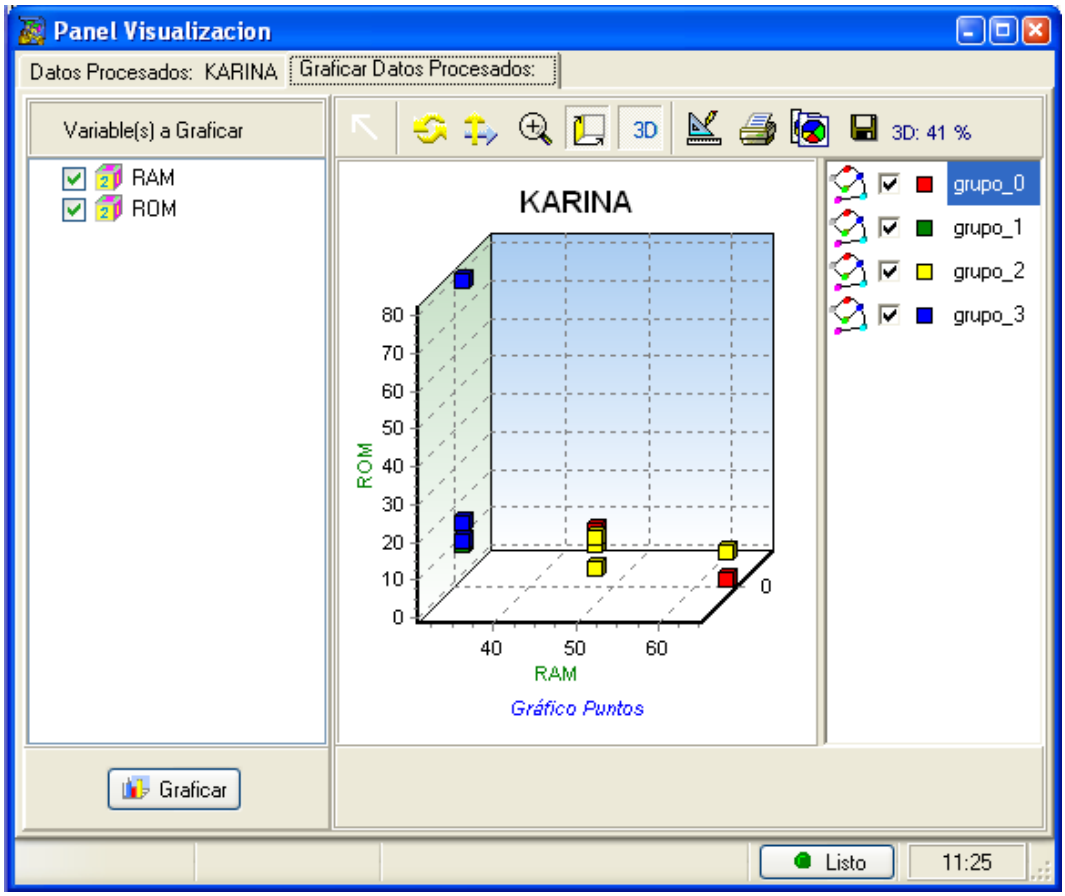
Al seleccionar alguna de éstas formas en las diferentes secciones de la sección de Análisis gráfico en el Panel de Opciones el sistema muestra la página de Graficar Datos Procesados en el Panel de visualización, donde se presenta un cuadro de gráficos al lado derecho y las variables procesadas o grupos formados dependiendo de la sección, además la página de Datos Procesados queda activa permitiendo así retroceder a los datos procesados.

Figura A.29. Análisis Gráfico.



al seleccionar las variables o grupos y presionar el botón graficar aparecerá la página de Graficar Datos procesados con la grafica de las variables o grupos seleccionados.

Figura A.30. Análisis Gráfico Listo.




Presione el botón listo para que el sistema acepte el gráfico obtenido y sea agregado al navegador de proyecto.