Universidad Central “Marta Abreu” de Las Villas

Facultad de Matemática, Física y Computación

Trabajo para optar por título de

Máster en Ciencia de la Computación

Título

Adecuación a un procedimiento de minería de datos para guiar la categorización no supervisada

Autor

Lic. Ciro Rodriguez León

Tutora

Dra. María Matilde García Lorenzo

2016

Resumen

Cada día la cantidad de datos generados por las sociedades modernas aumenta extraordinariamente. Las potencialidades de su correcto procesamiento y conversión en información relevante, son enormes. Con el objetivo de guiar este proceso y hacerlo más sencillo se han creado varios procedimientos de minería de datos. Estos procedimientos son de propósito general, por tanto, están diseñados para ser usados en una amplia gama de problemas y ninguno contiene técnicas y/o algoritmos que se ajusten a circunstancias específicas. Entonces, procesos importantes como la categorización de instancias no supervisadas, en conjuntos de datos tipo atributo-valor, son todavía complejos.

En la presente investigación las fases de un procedimiento de minería de datos conocido como CRISP-DM fueron adecuadas para la categorización de instancias no supervisadas de conjuntos de datos tipo atributo-valor. CRISP-DM fue elegida sobre otros procedimientos, como el proceso KDD y SEMMA, por ser de libre distribución, independiente de la aplicación y la más usada por los expertos en el campo.

Por último, las fases adecuadas de CRISP-DM fueron validadas mediante un caso de estudio relacionado con la diabetes mellitus tipo 2 en la provincia de Cienfuegos. Después de un estudio inicial, los pacientes fueron analizados, independientemente, por género. Los resultados mostraron tres grupos para los pacientes masculinos y cuatro para los femeninos; todos los grupos fueron interpretados como niveles de riesgo de complicaciones futuras de la enfermedad.

Abstract

Each day the amount of data generated by modern societies increases massively. The potential for its correct processing and conversion into relevant information is enormous. With the goal of guiding this process and making it simpler, several data mining proceeding have been proposed. These procedures are of general purpose; therefore, they are designed to be used in a wide range of problems and none of them contains techniques and/or algorithms that fit specific situations. So, important processes such as the categorization of unsupervised instances, in datasets of type attribute-value, are still complex.

In the present research, the phases of a data mining procedure known as CRISP-DM were particularized for the categorization of unsupervised instances from datasets of type attribute-value. CRISP-DM was picked over other existing procedures, such as the KDD process and SEMMA, for being of free distribution, independent of the application and the most used by experts in the field.

Finally, the particularized phases of CRISP-DM were validated with a study case concerning type-2 diabetes mellitus in the province of Cienfuegos. After an initial study, the patients were analyzed, independently, by gender. Results showed three clusters for male patients and four for female patients; all clusters were interpreted as risk levels for future disease complications.

Dedicatoria

Agradecimientos

Tabla de contenidos

Introducción

La cantidad de información continua creciendo; sin embargo, la habilidad de los humanos para procesarla y asimilarla permanece constante (Olson and Delen, 2008). Además, la información en sí misma tiene pocas ventajas, su sistematización, incorporación y utilización son los elementos que aportan su valor añadido: el conocimiento. Es necesario crear sistemas que generen conocimiento, para asegurar el uso productivo de la información y guiar una toma de decisiones óptima (Canals et al., 2003, Dalkir, 2005).

Hoy día, tanto las comunidades científicas, organizaciones y gobiernos, invierten en función del desarrollo de la gestión de la información y el conocimiento, a través de proyectos, congresos, postgrados y desarrollo de sistemas con este fin. Algunos ejemplos son: las políticas trazadas por la Unión Europea para incrementar la competitividad de una economía basada en el conocimiento[1], las acciones realizadas por la OPS/OMS en países en desarrollo[2] y las facilidades brindadas por la UNESCO para desarrollar software para el procesamiento de la información[3].

Cuba no está exenta del desarrollo de investigaciones que contribuyan a la gestión de la información y el conocimiento. Ejemplos de ello lo constituyen los trabajos realizados en el Centro de Aplicaciones de Tecnologías de Avanzada (CENATAV) de la Ciudad de La Habana, en el Centro de Reconocimiento de Patrones y Minería de Datos de la Universidad de Oriente en Santiago de Cuba, así como las investigaciones realizadas en el Centro de Investigaciones Informáticas (anteriormente CEI) de la Universidad de las Villas y la labor desarrollada por la Asociación Cubana de Reconocimiento de Patrones (ACRP)[4], entre otros muchos centros de estudios existentes en las universidades cubanas.

---

[1] Comunicación de la comisión de comunidades europeas al parlamento europeo “Información científica en la era digital”. Bruselas. 2007. http://ec.europa.eu/research/science-society/document_library/pdf_06/communication-022007_en.pdf.
[2] Newsletter BVS 056. 10 de agosto de 2006. http://newsletter.bireme.br/new/index.php?lang=pt&newsletter=20060810.
[3] CDS/ISIS de UNESCO http://hrueda-isis.blogspot.com/2007_11_01_archive.html
[4] http://acrp.cenatav.co.cu

Introducción

La limitación humana y la necesidad de técnicas inteligentes para deducir nuevo conocimiento a partir de grandes volúmenes de datos condicionan el surgimiento de un importante campo, la Minería de Datos. La Minería de Datos (MD) o KDD (*Knowledge Discovery in Databases*) como se le comenzó a llamar a inicios del año 1996, se define como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, de grandes cantidades de datos almacenados en distintos formatos. La MD se desarrolla por el reconocimiento de un nuevo potencial: el valor de la gran cantidad de datos almacenados informáticamente en los sistemas de información de instituciones, empresas, gobiernos y particulares (Gorunescu, 2011).

La MD permite que los datos pasen de ser un "producto" a ser una "materia prima" que hay que explotar para obtener el verdadero "producto elaborado", el conocimiento (Cios et al., 2007). Dado que la MD excede la capacidad humana para el análisis de grandes volúmenes de datos, la utilización plena de los datos almacenados depende del uso de técnicas del descubrimiento del conocimiento (Molina and García, 2006), entre las que se encuentran las de Inteligencia Artificial.

Probablemente en pocos años, el uso de la MD se haya extendido a todas las actividades humanas complejas en las que interviene gran cantidad de datos y variables. Se podrán analizar y comprimir datos para tomar decisiones de poca importancia y servir como medio de apoyo para decisiones complejas o de gran trascendencia.

De acuerdo con la definición dada por Michalski en 1986, aprender es la habilidad de adquirir nuevo conocimiento, desarrollar habilidades para analizar y evaluar problemas a través de métodos y técnicas, así como también por medio de la experiencia propia; siendo un requisito que el resultado del aprendizaje sea entendible para el hombre (Michalski, 1986).

Dependiendo del esfuerzo requerido por el aprendiz (o número de inferencias que necesita sobre la información que tiene disponible) han sido identificadas varias estrategias de aprendizaje. Las más estudiadas y conocidas de estas clasificaciones son (Mitchell, 1997, Russell and Novig, 2009):

- **Aprendizaje por instrucción**: el sistema de aprendizaje adquiere el nuevo conocimiento a través de la información proporcionada por un maestro.
- **Aprendizaje por deducción**: partiendo del conocimiento suministrado y/o poseído, se deduce el nuevo conocimiento
- **Aprendizaje por inducción**: el sistema de aprendizaje aplica la inducción a los hechos u observaciones suministrados, para obtener nuevo conocimiento. Hay dos tipos de aprendizaje inductivo: aprendizaje con ejemplos o Supervisado y Aprendizaje por observación y descubrimiento o no Supervisado.

Luego, teniendo en cuenta la disponibilidad o no de clasificación para los casos que componen los conjuntos de datos existen dos principales estrategias: aprendizaje supervisado y aprendizaje no supervisado.

Se entiende por **aprendizaje supervisado**: cuando un algoritmo produce una función que establece una correspondencia entre las entradas y las salidas deseadas del sistema. Por tanto, las bases de conocimiento de estos sistemas están formadas por las características (rasgos) de los ejemplos y la clasificación (clases o categoría) a la que éstos pertenecen.

Por otro lado, **aprendizaje no supervisado**: es cuando todo el proceso de modelado se lleva a cabo sobre un conjunto de ejemplos formado tan sólo por entradas al sistema, es decir, sus rasgos. No se tiene información sobre las categorías o clasificación de esos ejemplos. Constituye un tipo de aprendizaje por observación y descubrimiento, donde el sistema de aprendizaje analiza una serie de entidades y determina que algunas tienen características comunes, por lo que pueden ser agrupadas formando un concepto.

Un objetivo del aprendizaje no supervisado es utilizar los datos ya clasificados o etiquetados para confeccionar herramientas o modelos que sin esta catalogación no hubiesen tenido sentido. O bien la propia división en clases de los ejemplos puede representar un resultado sustancial, debido a que pueden llevar a conclusiones a especialistas en el tema. Por ejemplo, en un grupo de pacientes que padecen la misma enfermedad, luego de un correcto agrupamiento, pueden extraerse factores de riesgo o relaciones entre síntomas que antes no se habían sospechado.

El proceso de MD involucra numerosos pasos e incluye muchas decisiones que deben ser tomadas por el usuario. Entre ellas, se dice que la adecuación de los datos para que las técnicas de descubrimiento puedan utilizarlos demanda el 70% del esfuerzo. Para organizar este proceso han surgido varias metodologías o procedimientos que lo guían. La primera de ellas fue el Proceso KDD (Fayyad et al., 1996), estructurado en las siguientes etapas:

- Comprensión del dominio de la aplicación.
- Limpieza y pre-procesamiento de los datos.
- Elección de la tarea de minería de datos.
- Elección del algoritmo(s) de minería de datos.
- Minería de datos.
- Interpretación de los patrones encontrados.

Otra de las metodologías más seguidas y referenciadas es CRISP-DM (Chapman et al., 2000). Lo cual se corrobora por una encuesta realizada por el portal para análisis de datos KDnuggets en octubre del 2014 (2014). Esta metodología es de distribución libre y resulta independiente de la herramienta que se utilice para realizar la MD.

CRISP-DM estructura el proceso en seis fases generales: análisis del problema, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue (Chapman et al., 2000). Las fases de esta metodología se descomponen en tareas generales de segundo nivel y estas a su vez en tareas específicas, aunque no especifica cómo realizar estas, es decir, no se detallan las técnicas y algoritmos a utilizar en cada problema particular que se presente, puesto que es de propósito general.

Por otro lado se tiene, con objetivos similares, el procedimiento SEMMA desarrollado por el instituto SAS[5] (*Statistical Analysis Systems*: Sistemas de Análisis Estadístico). Las siglas SEMMA vienen de las palabras en inglés: *Sample, Explore, Modify, Model, Assess* (Muestreo, Exploración, Manipulación, Modelado y Evaluación); y se refieren a su proceso de conducción al realizar la minería (Azevedo and Santos, 2008). El instituto SAS considera un ciclo con cinco etapas para el proceso:

[5] www.sas.com

1. **Muestreo**: consiste en la extracción de la mayor porción significativa de estos de manera que se contenga en ellos una cantidad significativa de información.
2. **Exploración**: consiste en detectar errores indeseados y anomalías con la meta de ganar en entendimiento.
3. **Manipulación**: consiste en la modificación de los datos mediante la creación, selección y transformación de las variables de mayor interés.
4. **Modelación**: en esta se modelan los datos por medio del software, buscando de forma automática la combinación de datos que de forma más veraz representan la salida deseada.
5. **Evaluación**: en esta última etapa se evalúa la usabilidad y confiabilidad de la información encontrada durante la minería y se estima cuán bien funciona.

Aunque el proceso SEMMA es independiente de la herramienta de MD que se escoja, está muy unido al software de SAS: *Enterprise Miner* que pretende guiar a los usuarios en la implementación de aplicaciones de MD (Azevedo and Santos, 2008). Al igual que el proceso KDD y CRISP-DM, esta metodología es de propósito general y no especifica en ninguno de sus etapas que métodos, técnicas y algoritmos son más apropiados.

Debido a que las metodologías y procedimientos de MD estudiados en la literatura son de propósito general, no contienen especificaciones como las técnicas y algoritmos a utilizar en cada problema específico; además, valorando la importancia del proceso de agrupamiento de instancias sin previa clasificación se propone como **problema científico**: las metodologías de minería de datos existentes, por su carácter general, dificultan su aplicación en la categorización de conjunto de datos en problemas no supervisados.

Para la solución de este problema se dará respuesta a las siguientes **preguntas de investigación**:

1. ¿Cuál de los procedimientos de minería de datos existentes es más idónea para adecuar a los problemas no supervisados tipo atributo-valor?
2. ¿Qué métodos de minería de datos serán más idóneos proponer para cada paso de la adecuación del procedimiento?
3. ¿Cómo validar la adecuación realizada?

Introducción

Para dar solución al problema científico se plantea como **objetivo general** de esta tesis: Adecuar un procedimiento de minería de datos que guíe la categorización de conjuntos de datos no supervisados tipo atributo – valor.

Este objetivo general fue desglosado en los siguientes **objetivos específicos**:

1. Identificar qué procedimientos de minería de datos de las existentes es más idónea para realizar su adecuación.
2. Adecuar el procedimiento escogido detallando los algoritmos y técnicas a aplicar en cada paso.
3. Validar la adecuación aplicándola a un caso de estudio.

Después de realizar el marco teórico se enuncia la siguiente hipótesis de investigación:

H1: La adecuación de un procedimiento de minería de datos para la categorización de conjuntos de datos no supervisados tipo atributo – valor provee a los investigadores de una guía que facilita los procesos de minería de datos en estos tipos de problemas.

La tesis consta de tres capítulos. El capítulo uno aborda aspectos relacionados con el marco teórico y trata los aspectos más importantes relacionados con los procedimientos de minería de datos, métodos de agrupamiento, ejemplos de distancia, métodos de pre-procesamiento, medidas de evaluación del agrupamiento y herramientas de minería de datos. El capítulo dos está dedicado a detallar la adecuación del procedimiento CRISP-DM a los problemas de categorización no supervisados en los que se enmarca la presente investigación. En el capítulo tres se presenta un caso de estudio, relacionado con pacientes que padecen diabetes tipo 2 de la provincia de Cienfuegos. El documento continúa con las conclusiones, recomendaciones para trabajos futuros, referencias bibliográficas y terminas con los anexos.

Capítulo I. Marco teórico

1 Capítulo I. Marco teórico

A lo largo de este capítulo se hace un estudio del estado del arte alrededor de las metodologías de MD y las principales técnicas y algoritmos utilizados para trabajar sobre conjuntos de datos no supervisados tipo atributo-valor.

Primero se hace un estudio comparativo sobre las tres metodologías de MD más usadas en la comunidad científica para decidir cuál de ellas será la más adecuada para realizar la adecuación. Luego se analizan los principales métodos de agrupamiento divididos en particionales, jerárquicos y basados en redes neuronales artificiales. Además, se ven ejemplos de las principales funciones de distancia, métodos de pre-procesamiento, medidas de evaluación interna del agrupamiento y, por último, algunas herramientas para realizar MD.

1.1 Análisis de metodologías de minería de datos existentes

Con el objetivo de dirigir sistemáticamente el proceso de MD se sigue una metodología o procedimiento, se llaman de estas dos formas dependiendo del contexto o bibliografía que se consulte. Existen metodologías estándares para cumplir este objetivo, tres de las más utilizadas serán analizadas en este epígrafe.

1.1.1 Proceso KDD

El proceso KDD es iterativo e interactivo como puede apreciarse en la Figura 1-1. Se dice iterativo ya que la salida de algunas fases puede retornar a pasos anteriores y porque a menudo son necesarias varias iteraciones para extraer conocimiento de alta calidad. Se habla de interactivo porque el usuario, o generalmente el experto del dominio del problema, deben ayudar en la preparación de los datos, validación del conocimiento extraído, entre otros. (Hernández-Orallo et al., 2004)

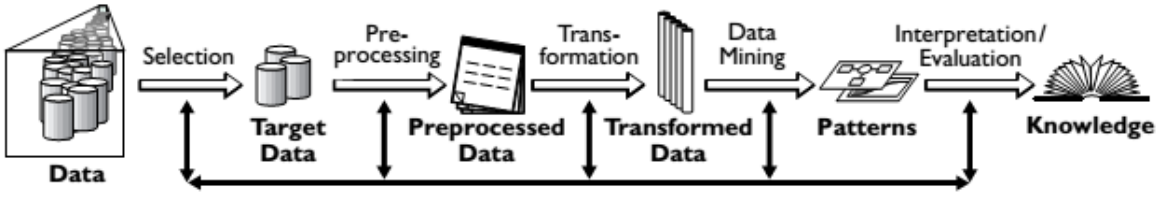


Figura 1-1. Fases del proceso KDD (Fayyad et al., 1996)

El proceso se organiza en torno a cinco fases:

- **Selección**: es donde se determinan las fuentes de información que pueden ser útiles y donde conseguirlas. Dado que los datos provienen de diferentes fuentes, pueden contener valores erróneos o faltantes.
- P**re-procesamiento**: se eliminan o corrigen los datos incorrectos y se decide la estrategia a seguir con los datos incompletos.
- T**ransformación**: se realizan trasformaciones a los datos usando métodos de transformación o reducción de dimensiones.
- M**inería de datos**: se decide cuál es la tarea a realizar (clasificar, agrupar, etc.) y se elige el método que se va a utilizar para buscar los patrones de interés.
- E**valuación e interpretación**: se valoran los patrones y se analizan por los expertos, y si es necesario se vuelve a las fases anteriores para una nueva iteración. Esto incluye resolver posibles conflictos con el conocimiento que se disponía anteriormente.

Una descripción más detallada del proceso KDD puede encontrarse en (Fayyad et al., 1996), donde los autores profundizan un poco más en las tareas a realizar en cada una de las fases, aunque con un enfoque genérico para abarcar cualquier tipo de problema de MD.

1.1.2 CRISP-DM

La metodología CRISP-DM fue creada en el 2000 por el grupo de empresas SPSS, NCR y Daimer Chrysler. Es de distribución libre lo que permite que esté en constante desarrollo por la comunidad internacional. Además, resulta independiente de la herramienta que se utilice para llevar a cabo el proceso de MD.

CRISP-DM (*Cross-Industry Estándar Process for Data Mining:* Procedimiento Industrial Estándar para realizar Minería de Datos) es ampliamente usado por los miembros de la industria. El modelo consiste en seis fases definidas de manera cíclica (Figura 1- 2) (Olson and Delen, 2008, Chapman et al., 2000):

- **Análisis del problema**: también llamada comprensión del negocio, es donde se identifica la expectativa del cliente con el proceso de MD. Durante esta etapa CRISP-DM propone el cumplimiento de tareas como la determinación de los objetivos del negocio, evaluación de la situación, determinación de los objetivos de la minería de datos y producción del plan del proyecto.
- **Comprensión de datos**: se obtiene una visión más realista del conjunto de datos del que se extrae el conocimiento. Esto es posible mediante la ejecución de las tareas: recolección de los datos iniciales, describir y explorar los datos y verificar su calidad. Como resumen de esta fase se puede realizar un esbozo gráfico de las variables categóricas con el cual se pueden llegar a conclusiones.
- **Preparación de datos**: una vez conscientes de la condición de los datos se requiere realizar acciones concretas para alistarlos para la etapa de minería. Para esto se realizan las tareas: seleccionar, limpiar, construir, integrar y formatear los de datos. La limpieza y la transformación de los datos es necesaria para luego realizar su modelado. Se debe ejecutar una exploración de los datos a un nivel más profundo que en la pasada fase.
- **Modelado**: también llamada minería de datos, es donde se cumple esta función realizando las tareas: seleccionar las técnicas de modelado, generar la prueba de diseño, construir y evaluar el modelo. Inicialmente se pueden utilizar herramientas de visualización (graficar datos y encontrar relaciones) y análisis de conglomerados (para identificar qué variables están relacionadas). Al tener una mayor comprensión de los datos, más específicos pueden ser los modelos que se utilicen.
- **Evaluación**: el resultado de los modelos debe ser evaluado en el contexto de los objetivos del negocio establecidos en la primera fase. Esto llevará a la identificación de otras posibles necesidades, que muchas veces provoca el retornar a fases anteriores. Ganar en el entendimiento del negocio es un proceso iterativo en la MD.
- **Despliegue**: la MD se utiliza tanto para comprobar hipótesis como para descubrir conocimiento (identificando inesperadas y valiosas relaciones). Mediante el

descubrimiento de conocimiento en las etapas anteriores se pueden obtener modelos que pueden ser aplicados a las operaciones del negocio para muchos propósitos, lo que incluye la predicción o identificación de situaciones claves. Estos modelos necesitan ser monitoreados pues las condiciones de operación pueden cambiar con el tiempo y el modelo puede precisar ser cambiado. Por lo que es sabio documentar todo el proceso de MD para su uso en estudios posteriores.

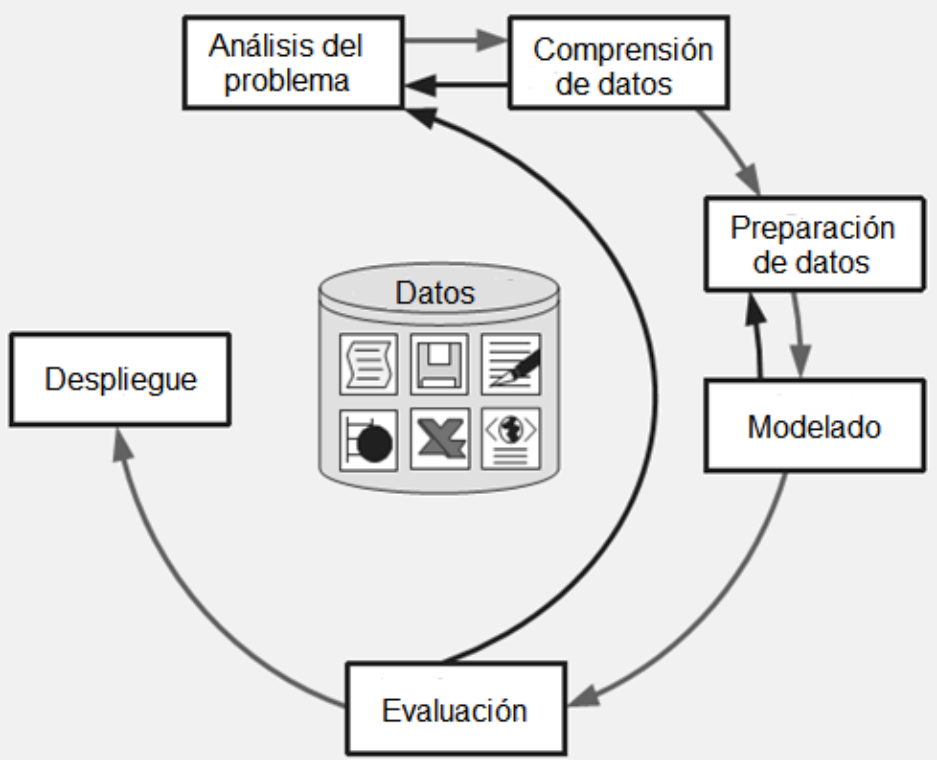


Figura 1- 2. Fases del proceso CRISP-DM (Olson and Delen, 2008)

Las seis fases no son rígidas en procedimiento, de hecho, muchas veces existe retroalimentación entre diferentes fases. Depende en gran medida la salida de una fase para saber cuál etapa o tarea de la etapa se va a seguir a continuación. Además, los analistas experimentados no necesitan seguir cada etapa del problema en cuestión. (Olson and Delen, 2008, Chapman et al., 2000)

1.1.3 SEMMA

Comenzando con una representación estadística de los datos, SEMMA pretende facilitar la exploración estadística, las técnicas de visualizar, seleccionar y transformar las variables más significativas en la predicción, modelar las variables para predecir salidas y finalmente confirmar la precisión del modelo. Una representación gráfica de este proceso se puede ver en la Figura 1- 3.(Olson and Delen, 2008)

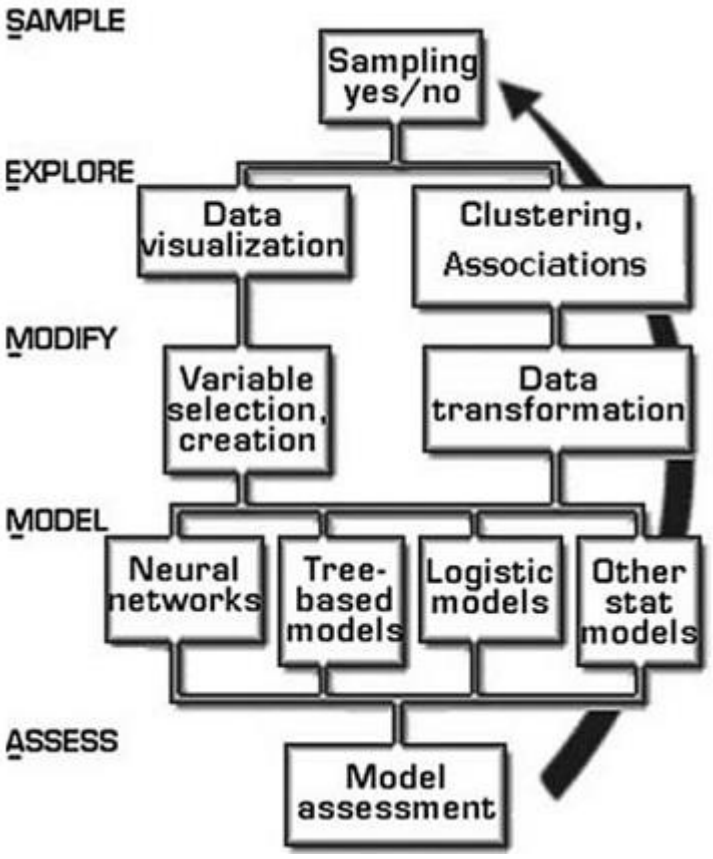


Figura 1- 3. Fases del procedimiento SEMMA.(Olson and Delen, 2008)

Evaluando la salida de cada etapa en el proceso SEMMA se puede determinar cómo modelar nuevas interrogantes surgidas por los resultados anteriores. Luego, se procederse a la fase de exploración para un refinamiento adicional de los datos. A continuación, se muestra lo que realiza SEMMA en cada una de sus fases, de forma abreviada: (Olson and Delen, 2008)

- **Muestreo**: es donde una porción de los datos (suficientemente grande como para contener información significante pero lo adecuadamente pequeños para poder manipularla rápidamente) es extraída. Para tener un costo computacional óptimo algunos (incluyendo el Instituto SAS) defienden esta estrategia, que en resumen es realizar una representación estadística del total de los datos.
- **Exploración**: es donde se buscan tendencias y anomalías que no se habían anticipado con el objetivo de ganar en entendimiento del conjunto de datos. Esta etapa permite detectar, visual o numéricamente, patrones o agrupamientos y ayuda a refinar y redirigir el proceso de descubrimiento de información.
- **Manipulación**: es donde se crean, seleccionan y transforman las variables sobre las que se concentrará el proceso de construcción de modelos. Basado en la salida de la fase anterior se puede necesitar manipular los datos para incluir información del agrupamiento o la introducción de nuevas variables. También puede ser necesario encontrar valores extremos y reducir el número de variables para solo dejar las más significativas.
- **Modelación**: en esta etapa es donde se buscar una combinación de variables que prediga confiadamente la salida deseada. Una vez que se han preparado los datos, se está listo para construir modelos que muestren los patrones en los datos. La cuestión está en encontrar el modelo adecuado para las características de los datos que se tienen.
- **Evaluación**: aquí es donde se evalúa la utilidad y la relevancia de los hallazgos del proceso de MD. En este paso final se examina todo el proceso para estimar cuan bien se desempeña. Una forma muy común de hacer esto es dejar una parte del conjunto de datos aparte, luego analizar si el modelo es válido también para este subconjunto no utilizado en su construcción.

Este procedimiento de MD ofrece un proceso fácil de entender, permitiendo un organizado y adecuado desarrollo y mantenimiento del proceso de MD. (Azevedo and Santos, 2008)

1.1.4 Comparación

Al hacer un análisis del uso, a nivel mundial, de las metodologías de MD se obtiene que CRISP-DM es la más seguida de todas. Esto se corrobora en los resultados de una encuesta mostrada por el portal para análisis de datos KDnuggets en el 2014 (2014), ver Figura 1- 4.

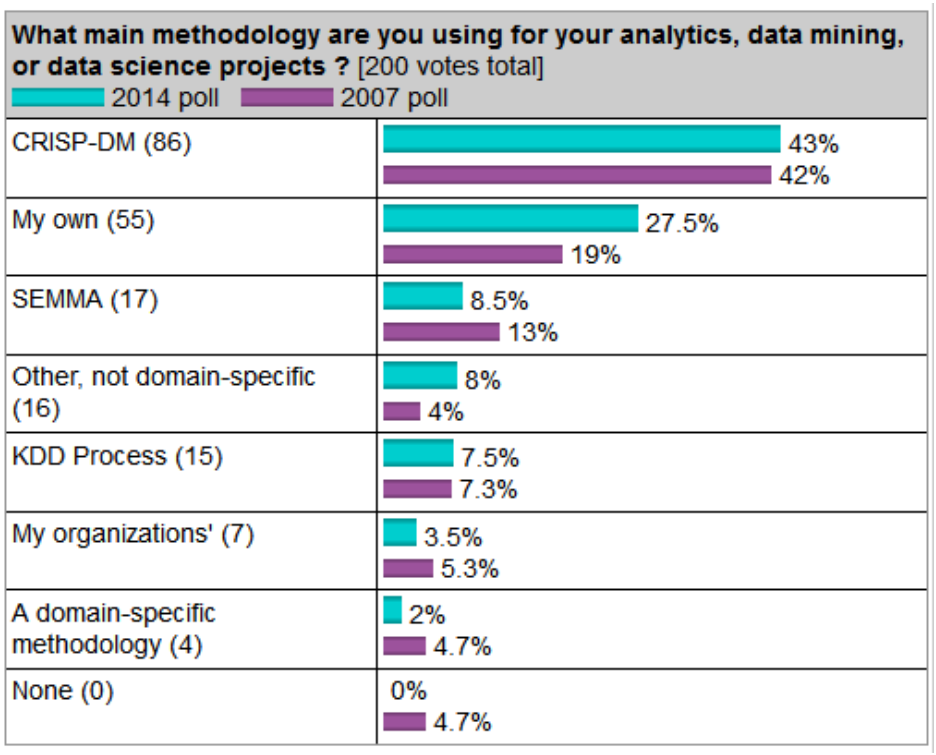


Figura 1- 4. Principales metodologías empleadas para realizar procesos de MD según una encuesta realizada por KDnuggets en el 2014.

Comparando particularmente las etapas del proceso KDD con las de SEMMA se puede, inicialmente, afirmar que son equivalentes:

- *Muestreo* puede identificarse con la de *selección*.
- Exploración puede identificarse con Pre-procesamiento.
- *Modificación* puede identificarse con *Transformación*.
- Modelación puede identificarse con Minería de Datos.
- Evaluación puede identificarse con evaluación e interpretación.

Examinando más profundamente puede afirmarse que las cinco etapas de SEMMA pueden ser vistas como una implementación práctica de las cinco fases del proceso

KDD, lo que en este caso está directamente ligado al software de SAS Enterprise Miner. (Azevedo and Santos, 2008)

Por otro lado, si se comparan las fases del proceso KDD con las de CRISP-DM estas no están en completa correspondencia como en el caso de SEMMA. Sin embargo, primeramente se observa que la metodología CRISP-DM incorpora pasos que, como ya se explicó, pueden preceder y suceder a KDD (Fayyad et al., 1996):

- La fase de análisis del problema puede ser identificada con el desarrollo del entendimiento del dominio de aplicación, el conocimiento relevante y los objetivos a seguir.
- La etapa de despliegue puede identificarse con la consolidación de este conocimiento a un sistema.

Y lo relacionado con las restantes etapas decir que:

- La fase de *comprensión de datos* puede ser identificada como la combinación de las etapas *selección* y *pre-procesamiento.*
- La fase de *preparación de datos* puede ser identificada con la de *transformación*.
- La de *modelado* puede identificarse con la de *minería de datos*.
- La etapa de evaluación puede identificarse con la de *evaluación* e *interpretación.*

Se concluye, entonces, que CRISP-DM y SEMMA pueden ser vistas como implementaciones del proceso KDD. Con un primer acercamiento se puede pensar que CRISP-DM es más completa que SEMMA, pero haciendo un análisis más profundo se concluye que las etapas que están presentes en la primera pueden ser vistas como fases implícitas dentro de la metodología SEMMA.

Las metodologías analizadas para realizar procesos de MD son de propósito general; no especifican cuáles métodos, algoritmos y/o técnicas se deben utilizar en situaciones particulares. Sin embargo, se considera a CRISP-DM más oportuna para realizar una adecuación a problemas no supervisados tipo tributo - valor pues es de libre distribución y no es dependiente de la herramienta que se utilice; además es la más usada dentro de la comunidad de científica de que realiza MD.

Una fase común en todas las metodologías es el proceso donde las técnicas de MD se ejecutan, llamadas Minería de Datos o Modelado. Muchas de estas técnicas son estudiadas bajo el nombre de *Clustering* (agrupamiento). En la próxima sección se abordan algunas de las más representativas de este dominio.

1.2 Métodos de agrupamiento

En esta sección se estudian algunos de los algoritmos de agrupamiento más conocidos y utilizados por la comunidad científica de este dominio. Se abordan sus fortalezas y debilidades, tipos de problemas donde son más idóneos y cuáles son sus posibles resultados. Para una mejor comprensión de la clasificación de los métodos analizados se puede observar la Figura 1- 5.

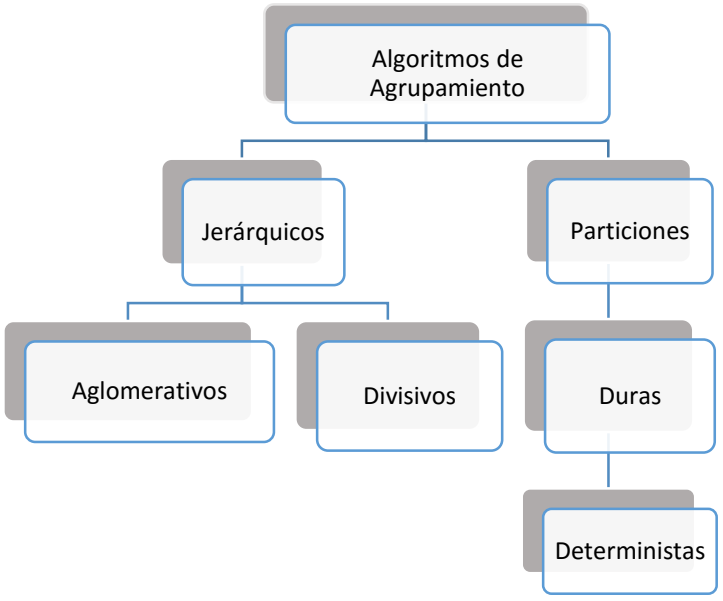


Figura 1- 5. Clasificación de los métodos de agrupamientos a estudiar.

1.2.1 Particionales

En esta sección se centra la atención en los métodos de agrupamiento que crean particiones duras y deterministas, puesto que al culminar su ejecución cada objeto del conjunto de datos pertenece a un único grupo de los conformados. Es decir, si se denota $O = O_{1,}O_{2,} \dots, O_n$ al conjunto de n objetos, se trata de dividir en k grupos $Cl_{1,}, \dots, Cl_k$ de tal forma que:

- $\bigcup_{j=1}^{k} Cl_j = O$
- $Cl_j \cap Cl_j = \emptyset \; \forall i \neq j$

Tratando de que los objetos que pertenecen a un mismo grupo sean semejantes entre sí y a la vez disjuntos a los restantes.

El método más clásico de agrupamientos es el **K-medias** (MacQueen, 1967), se considera uno de los pilares de los métodos de agrupamiento (Fielding, 2007), debido en gran medida a su fácil implementación. Funciona bien para problemas prácticos, especialmente cuando la forma de los conglomerados resultantes es compacta e híper-esférica. La complejidad computacional de este algoritmo es $O(NKdT)$, donde T es el número de iteraciones. Al ser K, d y T mucho menores que N, la complejidad temporal del método es lineal. Por lo que, K-medias es una buena opción para hacer agrupamiento a conjuntos de datos de grandes escalas, según (Xu and Wunsch, 2009).

Las limitantes de k-medias vienen dadas por su convergencia puesto que el procedimiento iterativo no garantiza convergencia a un óptimo global. Como el método puede converger a un óptimo local, diferentes puntos iniciales generalmente llevan a diferentes centroides, lo que hace importante el hecho de comenzar con una razonable partición inicial con el objetivo de encontrar soluciones de gran calidad. Sin embargo, no existe un método eficiente y universal para determinar cuál será esa partición inicial. (Fielding, 2007)

Otro detalle importante es el número K, ya que el algoritmo asume que es conocido de antemano por el usuario, lo que no es usualmente cierto en la práctica. Relacionado con la robustez se dice que el k-medias es sensible al ruido y a los objetos o puntos atípicos (*outliers*). (Xu and Wunsch, 2009).

Otro método que emplea una estrategia de partición similar al K-medias es el **K-medoides** (*k-medoids*). Los K-medoides son los puntos del conjunto de datos más representativos de cada grupo. Esta representación tiene dos ventajas: no tiene limitaciones sobre el tipo de atributo y la selección de los medoides se hace según su localización en una porción significante de un grupo, por tanto, es menos sensible al ruido que K-medias. Los algoritmos PAM (*Partitioning Around Medoids*), CLARA

(*Clustering LARge Applications*) y CLARANS (*Clustering Large Applications based upon RANdomized Search*) emplean este método. (González, 2010)

PAM (Kaufman and Rousseuw, 1990) emplea una alternativa diferente a la de centroides, en su lugar toma una instancia real perteneciente a la base de conocimiento a la que llama medoide. Para seleccionar los medoides de los grupos emplea una función de optimización. Obtiene mejores resultados que K-medias porque minimiza una suma de distancias en lugar de una suma de cuadrados; desarrolla la misma estrategia de ubicación de los puntos que K-medias, siendo más robusto ante el ruido y datos atípicos, pero es lento en conjuntos de datos grandes, lo que originó la aparición del algoritmo CLARA. (González, 2010, Rao et al., 2005)

CLARA (Kaufman and Rousseuw, 1990) se basa en muestreos. Los medoides son escogidos de la muestra usando PAM, minimiza la función de disimilaridad promedio del agrupamiento para retener los medoides en una de las muestras, de entre todas las muestras seleccionadas. (González, 2010)

CLARANS (Ng and Han, 1994) es una mezcla de PAM y CLARA, trabaja sobre muestras de la base de conocimiento. Para disminuir la complejidad, considera los vecinos de los medoides como candidatos a ser nuevos medoides e itera varias veces tomando distintas muestras en cada ciclo, con el objetivo de evitar la posible selección de malas muestras. Emplea un grafo cuyos nodos son el conjunto de k medoides, y dos nodos se conectan si difieren en exactamente un medoide. La complejidad es $O(N^2)$. (González, 2010)

Las estrategias CLARA y CLARANS tienen entre sus limitaciones la dependencia del resultado del agrupamiento del orden en que se presentan los objetos y tienden a crear grupos esféricos al igual que el K-medias.

En (Sehgal and Garg, 2014) se describen los métodos Farthest First, Basado en densidad y DBSCAN; a continuación se describen los mismos.

Farthest First (El más lejano primero) es una modificación al K-medias que localiza el centro del conglomerado en torno a un punto más que en el centro del propio grupo. Este punto debe encontrarse en el área de los datos. Este procedimiento aumenta la velocidad

del agrupamiento en la mayoría de los casos mientras menos reasignaciones y modificaciones se necesite. Las limitantes de este método son las mismas que las del K-medias (Zafar and Ilyas, 2015).

El algoritmo **basado en densidad** trata de determinar grupos basándose en la densidad de los objetos de una región. La idea clave es que por cada instancia de un grupo existe una vecindad, determinada por un radio, que debe contener al menos una cantidad mínima de instancias (Sehgal and Garg, 2014). La densidad de los puntos pertenecientes a un grupo se considera mayor que la de los que están fuera de este. Tiene la ventaja de generar grupos con una forma arbitraria y buena escalabilidad.

El algoritmo **DBSCAN** (*Density Based Spatial Clustering of aplication with Noise*, Agrupamiento Espacial Basado en Densidad de aplicación con Ruido) (Ester et al., 1996) implementa el concepto de densidad-alcanzabilidad y densidad-conectividad para definir grupos.

Dentro de los algoritmos de agrupamiento particionales existen algunos que, por su forma particular de funcionamiento, infieren o estiman la cantidad de grupos que deben conformarse. A continuación, se hace mención de algunos de ellos.

Mountain clustering (agrupamiento montañoso) (Yager and Filev, 1994) es un enfoque para estimar el centroide de los grupos sobre las bases de una función montaña. Puede ser útil para encontrar los centroides y cantidad de grupos. El método cuadricula el espacio de los datos y calcula la potencialidad de cada casilla basada en su distancia a los objetos reales. Cada cuadrícula es un centroide potencial; la cuadrícula con mejores valores es seleccionada como el primer centro y luego el valor potencial de las demás es disminuido en relación con su distancia al primer centro encontrado. Esta estructura hace que la complejidad crezca exponencialmente con el aumento de la dimensión del problema.

El algoritmo ***Expectation Maximization*** (Dempster et al., 1977) (EM) es un método iterativo para hallar la máxima similitud o máxima posteriori estimación de parámetros en un modelo estadístico, este modelo depende de variables latentes no observadas. Sus iteraciones alternan entre el paso ejecutar una expectación (E), que crea una función para

la estimación de la similitud logarítmica evaluada usando los parámetros actuales, y el paso de maximización (M), que calcula los parámetros de forma que se maximice la función hallada en el paso E.

Las mayores desventajas del EM son la sensibilidad a la selección de los parámetros iniciales, los efectos de una sola matriz de covarianza, la posibilidad de convergencia a un óptimo local y su lenta tasa de convergencia. (Xu and Wunsch, 2005)

El **Conglomerado en dos fases** es un algoritmo con escalabilidad, diseñado para manejar conjuntos de datos muy grandes. Es capaz de manipular variables continuas y discretas. En la primera fase del procedimiento, se pre-agrupan los casos en muchos pequeños sub-grupos. Luego, agrupan los sub-grupos del paso anterior en el número de grupos deseados. Si el número de grupos es desconocido, el "Conglomerado en dos fases" encontrará el número apropiado automáticamente utilizando dos criterios de conglomeración.(Bacher et al., 2004)

1.2.2 Jerárquicos

Los métodos jerárquicos crean una descomposición jerárquica de los objetos del conjunto de datos de entrada formando un dendograma. Esto no es más que un árbol que divide la base de conocimiento en muestras más pequeñas recursivamente. Este gráfico puede ser formado de dos formas, de abajo hacia arriba o desde arriba hacia abajo. Estos algoritmos combinan o dividen los grupos existentes creando una estructura escalonada que refleja el orden en el que los grupos van siendo tratados.

El esquema de abajo hacia arriba (***bottom up***), también llamado método aglomerativo, comienza con cada objeto siendo parte de un grupo independiente. Luego va mezclando secuencialmente los objetos o grupos de acuerdo con alguna medida. El otro esquema (***top down***), también llamado divisor, comienza con todos los objetos en un grupo. En cada iteración sucesiva los grupos son divididos en conglomerados más pequeños de, acuerdo a alguna condición de parada. (Sehgal and Garg, 2014)

Cobweb (Fisher, 1987) genera un dendograma de agrupamiento llamado árbol de clasificación que caracteriza cada grupo mediante una descripción probabilística. Usa una medida de evolución heurística llamada categoría de utilidad para guiar el proceso

de construcción del árbol. La técnica va incorporando incrementalmente objetos al árbol de clasificación con el objetivo de obtener valores más altos de la categoría de utilidad.

BIRCH (*Balanced Iterative Reducing and Clustering using Herarchies*, balanceo iterativo para reducir y agrupar usando jerarquías) es un algoritmo de agrupamiento especialmente adecuado para grandes bases de conocimiento. Este método incremental y dinámicamente agrupa puntos de entradas multidimensionales para tratar de producir la mejor calidad de agrupamiento con los recursos disponibles (memoria disponible y restricción de tiempo). (Zhang et al., 1996)

BIRCH normalmente puede encontrar buenos grupos con solo un escaneo de los datos y con unos cuantos más mejora la calidad de estos. Además, maneja el problema del ruido de manera efectiva. La idea de funcionamiento de BIRCH es concentrarse en las porciones densamente ocupadas y usar un resumen compacto. Estudios experimentales han demostrado un mejor desempeño de este método con respecto al algoritmo CLARANS. (Zhang et al., 1996)

1.2.3 Redes neuronales artificiales

Las redes neuronales artificiales se han desarrollado en analogía con el funcionamiento de las redes biológicas del cerebro humano. Se pueden definir como un grafo dirigido donde a cada nodo se le asocia una variable de estado y una función de activación para determinar su salida; a cada arista se le define un peso. Existen varias arquitecturas como monocapas, multicapas y recurrentes. También pueden clasificarse según su aprendizaje en supervisadas y no supervisadas. A continuación, se exponen algunas de las más empleadas en las tareas de agrupamiento.

***Learning Vector Quantization* - LVQ** (Kohonen, 1986). Se introduce cada uno de los ejemplos de entrenamiento en la red. Para cada uno se calcula la neurona ganadora de la capa de competición. En esta capa están los prototipos de las clases que producirá como salida la red. Así pues, la neurona ganadora se corresponderá al prototipo asignado a la instancia de entrada, que es la clase a la que se le hará pertenecer. Luego se realizará el aprendizaje para esta neurona ganadora y las de su vecindad. Finalmente, al alcanzar

algún criterio de parada se agrupan los objetos usando por su cercanía a los prototipos creados. (Hernández-Orallo et al., 2004, Xu and Wunsch, 2009)

***Self-Organizing Maps* - SOM** (Kohonen, 1995) El espacio de características se basa en la "disposición física" de las neuronas de salida para modelar algunas características del espacio de entrada. En particular, si dos entradas, x_1 y x_2 están próximas entre sí con respecto a alguna medida en el espacio de entradas, y causan la activación de las neuronas de salida y_a e y_b respectivamente, entonces y_a e y_b deben estar próximas entre sí respecto a algún tipo de composición o disposición de las neuronas de salida. Más aún, puede decirse que lo opuesto se debe mantener. Es decir, si y_a e y_b están próximas en la capa de salida, entonces las entradas que las producen deben estar también próximas en el espacio de entrada (Hernández-Orallo et al., 2004).

Se recomienda la normalización de los datos puesto que el vector de referencia resultante tiende a tener el mismo rango dinámico. Esto puede mejorar la precisión numérica. Sus mayores potencialidades están en el procesamiento de datos de alta dimensión. Entre sus mayores desventajas está que no es garantizada su convergencia y que muchas veces su resultado es dependiente de la secuencia de los datos de entrada. (Du, 2010)

Teoría de resonancia adaptativa (*Adaptive resonance theory*, ART) (Carpenter and Grossberg, 1987b, Carpenter and Grossberg, 1988) La mayor ventaja de este modelo es su habilidad para adaptarse sin olvidar lo aprendido en el pasado, para de esta forma sobreponerse al llamado dilema de la estabilidad plástica de las demás redes competitivas. (Du, 2010)

ART1 (Carpenter and Grossberg, 1987b) que solo maneja patrones binarios como entrada. Esta puede ser extendida a entradas arbitrarias utilizando un mecanismo de codificación. **ART2** (Carpenter and Grossberg, 1987a) extiende la aplicación a entradas con patrones analógicos. Por último **ART3** (Carpenter and Grossberg, 1990) introduce un mecanismo organizado del proceso biológico para lograr más eficiencia en la búsqueda paralela en las estructuras jerárquicas

Diferentes métodos de agrupamiento la base de su funcionamiento se basa en buscar los objetos similares del conjunto de datos, de ahí que resulte necesario definir la distancia o disimilitud más apropiada. Estas funciones de distancia reflejan el grado de cercanía o separación de los objetos y deben corresponderse con las características que se creen distinguen a los grupos contenidos en los datos. En muchos casos esta característica es dependiente de los datos o el contexto del problema en cuestión, y no existe una función que universalmente sea la mejor para todo tipo de problema de agrupamiento. Por lo que es crucial, para el buen desempeño de una técnica de este tipo, la elección de una correcta función de distancia. (Huang, 2008, Pandit and Gupta, 2011)

1.3 Ejemplos de distancias

1.3.1 La distancia euclidiana (Pandit and Gupta, 2011).

$$d(i,j) = \sqrt{\sum_{l=1}^{n} (x_{il} - x_{jl})^2} \; ; \; n \in N$$

Se dice que los grupos formados con esta distancia son imposibles de trasladar o rotar en el espacio de rasgos. Además, si las variables son medidas en unidades que son muy diferentes los rasgos con valores y varianza altos tenderán a dominar sobre los demás. Transformaciones lineales o de otro tipo pueden causar también distorsión en la relación de distancia. Una posible solución a este problema es la normalización de los datos haciendo que cada variable contribuya igualmente a la función de distancia (Xu and Wunsch, 2009).

1.3.2 Manhattan

Se dice que tiende a crear un agrupamiento híper-rectangular (Xu and Wunsch, 2005). Su definición es:

$$d(i,j) = \sum_{l=1}^{n} \left| x_{il} - x_{jl} \right| \; ; \; n \in N$$

1.3.3 Minkowski

Cuando es utilizada hay una tendencia a prevalecer los de los rasgos con valores y varianza elevados (Xu and Wunsch, 2005). Su definición es:

$$d(i,j) = \sqrt[p]{\sum_{i=1}^{n}(x_{il} - x_{jl})^p} \; ; \; n, p \in N$$

Cuando $p = 1$ es la distancia de Manhattan y cuando $p = 2$ es la euclidiana.

1.3.4 Tchebyschev

Esta simplemente busca la discrepancia más grande entre una de las dimensiones (Hernández-Orallo et al., 2004).

$$d(i,j) = max_{l=1,\dots,n}\left|x_{il} - x_{jl}\right|; n \in N$$

1.3.5 Coseno

Evalúa la similitud entre dos vectores con el valor del coseno del ángulo comprendido entre ellos. Es la más común utilizada en minería de textos como un indicador de cohesión de los grupos, puesto que es independiente del tamaño de dicho vector. Su definición es:

$$S(i,j) = \frac{X_i^T X_J}{\left||X_i|\right| \left||X_j|\right|}$$

(Singhal, 2001, Tan et al., 2006).

1.3.6 Mahalanobis

Esta distancia, considerada una métrica, es robusta pues utiliza una matriz de covarianzas. Supone una misma distribución probabilística a las variables u objetos que mide (Hernández-Orallo et al., 2004, Portillo and Mendoza, 2008). Su definición es:

$$d(i,j) = \sqrt{(X_i - X_j)^T S^{-1}(X_i - X_j)}$$

Donde S es la matriz de covarianza.

1.3.7 Hamming

La distancia de Hamming es la cantidad de posiciones, en una cadena de igual longitud, donde los símbolos correspondientes sean diferentes. Es recomendado para variables con valores discretos y preferentemente binarios. Su definición formal es:

Sean x, y dos vectores de valores discretos, la distancia de Hamming denotada como $dH(x, y)$, el número de lugares donde x, y son diferentes (Pandit and Gupta, 2011).

Puede ser interpretada como el número de caracteres que deben ser cambiados de una cadena para ser convertida en la otra.

1.3.8 Coeficiente de correlación de Pearson

Es una medida de la relación lineal entre dos variables aleatorias cuantitativas.

$$r_{ij} = \frac{\sum_{i=1}^{n} x_i y_{i-} \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\sum_{i=1}^{n} {x_i}^2 - (\sum_{i=1}^{n} x_i)^2}\sqrt{\sum_{i=1}^{n} {y_i}^2 - (\sum_{i=1}^{n} y_i)^2}}$$

Pero cuando este coeficiente es usado como medida de distancia lo importante no es el signo sino el valor modular

$$D(x_i, x_j) = (1 - r_{ij})/2$$

En (Giancarlo et al., 2010) se dice que esta medida es muy apropiada para usar con los mapas auto-organizados.

1.3.9 Índice de Jaccard

El Índice de Jaccard, también conocido como coeficiente de similitud de Jaccard, entre dos conjuntos muestra es el radio del tamaño de su intersección con el de su unión (Gorunescu, 2011):

$$J(A, B) = \frac{||A \cap B||}{||A \cup B||}$$

Por su semántica es lógico pensar que su uso más adecuado es para los casos donde los rasgos son discretos.

1.3.10 Aprendiendo la función de distancia

En (Brown et al., 2012) se confecciona un procedimiento para a partir del comportamiento de los datos confeccionar una función de distancia que sea la que mejor se ajuste a estos. El proceso consta primeramente del análisis del gráfico bidimensional de los datos, luego el experto elige algunas instancias que son semejantes entre sí y algunas que no lo son, utilizando alguna métrica ya existente. Con esta información se hace un aprendizaje de una nueva función de distancia más acorde al problema.

Esta opción de construcción de funciones de distancia o similitud es el tema de muchas investigaciones como son los casos de (Bar-Hillel et al., 2005), (Rosales and Fung, 2006), (Weinberger et al., 2005), (Weinberger and Saul, 2008), (Yang and Jin, 2006), (Ying et al., 2009) y (Ying and Li, 2012).

Existen muchas otras funciones de distancias, solo se han analizado algunas de las más comunes y utilizadas por las técnicas agrupamiento. En (Deza and Deza, 2006) se expone un compendio detallado de una gran cantidad de funciones de distancias con sus definiciones matemáticas, semántica, entre otros detalles.

1.4 Métodos de pre-procesamiento

Hoy en día los conjuntos de datos del mundo real son altamente susceptibles al ruido, los valores perdidos y la inconsistencia. Esto debido los enormes tamaños de los volúmenes de datos y/o a que su origen, generalmente, proviene de múltiples fuentes. Datos de baja calidad conllevaran a un proceso de minado de baja calidad. La calidad de un conjunto de datos es afectada por el número de valores perdidos, la inconsistencia, el grado de imprecisión de las fuentes, entre otros. Por lo que existen varias técnicas de pre-procesamiento, a continuación, se analizan algunas de las más relevantes relacionadas con los problemas no supervisados tipo atributo-valor.

1.4.1 Tratamiento de valores perdidos

Existen varias opciones y tratamientos en caso de que la muestra tener valores perdidos. Cuando una variable presenta un porciento elevado de estos puede valorarse eliminar esta variable de la muestra; el mismo tratamiento puede hacerse en el caso de las

instancias; esta es llamada: **variante de eliminación** (*case deletion CD*). Otra alternativa es reemplazar los valores perdidos por la media, en el caso de variables continuas, o por la moda, en el caso de las nominales; llamada **imputación de la media** (*Mean Imputation, MI*)(Cios et al., 2007). Muy parecida a esta última es la imputación de la mediana (*Median Imputation, MDI*) sustituyendo por la mediana el valor ausente. (Acuna and Rodriguez, 2004).

Existen además métodos que estiman estos valores perdidos como la **imputación de KNN** (*KNN Imputation, KNI*) (Batista and Monard, 2002) y el ***EMImputation*** descrito en (Schafer, 1997).

1.4.2 Datos ruidosos

Los valores ruidosos o extremos pueden presentarse de varias formas en los datos, en dependencia de ello se clasifican en:

- Tipo A (Punto extremo): un punto puede ser considerado anómalo respecto al resto de los datos. Es una de las formas más simples de valores atípicos.
- Tipo B (Extremo contextual): si una instancia de los datos es una rara ocurrencia respecto a un contexto específico de los datos y es normal respecto a otro.
- Tipo C (Extremo colectivo): es el caso en el que una instancia de los datos individualmente no es anómala, pero de conjunto con la totalidad de los datos sí es un extremo.
- Tipo D (Extremo real): estas son las observaciones ruidosas que son de interés en el sistema que se analice. Tienen alguna utilidad que ayuda a los analistas a hallar algo nuevo o innovador y si se eliminan, de alguna forma, se pierde esta posibilidad. Por tanto, no se deben tratar bajo el concepto de ruido sino el de extremos reales.
- Tipo E: (Extremo de error): es el caso en el que alguna observación es denominada incorrectamente como extremo, debido a algo inherente al problema en cuestión o fallo. En otro análisis saldrían como parte normal de los datos.

Existen dos partes en el proceso de lidiar con los datos ruidosos: encontrarlos y tratarlos.

Para encontrarlos existen varias estrategias: (Hodge and Austin, 2004, Malik et al., 2014)

- Estadística: es la primera para hallar valores atípicos (Bamnett and Lewis, 1994). Algunos de las iniciales solo son usadas para conjuntos de datos de dimensionalidad simple como (Grubbs, 1969) y (Rousseeuw and Leroy, 1996). Probablemente la más usada sea la descrita en (Laurikkala et al., 2000), esta usa una caja de graficado informal encontrar los extremos tanto en datos univariados como multivariados.
- Basadas en proximidad: son simples de implementar y no asumen previamente distribución alguna en los datos. Sin embargo, sufren de un crecimiento exponencial en la complejidad de los datos pues su principio está fundado en la distancia entre las instancias, por lo que variantes del algoritmo k-NN (k vecinos más cercanos, (Altman, 1992)) son usadas para su funcionamiento. En (Ramaswamy et al., 2000) se introduce un k-NN optimizado para producir una lista ordenada de posibles extremos.
- Paramétricas: permiten al modelo evaluar muy rápidamente nuevas instancias y son usables en grandes conjuntos de datos, la complejidad solo aumenta en proporción al modelo no al tamaño de los datos. Sin embargo, están limitadas a que los datos deben ajustarse a una determinada distribución; en caso de saberlo con antelación son altamente exactas. Ejemplo de ello lo es la estimación *Minimum Volume Ellipsoid* (MVE) (Rousseeuw and Leroy, 1996), que ajusta la menor volumen elipsoide permisible alrededor de la distribución de los datos.
- No paramétricas: en contraste con la anterior, en esta técnica no se necesita asumir distribución alguna de los datos. El enfoque más importante en esta categoría lo tienen los histogramas y la función de densidad nuclear de (Fujimaki et al., 2005).

En (Hodge and Austin, 2004) y (Malik et al., 2014) se hace un estudio más detallado de estas técnicas, también se argumentan sobre otras.

Luego para el tratamiento de estos el procedimiento es muy parecido al caso de los valores perdidos, se pueden: (Hernández-Orallo et al., 2004)

- Ignorar (dejar pasar): algunos algoritmos son robustos a datos anómalos.
- Filtrar (eliminar o reemplazar) la columna: solución extrema, pero a veces existe otra columna dependiente con datos de mayor calidad. Preferible a eliminar la columna es reemplazada por una columna discreta, diciendo si el valor era normal u erróneo (por encima o por debajo). En el caso de los anómalos se puede sustituir por no anómalo, anómalo superior o anómalo inferior.
- Filtrar la fila: puede sesgar los datos porque muchas veces las causas de un dato erróneo están relacionadas con casos o tipos especiales.
- Reemplazar el valor: por el valor nulo si el algoritmo lo trata bien o por máximos o mínimos, dependiendo por dónde es el anómalo, o por medias. A veces se puede predecir a partir de otros datos, utilizando cualquier técnica de aprendizaje automático supervisado.
- Discretizar: transformar un valor continuo en uno discreto (por ejemplo: muy alto, alto, medio, bajo, muy bajo) hace que los anómalos caigan en 'muy alto' o 'muy bajo' sin mayores problemas.

1.4.3 Selección o extracción de rasgos

Los datos a menudo son representados por una alta dimensionalidad de atributos en muchas áreas (Wang et al., 2009a, Wang et al., 2009b), el problema que se trata en la presente investigación no es una excepción. En la práctica no todos los rasgos son relevantes e importantes para la tarea de aprendizaje, muchos de ellos son a menudo redundantes, correlacionados e incluso ruidosos en ocasiones (Liu et al., 2011), lo que puede traer consigo efectos negativos como sobre entrenamiento, baja eficiencia y desempeños. Se estudian a continuación algunos métodos para dar solución a esta dificultad.

1.4.3.1 Análisis de componentes principales

Una técnica muy utilizada para reducir la dimensionalidad por transformación es denominada así (*principal component analysis*, PCA) también llamada método Karhunen-Loeve. Consiste en transformar los atributos o variables originales $x_1, x_2, \dots, x_m$ en otro conjunto de estos $f_1, f_2, \dots, f_p$ donde $p \leq m$. Este proceso se puede

ver geométricamente como un cambio de ejes en la representación. Lo relevante de esta transformación es que los nuevos rasgos se generan de tal manera que sean independientes entre sí y, además, que los primeros f tengan más relevancia (contenido informacional) que los últimos. (Han et al., 2011)

1.4.3.2 Laplacian Score

Este es un método para la selección de rasgos independiente del algoritmo de aprendizaje que se utilice y puede usarse tanto para problemas supervisados como para no supervisado. El fundamento del procedimiento está en la observación de que, en la mayoría de los problemas reales de clasificación del mundo real, los datos de una misma clase están más cercanos unos de otros. La importancia de un rasgo se basa entonces por su poder de preservación local, lo que es llamado en este caso, Laplacian Score. (He et al., 2005)

1.4.3.3 Selección de rasgos multi-grupo

La selección de rasgos es en esencia un problema de optimización combinatoria, la que es computacionalmente costosa. Tradicionalmente los métodos no supervisados de selección tratan este problema seleccionando los mejores atributos de un ordenamiento, basado en alguna medida calculada independientemente para cada variable. Este proceder obvia la posible correlación entre diferentes atributos y por lo tanto no puede producir un subconjunto óptimo de rasgos. Por lo que el presente método, basado en análisis espectral de los datos y el modelo L1-regularizado, selecciona los atributos tales que la estructura multi-grupos pueda ser mejor preservada. (Cai et al., 2010)

1.4.3.4 Selección de rasgos discriminativa no supervisada

Este método trabaja bajo la suposición de que las etiquetas de las clases de los datos de entrada pueden ser predichas por un clasificador linear. Se incorpora el análisis discriminante y la $l_{2,1}$-norma de minimización a un marco de trabajo para realizar la selección de rasgos. Este algoritmo logra seleccionar los atributos más discriminantes de todo el conjunto en segundo plano. (Yang et al., 2011)

1.4.3.5 Selección de rasgos discriminativa no negativa

NDFS, por sus siglas en inglés, para utilizar la información discriminativa en los escenarios no supervisados lleva a cabo un agrupamiento espectral para encontrar las etiquetas en los datos de entrada, durante el cual realiza la selección de rasgos de manera simultánea. Este procedimiento le permite al método seleccionar los atributos más discriminantes del conjunto total. Además utiliza la $l_{2,1}$-norma de minimización para reducir los atributos redundantes e incluso con ruido. (Li et al., 2012)

1.4.3.6 Selección de rasgos robusta no supervisada

Diferente a los métodos tradicionales de selección de atributos no supervisados, este usa un etiquetado realizado con pseudo-agrupamiento, este aprendizaje es logrado por la vía local mediante una matriz de factorización no negativa. Durante este proceso de etiquetado la selección de rasgos es ejecutada simultáneamente por norma de minimización $l_{2,1}$. Posee todas las ventajas de los métodos anteriores al usar la $l_{2,1}$-norma de minimización, pero además es escalable gracias al diseño de un algoritmo iterativo para resolver los problemas del método en cuanto a complejidad espacial y temporal. (Qian and Zhai, 2013)

1.5 Medidas de evaluación del agrupamiento

Son muchas las técnicas desarrolladas para la validación de los conglomerados. Estas estrategias son denominadas índices y se pueden clasificar en externos o internos. Los que utilizan una partición de referencia obtenida de manera independiente son los externos; los internos utilizan información que se obtiene a partir del propio proceso de agrupamiento.

Dado que en la problemática que define la presente investigación nunca se tiene presente una clasificación de referencia de los datos, se estudiaran los índices de validación internos. Estos permiten verificar si la estructura de los grupos producido por un algoritmo colocan adecuadamente los datos, pero usando solamente información inherente a la base de casos (Toledo, 2005). A continuación, se exponen algunos de los más conocidos y usados.

1.5.1 Davies-Bouldin

Este índice de validación trata de maximizar la distancia entre grupos a partir de minimizar la distancia entre los elementos de cada conglomerado y su centroide. Su fórmula de cálculo es: $1/k \sum_{i=1}^{k} R_i$ donde $R_i$es el máximo valor de R_{ij} (cierta medida de similitud entre los grupos C_i y C_j) para $i \neq j$ y su fórmula es $(SSW_i + SSW_j)/DC_{ij}$ y DC_{ij} es la distancia entre el centroide i y el centroide j. El **mínimo** valor es tomado como el número de grupos más apropiado.(Davies and Bouldin, 1979)

1.5.2 SDbw

Este método basa su fundamento en un criterio ampliamente utilizado para realizar la conglomeración, la compactación de los elementos de un mismo grupo (varianza intra-grupo) y la separación de los que están en otro (densidad inter-grupo). Sin embargo, en este caso, se trata de satisfacer este punto basándose en alguna suposición inicial (la localización inicial de los centros de los grupos) o parámetros de entrada (número de grupos, diámetro mínimo o número de puntos por conglomerado). Su salida es un valor real y el valor **mínimo** de este índice indicará el mejor agrupamiento. Su definición matemática y validación teórica y experimental se encuentran en (Halkidi and Vazirgiannis, 2001).

1.5.3 Dunn

Dunn (Dunn, 1974) corresponde al radio de la distancia más pequeña entre las observaciones de diferentes grupos y la distancia inter-grupo más grande. Tiene valores entre cero e infinito y debe **maximizarse** para que la agrupación sea la óptima. Semánticamente puede decirse que mide cuán compactos y separados están los conglomerados entre ellos. Es definido como: (Halkidi et al., 2001)

$$D = \min_{i=1,\ldots,m} \left\{ \min_{j=i+1,\ldots,m} \left(\frac{d(C_i, C_j)}{\max_{k=1,\ldots,m} diam(C_k)} \right) \right\}$$

donde la función de distancia entre dos grupos C_i y C_j es $d(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x, y)$ y el diámetro de un conglomerado C es definido como $diam(C) = \max_{x,y \in C} d(x, y)$

1.5.4 Calinski y Harabasz

Creado por (Calinski and Harabasz, 1974), está basado en la suma cuadrática interna y entre grupos. Utiliza las matrices de dispersión y el valor que **maximice** esta función será el candidato para especificar el número de conglomerados que se usará para clasificar los datos. Si se tiene una partición con k grupos el índice de Calinski y Harabasz se calcula de la siguiente manera (Liu et al., 2010):

$$CH(k) = \frac{tr(S_B)/(k-1)}{tr(S_W)/(n-k)}$$

S_B y S_W son las matrices de dispersión externa e interna respectivamente:

$$S_W = \sum_{k=1}^{k} \sum_{i \in P_k} (x_i - \bar{x}_k)^T (x_i - \bar{x}_k) \quad S_B = \sum_{k=1}^{k} n_k (x_i - \bar{x}_k)^T (x_i - \bar{x}_k)$$

1.5.5 Ball and Hall

Introducido por (Ball and Hall, 1965), se basa en las matrices de dispersión interna y externa. Su fórmula de cálculo es: SSW/k donde k es el número de grupos y SSW es la suma de cuadrados dentro de estos. El **máximo** valor de las segundas diferencias respecto al de la izquierda es tomado como el número de conglomerados apropiado. (Dimitriadou et al., 2002)

1.5.6 RMSSTD

Root Mean Square Standard Deviation (Raíz Media Desviación Estándar Cuadrada) (Sharma, 1996) es la varianza que mide la homogeneidad de los grupos, formalmente definida por:

$$\sqrt{\frac{\sum_{\substack{i=1..nc \\ j=1..d}} \sum_{k=1}^{n_{ij}} (x_k - \overline{x_j})^2}{\sum_{\substack{i=1..nc \\ j=1..d}} (n_{ij} - 1)}}$$

donde nc es el número de grupos, d es el número de dimensiones, n_{ij} es el número de elementos en el conglomerado i y dimensión j y $\overline{x_j}$ es el valor esperado en la dimensión j. Como el objetivo del proceso de agrupamiento es identificar grupos homogéneos el valor de RMSSTD más **pequeño** indica la mejor partición. (Halkidi et al., 2001)

1.5.7 RS

R Squared (R Cuadrado) (Halkidi et al., 2001, Sharma, 1996) es la medida de la diferencia de los grupos. Formalmente mide el grado de homogeneidad entre estos. Los valores de RS están en un rango de 0 a 1, donde 0 significa que no hay diferencias y 1 que existen diferencias significativas entre estos. (Halkidi et al., 2001, Liu et al., 2010)

1.5.8 Hartigan

Esta medida fue introducida en (Hartigan, 1975) inicialmente para detectar el óptimo número de grupos para el algoritmo K-medias. Su fórmula de cálculo es:

$$H(k) = \gamma(k)\frac{W(k) - W(k+1)}{W(k+1)}, \gamma(k) = N - k - 1$$

siendo $W(k)$ la dispersión entre grupos, definida como la suma total de los cuadrados de las distancias de los objetos a los centroides de los grupos a los que pertenecen. El parámetro γ es introducido con el objetivo de evitar el aumento de la monotonía con el incremento de k. Un **máximo** valor del índice indica mejores valores de la cantidad de grupos. (Albalate and Suendermann, 2009)

Cuando se adentra en un proceso de MD también se debe escoger una o varias herramientas donde los algoritmos y funcionalidades vistos hasta ahora estén implementados. A continuación, se analizan algunas de las más acordes al presente problema y usadas por la comunidad.

1.6 Herramientas de minería de datos

1.6.1 Orange

Es una suite de software para minería de base de datos y aprendizaje automático basado en componentes que cuenta con una fácil y potente, rápida y versátil interfaz de programación visual para el análisis exploratorio de datos y visualización. Contiene un completo juego de componentes para pre-procesamiento de datos, característica de puntuación y filtrado, modelado, evaluación del modelo, y técnicas de exploración. Está escrito en C++ y Python, bajo la licencia GPL y su interfaz gráfica de usuario se basa en la plataforma cruzada del framework Qt. (Demsar et al., 2013)

1.6.2 RapidMiner

Cubre un amplio rango de minería de datos. Además de ser una herramienta flexible para aprender y explorar la minería de datos, la interfaz gráfica de usuario tiene como objetivo simplificar el uso para las tareas complejas de esta área. Es de código abierto con licencia GNU GPL, basado en java. Posee alrededor de 400 operadores que pueden ser combinados y utiliza el lenguaje XML para describir estos y su configuración. Posee una gran cantidad de extensiones y la capacidad de construir grandes y complejos árboles de operadores.(Huang and Wu, 2008)

1.6.3 WEKA

Escrito en Java, WEKA (*Waikato Environment for Knowledge Analysis*, Entorno de Waikato para el Análisis del Conocimiento) es una conocida suite de software, de la universidad de Waikato de Nueva Zelanda. Soporta varias tareas típicas de minería de datos, especialmente pre-procesamiento de datos, agrupamiento, clasificación, regresión, visualización y características de selección. Su interfaz de usuario principal es el Explorer, pero la misma funcionalidad puede ser accedida desde la línea de comandos o a través de la interfaz de flujo de conocimientos basada en componentes. (Bouckaert et al., 2011)

Es software libre desarrollado bajo licencia GPL lo cual ha impulsado que sea una de las suites más utilizadas en el área en los últimos años. Además de su facilidad para añadir

extensiones y modificar métodos gracias a su filosofía de paquetería para lograr esto sin la necesidad de modificar el núcleo de la aplicación.

1.6.4 Knime

Es una plataforma de código abierto de fácil uso y comprensible para integración de datos, procesamiento, análisis, y exploración. Ofrece a los usuarios la capacidad de crear de forma visual flujos o tuberías de datos, ejecutar selectivamente algunos o todos los pasos de análisis, y luego estudiar los resultados, modelos y vistas interactivas. Está escrita en Java y está basada en Eclipse, hace uso de sus métodos de extensión para que los usuarios puedan añadir módulos de texto, imagen y procesamiento de series de tiempo.

1.7 Conclusiones parciales

Luego de un estudio de las principales metodologías para hacer Minería de Datos se obtiene que CRISP-DM es la más oportuna para realizar su adecuación a problemas no supervisados tipo atributo-valor. Pues se nota un alto grado de equivalencia entre las examinadas, pero la escogida sobresale por ser de distribución libre, independiente de la herramienta que se utilice y la más usada.

Se estudian técnicas de agrupamiento como el k-medias recomendado para grandes volúmenes de datos y las variantes del k-medoides que soluciona muchas de las desventajas del k-medias. Otras recomendadas para estimar la cantidad de grupos como el Mountain clustering, Expectation Maximization y el Conglomerado en dos fases. Además, algoritmos jerárquicos que con la creación del dendograma muestran el comportamiento del agrupamiento a varios niveles. También se analizaron algoritmos basados en Redes Neuronales Artificiales.

Además, se analizan métodos de pre-procesamiento de los datos, haciendo énfasis en el enfoque no supervisados. También se estudian medidas de evaluación internas de este agrupamiento como Dunn que mide la compactación entre los grupos, RMSSTD y RS que estiman la homogeneidad entre los agrupamientos, entre otros. Para determinar qué medidas de validación usar en cada problemática es importante saber el comportamiento

que se quiere lograr con el etiquetado de los datos, si se tiene se pueden usar entonces solo los índices que mejor midan esa característica (compactación, homogeneidad, etcétera) o ponderarlas por encima de los demás; si no se posee dicha información es recomendable entonces usar la mayor gama de validación posible.

Así mismo se analizaron ejemplos de funciones de distancia que pueden implementar los métodos de agrupamiento. Tales como las muy conocidas Euclidiana, Manhattan, Minkowski y Tchebyschev, dando algunas de ellas representaciones geométricas distintivas a los grupos que ayudan a formar. La función coseno muy utilizada en minería de texto, la muy robusta Mahalanobis, el Coeficiente de correlación de Pearson recomendada a usar por los mapas auto-organizados y el Índice de Jaccard adecuada para rasgos discretos son también tratadas.

Capítulo II.

Adecuación de la metodología CRISP-DM

2 Capítulo II. Adecuación del procedimiento CRISP-DM

En este capítulo se aborda la adecuación del procedimiento CRISP-DM para aplicar a problemas no supervisados tipo atributo-valor. Esta consta de seis fases: análisis del problema, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue; aunque es importante recordar que las fases no son rígidas y que pueden existir retroalimentación entre ellas.

En (Chapman et al., 2000), manual de usuario de la metodología CRISP-DM, se describe esta paso por paso. En ese documento se observa que la guía está enfocada a toda la amplia gama de problemas de MD. Por tanto, en cada fase los autores solo describen los procedimientos a realizar de manera general, pudiendo usarlos en problemas supervisados, no supervisados, de minería de texto, etcétera.

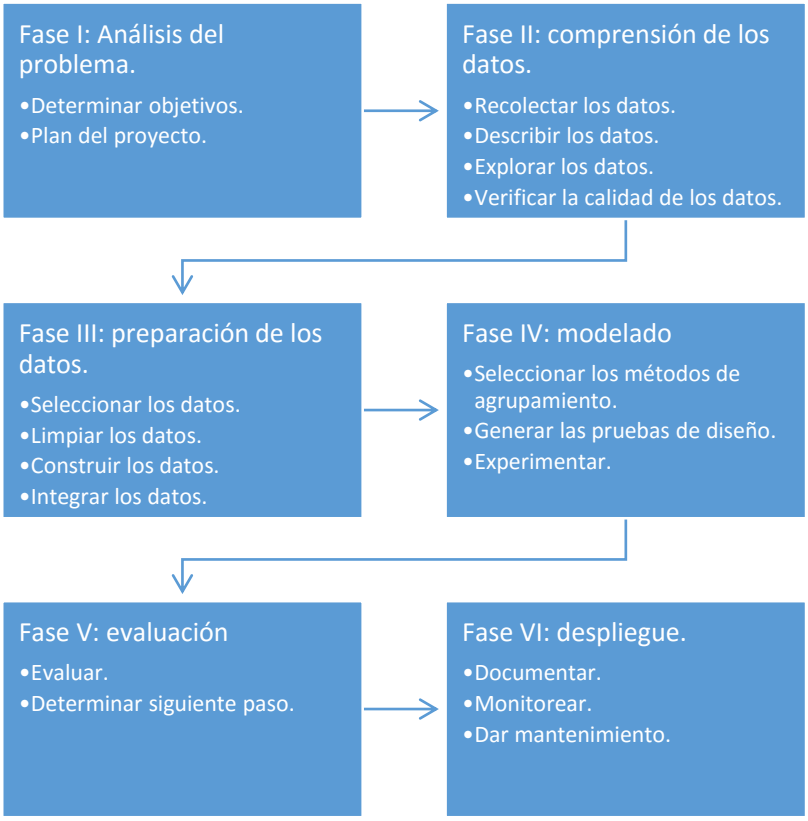


Figura 2-1. Esquema de las fases y tareas a adecuar de CRISP-DM.

Se desarrolla a continuación el contenido de cada una de las fases particularizadas para los problemas no supervisados tipo atributo-valor, en la Figura 2-1 se muestra un esquema del orden y las tareas que serán adecuadas. Se llega a un grado de especificidad más profundo en las fases II, III, IV y V; pues en las fases I y VI no se aplican técnicas o algoritmos que sean particulares para estos tipos de problemas.

2.1 Fase I: análisis del problema

En esta fase, se trata de comprender el negocio, donde se identifica la expectativa del cliente con el proceso KDD. Se **determinan los objetivos** y la producción del **plan del proyecto**. Para determinar los objetivos se debe distinguir los beneficios que se brindarán a la institución, organismo, empresa o cliente de forma general, que desea obtener información útil a partir de datos no supervisados. Es por ello, que en esta parte del proceso es imprescindible el diálogo con los expertos del área del conocimiento en que se encuentre, explicando lo que se necesita para lograr lo que ellos esperan; así como las potencialidades de los resultados a obtener.

Es necesario precisar en el comienzo de esta fase I si se tiene alguna idea de cómo quedarán conformados lo grupos o su cantidad, por hipótesis de variables que son relevantes, problemas similares que se han solucionado, etcétera; para entonces encausar el proceso hacia ese objetivo. Pero en muchas ocasiones solo se tiene la "materia prima" pero no se sabe que puede sacarse de ella, teniendo en este caso metas mucho más amplias.

Se debe definir desde el principio cómo será utilizado el conjunto de datos una vez realizado todo el proceso y categorizadas las instancias. Si con el solo hecho de haberlos dividido en grupos es ya suficiente como para brindar información relevante o esto, de forma independiente, no es suficiente y es necesario utilizarlos para crear un modelo computacional utilizando técnicas estadísticas, de aprendizaje automático, etcétera.

Con los objetivos anteriores se hace necesario identificar los recursos de los que se dispone para el proceso de MD. Evaluar la situación ayudó a tener conciencia de la realidad en la que se llevaría a cabo el proceso, las salidas de esta tarea son los recursos disponibles, tanto humanos como materiales y digitales, para realizarlo.

2.2 Fase II: comprensión de datos

Durante la comprensión de los datos se obtiene una visión más realista del conjunto de casos del que se extraerá el conocimiento. En (Chapman et al., 2000) esto se hace posible mediante la ejecución de tareas como: la recolección de los datos iniciales, describir y explorar los datos y verificar su calidad. Se analiza a continuación como realizar estas actividades en los problemas no supervisados tipo atributo-valor.

Una vez identificadas las líneas a seguir, el primer paso la **recolección de los datos iniciales**. Ya sea que se encuentren sin digitalizar, en el caso que habrá que transformarlos a este formato; o si se encuentran en algún formato digital pero que no sea en forma de tabla (cada columna un atributo y cada fila un caso) se debe llevar a este estilo.

El segundo paso es **describir los datos** de modo que se pueda identificar de manera general sus características. Es recomendable para ello la creación de una estructura como la mostrada en la Tabla 2-1, donde de conjunto con los especialistas del dominio de aplicación se pueda realizar una primera selección de las variables importantes para el problema en cuestión. Atributos muchas veces presentes como identificadores, nombres propios, direcciones, fechas, etcétera, en la mayoría de los casos carecen de relevancia para el problema. Dejando solamente las variables que evidentemente pueden influir en los patrones que se desean reconocer. De igual forma el tipo de las variables es importante para el tipo de técnica que se le pueden aplicar a la base de conocimiento que se está creando y además para saber las transformaciones que pueden hacerse más adelante a estos datos.

Identificador	Descripción	Tipo	Relevancia
Variable 1	Descripción 1	Nominal	Relevante
Variable 2	Descripción 2	Numérica	No relevante
…	…	…	…
Variable n	Descripción n	Nominal	Relevante

Tabla 2-1. Ejemplo de la descripción de las variables

También se realiza una **exploración de los datos** mediante análisis de visualización, a través de representaciones espaciales y analizar gráficos de dispersión, histogramas, por ejemplo, de algunas variables. Esta es, igualmente, una vía de **verificar la calidad de los datos**, examinando si el conjunto sobre el que se trabaja concuerda con la teoría del dominio y no posee errores. Además, analizar estadísticas como la media y la moda, de las variables declaradas como más relevantes por los especialistas, pueden indicar el si se está en el camino correcto.

Se hace necesario verificar la calidad de los datos para aplicarles técnicas inteligentes en próximas fases y proponer en la medida de lo posible soluciones que mejoren la misma.

2.3 Fase III: preparación de los datos

En esta fase se seleccionan, limpian, construyen, integran y da la forma final a los datos. Los criterios para esta selección incluyen: la importancia para el cumplimiento de los objetivos de la MD, la calidad y las restricciones técnicas como límites sobre el volumen de datos o los tipos de datos. Este proceso incluye no solo la selección de las variables sino también de los casos.

Para llevar a cabo la tarea de **selección de los datos** pueden tenerse en cuenta los métodos estudiados en el apartado 1.4.3.

Realizar una **limpieza de los datos** es la siguiente tarea; es pertinente realizar un análisis sobre los valores perdidos en la muestra, para ello se puede seguir el análisis efectuado en la sección 1.4.1. Es el momento además de aplicar filtros que modifiquen la muestra tratando que la obtenida pueda ser mejor generalizada por los métodos de agrupamiento. Los análisis para trabajar con las instancias pueden incluir buscar en la muestra valores atípicos y extremos. Una descripción de cómo tratar estas dificultades se encuentran en el epígrafe 1.4.2.

La **construcción** e **integración los datos** son tareas muy relacionadas y que involucran varias actividades. Primero se recomienda nivelar el comportamiento de las variables, para ello se estandarizan los datos numéricos, en caso de que determinados rasgos puedan tener un peso mayor que otros, simplemente porque la unidad de medida en que aparecen da lugar a puntuaciones con valores relativamente altos en comparación con las de los otros.

Pero cuando los rasgos estén en la misma escala, tales como, puntuaciones de ítem en un cuestionario o porcentajes, no es aconsejable la estandarización. Es apropiado darles un orden aleatorio a los ejemplos de la muestra pues algunos algoritmos de agrupamiento que son sensibles al orden de entrada de los datos.

Cuando se tiene el caso de variables nominales cuyos dominios tienen múltiples valores, modificar esta situación codificando valores y creando nuevas variables, siempre que en el problema no se necesiten todos estos valores e incluso pueda oscurecer lo que realmente se desea representar. Por ejemplo, se puede tener una variable nominal con cinco posibles valores, uno es que la característica no está presente y los demás cuatro son diferentes atenuaciones de esa característica cuando está presente; es posible que solo se necesite saber si el rasgo está presente o no. Entonces es el momento de codificar esta variable como una nueva con solo dos valores: presente la cualidad y no presente.

Existen casos donde debe realizarse recodificación de variables continuas, por ejemplo, es posible que varios atributos estén presentes en los datos solo porque son utilizados para calcular un valor que es realmente el centro de atención. Es recomendable entonces calcular ese valor para cada caso, añadirlo como una columna y eliminar las demás menos representativas.

Un último paso en el pre-procesamiento es la **adecuación de los datos** a los requerimientos sintácticos de las herramientas a utilizar en la etapa de modelado. Debe quedar representada la forma en que las aplicaciones representan los valores perdidos, el separador de un dato de otro y la sintaxis de los datos de forma general. Es muy recomendable que para la siguiente fase pasen varios conjuntos de datos con cada uno de los filtros que se hayan aplicado e incluyendo en estos uno con todas las variables clasificadas como relevantes; puesto que en ocasiones el filtrado, selección y extracción de casos y variables puede traer consecuencias negativas en lugar de positivas debido a pérdida de información. Experimentando luego con toda la gama de datos se está garantizando una mayor variabilidad en los resultados finales y por ende mayor calidad en la solución que se encuentre.

2.4 Fase IV: modelado

Esta constituye la cúspide en el proceso de descubrir el conocimiento. Es aquí donde los algoritmos "cavarán" y se adentrarán en los casos preparados y extraerán el preciado producto: el conocimiento. Se cumple esta función realizando tareas como: seleccionar las técnicas de modelado, generar la prueba de diseño, construir y evaluar el modelo. (Chapman et al., 2000)

La **selección de los métodos de agrupamiento** que se van a utilizar en el proceso de MD es de las tareas más difíciles del proceso y de las más determinantes también. Dependen en gran medida de la experiencia del ingeniero del conocimiento y de lo bien que se conozcan los datos; por ello si hasta el momento no se tiene idea alguna de cómo se comportan estos últimos, los análisis como graficado, análisis de distribuciones estadísticas y de cuáles variables son más influyentes son recomendables. Es válido aclarar que no hay una forma determinista de elegir las técnicas a utilizar, la variedad en la experimentación y la correcta evaluación es la que dirá la última palabra. No obstante, en esta sección se dan algunas sugerencias de cómo escoger estos métodos:

- Si no se tiene idea de la cantidad de grupos que se quieren obtener, es recomendable utilizar las técnicas que los estiman como: *mountain clustering*, *Expectation Maximization* o el Conglomerado en dos fases, revisados en la sección 1.2.1. La mayoría de estos métodos lo que hacen es ir variando el número de grupos, dentro de un rango lógico o dado por el usuario, y evaluando el agrupamiento con un índice de validación interno (1.5) en cada caso. Luego es escogido el que mejor evaluado haya sido. Por lo que, si se dispone de la implementación de medidas de evaluación internas del agrupamiento, puede lograrse con cualquier algoritmo el mismo comportamiento de manera manual, o la implementación para que este funcione de esta forma no debe ser compleja. Si las predicciones de estas técnicas no coinciden en el mismo número de grupos puede escogerse el rango de posibilidades entre la menor estimación y la mayor para comenzar a experimentar.
- Si el volumen de los datos es elevado se puede escoger algunos métodos de los mencionados en la sección 1.2 como el K-medias, BIRCH y los *Self Organizing Maps*.

- Si con las técnicas anteriores los resultados no son buenos atendiendo a las medidas de evaluación internas, una selección o extracción de rasgos más profunda puede mejorar los resultados.
- Escoger la función de distancia para los métodos de agrupamiento que mejor se ajuste, con la característica de los datos y los resultados que se desean alcanzar con el agrupamiento, es clave para cumplir los objetivos de la MD. En la sección 1.3 se abordan una variedad de funciones de distancia, especificando en cada una el tipo de agrupamiento para el que se suele usar y la forma en la que suelen quedar creado los grupos. Si luego de una amplia experimentación utilizando varias funciones de distancia y similitud los resultados no son satisfactorios, puede valorarse entonces la opción de aprender una nueva función de distancia que se ajuste mejor a los datos con los que se trabaja, varios trabajos sobre este tema se encuentran descritos en la sección 1.3.10.
- Si con los primeros acercamientos al agrupamiento los resultados, evaluados con las técnicas pertinentes, son convincentes, quizás sea momento de escoger uno o varios agrupamientos con la consideración de los expertos. Pero si no obtiene resultados promisorios se debe probar una mayor variabilidad en los parámetros y de métodos de forma general hasta alcanzar un resultado satisfactorio.

Los resultados de las técnicas seleccionadas serán, desde un acercamiento inicial, grupos arbitrarios. Algunos de ellos estarán mejor formados que otros. Por ello es necesaria la **generación de pruebas de diseño**; una manera de realizarlas es aplicar a los conglomerados formados los índices de validación explicados en la sección 1.5. La función que ejecuta los índices internos tiene como parámetros principales: una matriz de datos (donde cada columna representa un rasgo y las filas las instancias) y un vector con las etiquetas de las clases conformadas por los algoritmos.

Luego de ejecutada la validación, se debe ordenar; una forma de hacerlo es poner en las primeras posiciones los resultados con mayor cantidad de índices en primer lugar. También se puede realizar un ordenamiento por el promedio del ranking que le da cada índice de validación a cada agrupamiento. La idea es obtener, de alguna forma, una muestra del total de la experimentación como más prometedora, para mostrársela a los expertos del negocio.

Solo falta **experimentación con las técnicas y datos seleccionados**, en esta tarea es importante trabajar con cada algoritmo lo posible dentro de los parámetros lógicos que el problema muestre. Esto se logra básicamente variando los parámetros de cada una de las técnicas, para obtener varios agrupamientos con una misma elección de algoritmo. Luego seleccionar las muestras más prometedoras, como se explicó en la sección anterior, utilizando los índices de validación internos.

2.5 Fase V: evaluación

En esta fase se hace la **evaluación** de los agrupamientos seleccionados como más promisorios en la fase anterior, no debe ser únicamente el que queda en primer lugar luego del ordenamiento según los índices de validación, sino una muestra de estos. De conjunto con los especialistas en el dominio se seleccionará el que mejor solucione las necesidades existentes y por tanto se compruebe sea el más útil de la modelación realizada. Para ello pueden realizarse análisis de los centroides de los grupos, comportamientos de las variables dentro de un mismo conglomerado, sus distribuciones, etcétera.

En este punto se **determina el siguiente paso**, pueden descubrirse cuestiones importantes que pueden hacer que el proceso de MD que se viene realizando regrese a la fase de modelado, o incluso anteriores, con un enfoque diferente. Por ejemplo, puede que luego de todos los análisis realizados se tenga una idea más clara de la cantidad de grupos a construir. O que los agrupamientos encontrados no brindan información novedosa o útil al cliente. En casos donde el agrupamiento es realizado por la predominancia de una variable y los grupos creados sean muy triviales, puede pensarse repetir el proceso, ahora tomando los conglomerados construidos como conjuntos de datos independientes. También puede considerarse, en caso de tener grupos desbalanceados, el unir dos o más de estos para conformar un agrupamiento con más sentido.

2.6 Fase VI: despliegue

Con el objetivo de desplegar los resultados de la MD en esta etapa se toman todos los resultados y se selecciona estrategia que se seguirá. Es el momento de documentar todo lo realizado y unir todos los reportes que el equipo de trabajo ha realizado independientemente. Todo esto es necesario para trazar una estrategia de monitoreo y

mantenimiento del proceso. Pues pasado un determinado tiempo pueden comenzarse a cometer errores con los resultados de la minería pues el comportamiento de los datos puede variar.

2.7 Conclusiones parciales

Se realizó la adecuación del procedimiento CRISP-DM para aplicarla a los problemas no supervisados tipo atributo-valor. Se analizaron cada una de las fases del procedimiento y se particularizaron, basándose en su guía, cada una de ellas.

Durante la fase I es determinante la comunicación con los expertos y a través de ellos obtener información útil para el proceso como puede ser la hipótesis de la cantidad de grupos o cómo quiere usarse los resultados de la minería. En la fase II son relevantes actividades como una primera selección de las variables relevantes y visualizaciones de estos atributos para comprobar que sus características coincides con las teorías del dominio.

En la fase III se aplican técnicas de filtrado de las instancias y atributos a los datos originales para obtener nuevas bases de conocimiento que puedan ser mejor generalizadas por las técnicas de agrupamiento. Para ello se realizan análisis de valores perdidos (1.4.1), búsqueda de valores atípicos y extremos (1.4.2) y selección y extracción de rasgos (1.4.3). Se recomienda como salida de esta fase varios conjuntos de datos con varias combinaciones de los filtros anteriores para lograr variabilidad en la experimentación.

La fase IV es la cúspide del proceso de MD, su principal tarea es seleccionar las técnicas de modelado. Se dan algunas recomendaciones sobre cómo realizar los acercamientos iniciales utilizando los análisis realizados en la secciones 1.2, 1.3 y 1.5, recomendando una mayor variabilidad de los parámetros y cantidad de métodos en caso de no obtener resultados convincentes para los expertos. Se recomienda como pruebas de diseño validar los agrupamientos usando las medidas de evaluación del agrupamiento revisadas en 1.5 y luego hacer un ordenamiento por promedio del ranking.

En la fase V de evaluación se analizan los resultados de conjunto con los especialistas del dominio y se determina el resultado más satisfactorio. Si no sucede así, se debe volver a etapas anteriores del proceso, con hipótesis ya formuladas debido a los resultados parciales,

entonces realizar cambios para obtener una salida más pertinente al problema en cuestión. Por último, en la fase VI se documenta todo el proceso con vistas a futuras actualizaciones.

Capítulo III. Caso de estudio

3 Capítulo III. Caso de estudio

En este capítulo se aplica la adecuación realizada al procedimiento CRISP-DM, a problemas no supervisados tipo atributo-valor, a un caso de estudio real, consistente en un conjunto de pacientes diabéticos tipo 2 de la provincia de Cienfuegos. Se muestra cómo se procede en cada una de las fases tomando como guía lo expuesto en el capítulo II de este trabajo.

3.1 Fase I: análisis del problema

Como ya fue analizado, en esta fase, se trata de comprender el negocio y se identifica la expectativa del cliente con el proceso de MD. Para ello primeramente se determinan cuáles son los objetivos principales. En este caso está muy relacionado con saber cómo se puede contribuir a la problemática de la Diabetes Mellitus tipo 2 (DM-tipo2) de la provincia de Cienfuegos. Se debe entonces lograr una comprensión de cómo se manifiesta este fenómeno en el mundo y nuestro país.

La DM-tipo2 es una enfermedad que se caracteriza por el aumento de los niveles de glucosa en la sangre. Las personas con diabetes tienen una esperanza de vida reducida y una mortalidad dos veces mayor que la población general. En el mundo, la diabetes afecta a 366 millones de personas, y se estima que para el 2030 el número de afectados ascenderá a 552 millones. Esta enfermedad es la cuarta causa de muerte a nivel mundial y se supone que al menos el 50% de las personas diabéticas ignoran que lo son. Cuba no es ajena a este problema de salud mundial. La prevalencia de esta enfermedad en la población cubana se ha acrecentado en los últimos 20 años y en el año 2000 existían 263 808 diabéticos documentados y se estimaba una cantidad similar sin diagnosticar para un total de 40 por cada 1000 habitantes.

Los esfuerzos se han dirigido a la disminución de la mortalidad por diabetes, a reducir la frecuencia y severidad de las complicaciones agudas y crónicas y a mejorar la calidad de vida de los diabéticos. Además, a mejorar el conocimiento de la magnitud del problema en Cuba, desarrollar metodologías educativas para la población en general, disminuir los

costos de esta enfermedad a la sociedad y apoyar investigaciones destinadas a la prevención y control de la diabetes mellitus.

Por tanto, se propone como objetivo de minería apoyar el proceso de diagnóstico de la DM-tipo2, permitiendo que este se haga en etapas tempranas de la enfermedad o incluso antes de su debut, en las áreas de atención primaria de Cienfuegos.

Una vez determinado el objetivo del negocio es preciso expresar la meta que regirá la MD en los diabéticos tipo 2. Esta consiste en identificar grupos existentes dentro de los datos, para luego de categorizados analizar comportamientos comunes de los pacientes en cada conglomerado. Además, con el nuevo conjunto de datos, ya clasificado, pueden construirse modelos de predicción utilizando el propio resultado del agrupamiento o técnicas supervisadas.

Con los objetivos anteriores en mente se hace necesario evaluar la situación para tener una idea concreta de cómo se llevará a cabo el proceso de MD e identificar los recursos de los que se dispone para ello. Algo indispensable es el conjunto de casos, para este proyecto se obtienen de una hoja de cálculo Excel disponible en el Centro de Atención y Educación en Diabetes (CAED), donde se encuentra registrada la información de las historias clínicas de los pacientes que han ingresado en centro.

Otro aspecto importante es la elección de la herramienta de software para realizar la MD. En este caso se utiliza WEKA, pues como se ha visto en secciones anteriores es considerada idónea para realizar tareas de este tipo y además de fácil extensión, potencialidad que es usada más adelante para añadir funcionalidades necesarias en el proceso de MD.

3.2 Fase II: comprensión de datos

La primera tarea a realizar en esta fase es la recolección de datos iniciales, en este caso de los pacientes. Esto se hace a partir de una hoja de cálculo de Microsoft Office Excel 2003 donde están registradas 77 características, correspondientes a 1951 pacientes, recogidas de las historias clínicas (Anexo 1, Anexo 2 y Anexo 3) de los diabéticos que han ingresado en el CAED.

Luego se pasa a la tarea de describir los datos de una forma en la que se pueda comprender sus características principales. Para ello primeramente se realiza una descripción de las variables que se muestra en el Anexo 4. De cada característica se especifica el identificador, una breve descripción de la información que aporta a los médicos, el tipo de datos: binario (la característica está presente o no), nominal (existe una cantidad finita de valores posibles para el atributo), numérico (el valor de la característica está dentro de un rango de números reales) y la importancia del atributo para la investigación (Relevante o Sin importancia). Este último acápite fue decidido de conjunto entre los expertos del dominio y los ingenieros del conocimiento.

Del total de las variables se identificaron 37 que tienen relevancia para la investigación, básicamente son aquellos factores que los médicos identifican como de riesgo, algunos análisis que son comunes en la atención primaria y las complicaciones que presentan los pacientes. Las 40 restantes no tienen importancia para los objetivos del proceso de MD. Entre estas se encuentran variables que no aportan información sobre el debut y aquellas que son segundos valores de una misma variable (peso final, glicemia final, ppd final). El caso específico de la variable HDL-c es considerada sin importancia, aunque es relevante para los expertos por los problemas de calidad que serán posteriormente explicados.

Con la descripción de los datos se obtuvieron nociones globales sobre la muestra. Un paso que profundiza en el conocimiento de la muestra es la exploración de los datos. Durante esta tarea se utilizan las gráficas y visualizaciones de los datos, para describirlos mejor utilizando las potencialidades de la herramienta seleccionada. Este análisis se realizará sobre algunas variables consideradas como factores de riesgo de padecer la enfermedad. El principal objetivo de esta labor es comprobar la concordancia de las características del conjunto de datos con lo que se expone en la teoría y la práctica médicas para luego pasar a las siguientes fases.

Utilizando los histogramas generados con WEKA se realizó el análisis de las variables Edad (Anexo 5), donde se observa un predominio de un rango de edad entre los 40 y los 69 años. Esto concuerda con el criterio de experto que plantea que la mayoría de las personas que padecen diabetes tipo 2 son mayores de 40 años (Díaz, 2007). Estudios

similares se realizaron con las variables sexo, hábito de fumar, obesidad, antecedentes familiares de DM-tipo2 y padecimiento de hipertensión arterial (HTA) (ver del Anexo 6 al Anexo 10).

Las observaciones resultantes indican que existe un predominio del sexo femenino lo que es corroborado por la teoría médica, puesto que existen situaciones obstétricas desfavorables que constituye factores de riesgo en las mujeres y aumenta su propensión a padecer la enfermedad (Díaz, 2007). En cuanto al hábito tóxico de fumar, no se observa en los datos que abunden los fumadores; aunque los malos hábitos son un aspecto preocupante para los médicos. La obesidad, los antecedentes familiares de diabetes mellitus y el padecimiento de HTA son rasgos que predominan en el conjunto de pacientes. Esto concuerda con el criterio médico de que estos son factores de riesgo de padecer DM-tipo2.

Al concluir esta tarea se evidencia una concordancia entre el comportamiento de las variables en los datos y lo que la teoría médica expresa.

Los siguiente es verificar la calidad de los datos, pues pese a los resultados antes obtenidos, se identifican varios problemas. A primera vista se nota que existen valores perdidos en la muestra. En algunos casos la cantidad es considerable como en los años 2005 y 2006 que variables como “Modo de debut”, “Obesidad al debut”, “Antecedentes obstétricos”, “APF DM”, “APP”, “APF Otras”, “Complicaciones” y “Pie con riesgo” están completamente en blanco. Por esta razón se omiten los pacientes que ingresaron en estos años.

En cuanto a las variables, “HDL-c” solo es tomada en el año 2010 en algunos pacientes, por ser un análisis cuyo reactivo escasea con frecuencia, esto representa el 2% de la muestra y por tanto se descarta para la investigación. En una situación similar se encuentra “Reingreso” ya que solo tienen valores en esta variable en los años 2005 y 2006 para un 0.26% del total. En “Sexo” se identificaron pacientes con nombres de mujer y con valores en los antecedentes obstétricos, que tenían especificado sexo masculino; por tanto, se les sustituye el valor del sexo por femenino.

Se detectan dificultades además en las variables continuas, por los errores propios que genera la manera de registrar los datos, es la forma de delimitar las cifras decimales, ya que estas se delimitan en algunas ocasiones con una coma y otras con un punto. Esto, sumado a dejar espacios innecesarios entre la unidad y las cifras decimales. Para solucionar los problemas de calidad anteriores se corrigen sustrayendo los espacios y estableciendo como carácter delimitador de números decimales el punto.

En el caso específico de la talla se encuentran 23 celdas con algunas imprecisiones, las cuales fueron solucionadas sustituyendo los valores por aquellos que se consideraron lógicos (Anexo 11). En las variables relacionadas con el peso también se encontraron imprecisiones en peso inicial y peso final, los cuales fueron sustituidos teniendo como referencia la variable “IMC”. Todos los valores sustituidos se exponen en el Anexo 12 y Anexo 13.

Al concluir esta fase se cuenta con una idea clara de la información contenida y de las dificultades que presentan los datos. Haciéndose posible la corrección de algunas de estas dificultades, con el propósito de mejorar su calidad. No obstante, a estos problemas, es relevante la correspondencia entre el comportamiento en los datos y la teoría médica, de las variables consideradas significativas en estos pacientes.

3.3 Fase III: preparación de los datos

Una vez conscientes de la condición de los datos, se requiere realizar acciones concretas para alistarlos. En este punto del pre-procesamiento, donde ya se han identificado problemas de calidad y se conocen las características de las variables, se está en condiciones de decidir qué datos serán usados para el análisis: selección de los datos.

Las variables que se deciden utilizar para la solución del problema son aquellas que en la descripción de los datos se clasificaron como relevantes. También hay que descartar algunas instancias como los pacientes que presentan un tipo de diabetes diferente al tipo 2.

Es necesario hacer una limpieza de los datos para así elevar la calidad de los seleccionados al nivel requerido por las técnicas de agrupamiento. Para esto se solucionan los problemas de calidad descritos en la fase anterior.

Uno de los principales problemas es la cantidad de valores perdidos en la muestra, solo cerca de un 30% de los casos seleccionados tienen todos sus atributos sin valores ausentes. Es necesario, por tanto, mejorar los datos restantes para no descartarlos, teniendo en cuenta la necesidad de las técnicas inteligentes de tener una cantidad considerable de ejemplos para su entrenamiento. WEKA tiene varios filtros no supervisados de atributos que ayudan a mejorar esta situación; para el caso de las variables numéricas está el *EMImputation* que utiliza el algoritmo de maximización esperada para sustituir los valores perdidos por los que considera más adecuados. Para los atributos discretos se usa el *ReplaceMissingValues* que, en este caso, sustituye los valores ausentes por la moda.

Otro elemento que puede afectar el buen desempeño de las técnicas es la presencia en los datos de valores atípicos. Como se pretende hacer un análisis teniendo en cuenta el comportamiento típico de los diabéticos tipo 2, estos valores muy diferentes del comportamiento predominante de los datos se consideran como ruido y se descartan del análisis. Para ello se utiliza el filtro *InterquartileRange* y se eliminan del conjunto de datos.

Por último, se pasa a estandarizar las variables mediante el filtro *Standardize*, cuyo resultado es que todos los atributos numéricos pasan a tener media cero y desviación estándar uno. En este punto se cuenta con un conjunto de datos con una mejor calidad para aplicarles las técnicas de agrupamiento durante la fase de modelado. Aunque es válido aclarar que otros filtros y análisis pueden hacerse en esta etapa, pero con los realizados hasta el momento son suficientes.

Se pasa ahora a la tarea de construir los datos, donde es posiblemente necesario codificar los valores de algunas variables y crear algunas nuevas. Como es el caso de “APF DM” que contiene aquellos familiares del paciente que padecen DM-tipo2, lo que supone existan disímiles combinaciones de valores que dificultan el análisis. El factor de riesgo consiste en que el paciente presente antecedentes familiares de la enfermedad, no en quien específicamente la padece. Una transformación que soluciona el problema es codificar la variable para que tome solo dos valores: 1 si el paciente tiene al menos un familiar que padece DM-tipo2 y 0 en el otro caso.

Otras variables que presentan dificultades similares son las relacionadas con las patologías que padece el paciente (“APP”) y las que padecen su familia (“APF”), a parte de la DM-tipo2. Estas contienen para cada paciente un listado de estas enfermedades, que como en el caso anterior dificulta la correcta interpretación de la información por parte de los algoritmos. Crear nuevas variables dicotómicas, para cada una de las enfermedades que propone la historia clínica (Anexo 1), y agrupar en una variable el resto es la solución para este problema.

Las variables que se crean son “APP HTA” (Hipertensión Arterial), “APP hiperlipoproteinemia”, “APP cardiopatía isquémica”, “APP clauditación intermitente” y “APP otros” y las mismas enfermedades, pero para los Antecedentes Patológicos Familiares (APF). Cada una de estas variables toma valor 1 si el paciente o su familiar la padece, sino toma valor 0. La variable “otros” es igual a 1 si el paciente o su familia padecen otra enfermedad diferente a las que se definieron anteriormente. La descripción de estas variables y la justificación de su inclusión se pueden ver en el Anexo 14.

3.4 Fase IV: modelado

Lo primero a realizar en esta fase es la selección de las técnicas de modelado, para definir qué algoritmos de la herramienta WEKA se utilizarán para determinar los grupos en los diabéticos tipo 2. La aplicación cuenta con varias técnicas implementadas, para un primer acercamiento al problema (si los resultados no son satisfactorios se realizarán análisis más profundos) se utilizarán:

- ***EM* (Simple Expectation Maximization):** para tener una estimación inicial de la cantidad de grupos a conformar. Su funcionamiento está descrito en la sección 1.2.1.
- ***CascadeSimpleKMeans*:** estima la cantidad óptima de grupos realizando varios agrupamientos (utilizando como base el algoritmo K-medias) y escogiendo el mejor al aplicarle la medida de validación interna Calinski y Harabasz descrita en 1.5.4.
- **SimpleKMeans**: implementación del algoritmo K-medias descrito en la sección 1.2.1.

- **FarthestFirst**: variante del k-medias descrito en la sección 1.2.1.
- **LVQ**: implementación de *Learning Vector Quantization* descrito en la sección 1.2.3.
- **SelfOrganizingMap**: implementación de los mapas auto-organizados de Kohonen descrito en la sección 1.2.3.

El resultado de las técnicas seleccionas son grupos de pacientes, algunos de los cuales estarán mejor conformados que otros, por lo que es importante en este punto generar pruebas para estos resultados. Para ello se utiliza una nueva funcionalidad añadida a WEKA: *ClusterValidation*.

3.4.1 Implementación de *ClusterValidation* a WEKA

Es muy difícil realizar un proceso de MD utilizando una única herramienta, muchas veces es necesario utilizar otras para realizar ciertas tareas. En el marco teórico de esta investigación se estudia WEKA y se dice que posee facilidades de extensión. En esta sección se describe cómo se le añaden a esta herramienta cinco índices de validación interna, Davies-Bouldin (1.5.1), Hartigan (1.5.8), Calinski y Harabasz (1.5.4), Dunn (1.5.3) y Ball and Hall (1.5.5) debido a la no existencia de estos en la aplicación. Además, se añade una nueva ventana al *Explorer* de la aplicación donde se realizará el trabajo de validación propuesto.

3.4.1.1 Estructura de un paquete

Toda nueva funcionalidad que se añada a WEKA debe crearse en forma de paquete. Este no es más que un archivo compactado (*.zip*) que debe ser extraído en el directorio que crea por defecto la herramienta al instalarse: *\wekafiles\packages*; su estructura puede observarse en el Anexo 15. El fichero *Description.props* contiene información básica, como el nombre, el autor, una breve descripción, las dependencias, etcétera. Algunos campos son obligatorios y otros opcionales. El fichero *build_package.xml* también contiene información del paquete.

En el directorio *doc* se debe poner la documentación necesaria para la comprensión de la nueva funcionalidad. Es importante también que el paquete tenga al menos un compilado (archivo con extensión *.jar*) en la raíz y si es necesario para el

funcionamiento del mismo alguna otra biblioteca, se ubicará en el directorio *lib* como un compilado igualmente. Al ejecutarse WEKA, se cargan todos estos ficheros de forma automática, conteniendo entonces las nuevas funcionalidades extendidas.

3.4.1.2 Implementación de los índices de validación internos

Para implementar los índices de validación se creó la clase *Validation,* como se muestra en el Anexo 16, en la cual se agrupan los índices cada uno como una función. Se definen como variables globales los datos que se utilizan por varios de los índices, son calculadas al llamar el constructor de la clase, este recibe como parámetro instancias (*Instances*). Los principales métodos son:

- *distance(Instance x, Instance y):* calcula la distancia entre los elementos.
- *min(double[] s):* calcula el mínimo valor del arreglo "*s"*.
- *max(double[] s):* calcula el máximo valor del arreglo "*s"*.
- *getAssignments(Instances data):* devuelve un arreglo de enteros con las asignaciones de las instancias *"data"*.
- *getCentroids(Instances data, int[] assignments):* calcula los centroides de los datos representados a partir de las asignaciones de los mismos.

3.4.1.3 Implementación de la nueva interfaz gráfica de usuario

WEKA no tiene ninguna sección para realizar validación interna de agrupamiento. Es por eso que, para poder utilizar los índices implementados, desde la interfaz gráfica, se debe añadir un nuevo componente.

Para añadir un nuevo componente a la interfaz gráfica de WEKA es imprescindible que nuestra clase herede de la clase *javax.swing.JPanel* e implemente la interfaz *weka.gui.explorer.Explorer.ExplorerPanel* (Anexo 17).

Además de los directorios descritos que conforman el paquete, dentro de la raíz del mismo debe estar un fichero nombrado *Explorer.props,* el cual se obtiene del directorio de instalación de WEKA, el cual y se modifica para añadirle el nombre de nuestro nuevo componente (ver Figura 3- 1). En este fichero aparece precedido del símbolo # es comentario, la línea que añade el nuevo componente es la que está precedida por

"*Tabs=*" y a continuación la ruta del paquete. La cláusula *standalone* significa que el componente estará activado sin haber cargado instancias previamente.

```
# Explorer.props file. Adds the ValidationPanel to the Tabs key.

Tabs=weka.gui.explorer.ValidationPanel:standalone,
TabsPolicy=append
```

Figura 3- 1 Ejemplo de fichero Explorer.props.

Una vez incorporado el paquete *ClusterValidation*, que es el que contiene todas las clases mencionadas, al ejecutar nuevamente la aplicación se puede observar la nueva ventana, como se muestra en la Figura 3- 2.

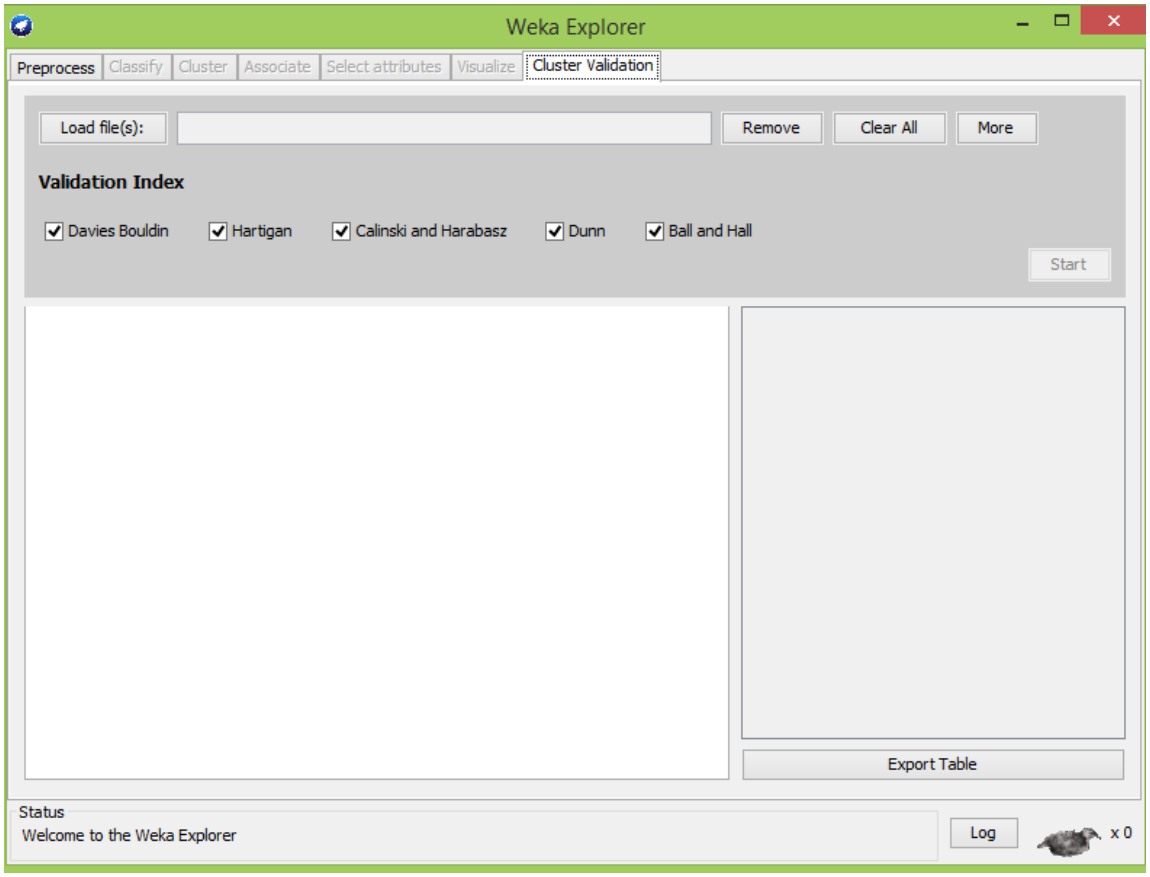


Figura 3- 2 Nueva Interfaz gráfica de usuario incorporada a WEKA.

3.4.2 Pruebas de diseño

La nueva funcionalidad añadida a WEKA evalúa los índices de validación interna mencionados a cada uno de los resultados. Luego se realizará un ranking de cada agrupamiento por cada una de las medidas y serán escogidos, para el análisis de los expertos, los que mejor promedio de este ranking hayan tenido.

El próximo paso es ejecutar las técnicas, para luego determinar la calidad de los resultados desde una mirada técnica. Para ello se realiza primeramente una experimentación con el total de los casos. Se aplica el algoritmo *EM* con el parámetro *numCLusters* en -1 para que estime la cantidad de grupos y dos variantes del *CascadeSimpleKMeans*, una utilizando la función de distancia euclidiana y la otra con la función de distancia manhattan. La estimación del primero son cuatro grupos y las dos variantes del segundo estiman dos grupos. Utilizando estos resultados se pasa a ejecutar las demás técnicas, en las que hay que especificarle manualmente la cantidad de grupos, se varía este valor entre dos y cuatro.

Luego de ejecutar los algoritmos, los resultados del ordenamiento del ranking por cada medida de validación obtenidos de se muestran en la Tabla 3-1. En la primera columna se observa el nombre del algoritmo y un número que representa la cantidad de grupos, que estima el método, si está entre paréntesis, o que se le introduce si está normal; en el caso de las técnicas en las que se varía su función de distancia, es especificado igualmente. En las restantes columnas se observan los ordenamientos del ranking que dan cada una de las medidas. Puede verse que el mejor comportamiento lo da el algoritmo *CascadeSimpleKMeans* que es el primero en la lista con un ranking promedio de 3.4 y luego el EM con uno de 4.6; en ambos casos la cantidad de grupos es dos.

P/I	Davies Bouldin	Hartigan	Calinski y Harabasz	Dunn	Ball y Hall	Promedio
CascadeSimpleKMeans - Manhattan - 2.arff	1	6	1	3	6	3.4

EM - 2.arff	2	9	2	5	5	4.6
EM - 3.arff	3	4	3	2	11	4.6
EM - 4.arff	5	2	9	1	16	6.6
SimpleKMeans - Euclidiana - 2.arff	4	13	6	11	2	7.2
CascadeSimpleKMeans - Euclidiana - 2.arff	6	12	5	10	4	7.4
CascadeSimpleKMeans - Euclidiana - 3.arff	9	10	7	4	9	7.8
SOM - 4.arff	7	1	8	7	17	8
CascadeSimpleKMeans - Manhattan - 3.arff	10	8	4	9	10	8.2
CascadeSimpleKMeans - Manhattan - 4.arff	15	3	10	6	14	9.6
CascadeSimpleKMeans - Euclidiana - 4.arff	12	7	12	8	13	10.4
SimpleKMeans - Euclidiana - 3.arff	11	11	13	12	8	11
FarthestFirst - 2.arff	8	17	15	15	1	11.2
SimpleKMeans - Euclidiana - 4.arff	17	5	11	13	15	12.2
FarthestFirst - 3.arff	16	15	16	17	7	14.2

FarthestFirst - 4.arff	14	14	17	16	12	14.6

Tabla 3-1. Resultados de la validación de los resultados con el total de los datos.

3.5 Fase V: evaluación

Al hacer un análisis de los agrupamientos que quedaron en primer y segundo lugar, su resultado consiste en poner en un conglomerado a los pacientes que son hombres y en otro a las mujeres. Incluso analizando los resultados que se encuentran en los lugares tercero y cuarto, donde los agrupamientos fueron de 3 y 4 grupos respectivamente, se observa que hay un gran desbalance de las clases obtenidas (ver Anexo 18 y Anexo 19), la mayoría de los elementos se agrupan en dos grupos y estos están determinados en su mayoría por el sexo igualmente.

Al presentar los resultados a los expertos se determina que este agrupamiento no aporta información útil para el diagnóstico temprano de la diabetes. Entonces, basado en lo estudiado en la presente investigación, que dice puede haber retroalimentación entre una fase posterior con una anterior; además con lo analizado en la sección 2.5, se determina dividir el conjunto de datos en dos, un conjunto de hombres y uno de mujeres. A estos nuevos conjuntos se les aplicará las fases IV y V, respectivamente, para tratar obtener clases útiles para el diagnóstico temprano de la DM-tipo2.

3.6 Fase IV: modelado (datos divididos por sexo)

Las técnicas seleccionadas, las pruebas de diseño para validar el agrupamiento y la experimentación con los datos se realizan de la misma forma que la pasada vez que se realizó esta fase.

Se realiza entonces el análisis con el grupo de pacientes **masculinos**. Para ello se creó un nuevo conjunto de datos donde solo están los hombres y entonces se eliminan las variables sexo y las correspondientes a los antecedentes obstétricos. Se aplican a los datos los algoritmos *CascadeSimpleKMeans* y *EM* con el propósito de determinar la cantidad óptima de grupos. El primero conformó dos grupos y el segundo estimó cuatro. Luego, al igual que el caso anterior, se aplican las demás técnicas variando la cantidad

de grupos entre dos y cuatro. Los resultados del ordenamiento realizado por los índices de validación interna se muestran en la Tabla 3-2.

P/I	Davies Bouldin	Hartigan	Calinski y Harabasz	Dunn	Ball and Hall	Promedio
SOM - 2.arff	3	5	1	1	8	3.6
SimpleKMeans - Euclidiana - 2.arff	5	9	2	3	2	4.2
SimpleKMeans - Manhattan - 2.arff	6	10	3	4	3	5.2
EM - 3.arff	1	23	8	2	1	7
LVQ - 3.arff	4	2	9	6	15	7.2
LVQ - 2.arff	7	6	10	8	9	8
LVQ - 4.arff	2	1	16	5	23	9.4
CascadeSimpleKMeans - Euclidiana(2).arff	17	17	4	12	4	10.8
EM - 2.arff	9	19	6	18	6	11.6
CascadeSimpleKMeans - Manhattan(2).arff	18	18	5	13	5	11.8
FarthestFirst - 2.arff	8	20	7	17	7	11.8
CascadeSimpleKMeans - Manhattan - 3.arff	12	12	11	10	14	11.8

CascadeSimpleKMeans - Euclidiana - 3.arff	11	13	12	14	10	12
SOM - 4.arff	10	4	18	7	22	12.2
EM(4).arff	15	3	17	11	16	12.4
CascadeSimpleKMeans - Manhattan - 4.arff	14	8	20	9	18	13.8
CascadeSimpleKMeans - Euclidiana - 4.arff	13	7	19	15	17	14.2
SimpleKMeans - Euclidiana - 3.arff	19	15	13	19	13	15.8
SimpleKMeans - Manhattan - 3.arff	20	16	14	20	12	16.4
SimpleKMeans - Manhattan - 4.arff	16	14	22	16	21	17.8
SimpleKMeans - Euclidiana - 4.arff	21	11	21	21	19	18.6
FarthestFirst - 3.arff	23	22	15	23	11	18.8
FarthestFirst - 4.arff	22	21	23	22	20	21.6

Tabla 3-2. Resultados de la validación de los resultados en los hombres.

El experimento en las mujeres se hace con los casos restantes y con el total de variables. De la misma forma se aplican los algoritmos *CascadeSimpleKMeans* y *EM* para tener un estimado de la cantidad de grupos. El primero estima dos grupos y el otro cuatro, por tanto, se procede a variar el valor de la cantidad de grupos, entre las demás técnicas, entre esos valores. Los resultados son mostrados en la Tabla 3-3. .

P/I	Davies Bouldin	Hartigan	Calinski y Harabasz	Dunn	Ball and Hall	Promedio
SOM - 2.arff	2	6	2	7	2	3.8
FarthestFirst - 2.arff	4	13	3	5	1	5.2
SOM - 4.arff	3	2	9	2	15	6.2
CascadeSimpleKMeans - Euclidiana(2).arff	6	14	4	10	5	7.8
LVQ - 4.arff	1	1	16	1	24	8.6
CascadeSimpleKMeans - Manhattan(2).arff	7	15	5	11	6	8.8
LVQ - 3.arff	5	3	17	3	17	9
CascadeSimpleKMeans - Euclidiana - 3.arff	11	12	12	8	9	10.4
EM(4).arff	8	4	18	6	16	10.4
CascadeSimpleKMeans - Manhattan - 3.arff	12	11	11	12	12	11.6
LVQ - 2.arff	24	5	1	24	8	12.4
SimpleKMeans - Euclidiana - 4.arff	10	7	19	9	20	13
SimpleKMeans - Euclidiana - 2.arff	19	22	6	16	3	13.2
EM - 3.arff	18	8	10	20	13	13.8

SimpleKMeans - Manhattan - 2.arff	20	23	7	17	4	14.2
SimpleKMeans - Manhattan - 3.arff	14	17	13	19	11	14.8
FarthestFirst - 3.arff	16	18	14	14	14	15.2
SimpleKMeans - Euclidiana - 3.arff	15	19	15	18	10	15.4
SimpleKMeans - Manhattan - 4.arff	13	10	21	15	19	15.6
EM - 2.arff	23	24	8	23	7	17
FarthestFirst - 4.arff	17	16	22	13	23	18.2
CascadeSimpleKMeans - Euclidiana - 4.arff	22	20	23	22	21	21.6
CascadeSimpleKMeans - Manhattan - 4.arff	21	21	24	21	22	21.8

Tabla 3-3. Resultados de la validación de los resultados en las mujeres.

3.7 Fase V: evaluación (datos divididos por sexo)

Nuevamente se comprueba si son útiles o no los modelos resultantes para el diagnóstico temprano de la DM-tipo2, verificando el cumplimiento de los objetivos de la MD. Durante la última experimentación se dividió el conjunto de datos para aplicar las técnicas a los pacientes masculinos y a los femeninos, obteniéndose conglomerados en ambos grupos. Por esta razón los resultados se analizan de manera independiente en los conjuntos.

3.7.1 Análisis en los hombres

Los tres primeros resultados del ordenamiento realizado por los índices de validación (SOM – 2, SimpleKMeans - Euclidiana – 2 y SimpleKMeans - Manhattan - 2) estiman

dos grupos. Pero pese a su buena calidad técnica, no reflejan características distintivas relevantes para los médicos entre los grupos, y por esta razón se desecha.

Sin embargo, el resultado logrado por el *EM* con 3 grupos, que se encuentra en el cuarto lugar, es el que propone el agrupamiento más adecuado para el diagnóstico de la DM-tipo2, de acuerdo al criterio de expertos. Este conformó tres grupos distribuidos como se muestra en la Figura 3- 3. Obesidad, 50 años de edad como promedio, antecedentes familiares de diabetes mellitus e HTA son características comunes de los hombres de estos grupos. No obstante, existen en los pacientes de cada uno de estos grupos diferencias significativas.

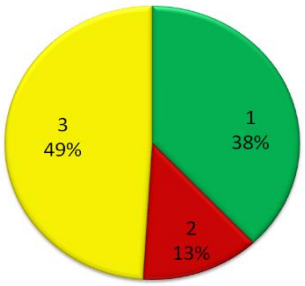


Figura 3- 3. Distribución de los hombres en los grupos por el EM con k=3.

En el primer grupo se encuentran los hombres con un peso promedio de 86.27 Kg y un IMC de 30.17, lo que indica un alto nivel de obesidad. Además, tienen el hábito de tomar café. Niveles de glicemia bajos y resultados de análisis médicos en niveles normales son características distintivas de los hombres de este grupo.

Un segundo grupo está conformado por pacientes tomadores de café, que con un peso promedio de 80 Kg y un IMC de 28.79, tienen glicemias por encima de los 10mmol/L y niveles de colesterol, triglicéridos y microalbuminuria muy altos.

El tercer grupo dentro de los hombres se compone de aquellos que con un IMC de 27.32, que implica niveles de obesidad bajos, y sin hábitos tóxicos presentan niveles de glicemia altos pero por debajo de los 10mmol/L. Además de tener el colesterol en niveles de riesgo y los triglicéridos un poco altos.

En los pacientes del grupo uno los médicos observan que, aunque son los hombres de mayor obesidad, el tener la glicemia y los análisis médicos en niveles normales son

factores favorables para la situación clínica de los pacientes. Por el contrario, en el grupo dos se encuentran aquellos, que, a consideración de los especialistas, debutan tóxicos. Ya que niveles de glicemia por encima de los 10mmol/L implican la posibilidad de complicarse y de presentar afectaciones en los órganos, lo que se evidencia en el resto de los análisis. El grupo tres caracteriza pacientes que presentan niveles de colesterol, triglicéridos y glicemias propios de una persona al debut, que, aunque no están en los niveles saludables, tampoco lo están en niveles críticos. Por lo anterior se observa en el agrupamiento diferentes niveles de riesgo en los pacientes, de complicarse o de tener afectaciones en los órganos.

Por tanto, se proponen como etiquetas para las clases encontradas en los hombres, tres niveles de riesgo: bajo, medio y alto. Que coinciden con los grupos uno, tres y dos respectivamente, propuestos por el EM con la cantidad de grupos igual a tres (Figura 3-4).

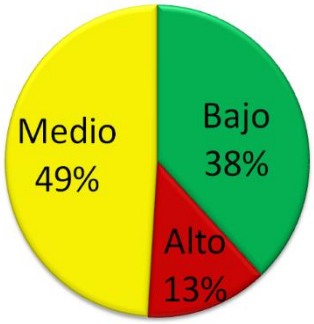


Figura 3- 4. Etiquetas de las clases en el conjunto de hombres.

Una vez identificadas las clases cabe preguntarse: ¿qué seguimiento proponen los expertos otorgar a las personas en cada uno de estos grupos? Se obtiene que, debido a que el seguimiento de un paciente depende de su situación clínica, se propone actuar de las siguientes formas para cada uno de los grupos:

- **Grupo de riesgo bajo**. Estas personas se encuentran con un cuadro clínico favorable, aunque factores agravantes son su obesidad, la HTA y los antecedentes familiares de DM-tipo2. El primer paso consiste en realizar una prueba de tolerancia a la glucosa. El segundo paso es modificar los estilos de vida del paciente, para lo que se le asigna una dieta y un plan de ejercicios

físicos. Estas personas deben tener seguimiento con el fin de observar el comportamiento de los factores de riesgo.

- **Grupo de riesgo medio**. Los hombres de este grupo ya debutaron con diabetes mellitus, pero al tener niveles de glucosa por debajo de los 10mmol/L los médicos proponen modificar su estilo de vida y, en caso de ser necesario, indicar un hipoglucemiante oral.
- **Grupo de riesgo alto**. A estos pacientes con un cuadro clínico tóxico los médicos no solo le modifican su estilo de vida, mediante una dieta y ejercicios, sino que le indican tratamiento medicamentoso. Este consiste, en muchos casos, en la indicación de insulina, durante una primera etapa, para lograr la estabilidad del paciente. Posterior a esto la persona puede mantenerse con hipoglucemiantes orales.

3.7.2 Análisis en las mujeres

En el caso de las mujeres sucede similar al de los hombres, los dos resultados que se encuentra en las primeras posiciones no ofrecen, según el criterio médico, un agrupamiento útil para cumplir los objetivos del proceso. Por su parte el *SelfOrganizingMap* que formó cuatro grupos es el próximo algoritmo, en la lista ordenada, que presenta mejores resultados en los índices. Al analizar los grupos conformados por esta técnica, desde el punto de vista de los médicos, se encuentran características interesantes en los diferentes grupos de mujeres.

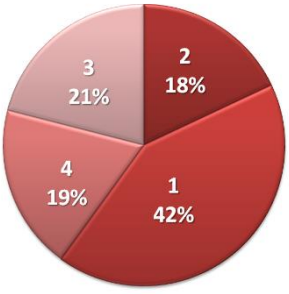


Figura 3- 5. Distribución de las mujeres en los grupos por el SelfOrganizingMap con k = 4.

Este algoritmo conforma cuatro grupos de mujeres distribuidos como se muestra en la Figura 3- 5. Es interesante que en ellos las glicemias se encuentren entre 7 y 10mmol/L, lo que es considerado por los médicos como glicemias altas, aunque no en niveles

críticos. Otra característica común es que las pacientes tienen antecedentes familiares de diabetes mellitus.

En el grupo uno se encuentran las mujeres que con 74.52Kg de peso tienen un IMC de 30.52, lo que indica niveles de obesidad elevados. Estas mujeres, muy obesas, toman café, son hipertensas y tienen los triglicéridos y el colesterol un poco altos.

El segundo grupo, identificado por la técnica, está formado por las mujeres de 47 años de edad aproximadamente, con 82.4 Kg de peso promedio y un IMC de 33.48. Lo que indica mujeres jóvenes muy obesas. Otras características interesantes son que padecen HTA y tienen los triglicéridos y la microalbuminuria altos.

El tercer grupo lo constituyen mujeres de 56 años, 64Kg y un IMC de 26.52. Estas pacientes no son obesas, no tienen hábitos tóxicos, no padecen HTA y sus análisis están normales.

El último grupo se compone de mujeres de 63 años, 64.24 Kg y 27.69 de IMC. Estas personas con niveles de obesidad bajos, son tomadoras de café, hipertensas y con los triglicéridos y el colesterol un poco elevados.

Al analizar las características de estos grupos se observa que las diferencias notables no están en los niveles de glicemia, como en los grupos de hombres, sino en la edad y el peso, específicamente los niveles de obesidad.

Según los médicos la obesidad es un factor agravante del riesgo que tienen los diabéticos de complicarse, por las consecuencias que esta tiene para la salud humana. La edad es otro agente que aumenta la posibilidad de riesgo ya que varias de las complicaciones de la diabetes mellitus aparecen después de padecer la enfermedad por algún tiempo. Además, cuando se debuta en edades tempranas es porque se tienen los factores de riesgo en niveles altos. Por lo anterior se pueden identificar en los grupos formados niveles de riesgo de complicación, ya no determinados por la glicemia, sino por estos factores igual de importantes en la prevención de la DM-tipo2.

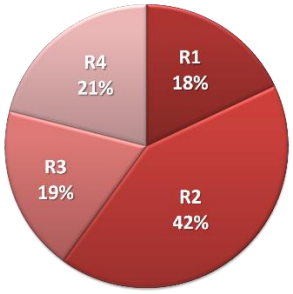


Figura 3- 6. Etiquetas de las clases en el conjunto de mujeres.

Por tanto se proponen como etiquetas, para las clases del conjunto de datos de mujeres, cuatro niveles de riesgo donde uno indica mayor riesgo y cuatro menor riesgo (Figura 3-6). La clasificación R1 corresponde al grupo dos conformado por la técnica, ya que estas son las mujeres más jóvenes y con más obesidad. R2 son las mujeres muy obesas pero menos jóvenes, agrupadas en el conglomerado dos. La clasificación R3 coincide con el cuarto grupo, donde las pacientes son obesas en menor medida. El grupo 3 se considera como R4 ya que estas mujeres no son obesas, ni hipertensas y sus análisis están en niveles saludables.

Debido a que en todos los grupos las mujeres tienen glicemias entre 7 y 10mmol/L el procedimiento a seguir es el mismo para los cuatro. Este consiste en modificar el estilo de vida de la paciente, mediante la dieta y el plan de ejercicios, e indicarle un hipoglucemiante oral, si es necesario. Pero para la indicación de la dieta los médicos tienen en cuenta el peso, la talla, el IMC, la edad, el sexo y la actividad física. Algunos de estos factores influyen también en el tipo de hipoglucemiante oral que le prescriben a la paciente. Por lo que, al existir diferencias notables en los grupos conformados de mujeres en estos factores las habrá también en el tipo de dieta y el tratamiento de cada una de estas pacientes.

El análisis del caso de estudio en cuestión ha llegado hasta este punto, puesto que la sexta fase, despliegue, está en proceso aún. Se han documentado los resultados, pero aún no se han creado las estrategias de mantenimiento y monitoreo del sistema para continuar desarrollando la aplicación. No obstante, los resultados obtenidos son suficientes para determinar que la adecuación de la metodología CRISP-DM a problemas no supervisados tipo atributo-valor del presente trabajo, usada para dar

solución a esta problemática real, guía correctamente y hace el trabajo más ordenado y fácil de realizar para los ingenieros del conocimiento.

3.8 Conclusiones parciales

La adecuación de la metodología CRISP-DM a problemas no supervisados tipo atributo-valor permitió organizar y guiar el proceso de minería de datos sobre los datos de los pacientes con diabetes mellitus tipo 2, haciéndolo más sencillo.

Dividir el conjunto de datos en hombres y mujeres resultó del análisis inicial, ya que esta categorización no cumplía lo objetivos iniciales de la minería de datos en el problema.

En un segundo análisis, independiente entre los pacientes masculinos y femeninos, la evaluación de los resultados, haciendo uso de las medidas de evaluación internas: Davies Bouldin, Hartigan, Calinski y Harabasz, Dunn, y Ball and Hall y el criterio de los expertos, dio como resultado tres grupos en los hombres y cuatro en las mujeres, encontrados por los algoritmos *EM* y *SelfOrganizingMap* respectivamente. Las etiquetas resultantes fueron interpretadas como niveles de riesgo de complicación de la diabetes mellitus tipo 2.

Conclusiones

Esta investigación se enmarca en el tema de la minería de datos sobre conjuntos de datos tipo atributo-valor. Con ella se arriban a las siguientes conclusiones:

1. Del análisis de las metodologías se concluye que CRISP-DM resulta la más adecuada por ser de libre distribución, independiente de la herramienta utilizada y la más usada, pero por su grado de generalidad debe ser adecuada a los problemas no supervisados tipo atributo-valor.
2. Se presenta un procedimiento resultante de la adecuación de la metodología CRISP-DM a los problemas no supervisados tipo atributo-valor facilitando el desarrollo de investigaciones en estos tipos de problemas.
3. El desarrollo del caso de estudio mostró que la adecuación realizada guía el proceso de minería de datos facilitando una correcta organización y documentación del mismo.

Recomendaciones

1. Construir un sistema basado en el conocimiento utilizando los resultados obtenidos en el capítulo III de la presente investigación, para de esa forma contribuir a un mejor diagnóstico y prevención de la diabetes mellitus tipo 2.
2. Profundizar en el estudio del proceso de ordenamiento de los agrupamientos, luego de ser validados por las medidas de evaluación internas, y de esta forma brindar soluciones más robustas.

Referencias bibliográficas

2014. *Data Mining Community's Top Resource* [Online]. http://www.kdnuggets.com/polls/2014/data-mining-applications.htm. [Accessed 7 de octubre 2014].

ACUNA, E. & RODRIGUEZ, C. 2004. The treatment of missing values and its effect on classifier accuracy. *Classification, clustering, and data mining applications.* Springer.

ALBALATE, A. & SUENDERMANN, D. A combination approach to cluster validation based on statistical quantiles. Bioinformatics, Systems Biology and Intelligent Computing, 2009. IJCBS'09. International Joint Conference on, 2009. IEEE, 549-555.

ALTMAN, N. S. 1992. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician,* 46**,** 175-185.

AZEVEDO, A. & SANTOS, M. F. 2008. KDD, SEMMA and CRISP-DM: a parallel overview. *In:* ABRAHAM, A. (ed.) *IADIS European Conf. Data Mining.* IADIS.

BACHER, J., WENZIG, K. & VOGLER, M. 2004. *SPSS TwoStep Cluster-a first evaluation*, Lehrstuhl für Soziologie Berlin, DE.

BALL, G. H. & HALL, D. J. 1965. ISODATA, A novel method of data analysis an pattern classification. *In:* NTIS (ed.). Menlo Park: Stanford Research Institute.

BAMNETT, V. & LEWIS, T. 1994. *Outliers in statistical data*.

BAR-HILLEL, A., HERTZ, T., SHENTAL, N. & WEINSHALL, D. 2005. Learning a mahalanobis metric from equivalence constraints. *Journal of Machine Learning Research,* 6**,** 937-965.

BATISTA, G. E. & MONARD, M. C. 2002. A Study of K-Nearest Neighbour as an Imputation Method. *HIS,* 87**,** 48.

BOUCKAERT, R. R., FRANK, E., HALL, M., KIRKBY, R., REUTEMANN, P., SEEWALD, A. & SCUSE, D. 2011. WEKA Manual for Version 3-7-5. *In:* WAIKATO, U. O. (ed.).

BROWN, E. T., LIU, J., BRODLEY, C. E. & CHANG, R. Dis-function: Learning distance functions interactively. Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on, 2012. IEEE, 83-92.

CAI, D., ZHANG, C. & HE, X. Unsupervised feature selection for multi-cluster data. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, 2010. ACM, 333-342.

CALINSKI, T. & HARABASZ, J. 1974. A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods,* 3**,** 1-27.

Referencias Bibliográficas

CANALS, A., BOISOT, M. & CORNELLA, A. 2003. Gestión del conocimiento. *Gestión: 2000.* Barcelona, España.

CARPENTER, G. & GROSSBERG, S. 1987a. ART2: Self - organization of stable category recognition codes for analog input patterns. *Applied Optics,* 26**,** 4919 – 4930.

CARPENTER, G. & GROSSBERG, S. 1987b. A massively parallel architecture for a self -organizing neural pattern recognition machine *Computer Vision, Graphics, and Image Processing,* 37**,** 54 – 115.

CARPENTER, G. & GROSSBERG, S. 1988. The ART of adaptive pattern recognition by a self - organizing neural network. *IEEE Computer,* 21 77 – 88.

CARPENTER, G. & GROSSBERG, S. 1990. ART3: Hierarchical search using chemical transmitters in self - organizing pattern recognition Architectures. *Neural Networks,* 3**,** 129 – 152

CHAPMAN, P., CLINTON, J., KERBER, R., KHABAZA, T., REINARTZ, T., SHEARER, C. & WIRTH, R. 2000. CRISP-DM 1.0 Step-by-step data mining guide.

CIOS, K. J., PEDRYCZ, W., SWINIARSKI, R. W. & KURGAN, L. A. 2007. *Data Mining, A Knowledge Discovery Approach*.

DALKIR, K. 2005. Knowledge Management in Theory and Practice. Burlington, USA: Elsevier.

DAVIES, D. L. & BOULDIN, D. W. 1979. A cluster separation measure. *Pattern Analysis and Machine Intelligence, IEEE Transactions on***,** 224-227.

DEMPSTER, A., LAIRD, N. & RUBIN, D. 1977. Maximum - likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B,* 39 1 – 38.

DEMSAR, J., CURK, T., ERJAVEC, A. & GORUP, C. 2013. Orange: Data Mining Toolbox in Python. *Journal of Machine Learning Research* 14**,** 2349-2353.

DEZA, E. & DEZA, M.-M. 2006. *Dictionary of Distances*.

DÍAZ, O. D. 2007. Programa de Atención al Diabético en Cuba. *In:* MINSAP (ed.). La Habana: Instituto Nacional de Endocrinología.

DIMITRIADOU, E., DOLNICAR, S. & WEINGESSEL, A. 2002. An examination of indexes for determining the number of clusters in binary data sets. *Psychometrika,* 67**,** 137-160.

DU, K.-L. 2010. Clustering: A neural network approach. *Neural Networks,* 23 89-107.

DUNN, J. 1974. Well separated clusters and optimal fuzzy partitions. *Journal on Cybernetics,* 4**,** 95-104.

Referencias Bibliográficas

ESTER, M., KRIEGEL, H., SANDER, J. & XU, X. A density - based algorithm for discovering clusters in large spatial databases with noise. Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD96), 1996 New York, NY. AAAI Press

FAYYAD, U., PIATETSKY-SHAPIRO, G. & SMYTH, P. 1996. The KDD Process for Extracting Useful Knowledge from Volumes of Data. *Communications of the ACM,* 39**,** 27-34.

FIELDING, A. H. 2007. *Cluster and classification techniques for the biosciences*, Cambridge University Press.

FISHER, D. H. 1987. Knowledge Acquisition Via Incremental Conceptual Clustering. *Machine Learning***,** 139-172.

FUJIMAKI, R., YAIRI, T. & MACHIDA, K. An approach to spacecraft anomaly detection problem using kernel feature space. Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, 2005. ACM, 401-410.

GIANCARLO, R., BOSCO, G. L. & PINELLO, L. 2010. Distance Functions, Clustering Algorithms and Microarray Data Analysis. *LNCS.* Berlin, Heidelberg: Springer-Verlag.

GONZÁLEZ, D. P. 2010. *Algoritmos de Agrupamiento basados en densidad y Validación de clusters.* Phd, Universitat Jaume I.

GORUNESCU, F. 2011. *Data Mining - Concepts, Models and Techniques*, Springer-Verlag Berlin Heidelberg.

GRUBBS, F. E. 1969. Procedures for detecting outlying observations in samples. *Technometrics,* 11**,** 1-21.

HALKIDI, M., BATISTAKIS, Y. & VAZIRGIANNIS, M. 2001. Clustering validity checking methods: Part II. *Journal of Intelligent Information Systems,* 17**,** 107–145.

HALKIDI, M. & VAZIRGIANNIS, M. Clustering validity assessment: Finding the optimal partitioning of a data set. Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on, 2001. IEEE, 187-194.

HAN, J., KAMBER, M. & PEI, J. 2011. *Data mining: concepts and techniques: concepts and techniques*, Elsevier.

HARTIGAN, J. A. 1975. Clustering Algorithms, New York: John Willey and Sons. *Inc. Pages113129*.

HE, X., CAI, D. & NIYOGI, P. Laplacian score for feature selection. Advances in neural information processing systems, 2005. 507-514.

HERNÁNDEZ-ORALLO, J., RAMÍREZ-QUINTANA, J. & FERRI-RAMÍREZ, C. 2004. *Introducción a la minería de datos.*, Pearson Education.

Referencias Bibliográficas

HODGE, V. J. & AUSTIN, J. 2004. A survey of outlier detection methodologies. *Artificial Intelligence Review,* 22**,** 85-126.

HUANG, A. Similarity Measures for Text Document Clustering. New Zealand Computer Science Research Student Conference, 2008 Christchurch, New Zealand.

HUANG, H. & WU, G. 2008. Introduce to Data Mining with RapidMiner. *In:* UNIVERSITY, S. (ed.).

KAUFMAN, L. & ROUSSEUW, P. J. 1990. Finding Groups in Data: An Introduction to Cluster Analysis. *Wiley*.

KOHONEN, T. 1986. Learning vector quantization for pattern recognition. Helsinky university of tecnology, department of technical physics.

KOHONEN, T. 1995. Self-Organising Maps. *Springer*.

LAURIKKALA, J., JUHOLA, M., KENTALA, E., LAVRAC, N., MIKSCH, S. & KAVSEK, B. Informal identification of outliers in medical data. Fifth International Workshop on Intelligent Data Analysis in Medicine and Pharmacology, 2000. Citeseer, 20-24.

LI, Z., YANG, Y., LIU, J., ZHOU, X. & LU, H. Unsupervised Feature Selection Using Nonnegative Spectral Analysis. AAAI, 2012.

LIU, H., WU, X. & ZHANG, S. 2011. *Feature selection using hierarchical feature clustering*.

LIU, Y., LI, Z., XIONG, H., GAO, X. & WU, J. 2010. Understanding of Internal Clustering Validation Measures. *2010 IEEE International Conference on Data Mining.* IEEE.

MACQUEEN, J. 1967. Some methods for classification and analysis of multivariate observations *5th Berkeley Symp,* 1**,** 281–297.

MALIK, K., SADAWARTI, H. & SINGH, K. G. 2014. Comparative Analysis of Outlier Detection Techniques. *International Journal of Computer Applications* 97.

MICHALSKI, R. S. 1986. Understanding the Nature of Learning. *Issues and Research Directions.* California.

MITCHELL, T. M. 1997. *Machine Learning*, McGraw-Hill Science/Engineering/Math.

MOLINA, J. M. & GARCÍA, J. 2006. Técnicas de análisis de datos aplicaciones prácticas utilizando Microsoft Excel y Weka.

NG, R. & HAN, J. Efficient and effective clustering methods for spatial data mining. 20th Conference on VLDB, 1994 Santiago de Chile 144 – 155.

OLSON, D. L. & DELEN, D. 2008. *Advanced Data Mining Techniques*.

Referencias Bibliográficas

PANDIT, S. & GUPTA, S. 2011. A comparative study on distance measuring approaches for clustering. *International Journal of Research in Computer Science,* 2**,** 29-31.

PORTILLO, M. T. E. & MENDOZA, J. A. S. P. 2008. P. Ch. Mahalanobis y las aplicaciones de sudistancia estadistica. *In:* UACJ, I. D. I. Y. T. (ed.) *CULCyT.*

QIAN, M. & ZHAI, C. Robust Unsupervised Feature Selection. IJCAI, 2013. Citeseer.

RAMASWAMY, S., RASTOGI, R. & SHIM, K. Efficient algorithms for mining outliers from large data sets. ACM SIGMOD Record, 2000. ACM, 427-438.

RAO, S., RODRIGUEZ, A. & BENSON, G. 2005. Evaluating distance functions for clustering tandem repeats. *GENOME INFORMATICS SERIES,* 16**,** 3.

ROSALES, R. & FUNG, G. Learning sparse metrics via linear programming. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, 2006. ACM, 367-373.

ROUSSEEUW, P. & LEROY, A. 1996. Robust Regression and Outlier Detection. *John Wiley & Sons*.

RUSSELL, S. & NOVIG, P. 2009. *Artificial Intelligence A modern Approach*, Pearson Education.

SCHAFER, J. L. 1997. *Analysis of incomplete multivariate data*, CRC press.

SEHGAL, G. & GARG, K. 2014. Comparison of Various Clustering Algorithms *International Journal of Computer Science and Information Technologies,* Vol. 5 (3)**,** 3074-3076.

SHARMA, S. 1996. Applied multivariate techniques. *USA, John Wiley& Sons*.

SINGHAL, A. 2001. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.,* 24**,** 35-43.

TAN, P.-N., STEINBACH, M. & KUMAR, V. 2006. *Introduction to Data Mining*, Addison-Wesley.

TOLEDO, M. D. G. 2005. *Una comparación de índices de validación de conglomerados.* Maestría en Ciencias, Universidad de Puerto Rico.

WANG, M., HUA, X.-S., HONG, R., TANG, J., QI, G.-J. & SONG, Y. 2009a. Unified video annotation via multigraph learning. *Circuits and Systems for Video Technology, IEEE Transactions on,* 19**,** 733-746.

WANG, M., HUA, X.-S., TANG, J. & HONG, R. 2009b. Beyond distance measurement: constructing neighborhood similarity for video annotation. *Multimedia, IEEE Transactions on,* 11**,** 465-476.

WEINBERGER, K. Q., BLITZER, J. & SAUL, L. K. Distance metric learning for large margin nearest neighbor classification. Advances in neural information processing systems, 2005. 1473-1480.

Referencias Bibliográficas

WEINBERGER, K. Q. & SAUL, L. K. Fast solvers and efficient implementations for distance metric learning. Proceedings of the 25th international conference on Machine learning, 2008. ACM, 1160-1167.

XU, R. & WUNSCH, D. 2005. Survey of Clustering Algorithms. *IEEE Transactions On Neural Networks,* Vol. 16.

XU, R. & WUNSCH, D. C. 2009. *Clustering,* Hoboken, New Jersey, John Wiley & Sons.

YAGER, R. R. & FILEV, D. P. 1994. Approximate clustering via the mountain method. *IEEE Transactions on Systems Man and Cybernetics,* 24.

YANG, L. & JIN, R. 2006. Distance metric learning: A comprehensive survey. *Michigan State Universiy,* 2.

YANG, Y., SHEN, H. T., MA, Z., HUANG, Z. & ZHOU, X. l2, 1-norm regularized discriminative feature selection for unsupervised learning. IJCAI Proceedings-International Joint Conference on Artificial Intelligence, 2011. Citeseer, 1589.

YING, Y., HUANG, K. & CAMPBELL, C. Sparse metric learning via smooth optimization. Advances in neural information processing systems, 2009. 2214-2222.

YING, Y. & LI, P. 2012. Distance metric learning with eigenvalue optimization. *The Journal of Machine Learning Research,* 13**,** 1-26.

ZAFAR, M. H. & ILYAS, M. 2015. A Clustering Based Study of Classification Algorithms. *International Journal of Database Theory and Application,* 8**,** 11-22.

ZHANG, T., RAMAKRISHNAN, R. & LIVNY, M. BIRCH: An Efficient Data Clustering Method for Very Large Databases. 1996 ACM SIGMOD International Conference on Magnament of Data., 1996 Montreal, Canada. ACM, 103-114.

Anexos

CENTRO DE ATENCIÓN Y EDUCACIÓN EN DIABETES CIENFUEGOS

HISTORIA CLINICA

EXPEDIENTE CLINICO: _____
CARNET DE IDENTIDAD: _____

1 ER APELLIDO	2DO APELLIDO	NOMBRE (S)	FECHA DE CONFECCION ___ ___ ___ DIA MES AÑO

DIRECCION:

Calle _____ No. _____ Entrecalles _____ Municipio. _____

Ciudad Publico _____ Provincia _____ Teléfono (T) _____ Teléfono (C) _____

OCUPACION:	ESCOLARIDAD	SEXO: F:___ M: ___	EDAD:	EDAD AL DEBUT.

TIEMPO DE EVOLUCION	MODO DE DEBUT:	OBESIDAD AL DEBUT
	___ SINTOMAS CLINICOS. ___ CHEQUEO SIN SINTOMAS. ___ CETOACIDOSIS (PROBADA). ___ DURANTE EL EMBARAZO. ___ OTROS. ___ NO SABE.	___ SI. ___ NO ___ NO PRECISADO.

ANTECEDENTES FAMILIARES DE DIABETES		HISTORIAOBSTETRICA
VIA MATERNA. ___ NADIE. ___ MADRE ___ ABUELO (A). ___ MADRE + ABUELO (A). ___ NO SABE.	VIA PATERNA. ___ NADIE. ___ PADRE ___ ABUELO (A). ___ MADRE + ABUELO (A). ___ NO SABE.	MENARCA FM : G___P___A___ MACROFETOS MALFORMACIONES MUERTES ANTICONCEPCION MENOPAUSIA EDAD

HERMANO:	HIJO.
___ SI. ___ NO. ___ NO SABE	___ SI. ___ NO. ___ NO SABE

ANTECEDENTES PATOLOGICOS	PERSONALES			FAMILIARES			HABITOS TOXICOS.
	SI	NO	NO PREC.	SI	NO	NO PREC.	___ NO FUMADOR. ___ EX FUMADOR (6 MESES) ___ FUMADOR. SI FUMADOR (# POR DIAS) ___ CIGARROS_____ ___ TABACO _____ ALCOHOL ___ NO ___ SI (GR/DIAS): _____ CAFÉ -3T ____ *3 T ____
HIPERLIPOPROTEINEMIA							
HIPERTENSION ARTERIAL							
CARDIOPATIA ISQUEMICA.							
CLAUDITACION INTERMITENTE.							
AVE							
OTROS (ESPECIFICAR).							

TRATAMIENTO AL INICIO	TRATAMIENTO ACTUAL	OTROS TTOS	SINTOMAS ACTUALES
____ SOLO DIETA: ____ COH. ____ INSULINA. ____ INSULINA + COH. ____ NO PRECISADO AÑOS DE COMIENZO _____	____ SOLO DIETA: ____ COH. ____ INSULINA. ____ INSULINA + COH. ____ NO PRECISADO AÑOS DE COMIENZO _____		

EXAMEN FISICO (EXAMENES POR APARATOS).	.
MUCOSAS	TALLA _____M. PESO _____ Kg. IMC: _____ PI_____ DIETA_____ CA_____ 0

Anexo 1- Historia clínica de los pacientes del CAED (a).

TCS

ARESP

FR
ACV

FC TA

ABDOMEN

EX CUELLO

EX ORAL

ESCALA NSS
-SENSACION.
QUEMAZON,ENTUMECIMIENTO,HORMIGUEO EN PIE (2 PTOS)
FATIGA,CALAMBRES ODOLORIMIENTO (1 PTO)
-LOCALIZACIÓN
PIES (2 PTOS)
PANTORRILLAS (1 PTO)
EN OTRO LUGAR (0PTO)
-TIENE SIEMPRE SINTOMAS AL DESPERTAR POR LA NOCHE
SI (1 PTO)
-RELACION HORARIA
EMPEORA POR NOCHE (2 PTOS)
PRESENTE DIA Y NOCHE (1 PTO)
PRESENTE DURANTE EL DIA (0 PTO)

-COMO MEJORAN LOS SINTOMAS
PASEANDO (2 PTOS)
DE PIE (1 PTO)
SENTADO O ACOSTADO O NO MEJORAN (O PTO)

EXISTE NEUROPATIA
NDS ES ≥6 O SI NDS ESTA ENTRE 3 Y 5 Y NSS ≥ 5

RESULTADO DE EDUCACION DIABETOLOGICA.

INICIO	FINAL
SUFICIENTE	SUFICIENTE
NO SUFICIENTE	NO SUFICIENTE

EXAMEN DE MIEMBROS INFERIORES.

ABDOMEN (AORTA) : NORMAL____ DILATADA____

SISTEMA VENOSO: VARICES____ MICROVARICES____

SISTEMA LINFÁTICO: LINFANGITIS____ LINFEDEMA____

DEFORMIDADES PODÁLICAS: PIE PLANO____ METATARSO CAIDO____

HALLUX VALGUS____ DEDOS EN GARRA____ PIE DE CHARCOT____

AMP. MENORES____ OTROS____

LESIONES DERMATOLÓGICAS: EPIDERMOFITOSIS____ FISURAS____ PÌEL SECA

Y CALLOSA____ HIPERPIGMENTACION____ DERMATITIS____

CELULITIS____ ÚLCERAS____ PALIDEZ____ CIANOSIS____ RUBICUNDEZ____

DEDOS CON NECROSIS____ OTRAS____

LESIONES A NIVEL DE LAS UÑAS: ONICOMICOSIS____ PARONIQUIA____

TIPO DE ANTEPIÉ: EGIPCIO____ GRIEGO____ 1 Y 2 IGUAL____ CUADRADO____

ESTANDAR____

EXAMEN ARTERIAL:

PULSOS	DERECHO	IZQUIERDO
FEMORAL:	_____	_____
POPLÍTEO:	_____	_____
TIBIAL POSTERIOR:	_____	_____
PEDEO:	_____	_____

SOPLO FEMORAL DERECHO_____ IZQUIERDO_____

IMPRESIÓN DIAGNOSTICA:

PIE DE RIESGO GRADO O ___,___

PIE DE RIESGO GRADO I___,___

PIE DE RIESGO GRADO II___,___

PIE DE RIESGO GRADOIII___,___

	DERECHO		IZQUIERDO
	___	PRESENTE	___
REFLEJOS	___	PRESENTE CON REFUERZO	___
AQUILEANO	___	AUSENTE	___

	DERECHO		IZQUIERDO
	___	NORMAL	___
PALESTESIA	___	REDUCIDA	___
(PUNTA DEL DEDO GORDO)			

	DERECHO		IZQUIERDO
TEMPERATURA	___	NORMAL	___
(DORSO DEL PIE)	___	REDUCIDA	___

	DERECHO		IZQUIERDO
	___	NORMAL	___
PERCEPCION DE	___	REDUCIDA	___
PUNTA ROMA			

PUNTUACION ESCALA NDS
0 A2 NO NEUROPATIA
3-5 LIG
6-8 MODERADA
9-10 GRAVE

Anexo 2- Historia clínica de los pacientes del CAED (b).

VISION BORROSA: SI ___ NO ___

AV: OD: _____ OI: _____

RD: OD: _____ OI: _____

OD: A: _____ SA: _____ M; _____ FO: _____

OI: A: _____ SA: _____ M; _____ FO: _____

RETINOPATIA DIABETICA. NO ___ NO PROLIFERATIVA ___ PROLIFERATIVA ___ NO PRECISADA ____

RETINOPATIA HIPERTENSIVA NO ___ GRADO 1 ____ GRADO 2 ___ GRADO 3 ____ GRADO 4 ____ NO PRECISADA ____

RETINOPATIA ARTEREO ESCLEROSIS NO ___ GRADO 1 ____ GRADO 2 ___ GRADO 3 ____ GRADO 4 ____ NO PRECISADA ____

HEMORRAGIA VITREA OD: NO ___ SI ____ NO PRECISADA ____

HEMORRAGIA VITREA OI: NO ___ SI ____ NO PRECISADA ____

MACULOPATIA OD NO ____ SI ____ NO PRECISADA

MACULOPATIA OI NO ____ SI ____ NO PRECISADA

CATARATA NO ____ METABOLICA ____ SENIL ____ OTRAS _____ NO PRECISADA _____

GLAUCOMA: NO ____ ANGULO ABIERTO _____ SECUNDARIO ____ ANGULO ESTRECHO ____ NO PRECISADO ____.

COMPLICACIONES: SI NO NO PRECISADO.
NEUROPATIAS REVERSIBLES:

MONONEUROPATIA

PARALISIS DE N CRANEALES

NEUROPATIAS POR COMPRESION

DSE

SEPSIS ESPECIFICAR: _____

DERMOPATIA

HIPERLIPOPROTEINEMIA ESPECIFICAR: _____

NEUROPATIA AUTONOMICA ESPECIFICAR: _____

OTRAS COMPLICACIONES ESPECIFICAR: _____

NEFROLOGIA

FGT_____
CREATININA____
UREA____
AC URICO_____
MICROALBUMINURIA
1____
2_____
3_____
ID:

DIAGNOSTICO DEFINITIVO

PACIENTE:

1ER APELLIDO	2DO APELLIDO	NOMBRE	H.C

Anexo 3- Historia clínica de los pacientes del CAED (c).

Identificador	Descripción	Tipo	Relevancia
No.(Número)	Identificador del paciente, coincide con la historia clínica	Numérico	Sin importancia
Grupo	Número del grupo al que el paciente perteneció al ingresar	Numérico	Sin importancia
Historia Clínica	Número de la historia clínica del paciente, esta queda archivada en formato físico en la clínica.	Numérico	Sin importancia
Nombres y Apellidos	Nombre y Apellidos del paciente	Nominal	Sin importancia
Edad	Edad en años del paciente al momento de ingresar	Numérico	Relevante
Sexo	Sexo del paciente	Nominal	Relevante
Dirección	Dirección de la residencia del paciente	Nominal	Sin importancia
Área	Área de salud del municipio Cienfuegos a la que pertenece el paciente	Nominal	Sin importancia
Municipio	Municipio en el que vive el paciente	Nominal	Sin importancia
Talla	Altura del paciente en metros	Numérico	Relevante
Peso Inicial	Peso en Kg. del paciente al	Numérico	Relevante

	ingresar		
Índice de Masa Corporal	El índice de masa corporal del paciente se calcula como la razón entre el peso de la persona (en Kg) y el cuadrado de la estatura (en m2). Da una medida de la cantidad de grasa corporal de una persona	Numérico	Relevante
Peso Final	Peso en Kg. del paciente al terminar el ingreso	Numérico	Sin importancia
Escolaridad	Nivel educacional más alto culminado por el paciente	Nominal	Sin importancia
Ocupación	Trabajo u oficio que realiza el paciente	Nominal	Sin importancia
Hábitos Tóxicos(Fuma)	Si el paciente ha fumado o fuma	Booleano	Relevante
Hábitos Tóxicos(Café)	Si el paciente toma café	Booleano	Relevante
Hábitos Tóxicos (Beb. Alcohólicas)	Si el paciente toma bebidas alcohólicas	Booleano	Relevante
Hábitos Tóxicos(Otros)	Si el paciente practica algún hábito tóxico diferente a los anteriores.	Booleano	Relevante
Modo de Debut	Describe la forma en la que el paciente debutó con la diabetes	Nominal	Sin importancia

Obesidad al Debut	Si el paciente es obeso o no cuando debuta con la enfermedad	Booleano	Relevante
Tiempo de Evolución de la Enfermedad	Tiempo que ha transcurrido desde el debut del paciente hasta la fecha en que ingresó a la clínica	Numérico	Sin importancia
Antecedentes Obstétricos (Menarca)	Edad de la primera menstruación de los pacientes femeninos	Numérico	Relevante
Antecedentes Obstétricos (Embarazos)	Cantidad de embarazos hasta el momento del ingreso	Numérico	Relevante
Antecedentes Obstétricos (Abortos)	Cantidad de abortos realizados a la paciente	Numérico	Relevante
Antecedentes Obstétricos (Malformaciones)	En algunos casos se toma la cantidad de niños con malformaciones que ha tenido la paciente, en otros solo si las ha tenido o no	Numérico o Booleano	Relevante
Antecedentes Obstétricos (Macrofetos)	En algunos casos se toma la cantidad de partos donde el niño ha pesado más de 9 libras, en otros solo si los ha tenido o no	Numérico o Booleano	Relevante

Antecedentes Obstétricos (Muerte Perinatal)	Si algún niño de la paciente ha muerto antes del parto	Booleano	Relevante
Antecedentes Obstétricos (Menopausia)	Edad en la que la paciente tuvo su última menstruación	Numérico	Relevante
APF DM (Antecedentes Patológicos Familiares de Diabetes Mellitus)	Qué antecesores del paciente padecen diabetes mellitus	Nominal	Relevante
APP (Antecedentes Patológicos Personales)	Qué patologías presenta el paciente a parte de la diabetes	Nominal	Relevante
APF Otras	Qué patologías diferentes de la diabetes mellitus presentan los ancestros del paciente	Nominal	Relevante
Tratamiento Inicial (Glibenclamida)	Si el tratamiento que le indican al paciente al debut es Glibenclamida	Booleano	Sin importancia
Tratamiento Inicial (Metformina)	Si el tratamiento que le indican al paciente al debut es Metformina	Booleano	Sin importancia
Tratamiento Inicial (Diabeton)	Si el tratamiento que le indican al paciente al debut es Diabeton	Booleano	Sin importancia

Tratamiento Inicial (Insulina)	Si el tratamiento que le indican al paciente al debut es Insulina	Booleano	Sin importancia
Tratamiento Inicial (Dieta)	Si el tratamiento que le indican al paciente al debut es Dieta	Booleano	Sin importancia
Tipo Dieta (Inicial)	Representa el tipo de dieta que presenta el paciente al inicio de su ingreso.	Numérico	Sin importancia
Tipo Dieta (Final)	Representa el tipo de dieta que presenta el paciente al final de su ingreso.	Numérico	Sin importancia
TGP	Resultado del análisis Transaminasa Glutámico Pirática.	Numérico	Relevante
Glicemia (Inicio)	Refleja la medida determinada en la Prueba de Tolerancia a la Glucosa en Ayunas al inicio del ingreso	Numérico	Relevante
Glicemia (PPD-Postprandial)	Refleja la medida determinada en la Prueba de Tolerancia a la Glucosa 2 horas después de ser alimentado al inicio del ingreso.	Numérico	Relevante
Glicemia (Final)	Refleja la medida determinada en la Prueba de Tolerancia a la Glucosa en Ayunas al final del ingreso	Numérico	Sin importancia

Glicemia (PPD-Postprandial)	Refleja la medida determinada en la Prueba de Tolerancia a la Glucosa 2 horas después de ser alimentado al final del ingreso.	Numérico	Sin importancia
Creatinina	Resultado del análisis que expresa la función renal	Numérico	Relevante
Micro albuminuria (1)	Medición de la cantidad de albumina excretada en la orina	Numérico	Relevante
Micro albuminuria (2)	Si la medición de la cantidad de albumina excretada en la orina es positiva o negativa.	Nominal	Sin importancia
Eritro	Resultado del análisis que expresa la velocidad de sedimentación de los eritrocitos (glóbulos rojos)	Numérico	Relevante
Hemoglobina	Resultado del análisis realizado para determinar los valores de hemoglobina en sangre.	Numérico	Relevante
Triglicérido	Refleja los valores de triglicérido en sangre	Numérico	Relevante
Acido Úrico	Representa el nivel de ácido úrico en sangre	Numérico	Relevante
Colesterol	Resultado del análisis realizado para determinar el	Numérico	Relevante

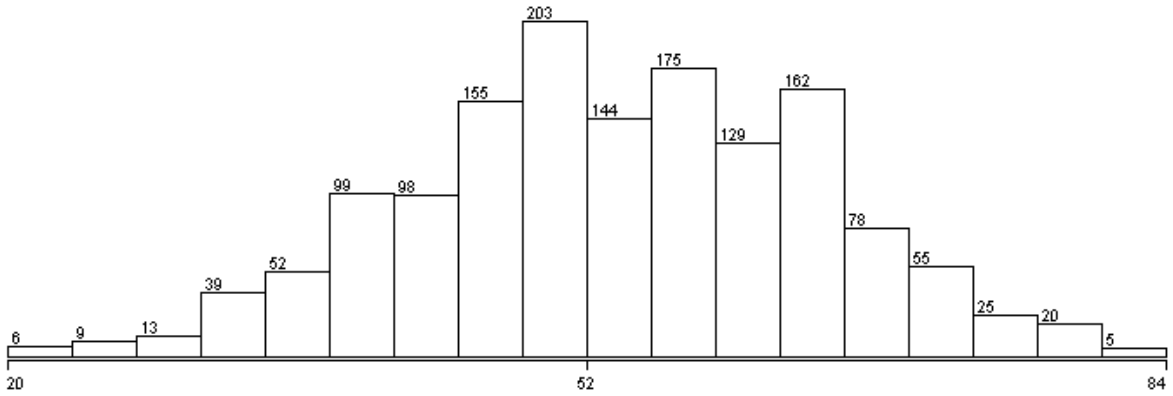
	nivel del colesterol en la sangre		
HDL-c	Refleja la medida de la cantidad de colesterol del tipo HDL del paciente.	Numérico	Sin importancia
Tratamiento Final (Glibenclamida)	Si el tratamiento actual del paciente es con Glibenclamida	Booleano	Sin importancia
Tratamiento Final (Metformina)	Si el tratamiento actual del paciente es con Metformina	Booleano	Sin importancia
Tratamiento Final (Diabeton)	Si el tratamiento actual del paciente es con Diabeton	Booleano	Sin importancia
Tratamiento Final (Insulina)	Si el tratamiento actual del paciente es con Insulina	Booleano	Sin importancia
Tratamiento Final (Dieta)	Si el tratamiento actual del paciente es con Dieta	Booleano	Sin importancia
Con Complicación Micro (RD)	El paciente presenta una retinopatía diabética	Booleano	Relevante
Con Complicación Micro (ND)	El paciente presenta una nefropatía diabética.	Booleano	Relevante
Con Complicación Micro (Neuro.D)	El paciente presenta una neuropatía diabética.	Booleano	Relevante
Con Complicación Macro (CI)	El paciente presenta una cardiopatía isquémica	Booleano	Relevante

Con Complicación Macro (AVE)	El paciente ha sufrido algún accidente vascular encefálico	Booleano	Relevante
Con Complicación Macro (IAP)	El paciente presenta alguna insuficiencia arterial periférica.	Booleano	Relevante
Pie con Riesgo	Se toma un número que representa el tipo del pie diabético que presenta el paciente	Numérico	Sin importancia
Tipo de Diabético	Tipo de diabetes que presenta el paciente. Se representa con un número.	Numérico	Relevante
Reingreso	Se señalan los pacientes que ya habían ingresado antes en el centro	Booleano	Sin importancia
Test de Inicio (Suficiente)	Si el resultado del test inicial realizado al paciente arroja que su conocimiento sobre DM-tipo2 es suficiente	Booleano	Sin importancia
Test de Inicio (Necesario)	Si el resultado del test inicial realizado al paciente arroja que su conocimiento sobre DM-tipo2 es necesario	Booleano	Sin importancia
Test de Inicio (Excelente)	Si el resultado del test inicial realizado al paciente arroja que su conocimiento sobre DM-tipo2 es excelente	Booleano	Sin importancia

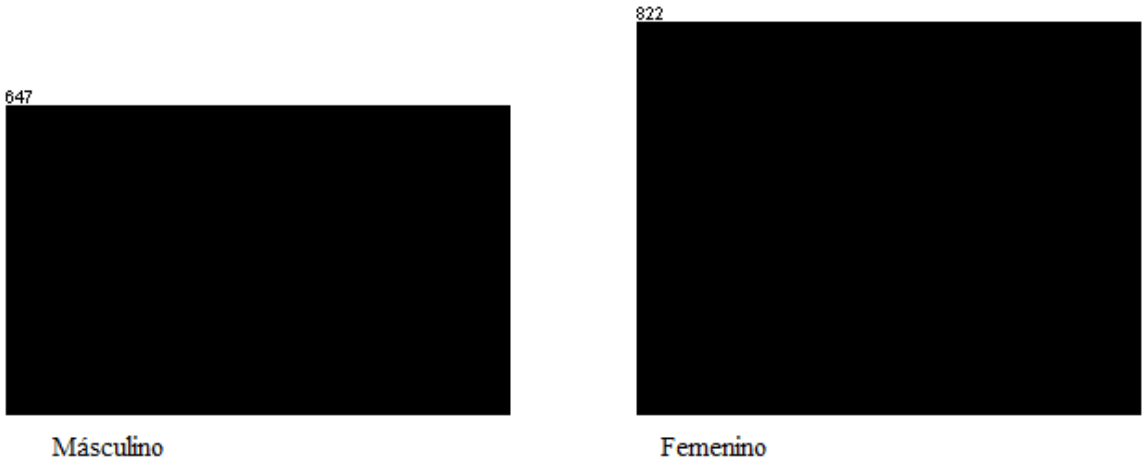
Test de Inicio (Notable)	Si el resultado del test inicial realizado al paciente arroja que su conocimiento sobre DM-tipo2 es notable	Booleano	Sin importancia
Test de Inicio (Insuficiente)	Si el resultado del test inicial realizado al paciente arroja que su conocimiento sobre DM-tipo2 es insuficiente	Booleano	Sin importancia
Test Final (Suficiente)	Si el resultado del test inicial realizado al paciente arroja que su conocimiento sobre DM-tipo2 es suficiente	Booleano	Sin importancia
Test Final (Necesario)	Si el resultado del test inicial realizado al paciente arroja que su conocimiento sobre DM-tipo2 es necesario	Booleano	Sin importancia
Test Final (Excelente)	Si el resultado del test inicial realizado al paciente arroja que su conocimiento sobre DM-tipo2 es excelente	Booleano	Sin importancia
Test Final (Notable)	Si el resultado del test inicial realizado al paciente arroja que su conocimiento sobre DM-tipo2 es notable	Booleano	Sin importancia
Test Final (Insuficiente)	Si el resultado del test inicial realizado al paciente arroja que	Booleano	Sin importancia

	su conocimiento sobre DM-tipo2 es insuficiente		

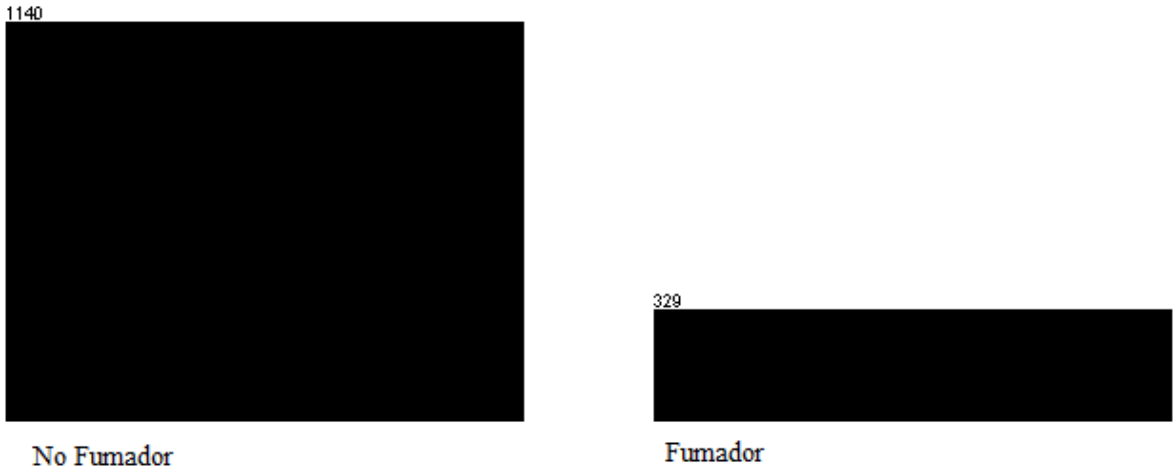
Anexo 4- Descripción de las variables.



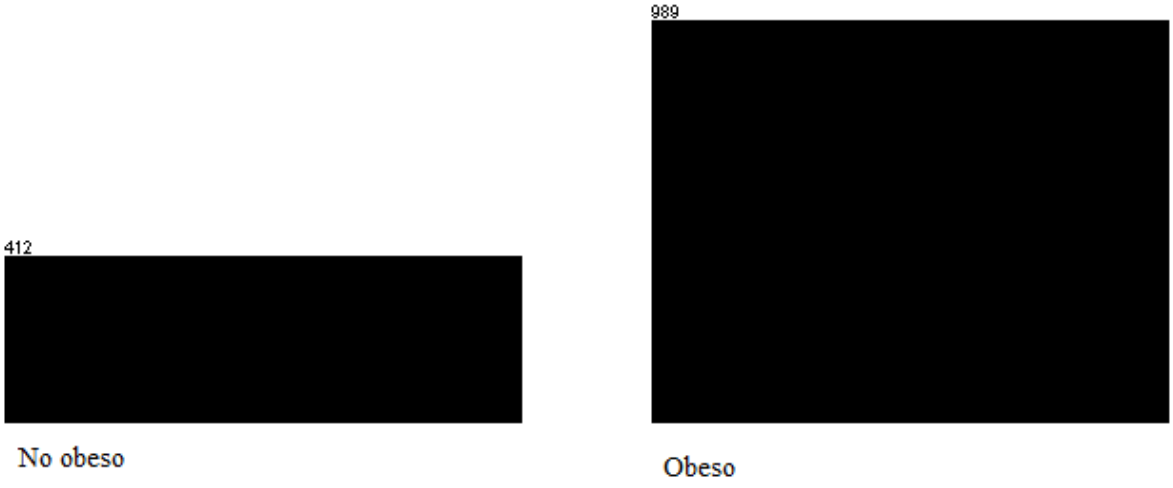
Anexo 5- Histograma de la variable edad.



Anexo 6- Histograma de la variable sexo.



Anexo 7- Histograma de la variable hábito de fumar.



Anexo 8- Histograma de la variable obesidad.

Anexo 9- Histograma de la variable antecedentes familiares de DM-tipo2.

Anexo 10- Histograma de la variable padecimiento de hipertensión arterial (HTA).

Historia Clínica	Talla	Talla sustituida
469	1.64.5	1.64
470	166	1.66
477	1.54.5	1.54
478	164.5	1.64
485	1.51.5	1.51
487	1.67.5	1.67
488	1.63.5	1.63
492	1.71.5	1.71
565	!,70	1.70
1121	17.6	1.76
1123	18.2	1.82

Historia Clínica	Talla	Talla sustituida
1141	17.3	1.73
1180	17.30	1.73
1183	17.3	1.73
1219	14.7	1.47
1364	161	1.61
1392	41.42	1.42
1414	6.9	
1415	162	1.62
1434	180	1.8
1722	179	1.79
2160	174	1.74

1135	15.4	1.54
2213	177	1.77

Anexo 11- Sustitución de valores imprecisos en la variable talla.

Historia Clínica	Peso inicial	Peso inicial sustituido	IMC
1329	10.1	101	33.7
1353	0.95	95	31
1547	77 1/2	77.5	26.3
1555	69 1/2	69.5	24.2
1567	82 1/2	82.5	32.9
1585	92 1/2	92.5	35.1
1596	11.5	115	39.7
1651	92 1/2	92.5	31.5
1697	90 1/2	90.5	30.7
1703	1.63	78.9	29.7
1826	66 1/2	66.5	25.1
1827	79 1/2	79.5	33.3

Anexo 12- Sustitución de valores imprecisos en la variable Peso inicial.

Historia Clínica	Peso final	Peso final Sustituido
1329	10.1	101
1513	94 1/2	94.5

1541	66 1/2	66.5
1544	73 1/2	73.5
1563	85 1/2	85.5
1565	101 1/2	101.5
1580	55 1/2	55.5
1585	91 1/2	91.5
1612	59 1/2	59.5
1638	84 1/2	84.5
1640	63 1/2	63.5
1642	92 1/2	92.5
1699	82 1/2	82.5
1827	78 1/2	78.5
1828	831/2	83.5
1830	95 1/2	95.5

Anexo 13- Sustitución de valores imprecisos en la variable peso final.

Identificador	Valor	Justificación
APP HTA	1: El paciente padece de hipertensión arterial. 0: El paciente no padece de hipertensión arterial	Es una de las enfermedades presentes en la historia clínica y además predomina en los pacientes (52,78%)

APP hiperlipoproteinemia	1: El paciente padece de hiperlipoproteinemia. 0: El paciente no padece de hiperlipoproteinemia	Aunque no predomina en la muestra (9,9%) es una de las enfermedades de la historia clínica.
APP cardiopatía isquémica	1: El paciente padece de cardiopatía isquémica. 0: El paciente no padece de cardiopatía isquémica	Aunque no predomina en la muestra (10,3%) es una de las enfermedades de la historia clínica.
APP clauditación intermitente	1: El paciente padece de clauditación intermitente. 0: El paciente no padece de clauditación intermitente	Esta variable no predomina en la muestra, solo la padece el 1,3% de los pacientes, pero es especificada por los médicos en la historia clínica.
APP otros	1: El paciente padece una de las siguientes enfermedades: ave, asma, problemas renales, hipercolesterolemia, neurosis, enfisema pulmonar, osteoporosis, glaucoma, hipotiroidismo, epilepsia y siclemia. 0: El paciente no padece ninguna de las enfermedades	En esta variable se añaden los pacientes que padecen ave porque aunque es una de las enfermedades explicitas en la historia clínica solo la padecen 3 personas y el resto son enfermedades que no están especificadas en la historia clínica y que no es considerable el número de pacientes que las padecen.

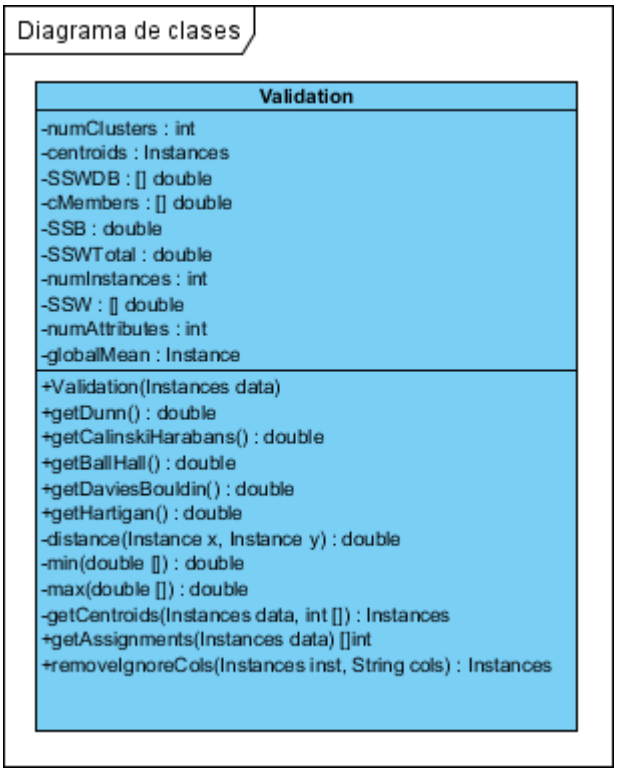
	mencionadas.	
APF HTA	1: El paciente presenta antecedentes familiares con hipertensión arterial. 0: El paciente no presenta antecedentes familiares con hipertensión arterial.	Es una de las enfermedades presentes en la historia clínica y es considerable la cantidad de pacientes que tienen este antecedente. (16,6%)
APF hiperlipoproteinemia	1: El paciente presenta antecedentes familiares con hiperlipoproteinemia. 0: El paciente no presenta antecedentes familiares con hiperlipoproteinemia	Esta variable solo es positiva para un 4,17 % de los pacientes pero es especificada en la historia clínica
APF cardiopatía isquémica	1: El paciente presenta antecedentes familiares con cardiopatía isquémica. 0: El paciente no presenta antecedentes familiares con cardiopatía isquémica	El 11,99% de los pacientes presentan este antecedente y es especificado en la historia clínica.
APF clauditación intermitente	1: El paciente presenta antecedentes familiares con cardiopatía isquémica. 0: El paciente no presenta antecedentes familiares con cardiopatía isquémica	Aunque solo presenta este antecedente el 1,85% de la muestra es una enfermedad que se especifica en la historia clínica
APF otros	1: El paciente presenta	La enfermedad AVE fue

	antecedentes familiares con ave, hipotiroidismo, asma, y dermatitis. 0: El paciente no presenta antecedentes familiares con estas enfermedades	incluida, aunque está en la historia clínica, porque solo un paciente presenta este antecedente. El resto son enfermedades no especificadas en la historia clínica y que no es considerable su cantidad en la muestra.

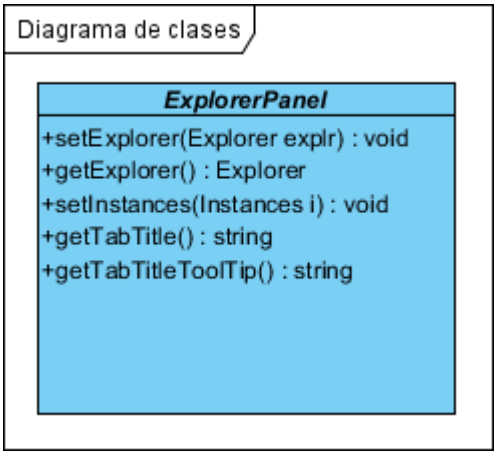
Anexo 14- Variables añadidas referente a los padecimientos personales y familiares

```
<current directory>
  +-DTNB.jar
  +-Description.props
  +-build_package.xml
  +-src
  |   +-main
  |   |   +-java
  |   |       +-weka
  |   |           +-classifiers
  |   |               +-rules
  |   |                   +-DTNB.java
  |   +-test
  |       +-java
  |            +-weka
  |               +-classifiers
  |                   +-rules
  |                       +-DTNBTest.java
  +-lib
  +-doc
```

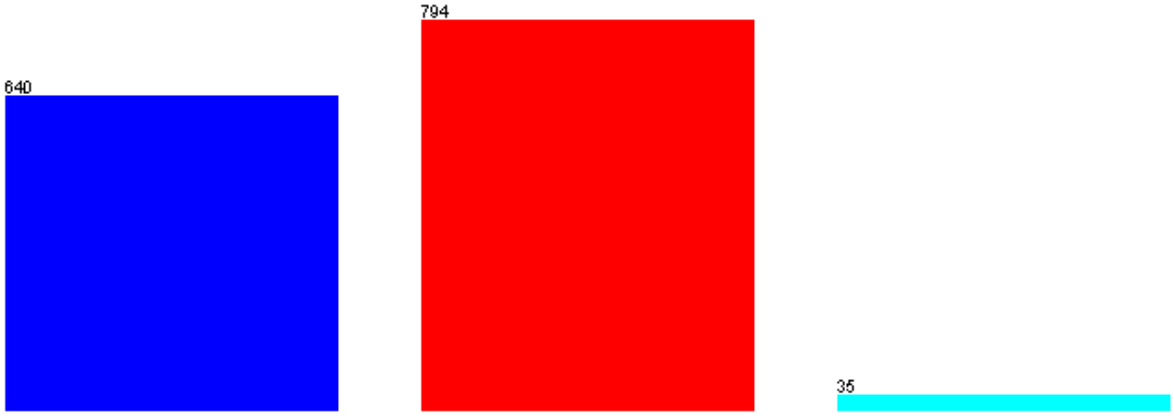
Anexo 15- Estructura de un paquete de WEKA.



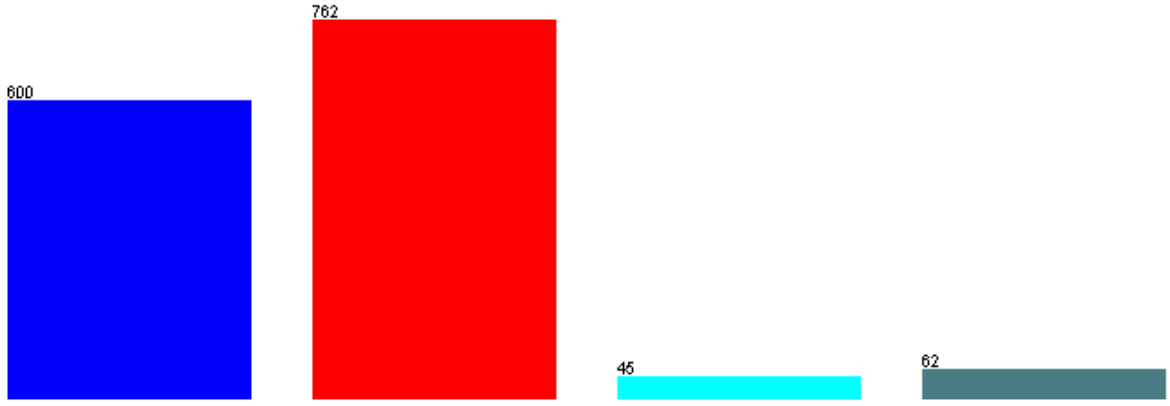
Anexo 16- Clase Validation.



Anexo 17- Clase ExplorerPanel.



Anexo 18- Grupos desbalanceados conformados por el algoritmo EM al experimentar con la totalidad de los casos (a).



Anexo 19- Grupos desbalanceados conformados por el algoritmo EM al experimentar con la totalidad de los casos (b).