CENTRO DE INVESTIGACIÓN CIENTÍFICA Y DE EDUCACIÓN SUPERIOR DE ENSENADA

PROGRAMA DE POSGRADO EN CIENCIAS DE LA COMPUTACIÓN

Extracción de reglas usando programación genética como base para un sistema de soporte a la toma de decisiones clínicas

TESIS

que para cubrir parcialmente los requisitos necesarios para obtener el grado de MAESTRO EN CIENCIAS

Presenta:

AMADO REYES VELÁZQUEZ MONTALVO

Ensenada, Baja California, México, Enero de 2011.

RESUMEN de la tesis de **Amado Reyes Velázquez Montalvo,** presentada como requisito parcial para la obtención del grado de MAESTRO EN CIENCIAS en Ciencias de la Computación. Ensenada, Baja California. Diciembre 2010.

Extracción de reglas usando programación genética como base para un sistema de apoyo a la toma de decisiones clínicas

Resumen aprobado por:

_____	_____
Dr. Fernando Rojas Íñiguez	Dra. Ana Isabel Martínez García
Codirector de Tesis	Codirector de Tesis

La toma de decisiones se da en todo tipo de procesos. En la práctica clínica es especialmente complejo tomar una decisión. Los sistemas computacionales que apoyan a la toma de decisiones en la salud (HDSS por sus siglas en inglés) han ido evolucionando desde sistemas que dan soporte a los diferentes procesos que utiliza un hospital hasta los que apoyan la toma de decisiones clínicas del médico con conocimiento nuevo.

El propósito de este trabajo de tesis es el descubrimiento de conocimiento nuevo sobre un padecimiento denominado el síndrome metabólico, aplicando un enfoque socio-técnico que guía todo el proceso de descubrimiento de conocimiento. En este enfoque se le da realce a entender la forma en que el médico maneja el síndrome metabólico. Producto de este análisis se determinó que se debía obtener una clasificación de riesgo del síndrome metabólico que permita al médico ubicar su manejo incluso en las etapas tempranas del padecimiento.

Se analizó un caso de estudio real en un hospital donde se identificó el manejo que se hace del síndrome metabólico por parte de los diferentes involucrados. Al mismo tiempo se recabó una serie de archivos generados en el área de epidemiología con el propósito de entenderles, limpiar y transformar los datos para obtener un conjunto de datos listo para ser minados (vista minable). Al analizar los procesos médicos relacionados con el síndrome se encontró un eje sobre el cual gira el padecimiento, la obesidad. Para encontrar la clasificación se seleccionó a la red neuronal semántica de Kohonen como el algoritmo de formación de conglomerados que mejor preserva las relaciones entre los datos relacionados con el síndrome y este eje. Una vez obtenidas las clases, se aplicaron dos algoritmos distintos para obtener las reglas que determinan la pertenencia en cada clase: el algoritmo C4.5 y la programación genética (GP por sus siglas en inglés). La GP permitió generar un conjunto de reglas con una alta exactitud predictiva, y para

mejorar sus resultados se utilizó una gramática restrictiva la cual asegura la congruencia en la conformación de las reglas. Se obtuvieron 25 reglas divididas en 5 clases de síndrome metabólico; las que se evaluaron con una toda la población de prueba encontrando una exactitud predictiva promedio del 90%.

En base a la clasificación de riesgo obtenida y otros elementos de apoyo que de ella se desprenden, se desarrolló el SATDSmet un CDSS que da apoyo al médico al tomar decisiones relacionas con el manejo del síndrome. Finalmente se realizó una evaluación cualitativa para conocer la percepción de utilidad de la clasificación y del CDSS, mediante pruebas realizadas con los médicos. Donde se pudo obtener un tendencia que marca que ambos son percibidos como útiles y bien aceptados por los médicos participantes.

Palabras clave: clasificación de riesgo del síndrome metabólico, minería de datos, proceso de descubrimiento de conocimiento con enfoque socio-técnico, red neuronal semántica de Kohonen, programación genética con semántica restrictiva, sistema de soporte a la toma de decisiones clínicas.

ABSTRACT of the thesis presented by **Amado Reyes Velázquez Montalvo** as a partial requirement to obtain the MASTER OF SCIENCE degree in Computer Science. Ensenada, Baja California, México December 2010.

Rules Extraction using Genetic Programming as a base for a Clinic Decision Support System

Take decisions happens in all kind of process. Take a decision is particularly complex in the clinical practice. The health decision support system (HDSS) has evolved from management hospital process till those than supports the clinical decisions with brand new knowledge for the medic.

The aim of this thesis is discover new knowledge about a disease named metabolic syndrome, applying a socio-technical approach who guides all the knowledge discovery process. This approach embosses the understanding of how the medic handles the metabolic syndrome. Using this approach we determined the need of a metabolic syndrome risk classification that makes possible handle the syndromes even in the early stages of develop.

We analyze a study case in a hospital where we identify the metabolic syndrome handling of several stake holders. At same time we collected a set of files generated by the epidemiology area. We understand, clean and transform this data to obtain a set ready to mining (the data mining view). Analyzing the medical process related with the syndrome we found an axis for the syndrome build up: the obesity. To find the classification we selected the semantic Kohonen neural network as the clustering algorithm that better preserves the relations between data involved with the syndrome and this axis. Once obtained the classification, we applied two different algorithms to obtain the rules that establish the membership in every class: the C4.5 and the genetic programming (GP). The GP made possible obtain a set of rules with high predictive accuracy, to improve this results we used a restrictive grammar to assure congruence in the rules construction. We obtained 25 brand new rules divided in five metabolic syndrome classes; these rules were evaluated with the complete test population founding a 90% of predictive accuracy.

Based on the new risk classification, and other support elements extracted from the same classification, we developed the SATDSmet a CDSS that supports medical decision related with the syndrome. Finally we made a qualitative evaluation to determine the classification and the CDSS utility perception. We found that both of them are useful and a have a good acceptance inside the medical evaluation team.

Keywords: Data Mining, Semantic Kohonen's Neural Network, Genetic Programming, extraction of medicine rules, socio-technical approach, Clinic Decision Support System.

Dedicatorias

A Beatriz mi amada esposa e Iker mi querido hijo, quienes siempre me han apoyado.

A mis padres a los que les hubiese gustado tanto acompañarme en este momento.

Y a mi entrañable hermano Roberto por su ejemplo, que nunca olvidaré.

Agradecimientos

A la Dra. Ana I. Martínez y el Dr. Fernando Rojas por sus consejos y su apoyo constante e incondicional.

Al Dr. Armando Paniagua por su invaluable asesoría y apoyo.

A mis compañeros de generación por su apoyo y compañerismo, sobre todo cuando más difíciles fueron las cosas.

A mis asesores de tesis por sus desveladas al revisar esta tesis.

Al doctor Sebastian Ventura por su ayuda.

Un agradecimiento al Conacyt por su apoyo económico, CVU 268181.

CONTENIDO

CONTENIDO (continuación)

CONTENIDO (continuación)

CONTENIDO (continuación)

Página

CONTENIDO (continuación)

Página

CONTENIDO (continuación)

Página

LISTA DE FIGURAS

Figura **Página**

LISTA DE FIGURAS (continuación)

Figura **Página**

LISTA DE FIGURAS (continuación)

Figura **Página**

LISTA DE FIGURAS (continuación)

Figura **Página**

LISTA DE TABLAS

Capítulo I

Introducción

I.1 Antecedentes

La toma de decisiones se da en todo tipo de procesos, sin embargo las que se realizan cuando un médico atiende un paciente son especialmente complejas y pueden llegar a ser extremadamente costosas.

Tomar una decisión se define como "*hacer una estimación con respecto a lo que se debería hacer en cierta situación después de haber deliberado en algunos cursos de acción alternativos"* (Ofstad, 1961).

Los sistemas computacionales dedicados al apoyo de la toma de decisiones se agrupan en el concepto denominado sistemas de apoyo a la toma de decisiones (Decision Support Systems-DSS por sus siglas en inglés), en el caso de la medicina existen los sistemas de apoyo a la toma de decisiones en salud (Health Decision Support System-HDSS por sus siglas en inglés) que abarcan gran variedad de tipos: el control de expedientes médicos, el control de información de laboratorio, el registro de admisión-contable, y finalmente los sistemas de apoyo a la toma de decisiones clínicas (Clinical Decision Support Systems-CDSS por sus siglas en inglés).

Existe un tipo de CDSS que está basado en la extracción de conocimiento, en el que es fundamental la forma en que se obtiene. Son dos enfoques principales de

cómo se extrae el conocimiento: extrayéndolo de los humanos y extrayéndolo de los datos (Tan, et al., 1998).

La idea central es descubrir conocimiento nuevo y útil, y no solo información. Desde este punto de vista la información, el conjunto de datos, es solo la materia prima de un proceso de análisis que permite entender las relaciones entre ellos y finalmente expresar conceptos, reglas, leyes, etc. Este producto final es el conocimiento.

El proceso implica analizar los datos registrados en la memoria organizacional desde dos puntos de vista distintos pero complementarios:

1. El de los procesos que generan dichos datos y
2. Desde los datos mismos.

Desde el punto de vista del primero la ingeniería de procesos (IP) aporta una serie de técnicas que permiten modelar los procesos capturando la forma en que se generan los datos y las decisiones que toman los participantes en el proceso; desde el punto de vista del segundo, existen el proceso de descubrimiento de conocimiento en bases de datos (Knowledge Discovery on Data Bases-KDD por sus siglas en inglés) que nos permite encontrar relaciones entre los datos por medio de las cuales podemos discernir conocimiento en apoyo a la toma de decisiones.

En este trabajo de tesis se presenta un caso real sobre el manejo de pacientes con un problema de salud llamado síndrome metabólico; esto se llevó a cabo con el Instituto Mexicano del Seguro Social (IMSS), analizado desde ambas perspectivas: la social, esto es tomando en cuenta la forma en que los médicos interpretan los datos y toman decisiones dentro del proceso aplicando la ingeniería de procesos; y desde el punto de vista técnico o duro que es en relación a los datos que quedan registrados en la memoria organizacional; esto último mediante el KDD. Lo anterior como parte de la metodología de análisis y apoyo a la toma de decisiones propuesta por Pacheco (2004).

El proceso KDD está constituido por 5 fases: integración y recopilación de la información, selección limpieza y transformación de los datos, minería de datos, evaluación e interpretación, difusión y uso (Hernández Orallo, et al., 2004). La fase de minería de datos es aquella donde se aplican una serie de algoritmos matemáticos de inteligencia artificial, que permiten extraer conocimiento de acuerdo al tipo de problema planteado. Los algoritmos evolutivos forman parte del cuerpo de algoritmos de los que hace uso KDD, y uno de ellos es la programación genética (Genetic Programming-GP por sus siglas en inglés) la que tiene como objetivo que una computadora evolucione programas de computadora por sí sola. Actualmente se utiliza para evolucionar muchos otros tipos de estructuras; como lo pueden ser las reglas de una clasificación, haciendo que al final del proceso evolutivo se encuentren reglas acertadas y robustas que puedan ser aceptadas y aplicadas por el médico en la consulta. En términos generales la GP se basa en la aplicación de los mecanismos evolutivos, de selección, cruzamiento, mutación y reemplazo, que al ser aplicados permiten encontrar una serie de soluciones candidatas altamente adaptadas para resolver el problema planteado; en este caso buscando extraer una serie de reglas robustas, nuevas y finalmente útiles para que el médico se apoye en ellas durante el manejo del síndrome metabólico.

Al utilizar la metodología para el análisis y apoyo a la toma de decisiones integrando el aspecto social y técnico, se integra al proceso KDD con criterios y métodos utilizados por el médico al manejar el síndrome metabólico.

Todo esto con el propósito de generar conocimiento útil que pueda ser utilizado para el manejo del síndrome metabólico.

I.2 Trabajo previo.

Son varios los trabajos de donde se aplica el KDD para descubrir conocimiento nuevo en medicina:

- Programación evolutiva y redes bayesianas para obtener conocimiento en bases de datos de fracturas de huesos y escoliosis (Leung Wong, et al., 1999).
- Programación genética para analizar bases de datos sobre enfermedades relacionadas con el dolor del pecho (Bojarczuk, et al., 2004).
- Uso de un algoritmo evolutivo en conjunto con programación genética, para analizar bases de datos de hepatitis y cáncer de mama (Ta-Cheng, et al., 2006).
- Uso de algoritmos genéticos para analizar bases de datos de cáncer de mama (Ta-Cheng, et al., 2006).
- Uso de un algoritmo híbrido que usa programación genética y máquinas de soporte vectorial, para analizar bases de datos de hepatitis, cáncer de mama, y diabetes (Tan, et al., 2009).

La mayoría de estos trabajos se caracterizan por el énfasis de la parte dura del proceso, aquella que tienen que ver con lo óptimo de las reglas encontradas, el desempeño del algoritmo, el establecimiento de una métrica que permita determinar un criterio novedoso o de utilidad en el conocimiento encontrado.

En el caso particular del trabajo de Bojarczuk (2004) se propone el uso de la GP para el descubrimiento de reglas de clasificación, estableciendo en sus conclusiones que las reglas resultantes fueron claras, simples y exactas. El énfasis de este trabajo se centra en la parte dura del proceso y en la comparación contra otro tipo de algoritmos. El problema que adolece este enfoque es que la evaluación de lo bueno de la regla se basa solo en la función de adaptabilidad como una forma de optimizar la exactitud predictiva de la regla, haciendo a un lado otros criterios que el médico toma en cuenta al aplicar una regla médica. El planteamiento de la presente tesis es que este tipo de criterios deben guiar a la programación genética y no solo usarse para establecer la función de adaptabilidad. Incluso se deben integrar otros algoritmos que cooperen con la GP

para cubrir todos los aspectos que el médico toma en cuenta y llevar la evaluación de la regla más allá de que tan simple o novedosa sea.

La tendencia actual de los CDSS es apoyar en la toma de decisiones del especialista aportando conocimiento que se integre rápidamente a su conocimiento previo y experiencia personal. En este sentido existe la necesidad de introducir la visión socio-técnica al KDD, con el objetivo de que el conocimiento sea bien aceptado por quienes finalmente le van a utilizar, los médicos. Las reglas que se generen deben ser robustas en tres sentidos, su exactitud al predecir el síndrome, cortas para hacerles más entendibles, finalmente novedosas y funcionales, para lo que se aplica la GP.

I.3 Planteamiento del problema.

Actualmente México ocupa el primer lugar a nivel mundial en obesidad infantil, y el 70% de la población mexicana tiene problemas de sobrepeso u obesidad (Téllez, 2010). Las enfermedades relacionadas con obesidad y sobrepeso se han vuelto una epidemia en México y gran parte del mundo. Aun cuando en México se han realizado diversos esfuerzos para manejar el problema de la obesidad y sus consecuencias, no se cuenta con información precisa para el manejo del síndrome metabólico; aún siendo este un elemento generalmente reconocido como causal de la obesidad y de las enfermedades que se relacionan con ella (Amy, 2007).

Además no existe en la literatura general ni en las fuentes de información médica más reconocidas como la organización mundial de la salud, american heart association, y entidades afines en Europa, una clasificación de riesgo sobre el síndrome tal que permita evaluar en una población, el nivel de riesgo que implica

el padecer el síndrome y ubicarle en diferentes etapas de la evolución del padecimiento.

Actualmente se cuentan con diversas herramientas de apoyo para el manejo de las enfermedades relacionadas con el síndrome, en ellas es difícil diferenciar los elementos de diagnóstico de las enfermedades relacionadas (comorbilidad[1]) del manejo propio del síndrome metabólico. Hacen falta herramientas de apoyo para tratar adecuadamente el síndrome metabólico, ya sea:

- Para prevenir las condiciones que provocan los padecimientos que lo componen
- Para mejorar las condiciones de salud de quienes ya sufren alguna enfermedad del síndrome metabólico y que pueden provocar otro de sus padecimientos

De aquí que existe la necesidad de proporcionar datos sustentados que permitan establecer una clasificación de riesgo sobre el síndrome metabólico que apoye a los médicos en el manejo del mismo.

I.4 Objetivos.

La extracción de reglas mediante la programación genética como base para el desarrollo a un sistema de apoyo a la toma de decisiones clínicas en apoyo al manejo del síndrome metabólico. Los objetivos específicos son los siguientes:

- Entender los elementos que conforman al síndrome metabólico
- Entender el funcionamiento y los elementos que forman los algoritmos que se van a usar

[1] La concurrencia de más de una patología en la misma persona (Kahl, 1990)

- Identificar en un caso de estudio real de práctica clínica, el manejo del síndrome metabólico
- Recolectar la memoria organizacional para generar un almacén de datos, enfocándose al síndrome metabólico
- Realizar un modelo socio-técnico del proceso de toma de decisiones
- Aplicar el proceso de extracción de conocimiento desde los datos
- Utilizar la programación genética como técnica de la minería de datos, generando reglas robustas
- Elaborar un sistema de apoyo a la toma de decisiones en base a las reglas descubiertas
- Evaluar los resultados con el especialista y desde el punto de vista de los algoritmos

I.5 Metodología de Investigación.

La metodología que permitió realizar este trabajo de tesis, integra en sus fases al proceso KDD, y las fases que le componen entre ellas la de minería de datos usando programación genética para obtener las reglas de clasificación del riesgo.

I.5.1 Revisión de la literatura.

Como se observa en la Figura 1 antes de poder extraer cualquier tipo de conocimiento se revisó la literatura para encontrar una metodología que permitiese establecer cuál es el problema de toma de decisiones a resolver, así como la forma en la que se le iba a resolver.

Otra revisión que se hizo fue la de varios libros donde se describe el proceso KDD y los pasos que le componen, se identificó cuál es el tratamiento previo que se le debe dar a la información, se revisó a fondo la fase de minería de datos en diferentes fuentes y se identificaron cuáles eran los algoritmo a utilizar y las pruebas que se deben hacer para evaluar la calidad de los resultados obtenidos.

También se analizó el contexto médico del síndrome metabólico en diferentes formas: primeramente se revisó la encuesta nacional de salud para dimensionar el tamaño del problema que representa el síndrome metabólico y su comorbilidad, se revisaron libros de fisiología humana para entender el síndrome, y finalmente se participó en la revisión que el cuerpo médico del hospital de zona 8 del IMSS hizo de las nuevas guías clínicas para el manejo de la Hipertensión Arterial (HTA) y la Diabetes Mellitus 2 (DM2), enfermedades que forman parte de la comorbilidad del síndrome.

I.5.2 Caso de estudio real.

Se realizó un caso de estudio real del proceso de manejo del síndrome metabólico, por parte del cuerpo de medicina familiar del hospital de zona 8 del IMSS, contando con el apoyo del departamento de epidemiología.

Para identificar y poder entender el problema de toma de decisiones que involucra el manejo del síndrome metabólico, se dio seguimiento a los actores participantes en el proceso (médicos familiares, epidemiólogos, asistentes, sicólogos, etc.) que participan en el proceso vistos desde una perspectiva de todas las áreas que

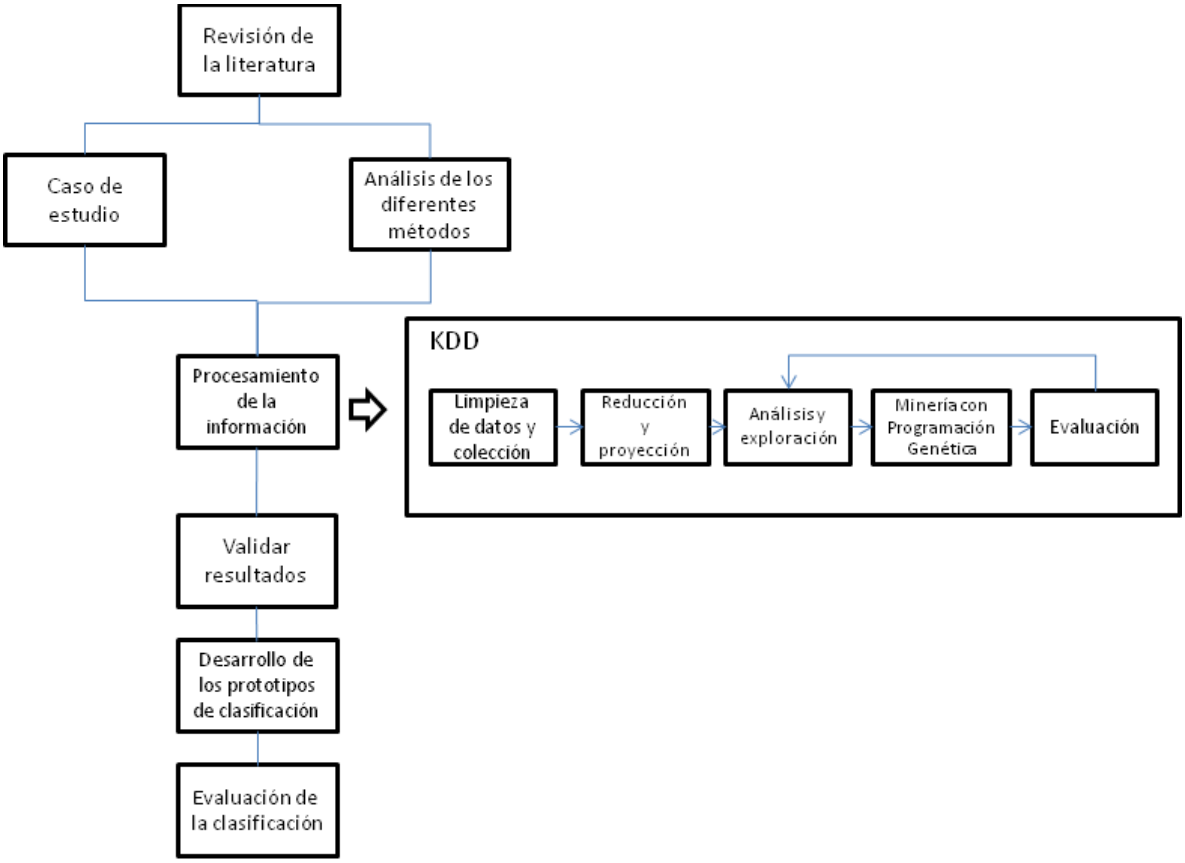


Figura 1. Esquematización de la metodología de investigación.

participan en el manejo que se hace del síndrome. Se recabaron una serie de entrevistas y notas, con las cuales se elaboraron modelos que permitieron identificar el problema de toma de decisiones a resolver, además de obtener una serie de archivos a los cuales posteriormente se les aplicó el KDD utilizando varios métodos complementando la GP.

Con la información obtenida se integró una base de conocimiento extraída directamente de las guía clínicas de donde se recabo todas las indicaciones del manejo del paciente con síndrome metabólico y los datos que se utilizan para su manejo (valores, escalas, claves, forma en que se obtuvo el dato, etc).

I.5.3 Análisis de los diferentes métodos.

Se aplicó la metodología de Pacheco (2004) que permitió delinear los procesos de toma de decisiones, así como los elementos que le componen dentro del caso de estudio.

Producto de esta etapa de la metodología, se obtuvieron una serie de modelos que dieron una visión socio-técnica del problema, donde se identificaron los actores participantes en los procesos de toma de decisiones, las decisiones a tomar y las incertidumbres que se tienen al manejar el síndrome.

El análisis realizado permitió resumir el problema de toma de decisiones como la falta de apoyo para el médico, tanto familiar como epidemiólogo, al determinar el riesgo que representa el síndrome metabólico en todas sus etapas, principalmente antes de que aparezca su comorbilidad; y el impacto que tiene el que se establezcan acciones, hasta que la comorbilidad del síndrome ha avanzado. Por lo que una herramienta básica a construir fue una clasificación del riesgo del síndrome, que fue la base para el apoyo que proporciona el CDSS.

Se realizaron diferentes estudios para determinar las versiones del tipo de algoritmos que se iban a tomar en cuenta para elaborar la clasificación de riesgo. Se estableció que se trata de un problema de clasificación donde no se cuenta a priori con las clases que la forman, y que la naturaleza médica de la información a trabajar es de carácter muy similar al análisis semántico que se hace al analizar textos, esto es, donde existe un contexto que aporta información importante sobre la clasificación y no solo la pura comparación de valores.
En una segunda etapa, ya contando con la clasificación, se requirió de algoritmos para extraer reglas que definen la pertenencia a cada grupo de la clasificación.

Los algoritmos seleccionados fueron: redes neuronales semánticas de Kohonen para conformar los grupos (Zurada, 1992), algoritmo C4.5 para extraer las reglas iniciales que determinan la clase (Witten, et al., 2005), y la GP para optimizar dichas reglas (Bojarczuk, et al., 2004).

I.5.4 Procesamiento de la información.

Es en esta fase que se llevó a cabo el proceso de KDD donde se realizaron las siguientes actividades:

- Pre-procesamiento de la información: En este paso se trabajó con un primer archivo de datos, el cual se analizó minuciosamente limpiando la información, entendiendo la naturaleza de los datos, los intervalos válidos de los valores que puede tomar el dato, y se elaboraron algunas estadísticas para identificar la composición de la muestra de datos obtenida, sobre todo por lo importante que es la variabilidad dentro de la muestra. Además, se realizó el cálculo del tamaño de la muestra basándonos en un método estadístico.
- Análisis y transformación de los datos: De acuerdo a la literatura analizada, se hicieron diferentes pruebas utilizando algunas aplicaciones que manejan minería de datos, identificando algunos aspectos importantes sobre las escalas utilizadas en cada dato, identificando las normalizaciones hechas por el propio epidemiólogo. Al final de este paso se obtuvo un conjunto de datos listos para la minería de datos.
- Minería de datos: Se obtuvieron los patrones que representan el conocimiento descubierto. Primeramente se trabajó con un grupo de datos de entrenamiento para identificar cada uno de los parámetros que manejan los algoritmos y cuáles son los valores más adecuados para cada

parámetro. Posteriormente se integraron los datos de prueba y se obtuvieron los conglomerados que forman la clasificación.

- Programación genética buscando obtener reglas más robustas.

I.5.5 Validar Resultados.

En esta fase de la metodología, se procedió a analizar en conjunto con el especialista epidemiólogo el conocimiento obtenido, validándole al compararle con los conocimientos previos que se tienen sobre el síndrome metabólico, la experiencia previa que tiene el especialista sobre el tema, y la congruencia que presentaban las reglas y las transiciones entre ellas.

I.5.6 Desarrollo de los prototipos de clasificación.

En base a las reglas y las transiciones se elaboran elementos que permiten un mejor manejo del conocimiento expresado en las reglas. En conjunto con el epidemiólogo se revisa el contexto médico de cada regla. Con las transiciones se identifica la severidad del síndrome de una clase a otra, lo que permite la jerarquización de las clases; además se define una tipología de los pacientes de acuerdo al perfil que establecen las reglas en cada clase, el objeto principal de este trabajo es re-expresar como se identifican las clases, de tal forma que basados en las reglas se establecen las característica que debe tener un paciente para pertenecer a la clase y esto es más fácil de asimilar para el médico. Por ejemplo si la regla dice que la persona tiene colesterol hasta 200 mg/dl se re-expresa diciendo que el paciente tiene el colesterol limítrofe, esto es que el valor

detectado en el nivel de colesterol está en los límites que marca la regla como punto de corte para decir que se padece una dislipidemia.

Contando con estos elementos se procedió a la construcción del CDSS para apoyar el manejo del síndrome de acuerdo a los resultados obtenidos; bajo dos enfoques, el del médico como tomador de decisiones al momento de la consulta de un paciente, y desde el punto de vista del epidemiólogo al manejar grupos de pacientes evaluando el riesgo cardiovascular.

Los elementos que se establecieron para integrar el conocimiento que sirve como base del sistema son: la clasificación de riesgo obtenida, la tipología delineada en conjunto con el epidemiólogo, las colindancias entre los conglomerados encontrados para determinar la posible evolución del síndrome, y una serie de acciones a ejecutar de acuerdo a las características de los indicadores que arroja la regla y su correspondencia con las guías médicas.

I.5.7 Evaluación de la clasificación.

Finalmente se procedió a la evaluación de los resultados obtenidos desde dos perspectivas:

- Para determinar la calidad desde el punto de vista del conocimiento previo del experto epidemiólogo de las reglas; este proceso fue iterativo buscando encontrar las reglas idóneas de acuerdo al criterio técnico del especialista.
- Desde el punto de vista de la utilidad y la facilidad de uso de las herramientas desarrolladas. En este caso se evaluó el CDSS mediante una serie de pruebas con un grupo variado de médicos familiares y el epidemiólogo.

I.6 Contenido del documento de tesis.

En este Capítulo se presenta la forma en que se estructuró el contenido del documento de tesis, así como una breve sinopsis del contenido de cada uno de los Capítulos.

Capítulo II. En este Capítulo de define el proceso de toma de decisiones, tanto desde un punto de vista general, como desde el punto de vista de la práctica clínica. Se presenta la metodología para analizar el proceso de toma de decisiones utilizada, para entonces introducir el concepto de sistema de soporte a la toma de decisiones y de decisiones clínicas. Finalmente se introduce el proceso de descubrimiento de conocimiento en bases de datos y su fase principal, la minería de datos.

Capítulo III. En este Capítulo se presentan a detalle los métodos de extracción de conocimiento utilizados. Se inicia definiendo el problema de clasificación, para entonces describir el aprendizaje no supervisado. Se presentan los algoritmos utilizados iniciando con el de K-Medias (base teórica de los algoritmos utilizados), la red neuronal de Kohonen y su producto principal, que son los mapas auto-organizados. Para finalizar presentando los algoritmos C4.5 que permitirá la primera extracción de reglas, y la programación genética como el algoritmo que optimizará las reglas.

Capítulo IV. En este Capítulo se presenta un análisis del contexto médico y del manejo del síndrome metabólico. El contexto médico permite conocer al síndrome, sus causas y efectos. Finalmente se presenta a detalle la aplicación de la metodología de análisis del proceso de toma de decisiones y la información más relevante que se usa en el manejo del síndrome metabólico.

Capítulo V. En este Capítulo se describe a detalle el proceso de generación de la clasificación de riesgo del síndrome metabólico, la formación de la vista minable, y los resultados de la red neuronal. En base al conglomerado obtenido, donde se establece a cual pertenece cada individuo, se presentan los resultados de la extracción de reglas tanto por C4.5, como por programación genética. Se finaliza presentando la clasificación de riesgo y la evaluación de calidad de las reglas que le componen.

Capítulo VI. En este Capítulo se presenta el diseño e implementación del sistema de soporte a la toma de decisiones clínicas. Se detalla el diseño arquitectónico, los casos de uso y los diagramas de secuencia. Al final se presenta el sistema implementado.

Capítulo VII. En este Capítulo se presenta la evaluación de la clasificación y del sistema de apoyo a la toma de decisiones clínicas; así como, una serie de experimentos donde se evaluaron de forma cualitativa la utilidad y la facilidad de uso percibida por el usuario.

Capítulo VIII. Se presentan una serie de conclusiones sobre el trabajo realizado y las aportaciones al conocimiento y el trabajo futuro propuesto.

Apéndice A. Presenta los formatos de los archivos preliminares utilizados para obtener la vista minable, las convenciones establecidas para la transformación de los datos dentro de la vista minable. La información que finalmente se registra en el sistema de apoyo a la toma de decisiones clínicas sobre el síndrome metabólico. Y finalmente el formato del archivo que recibe de entrada el sistema para trabajar poblaciones de pacientes.

Apéndice B. Base de reglas para determinar la comorbilidad del síndrome metabólico.

Apéndice C. Reglas de la clasificación del síndrome metabólico obtenidas por los diferentes algoritmos así como las de la clasificación final. También se presenta la evaluación de exactitud predictiva del conjunto de reglas finales y las tablas tetracóricas de cada regla obtenida.

Apéndice D. Protocolo de la entrevista para la evaluación de utilidad y facilidad de uso percibida.

Apéndice E. Gráficas ricas, diagramas IDEF, diagramas de influencias no presentadas en el cuerpo de la tesis.

Apéndice F. Casos de uso no presentados en el cuerpo de la tesis.

Capítulo II

Apoyo a la Toma de Decisiones en Salud usando Minería de Datos para la extracción de conocimiento

II.1 Introducción

Una decisión siempre implica cierto nivel de incertidumbre y un riesgo que conlleva un costo.
La mayoría de las decisiones importantes que se toman ya sea personales o dentro de alguna organización son complejas y afectan el entorno en el que se llevan a cabo (Pacheco Soto, 2004).

En el contexto médico las decisiones que se toman tienen las características arriba citadas, además de ser delicadas y costosas. Los expertos médicos toman decisiones clínicas, generalmente acertadas, pero en algunos casos pueden verse afectadas por la falta de una herramienta específica que ayude a tomar la decisión o al omitir algún concepto que pudiese haber sido valioso.

En la presente tesis se analiza la toma de decisiones relacionada con el manejo del síndrome metabólico.
El síndrome metabólico es un grupo de datos clínicos (cuadros) que ponen en riesgo de desarrollar una enfermedad cardiaca y diabetes. Estos datos clínicos (cuadros) son:

- *Hipertensión arterial*
- *Aumentos de los niveles de azúcar*
- *Niveles sanguíneos elevados de triglicéridos, un tipo de grasas*

- *Bajos niveles sanguíneos de HDL, el colesterol bueno*
- *Exceso de grasa alrededor de la cintura* (IDF, 2006).

Por lo tanto se trata de un desorden generalizado del organismo que está muy relacionado con enfermedades crónico degenerativas tales como: obesidad, hipertensión arterial (HTA), diabetes mellitus tipo 2 (DM2), dislipidemias (desordenes relacionados con grasa en la sangre).

El mal manejo de las decisiones tomadas con relación a este padecimiento ha hecho que en el 2008 México ya ocupase el segundo lugar a nivel mundial de obesidad solo por debajo de Estados Unidos. Que diecisiete millones de mexicanos mayores de 20 años padezcan obesidad, y de acuerdo a la tendencia observada en el crecimiento de la obesidad infantil se tenga la expectativa, en 10 años, de que el 90% de los mexicanos sean obesos (Rocha, 2010). Un ejemplo del impacto de las enfermedades relacionadas con el síndrome metabólico es que 90% de las camas de hospital de la Ciudad de México estén ocupadas por enfermos que padecen enfermedades crónico degenerativas (Universal, 2010).

De acuerdo a los indicadores, antes presentados, los problemas relacionados con la obesidad se han convertido en un asunto prioritario de salud pública en México.

Buscando atacar este problema instituciones tales como el Instituto Mexicano del Seguro Social actualmente cuentan con herramientas para el manejo de las enfermedades relacionadas con el síndrome metabólico. Igualmente el gobierno mexicano tiene una normatividad para el manejo de la obesidad, la diabetes mellitus tipo 2 y la hipertensión arterial. Sin embargo, aunque en todas ellas se menciona el síndrome metabólico, no existen herramientas específicas para su manejo.

De acuerdo a la literatura (IDF, 2006) el síndrome metabólico se determina si al menos tres de los datos clínicos (cuadros) arriba citados se cumplen en el paciente.

En términos generales ésta es la información con que se cuenta para evaluar la existencia o no del padecimiento. Podemos observar que no es una clasificación

jerárquica del riesgo que representa el padecimiento, por lo que no tiene mucha utilidad al momento de la toma de decisiones en la práctica clínica. Es esta la razón por la cual el síndrome metabólico solamente se utiliza como una referencia médica sin uso práctico al momento de tomar decisiones sobre dicho síndrome.
Uno de los objetivos de la presente tesis es obtener una serie de reglas que permitan determinar el nivel de riesgo de padecer el síndrome metabólico expresado en forma de una clasificación jerárquica útil en la práctica clínica.

Hasta hace poco tiempo la principal preocupación del médico clínico era la solución de problemas con respecto al diagnóstico, y en algunos casos, con respecto al pronóstico de sus pacientes. Actualmente, aunque sigue teniendo mucha importancia la medicina clínica, se hace gran énfasis en la capacidad para administrar grandes cantidades de información y el mayor número de recursos tecnológicos disponibles para el diagnóstico (Ruiz, et al., 2004).

En el presente Capítulo se analiza la forma en que los expertos médicos están tomando las decisiones relacionadas con el síndrome metabólico para entender cuáles son estas decisiones, que incertidumbres se tienen al tomarlas y con qué cursos de acción cuenta un especialista para manejar el síndrome metabólico.
Iniciamos presentando un análisis sobre el proceso de toma de decisiones en general y en el ámbito médico en particular (sección II.2). En la Sección II.3 se revisa una metodología para el apoyo a la toma de decisiones y después en la sección II.4 se presentan los llamados sistemas de apoyo a la toma de decisiones y los sistemas de apoyo de la toma de decisiones clínicas.

Ya se mencionó la importancia de poder analizar grandes cantidades de información y extraer conocimiento de ella, en este sentido la sección II.5 introduce el proceso de extracción del conocimiento de bases de datos, y la sección II.6 está dedicada a la fase de este proceso denominada Minería de Datos durante la cual se puede obtener conocimiento a través de la obtención e

interpretación de los patrones observados en los datos. Finalmente se presenta un resumen del Capítulo.

II.2 Toma de decisiones.

Tomar una decisión se define como hacer una estimación respecto a lo que se debería hacer en cierta situación después de haber deliberado en algunos cursos de acción alternativos (Ofstad, 1961).

De acuerdo a Pacheco Soto (2004) aun cuando las decisiones pueden ser de muy diversa naturaleza existen ciertos conceptos y términos comunes que permiten generalizar una metodología para su análisis, estos son:

- Complejidad: la mayoría de las decisiones importantes son complejas y afectan el entorno en que se llevan a cabo
- Estructura de los problemas de decisión: la forma de tomar decisiones puede referirse a problemas rutinarios y repetitivos para los cuales el proceso para obtener la mejor solución se conoce(estructurada); cuando varios de los elementos de un problema están mal definidos o simplemente no se conocen(no estructurada); y cuando se tiene ambas situaciones (semiestructurada).
- Datos e Información: ambos conceptos son diferentes entre sí, los datos son simples observaciones, mientras que la información es la que es obtenida cuando los datos relevantes son manipulados para dar soporte a la toma de decisiones.
- Incertidumbre: la incertidumbre es algo que es desconocido totalmente o no es conocido perfectamente.

- Objetivos: los objetivos en una decisión son lo que se busca al tomarla, frecuentemente estos objetivos pueden ser financieros, sociales, de salud, etc.
- Alternativas: son las diferentes opciones que se tienen en un problema de decisión, la falta de alternativas hace del proceso de toma de decisiones un proceso trivial.
- Probabilidad: cuando hay incertidumbre en las decisiones se usa la probabilidad para expresar explícitamente las opiniones acerca de la incertidumbre de eventos futuros o eventos los cuales han ocurrido pero no se tiene conocimiento del resultado, dicho de otra forma, la probabilidad es un juicio subjetivo acerca de la posibilidad que ocurra un evento futuro incierto (Skinner, 1999).
- Modelos: un modelo es una forma de representar un problema, este generalmente simplifica la visión que se tiene de la realidad

El análisis de toma de decisiones en medicina se define como: *un método de examen cuantitativo para comparar el valor relativo de varias decisiones, y para identificar la solución óptima ante condiciones de incertidumbre o cuando es necesario hacer concesiones entre diferentes enlaces* (Ruiz, et al., 2004).

De acuerdo a la definición anterior para realizar un análisis adecuado al tomar una decisión en la práctica clínica se siguen tres pasos:

1. Definir el problema de toma de decisiones en forme clara, estructuradamente y habiéndole delimitado.
2. Definir un tiempo determinado, por ejemplo si se usan antibióticos de forma profiláctica[2], hay que definir durante cuánto tiempo se suministrarán.
3. Enumerar las posibles alternativas de solución de consecuencias en un tiempo determinado.
4. Crear un árbol de decisión, este árbol de decisión es definido como modelo gráfico de las acciones y de las consecuencias de las mismas.

[2] Medidas que se toman para evitar enfermedades y su propagación.

Figura 2. Ejemplo de un árbol de decisión utilizado en el análisis de decisiones en medicina (Ruiz, et al., 2004).

Como podemos observar el proceso de toma de decisiones en la práctica clínica, aún con características muy particulares, sigue los patrones del modelo general de toma de decisiones que plantea Pacheco Soto (2004). En la Figura 2 se presenta el ejemplo de un árbol de decisión para el manejo de un tratamiento médico después de una cirugía, en él podemos observar varios tipos de reglas que se manejan, se ejemplifican dos de ellas:

- Primera rama: *SI* inicia con medidas preventivas (profilaxis bacteriana) y hay efectos y no hay infección *ENTONCES* se emprende una acción representada por el triángulo
- Última rama: *SI* no inicia con medidas preventivas (profilaxis bacteriana) y no hay infección *ENTONCES* se emprende una acción representada por el triángulo

Se puede observar como la estructura de árboles es adecuada para el tipo de reglas *SI-ENTONCES.*

II.3 Metodología de análisis para el apoyo a la toma de decisiones.

Dado que la toma de decisiones médicas involucra procesos médicos bien establecidos en conjunto con el criterio del médico que toma la decisión, la metodología a utilizar en este trabajo de tesis es la propuesta por Pacheco Soto (2004) denominada soporte a la toma de decisiones con enfoque a la ingeniería de procesos. La cual sirve como una base para el desarrollo de sistemas de soporte grupal y está basada en un proceso de apoyo socio-técnico, es decir tomando en cuenta los procesos de toma de decisiones médicas en conjunto con los criterios de los médicos en cuestión.

Se compone de un modelo que presenta las fases adecuadas para estudiar un problema de decisión grupal. Cada fase está compuesta de actividades que ayudan a identificar el soporte más adecuado al problema de decisión.

La metodología que guía el proceso de creación del CDSS de principio a fin y las fases que le componen son:

- Fase 1. Estudio del ámbito del problema y modelado de procesos
- Fase 2. Definición del problema de decisión
- Fase 3. Identificación de los componentes del problema de decisión y realización de cambios necesarios
- Fase 4. Establecimiento de los objetivos de la decisión.
- Fase 5. Análisis del modelo de decisión adecuado
- Fase 6. Estructuración del modelo del problema de decisión
- Fase 7. Estructuración de los elementos de coordinación y colaboración
- Fase 8. Desarrollo del soporte adecuado.

Fase 1 implica identificar cuál es el problema de decisión y brindar el apoyo adecuado. En esta fase se tiene como objetivo conocer el proceso o los procesos relacionados con el problema detectado y algunas de sus características. Se compone de una serie de actividades a realizar:

- La obtención de información del ámbito dentro del cual se encuentra el problema
    - Ubicar documentos de especificaciones existentes, documentos y diagramas de procesos
    - Observar el proceso en operación
    - Indagar con las personas involucradas en los procesos, por medio de entrevistas, cuestionarios o alguna otra técnica similar
    - Pláticas con uno o más de los expertos en el proceso
    - Crear una descripción hipotética y pedir a los involucrados en los procesos que la acerquen lo más posible a la realidad.
- Realizar una descripción textual del proceso
- Identificar los elementos que permiten tener una visión general del problema existente, que son
    - Actividades principales relacionadas con el proceso que se identificó como problema
    - Identificar a las personas involucradas
    - Interacciones entre ellas para llevar cabo el proceso
    - La información manejada

Esta fase se apoya mucho en la elaboración de diferentes diagramas de procesos, utilizando la gráfica rica y los diagramas de rol actividad (Rol Activity Diagram-RAD por sus siglas en inglés).

Una gráfica rica es una representación gráfica de la estructura de contextos de trabajo, por medio de las personas involucradas en el trabajo, representadas por los íconos de personas; los procesos, que aparecen como nubes transparentes;

los instrumentos de entrada/salida que se manejan, representados como rectángulos e íconos de objetos de uso común; sus relaciones indicadas por las flechas que señalan el flujo de los procesos, y las preocupaciones de los tomadores de decisiones, representadas por las nubes sombreadas, ver Figura 3.

Los RAD son diagramas de modelado de procesos que permite representar el comportamiento de las personas que realizan actividades para alcanzar una meta, en estos diagramas aparecen los roles y los actores que participan en estos roles de la toma de decisiones; las actividades que una persona realiza, representadas por los cuadrados; los cursos de acción alternativos, que son aquellos que son mutuamente excluyentes representados por los círculos consecutivos; las trayectorias paralelas, que son sub hilos del rol que pueden ser llevadas a cabo en cualquier orden representado por un círculo y triángulos consecutivos; las interacciones, representadas por las líneas que marcan el flujo del proceso; y otros símbolos que marcan el inicio y fin del proceso, la espera de eventos externos entre otros elementos. (Pacheco Soto, 2004).

En la **Fase 2** se realiza la definición del problema. Su objetivo es detectar claramente cuál es el problema a ser resuelto, así como entender las causas del mismo.
En esta fase, una vez recabada la información anterior, se pregunta si para resolver el problema existen una o varias situaciones de decisión claves; de tal forma que si el problema puede resolverse sin que haya una decisión trascendente, entonces no se trata de un problema de decisión.

En la **Fase 3** se entiende el problema de decisión, descomponiéndolo en sus piezas y sus relaciones para asegurar una mejor comprensión del mismo. Su

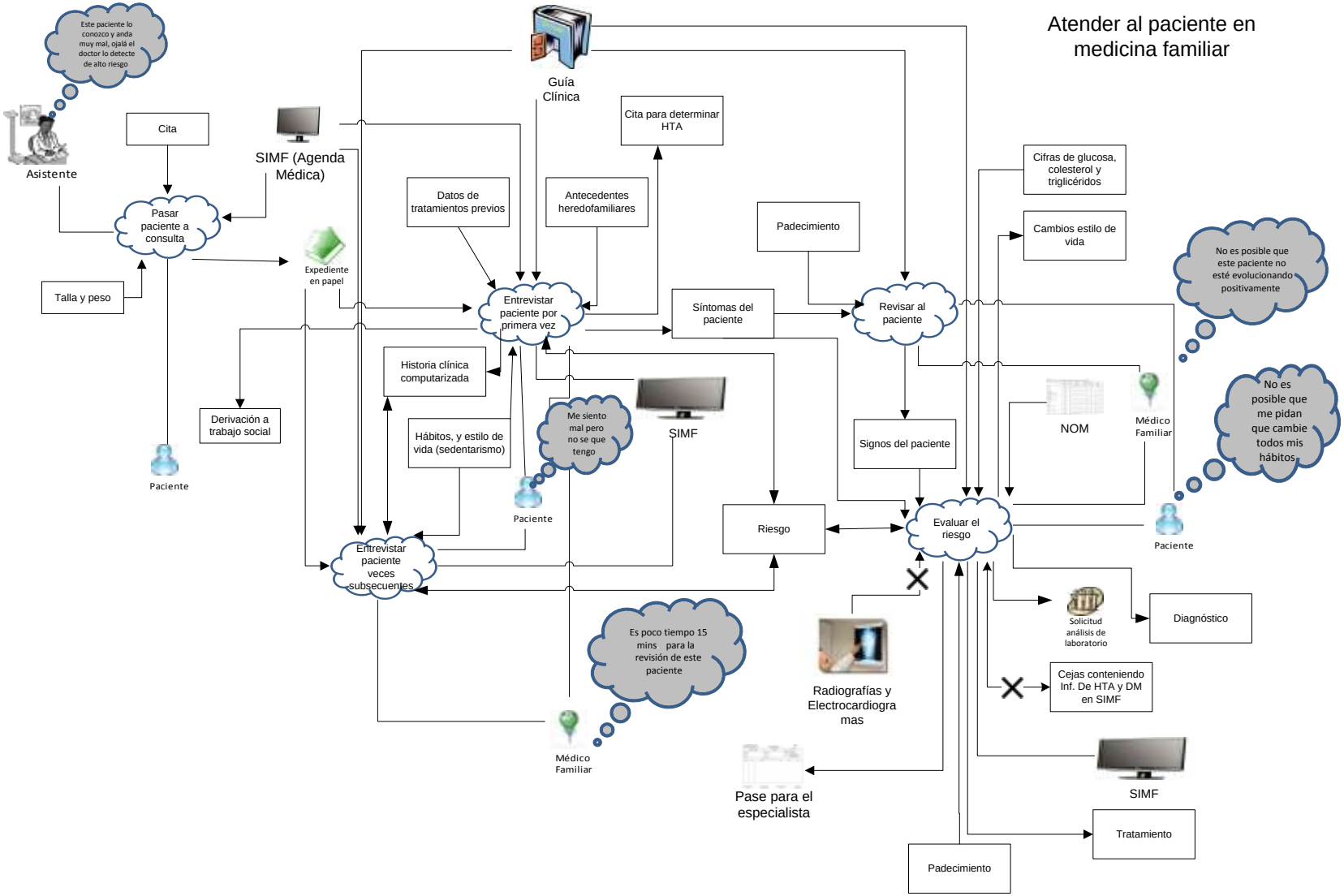


Figura 3 Ejemplo de gráfica rica.

objetivo es conocer cómo se llevan a cabo la decisión o decisiones del problema. Esta fase comprende dos actividades:

- Realizar el modelado de aspectos de decisión, principalmente se elaboran los diagramas de influencia, como el mostrado en la Figura 5 los cuales se utilizan para modelar la estructura interna de un problema de decisión.

  Los elementos que componen estos diagramas son: las decisiones, representadas por rectángulos de esquinas rectas; las incertidumbres, representadas por los círculos; los resultados finales, representados por los rectángulos con esquinas redondeadas; y las influencias; representadas por las flechas.

  En la Figura 5 se observa como el médico familiar, primeramente mediante simple inspección visual, tiene la incertidumbre de no percibir signos visibles. Después al realizar una revisión detallada del paciente hay más elementos de incertidumbre como lo son la falta de síntomas, el poco tiempo para realizar la revisión, etc. Además se observa como el nivel de riesgo de padecer el síndrome metabólico es un resultado final fundamental para la toma de decisiones.

- Análisis de las actividades internas a la decisión o actividad externa; estas se representan con RAD de actividades internas, los cuales permiten representar de forma más detallada las actividades y decisiones. En la
- Figura 4 se presenta el RAD del manejo del síndrome metabólico en medicina familiar; en ella se detallan cuáles son los cursos alternativos de acción que puede seguir el manejo del síndrome, y cuáles son los elementos con los cuales cuenta el médico familiar para decidir; así como el manejo de incertidumbres al tomar decisiones relacionadas con el síndrome.

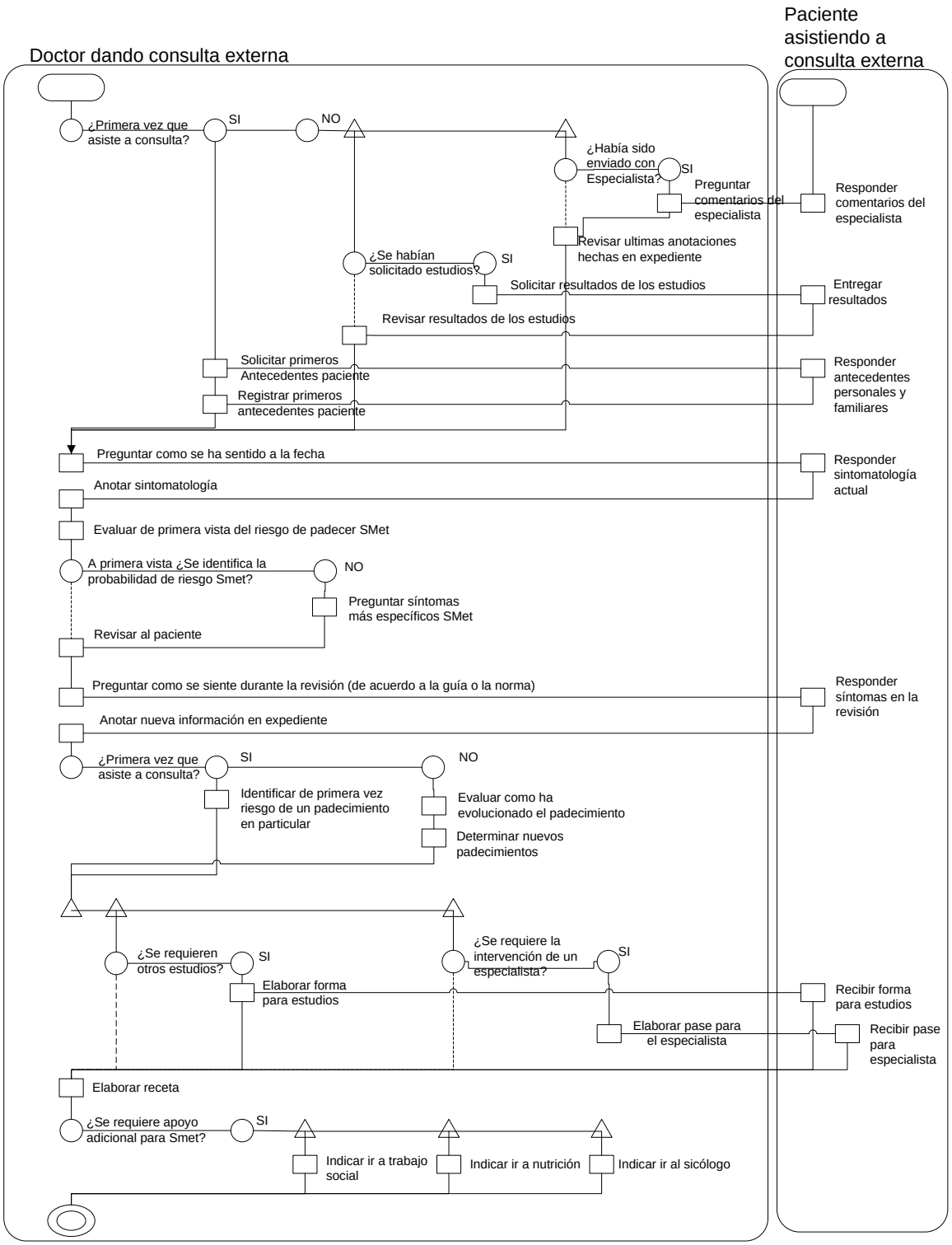


Figura 4. Ejemplo de un diagrama rol actividad (RAD): manejo que hace el médico familiar del síndrome metabólico.

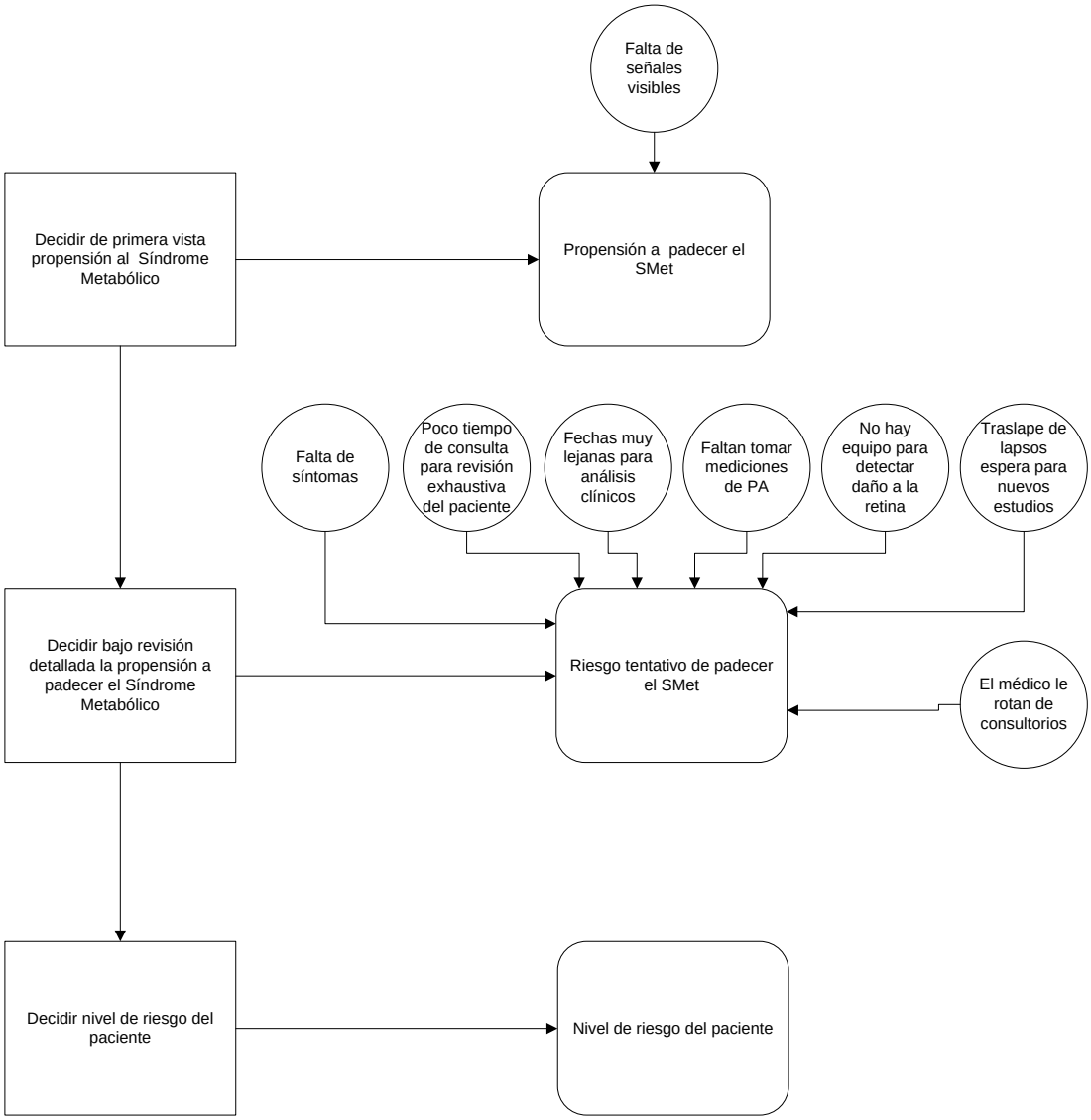


Figura 5. Diagrama de Influencias medicina familiar.

En la **Fase 4** se muestran los objetivos que se persiguen al tomar las decisiones. Los objetivos son importantes porque ellos forman la base para evaluar las alternativas. El procedimiento es el siguiente:

1. En un primer paso se identifican los objetivos

2. Determinar los objetivos que en realidad se toman en cuenta en el problema estudiado y cuáles no. Generalmente los problemas de toma de decisiones no son estructurados y las personas que intervienen en ellos tienen objetivos que en algunas ocasiones no son tomados en cuenta.
3. Habiendo clasificado los objetivos por decisión llevada a cabo y por roles, se identifica como pueden ser tomados en cuenta dichos objetivos al decidir. Ubicando además si existe la información suficiente para caracterizar a los objetivos y si el rol tiene la responsabilidad suficiente

En **la Fase 5** se determina el modelo de decisión apropiado a la situación en particular, ya que como se ha mencionado los problemas de decisión pueden ser de naturaleza muy variada, por lo que los modelos pueden tomar diferentes formas.

En la **Fase 6** se reúne toda la información que se conoce sobre el problema, y de acuerdo a los modelos, se analizan las alternativas de solución que se tienen. El producto de esta fase es un modelo analítico de decisión del problema.

En la **Fase 7** se obtienen los elementos de la coordinación y colaboración. El producto de ésta fase son los diagramas de transición de estados, diagramas de flujo de documentos, así como la coordinación de recursos. Esta última fase está más relacionada con la toma de decisiones en grupo.

Dado que esta es la metodología que se usó para identificar el problema de toma de decisiones y sus características, a lo largo de la presente tesis se hará referencia constantemente a esta metodología marcando en cada Capítulo la fase a la que pertenece el proceso que se presenta.

Enseguida se presenta el sistema de apoyo a la toma de decisiones.

II.4 Sistemas de Apoyo a la Toma de Decisiones.

Definir un DSS implica varios temas que van desde la estructura del problema, la decisión resultante y el control administrativo. Tomando en cuenta estos aspectos, un DSS se define como: *un sistema bajo el control de uno a más tomadores de decisiones que asisten en la actividad de realizar la toma de decisiones, proveyendo un conjunto organizado de herramientas pensados para imponer una estructura en una porción de la situación toma-de-una-decisión, y mejorar al final la efectividad de la decisión resultante* (Marakas, 2003).

Toda decisión en el contexto problema-solución requiere un razonamiento, mientras más estructurado sea el contexto de una decisión menor es el razonamiento que se necesita para obtener un resultado satisfactorio. Entendiendo como razonamiento el proceso por el cual se deriva nueva información partiendo de la combinación o combinaciones de la información existente o que previamente ha sido derivada. De esta forma el razonamiento permite descansar la toma de decisiones en información y en hechos.

A través de este proceso de razonamiento es que se adquiere conocimiento, de tal forma que este conocimiento es lo que se guarda dentro de un DSS, en la forma de reglas, heurísticas, límites, condiciones, resultados previos y cualquier otro tipo de información que pudiese haber sido programada en el DSS por sus diseñadores o habiéndose adquirido por el uso repetido del DSS.

El usuario de un DSS es la persona o personas responsables de proveer una solución al problema y en cuyas manos está el tomar las decisiones dentro del contexto para el cual el DSS fue diseñado.

Aun cuando los DSS pueden ser de naturaleza muy diversa, tienen características comunes entre ellos:

1. Se emplean en contextos de decisiones semi estructuradas o estructuradas
2. Pretenden apoyar a los tomadores de decisiones más que reemplazarles
3. Apoyan todas las fases del proceso de decisión
4. Se enfocan en la efectividad del proceso de toma de decisiones más que en su eficiencia
5. Está bajo el control del usuario del DSS
6. Utiliza los datos fundamentales y los modelos
7. Facilita el aprendizaje en la parte del tomador de decisiones
8. Es interactivo y amigable con el usuario
9. Generalmente se les desarrolla usando un proceso evolutivo, iterativo
10. Proveen soporte a todos los niveles desde los ejecutivos hasta los administradores de línea
11. Pueden proveer soporte a múltiples decisiones independientes o interdependientes
12. Proveen apoyo a un individuo, grupo, y en el contexto de decisiones tomadas en equipo

Como ya se mencionó anteriormente un DSS se basa en el conocimiento que es promovido por los expertos en un dominio específico, dirigido, organizado y formalizado con el propósito de transformarlo en una representación de soporte-computacional.

En el contexto médico, a los médicos frecuentemente se les exige proveer información confiable que apoye sus decisiones. Sin embargo, el conocimiento y la información que ellos requieren para responder este tipo de preguntas normalmente no son fáciles de obtener.

Un HDSS se define como *un sistema interactivo controlado por el usuario, generalmente basado en un software de computadora, que está diseñado específicamente para facilitar la toma de decisiones, usando datos, modelos y elementos de conocimiento para resolver decisiones semi estructuradas en problemas complejos de toma de decisiones* (Tan, et al., 1998).

Esta definición incluye varias categorías de sistemas, entre ellos herramientas de información para los administradores de hospitales, herramientas de computadora enfocadas a proporcionar información a los ejecutivos y soporte a sistema de apoyo a laboratorio que reporten valores anormales, y herramientas de computadora para la consulta-paciente. Una sub-clasificación de los HDSS son los CDSS, siendo este último del que trata esta tesis.

Los algoritmos y los enfoques utilizados en los CDSS varían considerablemente, incluyendo la aplicación de algoritmos clínicos, modelado patofisiológico, reconocimiento de patrones utilizando enfoques estadísticos bayesianos, sistemas de razonamientos expertos; y más recientemente redes bayesianas, sistemas basados en lógica difusa y redes neuronales artificiales.

Son diferentes los enfoques para el proceso de extracción del conocimiento de acuerdo al nivel del usuario, en el caso de la presente tesis son dos los niveles de experiencia a tomar en cuenta: el sub-experto del médico familiar, ya que éste tiene un enfoque solamente clínico del padecimiento; y el experto del médico epidemiólogo, quien además de tener el enfoque clínico hace estudios epidemiológicos sobre el síndrome metabólico y su comorbilidad.

La adquisición del conocimiento se ha convertido en un aspecto esencial de la inteligencia artificial y del desarrollo de los sistemas de soporte a la toma de decisiones. Desde el punto de vista del experto, la medicina se basa en el conocimiento declarativo que se refiere a la información basada en hechos, conceptos, y relaciones en el dominio; y el conocimiento procedimental, el cual se

refiere a como razona una persona al aplicar el conocimiento declarativo y que se obtiene en la práctica clínica.

Sin embargo, difícilmente un experto encontrará el tiempo suficiente para analizar grandes cantidades de datos en forma manual y, partiendo de ellos, extraer un conocimiento útil. Es en este sentido que el uso de técnicas de extracción de conocimientos basadas en los datos se han vuelto una herramienta fundamental para el desarrollo de HDSS.

II.5 El proceso de extracción del conocimiento basado en datos.

En la fase 5 de la metodología se determina el modelo de decisión apropiado al problema que se está trabajando. En este sentido, el proceso de KDD es una metodología que permitirá definir el modelo de decisión adecuado. Incluso será esta metodología la que guíe la recopilación de la información que finalmente llevará a un modelo analítico de decisión del problema (fase 6 de la metodología).

Generalmente un hospital, tiene múltiples sistemas que permiten el funcionamiento de la organización con bases de datos donde se registra gran cantidad de información. Adicionalmente existen otros asientos de datos donde áreas, como epidemiología, registran la información de las funciones de investigación que realizan sin hacer uso de un sistema institucional para ello. La institución se encarga de recolectar esta información posteriormente a su registro.

El KDD se define como *el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos* (Hernández Orallo, et al., 2004).

Inicialmente para tener una base de información integrada que conjunte diferentes fuentes, y que permita el desarrollo del proceso de descubrimiento de conocimiento, se conforman los denominados almacenes de datos (data warehouse en inglés) dentro de las organizaciones.

Los que se definen como *un conjunto de datos históricos, internos o externos, y descriptivos de un contexto o de una área de estudio, que están integrados y organizados de tal forma que permiten aplicar eficientemente herramientas para resumir, describir, y analizar los datos con el fin de ayudar en la toma de decisiones estratégicas* (Hernández Orallo, et al., 2004).

De acuerdo con Fayyad, et al (1996) las nueve fases que componen el proceso de descubrimiento de conocimiento en bases de datos son:

1. Desarrollo y entendimiento del dominio de la aplicación: esta fase incluye el aprendizaje del conocimiento relevante que se tenga a priori y los propósitos que tenga el usuario final para llevar a cabo la extracción de conocimiento
2. Creación del archivo objetivo: aquí se seleccionan el subconjunto de variables (atributos) y las entradas de datos (ejemplos) que serán usados para la tarea de descubrimiento. Este proceso generalmente incluye la revisión de la información existente buscando el subconjunto deseado.
3. Limpieza de datos y pre-procesamiento: en este paso se remueven los datos sesgados (outliners en inglés), lidiando con los datos ruidosos y los valores perdidos y la contabilización de información para secuencias de tiempo y cambios, de los que se tenga conocimiento.
4. Reducción de los datos y proyección: consiste en encontrar atributos útiles aplicando métodos de reducción y transformación, encontrando una representación invariante de los datos. Producto de este paso es la obtención de un conjunto de datos listo para usarse en la fase de minería de datos (paso 7) a los que se les denomina vista minable.

5. Selección de la tarea de minería de datos: es aquí donde el encargado de hacer la minería hace concordar los objetivos definidos en el paso 1 con un método en particular de la minería de datos.
6. Seleccionar el algoritmo de minería de datos, en este paso el encargado de la minería selecciona el método para la búsqueda de patrones dentro de los datos, y decide cuales modelos y parámetros son los más apropiados
7. Minería de Datos: en este paso se generan los patrones de una forma de representación en particular, tales como reglas de clasificación, árboles de decisión, modelos de regresión, tendencias, etc.
8. Interpretación de los patrones minados: aquí el analista analiza los patrones extraídos y los modelos; y se revisan los datos basándose en los modelos extraídos.

Consolidación del conocimiento descubierto: el paso final consiste en integrar el conocimiento descubierto en el proceso del sistema, documentar y reportar a las partes interesadas. Este paso puede incluir ubicar y corregir inconsistencias potenciales que pudiesen surgir entre el nuevo conocimiento y el conocimiento ya existente.

La Figura 6 presenta una esquematización del proceso de extracción de conocimiento. De las nueve fases descritas una de las más laboriosas es la minería de datos. De hecho antes de que se definiera el proceso de descubrimiento de conocimiento en bases de datos, la minería de datos era considerada como un proceso completo. En la siguiente sección se analiza más a detalle esta fase. De acuerdo a la metodología de soporte a la toma de decisiones con enfoque a la ingeniería de procesos la fase 6 establece que se debe estructurar un modelo del problema de decisión y es precisamente en la fase de minería de datos del proceso KDD donde se estructura dicho modelo.

II.6 La fase de minería de datos.

Los seres humanos tenemos la capacidad de identificar patrones de forma natural, lo hacemos al observar las nubes, las constelaciones, texturas en los materiales, etc.

Sin embargo, al elaborar algoritmos que permiten a una máquina descubrir patrones nos damos cuenta de lo complejo que son estos procesos. Además la

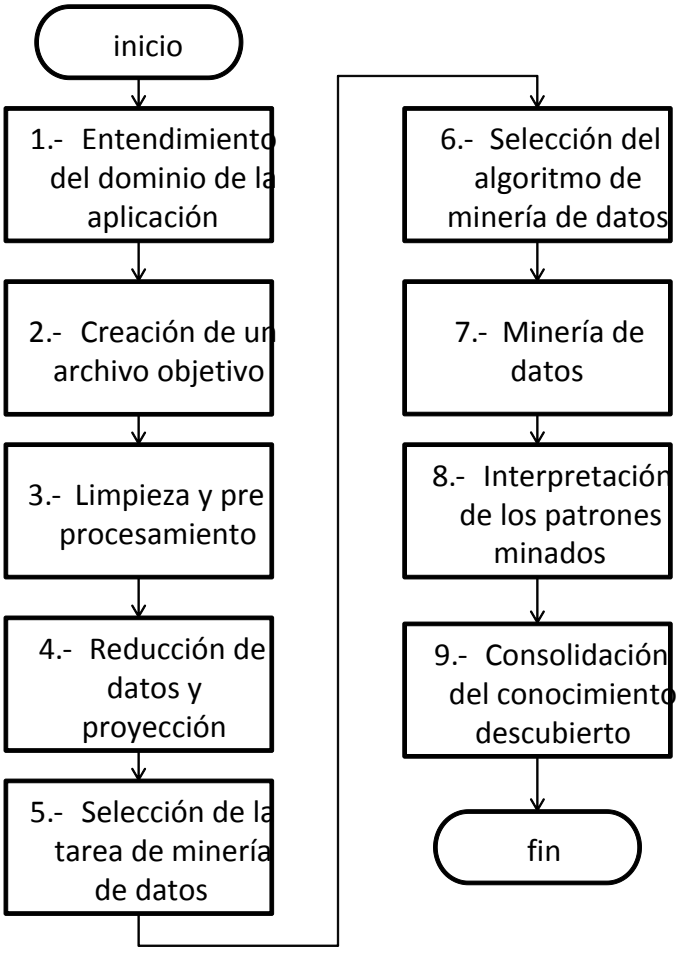


Figura 6. Esquematización del proceso de extracción de conocimiento en base de datos.

pregunta que sigue al saber que una máquina puede descubrir patrones es ¿qué tipo de patrones pude descubrir?

La intención principal de la minería de datos es *dar sentido principalmente a una gran cantidad de datos sin supervisar, dentro de algún dominio* (Cios, et al., 2007). La funcionalidad y los patrones que puede descubrir la minería de datos son los siguientes:

- Descripción de concepto/clase, caracterización y discriminación: los datos se pueden relacionar con conceptos y clases. Dichas descripciones se pueden derivar por:
    a) Caracterización de los datos, se resumen los datos de la clase que está bajo estudio, a la cual se le denomina clase objetivo, en términos generales.
    b) Discriminación de datos, donde lo que se hace es una comparación entre la clase objetivo y una o un conjunto de clases objetivo, normalmente llamadas clases restringidas.
- Minado de frecuencia de patrones, asociaciones y correlaciones: se trata de encontrar, como su nombre lo indica, los patrones que más frecuentemente se presentan dentro de los datos. Los patrones que generalmente se buscan son conjuntos de atributos, subsecuencias, y subestructuras. Los métodos que comúnmente se utilizan son el análisis de asociaciones y correlaciones entre los datos.
- Clasificación y predicción: se trata de encontrar un modelo o función que describa y distinga entre las diferentes clases de datos y conceptos; con el propósito de ser capaz, con el modelo, de predecir los objetos cuya clase de clasificación de entrada es desconocida.
- Análisis de conglomerados: a diferencia de la clasificación y la predicción, la cual analiza objetos de datos con una determinada etiqueta de clase dada a priori, en el análisis de conglomerados no existe dicha etiqueta. Los datos se agrupan de acuerdo a alguna medida de similitud usando el principio de

maximización de la similitud intra-clase, y minimizando la similitud inter-clase.

- Análisis de sesgos: una base de datos puede contener objetos de datos que no cumplen con el comportamiento general del modelo de los datos. Este tipo de análisis se utiliza para detectar comportamientos atípicos (por ejemplo fraudes, fugas, eventos extraños, etc.). Usan algunas pruebas estadísticas que suponen cierta distribución de probabilidad para los datos, o que usan alguna medición de distancia para determinar los datos sesgados.
- Análisis de evolución: este tipo de funcionalidad lo que hace es describir y modelar regularidades o tendencias entre los objetos cuyo comportamiento cambia en el tiempo, las cuales pueden incluir caracterización, discriminación, asociación y análisis de correlación, clasificación, predicción, o conglomerados de datos de tiempo relacionados, distinción de características producidas por análisis del tipo series de tiempo, concordancia de secuencia o periodicidad de patrones y similitudes.

Existen diferentes formas de representación de resultados utilizadas por los métodos de minería de datos: tablas de decisión, árboles de decisión, reglas de clasificación, reglas con excepciones, reglas que involucran relaciones, árboles de predicción numérica, representación basada en instancias, conglomerados. Entonces será la selección del método a utilizar la que determina así mismo la representación en los resultados obtenidos.
Los pasos 5 y 6 en la Figura 6 implican seleccionar cuál o cuáles de las funcionalidades arriba citadas son las que corresponden al problema de minería a trabajar.

En el caso particular de la medicina una forma generalmente utiliza para representar el conocimiento obtenido es mediante reglas del tipo *SI-ENTONCES* Por lo tanto ésta será una representación idónea para el conocimiento extraído.

II.7 Resumen.

En este Capítulo se establecieron las características de la toma de decisiones, para entonces presentar una metodología que permite determinar cuál es el problema de toma de decisiones que se quiere resolver. La metodología establece una serie de fases que guían completamente el proceso partiendo de la elaboración de un modelo que facilite el entendimiento del problema de decisión a resolver hasta ubicar cual es el apoyo adecuado a proporcionar.

En la fase 6 (estructuración del modelo del problema de decisión) de la metodología se integra otro proceso que permitirá el descubrimiento de conocimiento en bases de datos, el KDD. Al entender claramente la forma en que el médico maneja la toma de decisiones relacionadas con el síndrome metabólico se obtuvo una serie de elementos que se integraron en la preparación, ejecución y obtención de resultados de los diferentes algoritmos que conforman la fase de minería de datos del proceso KDD, en otras palabras se aplicó un enfoque socio-técnico en la fase de minería de datos. En el Capítulo IV se detallan a fondo los pasos seguidos por el médico en el manejo del síndrome metabólico.

Se presentó la clasificación de los DSS, los cuales nacen como producto de la evolución de los diferentes tipos de sistemas que apoyan la administración de una organización. Los DSS no pretenden reemplazar a los expertos, sino por el contrario apoyan a estos en el problema de toma de decisiones detectado. Se explicó que en el área médica existen los HDSS y una subclase de ellos, los CDSS que apoyan la toma de decisiones del médico durante la consulta.

También se introdujo el proceso KDD explicando brevemente cada una de sus clases, y detallando un poco más la fase de minería de datos. Se enlistó el tipo de patrones que se pueden descubrir a través de la minería de datos y las diferentes formas en que se puede representar el conocimiento extraído, haciendo hincapié

en que las reglas del tipo SI-ENTONCES son una representación idónea para el tipo de conocimiento médico, en el Capítulo III se profundiza en estos temas.

Lo que le da sustento a un CDSS es el conocimiento producido por un experto, en el Capítulo VI se detalla el sistema generado para apoyar en el manejo y detección del síndrome metabólico.

Capítulo III

Métodos de extracción del conocimiento

III.1 Introducción.

Hasta este momento se han explicado brevemente una serie de métodos que permitirán analizar los procesos y los datos relacionados con el manejo del síndrome metabólico. La finalidad de esta parte es entender cómo, desde el punto de vista del especialista médico, se maneja el síndrome.

De acuerdo a la fase 6 de la metodología se requiere reunir toda la información que se conoce del problema para finalmente obtener un modelo analítico del problema de decisión. En este Capítulo se ahondará en el proceso KDD viéndole desde el punto de vista de los datos y el conocimiento que queremos obtener de ellos; a esta parte se le conoce como la parte “dura” del proceso, esto es la que se centra en el análisis matemático de la información. Primeramente se conceptualiza el problema a resolver como un problema de clasificación, este problema será resuelto aplicando varios métodos basados en el aprendizaje supervisado, el cual permite que un agente, como un algoritmo, aprenda a través de una serie de ejemplos y a través de este aprendizaje pueda identificar los patrones que definen la clasificación que se está realizando.

En el Capítulo se describirán detalladamente una serie de métodos pertenecientes a la minería de datos y los conceptos teóricos en que se basan. Primeramente se introducirá el concepto de clasificación, para entonces definir lo que es el aprendizaje no supervisado y su relación con la formación de conglomerados de datos para su posterior clasificación (secciones II, III, IV).

Visto desde un punto de vista geométrico, los datos de entrada, esto es los vectores de datos que representan a un paciente con síndrome metabólico forman nubes de puntos en un espacio de búsqueda $d - dimensional$. Suponiendo que estos puntos provienen desde una distribución normal, lo que en la mayoría de los casos no se da, las estadísticas de dichas nubes nos darían mucha información sobre cómo se agrupan los datos y las relaciones que hay entre los individuos que pertenecen a un mismo grupo (conglomerado). Por ejemplo la media aritmética nos diría el punto donde se encuentra el centro de masa de la nube, siendo este punto el más representativo de la característica que definen a ese conglomerado de datos.

Encontrar los estimadores de dicha distribución, sin contar con algún conocimiento a priori de ella, es un proceso no trivial y matemáticamente bastante complejo.

Sin embargo existen procedimientos de generación de conglomerados que permiten trabajar este tipo de casos.

"Grosso modo los procedimientos de conglomerados producen una descripción de los datos en términos de los conglomerados o grupos de puntos de datos que poseen fuertes similitudes internas" (Duda, et al., 2001).

De acuerdo a lo expuesto existen algoritmos capaces de agrupar un conjunto de registros de entrada sin clasificar que están relacionados de alguna forma, en este caso con el síndrome metabólico, aun sin contar con ninguna información a priori sobre cómo clasificarles. Algunos de estos algoritmos trabajan formando conglomerados con los registros de entrada, de tal forma que al final del proceso que hace el algoritmo los parámetros(media, matriz de covarianza, etc.) encontrados en cada conglomerado son una buena aproximación de la distribución de probabilidad que define a los grupos que conforman la clasificación buscada. Habiendo identificado los conglomerados es posible entonces aplicar otro grupo de algoritmos para extraer las reglas de clasificación basadas en los datos con que se cuentan y no en un conocimiento ya existente.

En la sección III.5 se presenta el algoritmo para formar conglomerados denominado K-Medias. Enseguida en la sección III.6 se presentan los mapas auto-organizados donde se explica el concepto de red neuronal y se profundiza en la Red Neuronal de Kohonen que tienen como principal producto los Mapas Auto-Organizados.
En la sección III.7 Y III.8 se presentan los métodos C4.5 y la Programación Genética, cuya finalidad es el extraer las reglas de clasificación.
Al final se presenta un resumen del Capítulo.

III.2 El problema de clasificación.

Podemos definir a la clasificación como el proceso de análisis de datos cuyo objetivo principal es construir un modelo o clasificador que predice etiquetas categóricas para una instancia, las cuales nos indican la clase a la cual pertenece dicha instancia basándose en los valores de un atributo de predicción (Han, et al., 2006).
La diagnosis médica puede considerarse como un problema de clasificación, un registro es la información de un paciente, los atributos predictivos son todos los datos relacionados con él (incluyendo: síntomas, signos y señales, resultados de análisis clínicos que se le hayan practicado, de su historia clínica y de las entrevistas que se le han hecho). La clase es la diagnosis, enfermedad o condición clínica que el médico ha descubierto basándose en los datos del paciente.

De acuerdo a Fundación Internacional de la Diabetes (IDF, 2006) se presentan una serie de reglas para identificar el síndrome metabólico:
Una persona tiene síndrome metabólico si por lo menos tres de las siguientes frases se aplican a ella:

- *Está pasado de peso u obeso y la mayoría de su peso está en la mitad de su cuerpo. En el hombre, esto significa tener una cintura cuyo diámetro supera 40 pulgadas (101,6 cm). En la mujer, esto significa tener una cintura cuyo diámetro supera 35 pulgadas (88,90 cm).*
- *Tiene la presión sanguínea elevada: 130/85 mm de Hg o mayor*
- *Tiene un nivel elevado de azúcar en la sangre; su nivel de glucosa en ayunas equivale a 110 mg/dL o mayor*
- *Tiene un nivel elevado de grasa en la sangre; su nivel de triglicéridos equivale a 150 mg/dL o mayor*
- *Tiene un nivel de colesterol HDL (colesterol "bueno") bajo. En el hombre esto significa un nivel de colesterol HDL menor que 40 mg/dL. En la mujer esto significa un nivel de colesterol HDL menor que 50 mg/dL.*

Estas reglas establecen un solo tipo de riesgo: tiene o no tiene el síndrome metabólico. Sin embargo como se expuso en la introducción del Capítulo II se encontró que el manejo de dicha clasificación, para efectos prácticos, es mínimo al momento de la consulta. Muy probablemente por el bajo valor predictivo que tiene la clasificación para determinar la evolución de un paciente.

Además, en el manejo del síndrome metabólico, se tiene la peculiaridad de que generalmente el paciente ha dejado pasar muchos años, y el síndrome ha evolucionado a otras enfermedades ligadas a él, lo que provoca que la experiencia del médico familiar tienda más a la detección de las enfermedades crónico degenerativas que del síndrome per se.

Por lo tanto lo que se propone es aplicar los algoritmos descritos en la introducción de este Capítulo, para dividir en subgrupos la información de un número significativo de pacientes con diferentes niveles de riesgo de padecer el síndrome

metabólico, identificando claramente los patrones que describen la pertenencia de un paciente a cierto subgrupo y estableciendo las reglas para cada uno de ellos.
Se pretende que la clasificación sirva de herramienta en el pronóstico del nivel de riesgo de cada subgrupo, así como la posible evolución que puede llegar a tener el padecimiento.

Es en este sentido lo que se busca es aplicar el aprendizaje estadístico, y los métodos que de él se desprenden para proporcionar conocimiento que pueda ser utilizado por el médico familiar. A continuación se presenta un tipo de aprendizaje denominado aprendizaje no supervisado.

III.3 El aprendizaje no supervisado.

El aprendizaje estadístico juega un rol preponderante en diferentes áreas de la ciencia. Por ejemplo: el predecir si un paciente con cierto nivel de sobrepeso, al fumar incrementa las probabilidades de padecer enfermedades cardiovasculares, en comparación con otro que no fuma, es una predicción estadística basada en la demografía, la dieta, sus hábitos y la medición de otros indicadores clínicos del paciente.

En el caso del aprendizaje de máquina, el proceso de obtener conocimiento generalmente se relaciona con la acción de procesar cierto número de datos de entrenamiento, mediante los cuales el algoritmo va obteniendo información relevante sobre la estructura, patrones de correlación, indicadores estadísticos, etc. Se van integrando a una memoria del propio algoritmo y se van refinando mediante un proceso iterativo.

Podemos definir el proceso de aprendizaje inductivo de una máquina como la habilidad de un agente, tal como un algoritmo, para mejorar su propio desempeño basado en experiencias pasadas (Cios, et al. 2007).

La idea es entrenar a un algoritmo para que, mediante el aprendizaje inductivo, la máquina aprenda a clasificar correctamente nuevos individuos en la clase que les corresponde, en realidad lo que más nos interesa es medir la capacidad de desempeño o al menos el potencial de desempeño, en situaciones nuevas (Witten y Frank 2005).

Cuando a priori se cuenta con un atributo clasificador y en base a éste se pretende hacer que la máquina aprenda a clasificar a nuevos individuos, se dice que el aprendizaje es supervisado. Si se carece del atributo clasificador se dice que se trata de un aprendizaje no supervisado (Duda, et al., 2001), en la presente tesis este último caso es el que nos interesa, ya que en el proceso de determinar por primera vez una clasificación inexistente obviamente no se cuenta con una clase objetivo dada a priori lo que implica que no se puedan aplicar los algoritmos que requieren una supervisión basada en ejemplos.
Solo se cuenta con una serie de datos que requieren ser agrupados (formar conglomerados con ellos), conceptualizando a cada registro como un vector multi variado, y un método que tome en cuenta todas esas variables para el cálculo de distancias entre objetos de una misma clase. Para entonces, habiendo determinado los conglomerados, diferentes clases de objetos, y asignado a cada registro de entrada la clase objetivo, determinar el conjunto de reglas que mejor definan la pertenencia a cierta clase. Lo que establece que los métodos para minar la información sobre el síndrome metabólico sean el análisis de conglomerados y la extracción de reglas de clasificación y predicción.
Siendo este conjunto de reglas la información base para el paso 8 del proceso de extracción del conocimiento.

Formalmente en el aprendizaje no supervisado se tiene un conjunto de N mediciones $(x_1, x_2, \dots, x_n)$ de un vector aleatorio X teniendo la densidad de probabilidad conjunta $\Pr(X)$. El propósito es inferir directamente las propiedades de esta función de densidad, sin la ayuda del supervisor proveyendo respuestas

correctas o con un cierto grado-de-error para cada observación (Hastie, et al., 2001).

Hay algunos métodos no supervisados que proveen una forma de “extracción inteligente de características”, independiente de los datos, lo cual puede ser valioso para trabajar un análisis exploratorio y ganar algunos indicios sobre la naturaleza de la estructura de los datos.

El descubrimiento de diferentes subclases, esto es conglomerados o grupos de patrones cuyos miembros son más similares entre ellos que a los miembros de cualquier otro subgrupo, o de las características principales de los subgrupos pueden sugerirnos un acercamiento alternativo al problema de clasificación.
A continuación se procederá a explicar la formación de conglomerados bajo tres métodos diferentes usados en la presente tesis.

III.4 Análisis de conglomerados.

Supóngase que contamos con un conjunto de datos de los cuales no se conoce, a priori, la clase a la cual pertenece cada uno de ellos y lo que se quiere es poder obtener grupos dentro de la muestra, y los patrones que definen la pertenencia a cada grupo, para entonces poder predecir la clase a la cual pertenecen nuevos individuos sin clasificar.

Un objeto se puede describir por medio de un conjunto de mediciones, o por su relación con otros objetos. *El objetivo siempre es organizar los conglomerados dentro de una jerarquía natural. Lo que involucra el agrupamiento sucesivo de los conglomerados en sí mismos tal que, en cada nivel de la jerarquía, los conglomerados sean más similares entre ellos que entre los otros grupos* (Hastie, et al., 2001) .

Un punto central en esta definición es la noción de grado de similitud, o disimilitud según sea el caso, entre los individuos que están siendo agrupados.
Cada método de análisis de conglomerados intenta agrupar los objetos basándose en la definición de similitud que el mismo método provee. El establecimiento de dicha medida de similitud es totalmente subjetivo, y puede haber muchas formas de definirle, sin poder considerar a alguna de ellas mejor que las otras. Generalmente se utilizan medidas de distancia tales como la distancia euclidiana u otras similares.

De acuerdo a Han et al (2006) los principales métodos de generación de conglomerados pueden ser clasificados de la siguiente forma:

- Métodos que generan particiones: dado un conjunto de n datos de objetos de entrada, se procede a construir k particiones de datos, donde cada partición representa un conglomerado, siendo $k \leq n$. Se debe satisfacer la condición de que cada conglomerado al menos contenga un objeto y que cada objeto solo puede pertenecer a un solo subgrupo.
- Métodos jerárquicos: se trata de generar una secuencia de particiones tales que, partiendo de n muestras de entrada, la primera de estas particiones tiene exactamente n conglomerados, esto es cada conglomerado tiene un único miembro.

  La siguiente partición tiene $n-1$ conglomerados, la que sigue $n-2$, y así se continúa hasta llegar al enésimo conglomerado.
- Métodos basados en la densidad: la mayoría de ellos se basa en la distancia entre objetos. La idea general es continuar creciendo un conglomerado hasta que la densidad, esto es su número de objetos en el vecindario, exceda cierto límite. Para cada punto de datos dentro del conglomerado se tiene un vecindario determinado por un radio, partiendo del centro del conglomerado.
- Métodos basados en una malla: estos métodos cuantizan el espacio de los objetos en un número finito de celdas que forman una malla. Todas las operaciones de cuantización son realizadas en el espacio de la malla. Una

ventaja importante de estos métodos es que son independientes del número de objetos de entrada y solo dependen del número de celdas que dimensionan el espacio cuantizado.

- Métodos basados en el modelo: estos métodos construyen una hipótesis en forma de un modelo para cada uno de los conglomerados teniendo como objetivo final encontrar el modelo que mejor encaje con los datos.
- Métodos de generación de conglomerados basados en condiciones establecidas por el usuario: una condición expresa una expectativa determinada por el usuario o describe las propiedades deseables del conglomerado resultante.

.

En la presente tesis se utilizan varios algoritmos con diferentes paradigmas.
El primero que se analiza es el popular algoritmo de K-Medias que trabaja bajo el paradigma de generar particiones.

III.5 Algoritmo K-Medias.

Como se mencionó al inicio del Capítulo, la base estadística que sustenta a diversos algoritmos de generación de conglomerados es el principio de máxima similitud.

Supóngase que se tienen una colección de c archivos de muestras $\mathcal{D}_1, \dots, \mathcal{D}_c$ y que cada ejemplo dentro de la muestra está compuesto de una serie de variables aleatorias independientes idénticamente distribuidas. Para la muestra $\mathcal{D}_j$ se ha delineado la ley de probabilidad $p(X, \omega_j)$ que establece la pertenencia a dicho grupo, además de que las diferentes muestras son mutuamente excluyentes. Se asume que $p(X, \omega_j)$ tiene una forma paramétrica θ, de tal forma que $p(X, \omega_j)$ es una aproximación a algún tipo de distribución probabilística que mejor describe el

comportamiento de la variable, por ejemplo $p(X, \omega_j) \sim N(\mu_j, \sum_j)$ donde θ_j consiste de los componentes $\mu_j \, y \, \sum_j$. Entonces lo que se pretende es encontrar los vectores de parámetros $\theta_1, \ldots, \theta_c$ asociados con cada categoría.

Dado que las muestras son mutuamente excluyentes se infiere que $\mathcal{D}_i$ no da información acerca de θ_j. Lo implica resolver c problemas separados. Entonces para una muestra $\mathcal{D}$ conteniendo n muestras $x_1, \ldots, x_n$ y dado que las muestras son independientes se tiene que la similitud está dada por la ecuación

$$\boldsymbol{p(\mathcal{D}|\theta) = \prod_{k=1}^{n} p(x_k|\hat{\theta})} \quad (1)$$

donde $\hat{\theta}$ es el estimador de parámetros que maximiza la similitud entonces el problema es determinar dicho estimador (Duda, et al., 2001).

Dado que hablamos de un aprendizaje no supervisado, es de suponer que uno o varios de los parámetros que definen la distribución de probabilidad a la cual pertenece nuestra función de densidad son desconocidos; en la Tabla I se muestran los diferentes casos que pueden darse en el momento de la formación de conglomerados dependiendo de cuales parámetros son conocidos y cuáles no.

Tabla I. Posibles casos que se pueden presentar en el proceso de formación de conglomerados, se omite la media μ_i ya que ésta siempre se desconoce (Duda, et al., 2001)**.**

Caso	$\sum_i$	$p(\omega_i)$	c
1	✔	✔	✔
2	?	?	✔
3	?	?	?

Donde $\sum_i$ es la matriz de covarianza, $p(\omega_i)$ es la probabilidad de que el vector de parámetros que se está validando sea similar al vector de parámetros buscado, y *c* el número de conglomerados.

Son diferentes las técnicas que existen para estimar soluciones desconociendo los parámetros arriba citados. Por ejemplo el algoritmo de expectativa y maximización prueba aleatoriamente distribuciones normales con diferentes parámetros θ_j buscando que sean los más parecidos a ω_j y maximizando la probabilidad de que x pertenezca solo a esa distribución y no a otra parecida.

Otra forma común de solucionar el problema es trabajar con un estimado de los parámetros, sin embargo esta solución tiene el problema del tiempo que se requiere para que el algoritmo converja.
Si por ejemplo el estimado se obtuvo de haber procesado una gran cantidad de datos etiquetados, el estimado será de buena calidad y la convergencia será rápida. Pero si no se contó con dicho pre-procesamiento de la información entonces el tiempo para converger puede que sea muy largo.

Existen otras técnicas que se pueden utilizar para simplificar los cálculos computacionales y asegurar la convergencia. Una de esas técnicas es el algoritmo K-Medias.

La idea es encontrar k vectores de medias $\mu_1, \mu_2, \dots, \mu_k$ donde k es el número de centroides de los conglomerados establecidos de forma anárquica a priori (equivalente a c). Entonces lo que se hace es simplemente calcular la distancia euclidiana[3]

$$\|\mathbf{x_k} - \hat{\boldsymbol{\mu}}_\mathbf{i}\|^2 \qquad \textbf{(2)}$$

donde

$\mathrm{x_k}$ es una muestra de entrada a asignar a un conglomerado

$\hat{\mu}_\mathrm{i}$ es cada uno de los vectores de medias

para entonces determinar el vector $\hat{\mu}_\mathrm{m}$ más cercano a $\mathrm{x_k}$. y aproximar la estimación $\hat{P}(\omega_i|x_k, \hat{\theta})$ de la siguiente forma

[3] La distancia euclidiana es aquella que se obtiene a través del Teorema de Pitágoras

$$\widehat{\boldsymbol{P}}(\boldsymbol{\omega_i}|\boldsymbol{x_k}, \widehat{\boldsymbol{\theta}}) \simeq \begin{cases} \mathbf{1} & \boldsymbol{Sí\ \ i = m} \\ \mathbf{0} & \boldsymbol{de\ otra\ forma} \end{cases} \quad (3)$$

En otras palabras asignando la muestra x_i al conglomerado m más cercano (Duda, et al., 2001).

El algoritmo de K-Medias es una forma estocástica del método hill climbing[4] en la función log-similitud. Que permite, partiendo de p puntos generadores y mediante un proceso iterativo, converger a los centros de masa de los conglomerados (μ_k) asegurando que cada dato de entrada pertenece a un solo conglomerado y a ninguno otro más.

Habiendo definido el número de conglomerados, el seudocódigo del algoritmo es el siguiente:

Algoritmo K-Medias

1. **comienza inicialización** $n, c, \mu_1, \dots, \mu_c$
2. **haz** clasifica n muestras de acuerdo a la μ_i más cercana
3. recalcula μ_i
4. **hasta que** no haya cambios en μ_i
5. **regresa** $\mu_1, \dots, \mu_c$
6. **fin**

[4] El método hill climbing es un algoritmo de optimización del tipo de búsqueda local permite escapar de los máximos locales dando saltos en el recorrido de la función y de esta forma logra encontrar el máximo global.

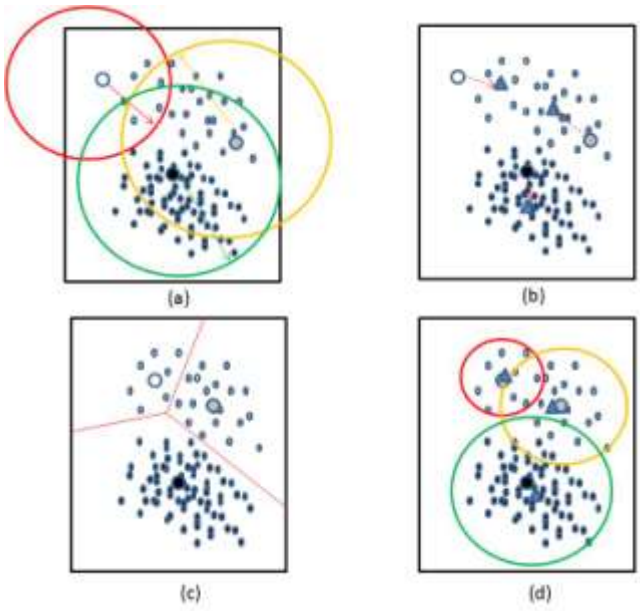


Figura 7. Proceso de generación de conglomerados utilizando K-Medias (a) selección de 3 puntos aleatorios como centros del conglomerado, (b) Diferencia entre el punto generador y la media del conglomerado (triángulo), (c) desplazamiento de los puntos generadores a la media del centroide (d) el proceso se repite

En la Figura 7 se observa la forma en que el algoritmo de K-Medias funciona. Primero en la parte (a) se toman tres puntos generadores al azar, y de acuerdo a

la distancia euclidiana se seleccionan los objetos de datos más cercanos a cada punto, produciéndose un radio que demarca el conglomerado formado en el punto generador. Los objetos de datos de entrada serán asignados al punto generador más cercano, conformando de esta manera los conglomerados como conjuntos mutuamente excluyentes, pudiendo cada objeto pertenecer a un único conglomerado.

En la parte (b) se pueden observar como existen diferencias entre los puntos generadores, representados por los círculos grandes y las medias de los conglomerados (los triángulos), en la parte (c) se representan las fronteras de los conglomerados formados y se reemplazan los puntos generadores por los vectores de medias de cada conglomerado, el proceso se repite parte (d) hasta

que los puntos generadores convergen con las medias de los conglomerados generados y el proceso se detiene.

Este procedimiento relativamente simple es la base de otros algoritmos que integran conceptos más elaborados para asegurar ciertas características dentro de los conglomerados formados. Una de estas características está relacionada con los valores que contiene cada entidad de datos. Cuando pensamos en un número, generalmente pensamos en su valor algebraico, con el que hacemos operaciones, sin embargo un argumento numérico puede contener más información que su puro valor algebraico.

Por ejemplo si se nos da una serie de coordenadas de un mapa topográfico difícilmente lo primero que pensaremos es cuanto suma latitud y longitud, en vez de ello pensamos en un lugar geográficamente hablando. Donde todas las coordenadas con valores numéricos similares estarán cercanas ha dicho lugar. En otras palabras los atributos latitud y longitud además de su información numérica guardan una relación topológica.
Ahora bien la información de un paciente con cierta enfermedad, además del valor numérico también tiene características similares a las descritas anteriormente. Un ejemplo muy simple es el siguiente: una persona con un alto grado de obesidad puede que tenga niveles de colesterol muy elevados similares a los de una persona que tiene problemas de grasa en la sangre (dislipidemias) aun cuando esta última no haya desarrollado ese nivel de obesidad.

Un algoritmo que aprovecha la información topológica contenida dentro de los datos de entrada son los llamados mapas auto-organizados (SOM por sus siglas en inglés) que se presentan en la siguiente sección.

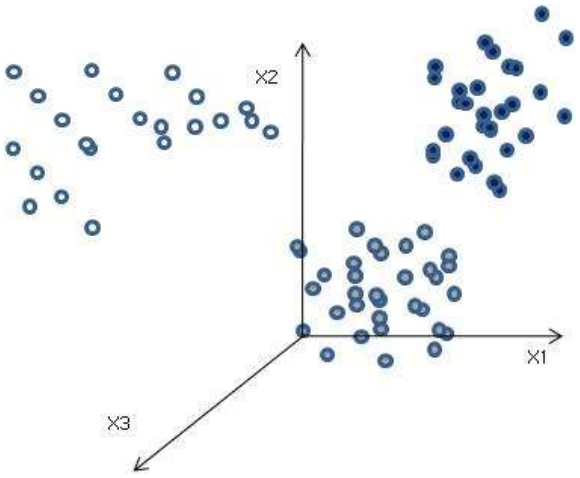


Figura 8. Datos simulados en un plano, semejando un agrupamiento de 3 conglomerados (negro, gris, blanco) por medio del algoritmo de K – medias.

III.6 Mapas auto-organizados (SOM).

De acuerdo a Hastie et al (2001) este método puede ser visto como una versión constreñida de K-Medias, en el cual los prototipos son forzados a representárseles en un mapa de una o dos dimensiones. Dicho autor hace una demostración de cómo, dado el conjunto de parámetros adecuados, el algoritmo SOM eventualmente se estabiliza en uno de los mínimos locales encontrados por K-Medias.

La Figura 8 muestra gráficamente la formación de tres conglomerados partiendo de una colección de datos de entrada y usando el algoritmo de K-Medias.
Podemos observar como los tres subgrupos formados están claramente separados y como los subgrupos no se traslapan, son mutuamente excluyentes entre ellos, de tal forma que un individuo solo puede pertenecer a un subgrupo.

Sin embargo el algoritmo SOM no se puede considerar simplemente otra versión del algoritmo de K-Medias, ya que tiene otras características relacionadas con el escalamiento multidimensional.

El objetivo principal de un mapa auto-organizado es representar todos los puntos en el espacio fuente por puntos en el espacio objetivo, tales que las relaciones de distancia y proximidad se preserven lo más posible (Duda, et al., 2001).

A esta representación del espacio de entrada a un espacio de salida se le considera un proceso de proyección (un mapeo). Dicha proyección también recibe el nombre de mapeo topológico constreñido.

Al preservar, de forma aproximada, las relaciones topológicas y métricas de los elementos de los datos originales, inherentemente se van conformando los conglomerados que forman los propios datos (Alhoniemi, et al., 1999).

Kohonen (1990) presentó el trabajo denominado The Self-Organizing Maps. En este trabajo se explica que un mapa auto-organizado es un modelo de red neuronal. Así mismo Kohonen narra cómo, conforme ha ido aumentando el conocimiento sobre cómo funciona el cerebro, se han planteado diferentes enfoques para algoritmos que emulan su funcionamiento.

Uno de estos enfoques trata sobre la forma en que ciertas regiones del cerebro están dedicadas a tareas específicas y como evidencias directas demuestran, mediante técnicas de desplegado, que la fuerza y la distribución espacial de las repuestas de las neuronas se circunscriben a grandes áreas de la corteza cerebral con una resolución de unos pocos milímetros (Kohonen, 1990).

En términos de las Ciencias de la Computación, una Red Neuronal es una representación algorítmica de este tipo de procesos. En su representación más simple se le puede considerar una caja negra que acepta una serie de entradas y produce una o más salidas (Zupan, et al., 1993) como se ha representado en la Figura 9.

De forma más detallada, una Red Neuronal se forma de señales de entrada, neuronas, conexiones (sinapsis), pesos, y dos funciones matemáticas que controlan el establecimiento de la conexión neuronal. Todos estos elementos se agrupan por capas.

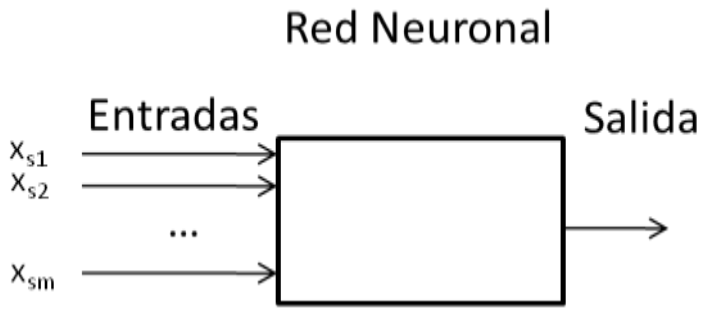


Figura 9. Representación básica de una red neuronal.

Existen diferentes tipos de redes neuronales entre ellas las redes neuronales competitivas, es a esta categoría a la que pertenece la red neuronal de Kohonen (Kohonen Neurol Network-KNN por sus siglas en inglés). Las células están específicamente sintonizadas a ciertas variaciones en las señales de entrada o clases de patrones a través de un proceso de aprendizaje no supervisado. En su versión básica solo una célula o un grupo de ellas, en un momento determinado, se activarán como respuesta a la entrada actual (Kohonen, 1990).

La KNN tiene su base en el aprendizaje competitivo, también denominado "el que gana se lleva todo", en él solamente una neurona de la capa activa es seleccionada después de que ocurre la entrada; no importa que tan cercanas estén otras neuronas a la ganadora, ellas estarán fuera de este círculo.
La red selecciona a la neurona ganadora(c) de acuerdo a uno de los siguientes criterios:

- La señal más grande en toda la red

$$\boldsymbol{out_c \leftarrow \max(out_j) = \max\left(\sum_{i=1}^{m} w_{ij} x_{si}\right)} \quad (4)$$

- El vector de pesos $W_j(w_{j1}, w_{j2}, \dots, w_{jm})$ más parecido a la señal de entrada $X_s(x_{s1}, x_{s2}, \dots, x_{sm})$

$$\boldsymbol{out_c \leftarrow \min\left\{\sum_{i=1}^{m}\left(x_{si} - w_{ji}\right)^2\right\} \quad j = 1, 2, \dots, n} \quad (5)$$

donde el índice j se refiere a una neurona en particular, n es el número de neuronas, m es el número de pesos en la neurona, s identifica a una entrada en particular, $\boldsymbol{x_{si}}$ son los vectores entrada a la red neuronal, $\boldsymbol{w_{ij}}$ es el vector de pesos asignado a cada neurona.

Habiendo seleccionado a la neurona ganadora se actualiza su peso, ya sea aumentando o disminuyendo sus valores, haciendo su respuesta mayor o menor según se desee. Los pesos de las neuronas vecinas también son actualizados generalmente escalándoles disminuyendo los pesos dependiendo de la distancia a c; es por esta razón que la función de escalamiento se le denomina topológicamente dependiente. En la Figura 10 se muestra esta retroalimentación local, las flechas curvas indican que solamente las dos neuronas que están más cercanas para cada una reciben la retroalimentación.

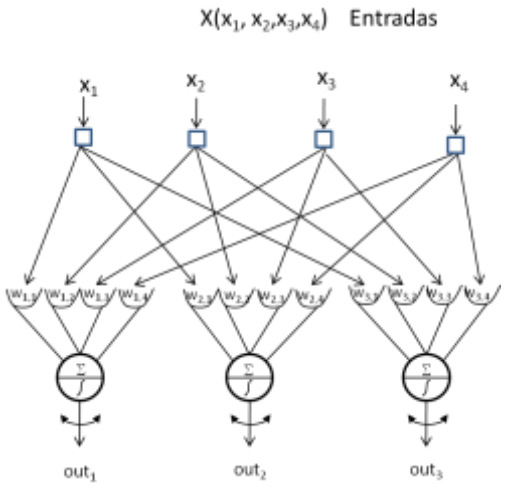


Figura 10. Red neuronal de Kohonen donde se aprecia como solamente las dos neuronas que están más cercanas reciben retroalimentación (Zupan, et al., 1993)**.**

Procediendo de esta forma una serie de vectores de entrada se ven reducidos a un número mucho menor de vectores, esto es se cuantizan, en este espacio de trabajo al cuál denominaremos espacio de la malla. Se le denomina malla porque se puede pensar en la red neuronal como un arreglo que conforma una malla de neuronas interconectadas.

Podemos definir al proceso de cuantización como aquel en que una serie de valores continuos descritos por una función $f(\cdot)$ son proyectados a ciertos valores discretos establecidos de antemano (Bovick, 2005).

Sin embargo, el simple proceso de cuantizar los vectores no es suficiente para preservar la topología de la información, cuando se piensa en datos normalmente pensamos en valores, magnitudes, signos, etc.; esto es un punto de vista algebraico del archivo, adicionalmente existe el punto de vista desde la ciencia de la información, la cual se enfoca en las relaciones entre los datos. Por ejemplo cuando se trabaja una imagen digital comprimida para reducir su almacenamiento

se da una pérdida de información y al trabajar con la imagen ya comprimida se debe lidiar con información perdida sin olvidar la posible relación que hay entre los datos que se obtienen de la imagen. Así cuando nos enfocamos en las relaciones entre datos, más que a sus atributos algebraicos se dice que está lidiando con la topología de la información (Zupan, et al., 1993).

El concepto de preservar la topología de la información es una característica esencial de los mapas auto-organizados, un ejemplo simple de cómo este concepto puede ser representado en un mapa auto-organizado lo tenemos en la Figura 11. En esta Figura se puede observar como las áreas presentadas en el mapa de la izquierda corresponde en términos del área y la colindancia de los elementos de la mano. Por lo que siendo una mano un objeto tridimensional, se ha obtenido la proyección de dicho objeto en una representación bidimensional que conserva las características topológicas del objeto representado.

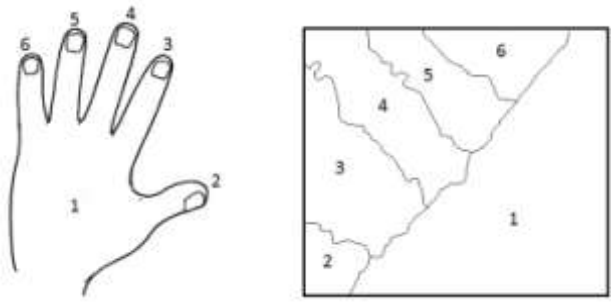


Figura 11. Representación de la topología de los cinco dedos y el dorso de la mano en forma de un mapa auto-organizado (Zupan, et al., 1993)**.**

Complementario a la cuantización de vectores, el aprendizaje competitivo maneja la preservación de la topología mediante la actualización de los pesos de las neuronas dentro de cierto vecindario alrededor de la neurona ganadora.

Sin embargo, este vecindario no se establece en base al criterio de similitud, distancia euclidiana por ejemplo, de los valores de sus vecinos, sino que se hace en base a las conexiones que se establecen con la neurona ganadora, análogo a como se da con las neuronas del cerebro.

La idea es proyectar en el mapa las señales similares a posiciones de neuronas similares (Zupan, et al., 1993).

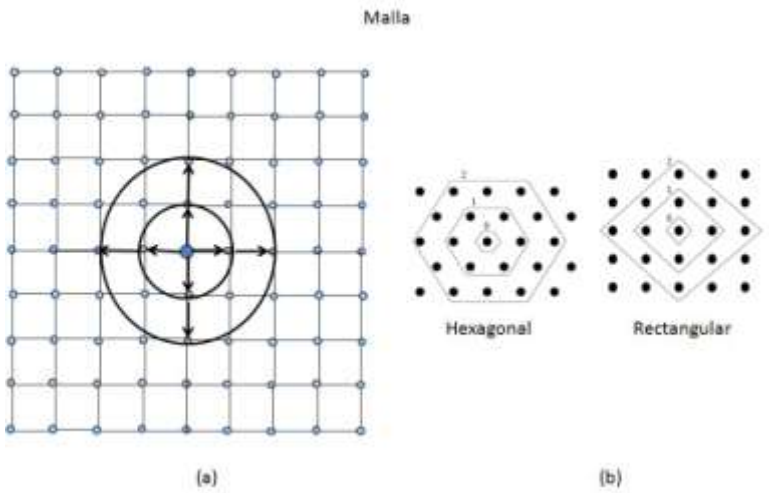


Figura 12. Tipos de vecindarios en la malla.

La Figura 12 (a) se presenta a una KNN como una malla donde cada círculo representa a una neurona y las líneas que les unen representan las distancias que hay entre una y otro neurona, además se presenta el concepto de vecinos más cercanos utilizando un vecindario de forma cuadrada, en este tipo de vecindario se tienen 4 vecinos más cercanos y no 8 como se pudiese pensar, dado que los vecinos de las esquinas son más lejanos que los otros cuatro. En la misma Figura se muestran dos niveles de vecinos más cercanos representados por los círculos concéntricos alrededor de la neurona ganadora.

En la Figura 12 (b) se presentan los dos diferentes tipos de vecindario que se pueden utilizar, el cuadrado y el hexagonal.

Habiendo definido cuales son las neuronas que se verán afectadas en la vecindad de la neurona ganadora, sus pesos deben ser reducidos dependiendo de la distancia a la neurona c, mediante una función de escalamiento dependiente de la topología, también llamada ventana de la función

$$\boldsymbol{a}(\cdot) = \boldsymbol{a}\,(\boldsymbol{d_c} - \boldsymbol{d_j}) \qquad (\mathbf{6})$$

donde $d_c - d_j$ es la distancia topológica entre la neurona central c y la neurona actual j, de tal forma que la extensión de la estimulación depende de $a(\cdot)$.

En la Figura 13 se pueden observar las funciones típicas que se utilizan para el escalamiento en los pesos del vecindario estas son la (a) constante, (b) triangular, y (c) la de forma de sombrero mexicano.
La ventana de la función es vital para el funcionamiento del algoritmo, garantiza que las neuronas dentro del vecindario tengan pesos que son similares y de esta forma aseguran la correspondencia con los puntos del espacio de entrada estableciendo de esta forma un vecindario topológico.

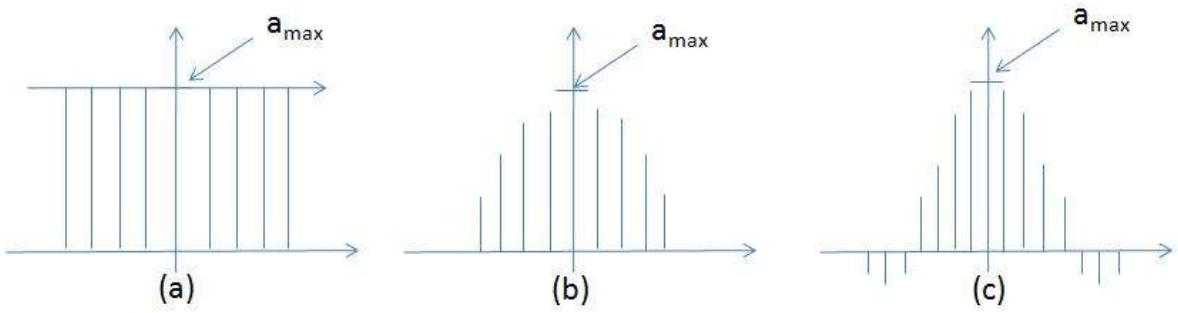


Figura 13. Diferentes formas que puede tomar la función a

La red neuronal es entrenada pasando una serie de datos muestra que mediante un proceso iterativo busca obtener la representación con el menor error posible.

El proceso por medio del cual se actualizan los pesos de las neuronas se describe mediante la siguiente ecuación

$$\boldsymbol{w_{ji}^{t} = w_{ji}^{t-1} + \eta(m) a\left(d_c - d_j\right)\left(x_i - w_{ji}^{t-1}\right)} \quad (7)$$

donde w_{ji} es el peso de una neurona dentro del vecindario en un momento t determinado, $a\left(d_c - d_j\right)$ es la ventana de la función, $\left(x_i - w_{ji}^{t-1}\right)$ es la diferencia entre el vector de entrada y la neurona actual, este valor puede ser positivo o negativo. En la ecuación (7) podemos observar que la ventana de la función es multiplicada por otra función monótona descendente representada por $\eta(m)$ denominada tasa de entrenamiento

$$\boldsymbol{\eta(m) = (a_{max} - a_{min})\frac{m_{max} - m}{m_{max} - 1} + a_{min}} \quad (8)$$

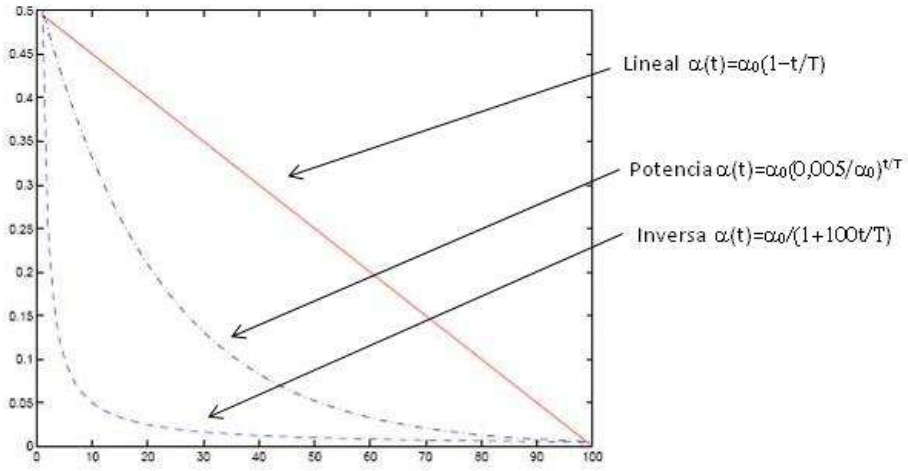


Figura 14. Diferentes funciones de entrenamiento.

También existen diferentes formas para la función de entrenamiento como se aprecia en la Figura 14, se hace notar que todas ellas son monótonas descendentes, en el caso particular de la ecuación (8) se tratan de funciones lineales.

El proceso de actualización de los pesos va provocando una distorsión en los pesos de la red neuronal que se muestra en la Figura 15.

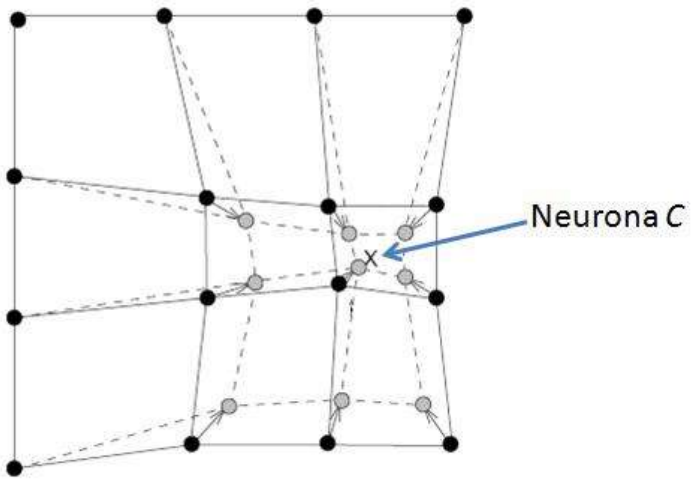


Figura 15. Distorsión de los pesos de las neuronas dentro del vecindario de la neurona ganadora

En la Figura 15 se representa cada neurona de la red como un punto y las rectas que les conectan son los pesos asignados a cada neurona, como dichos pesos están basados en la distancia a la neurona ganadora c podemos observar como las neuronas del vecindario se van plegando hacia aquella neurona que mejor les representa, con lo que se consigue la representación topográfica de los vectores de entrada.

Debido a las características especiales que tienen las salidas de una KNN no juegan un rol cuantitativo significativo como en otras redes, el único significado de la salida es ubicar topológicamente a la neurona con la salida más grande.

Con el propósito de mejorar el desempeño de la red es importante poner mucha atención en la normalización de los datos proporcionados o al menos hacer que estén escalados a algún valor razonable, esto independiente de que los pesos de las neuronas serán ajustados y normalizados de todas formas.
Adicionalmente hay que poner a punto los parámetros de la red, que son: el tamaño de la malla, el tipo de vecindario, el radio que determina el tamaño del vecindario, los valores iniciales de la ventana de la función y de la tasa de entrenamiento. Dichos parámetros son fundamentales y se definen mediante un proceso de entrenamiento de la red neuronal con datos de prueba, que se analizará a fondo en el Capítulo V.

Una característica importante de las KDD es que tiene la capacidad de manejar modelos semánticos, los cuales fueron utilizados en la presente tesis y que se procederá a explicar.

Hasta este momento hemos visto el caso en el cual al proyectar la geometría de un objeto, por ejemplo en el caso de la codificación de la partes de la mano presentada en la Figura 11, podemos asegurar que la topología responde a la codificación que se hizo.

Sin embargo en el campo de la representación lingüística uno de los primeros problemas que se enfrenta es encontrar relaciones métricas de distancia entre objetos simbólicos.

A diferencia de los casos que se han analizado, en los cuales las relaciones son derivables de las distancias mutuas en el espacio en el cual los vectores son representados, cuando se codifican símbolos no podemos asumir que tengamos ninguna relación con las características observables de los objetos correspondientes, por lo que entonces sería imposible representar la similitud lógica de los pares de objetos para entonces proyectarlas en un mapa auto-organizado.

La respuesta se encuentra en la "similitud lógica" lo que se logra al presentar los símbolos durante el proceso de aprendizaje en contexto. Así la similitud se reflejará a través de la similitud de los contextos.

Es evidente en el codificado de símbolos que las similitudes solo son derivables desde las probabilidades condicionales de su ocurrencia con otras codificaciones, independientemente del tipo de codificación.

Pongamos que tenemos dos vectores que nos representan a un objeto: por un lado el vector x_s representa la expresión simbólica y x_c la representación del contexto.

La representación más simple supone que ambos vectores están conectados en la misma neurona, i.e., la representación vector x (patrón) del objeto está formado por la concatenación de x_s y x_c tal que

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x_s} \\ \boldsymbol{x_c} \end{bmatrix} = \begin{bmatrix} \boldsymbol{x_s} \\ \boldsymbol{0} \end{bmatrix} + \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{x_c} \end{bmatrix} \quad (9)$$

La idea central es que las dos partes son ponderadas propiamente de tal forma que la norma de la parte contextual predomina sobre la parte simbólica durante el proceso de la auto-organización (Kohonen, 1990).

Ahora, en el contexto médico los componentes de los vectores contienen mediciones en los componentes que no corresponden al caso de la representación geométrica, tomando directamente la medición del colesterol y del hecho de fumar no podemos evaluar un indicador numérico del nivel de dislipidemia de un paciente. Por lo cual se necesita integrar un elemento que nos identifique el contexto de la misma forma como en la representación lingüística.

En análisis que se hizo con el especialista epidemiólogo se encontró que un eje central para determinar el contexto del síndrome metabólico son los niveles de obesidad, independientemente de los demás indicadores.
Por ello para cada objeto se definió un valor c_i para que cada objeto x_i se conformara de la siguiente manera

$$\boldsymbol{x_i} = \begin{bmatrix} \boldsymbol{x_c} \\ \boldsymbol{x_a} \end{bmatrix} \qquad (10)$$

donde la parte del contexto y la del atributo del vector de datos está representado por los vectores columna $[x_c\ 0]^t$ y $[0\ x_a\ ]^t$ respectivamente. Y los vectores de contexto están representados por

$$\begin{aligned} x_{s1} &= [c\ \ 0 \ldots 0]^t \\ x_{s2} &= [0\ \ c \ldots 0]^t \\ \boldsymbol{x_{sp}} &= [\boldsymbol{0}\ \ \boldsymbol{0} \ldots \boldsymbol{c}]^{\boldsymbol{t}} \end{aligned} \qquad (11)$$

donde c es el índice de obesidad y p es el número de componentes del vector de entrada. Es de esta forma como al final quedó la representación de los vectores que se suministraron a la KNN.

Alternativamente al algoritmo de KNN se seleccionó otro algoritmo de formación de conglomerados que se utilizó para tener una revisión cruzada de los resultados obtenidos. Este es el algoritmo de expectativa y maximización que se explica en la siguiente sección.

III.7 Algoritmo de clasificación C4.5.

El algoritmo C4.5 forma parte de los modelos aditivos basados en árboles de clasificación.

El principio de los modelos basados en árboles es segmentar el espacio de características de entrada en un conjunto de rectángulos, y entonces embonar un modelo, como una constante, en cada uno (Hastie, et al., 2001).

Para entender mejor este proceso consideremos el problema de regresión lineal con una respuesta continua Y y dos entradas x_1 y x_2, cada una tomando valores en el espacio unitario. En la Figura 16 (a) se puede observar la partición de dicho espacio por líneas paralelas a los ejes coordenados, también se observa como en cada partición se le puede modelar a Y con una constante diferente, lo cual se logra con una definición tan simple como $x_1 = c$. Sin embargo, se puede observar en dicha Figura que hay regiones que no son fáciles de describir. Con la intensión de simplificar el proceso el algoritmo se restringe a manejar particiones binarias recursivas tal y como se presentan en la Figura 16 (b). Primero se divide el espacio en dos regiones, y se modela la respuesta por la media de Y en cada región. Se selecciona la variable y el punto de división con el cual se obtenga la mejor representación. Entonces una o ambas de estas regiones se dividen en dos nuevas regiones, y este proceso continúa hasta que se cumple alguna regla que termina el proceso (Hastie, et al., 2001).

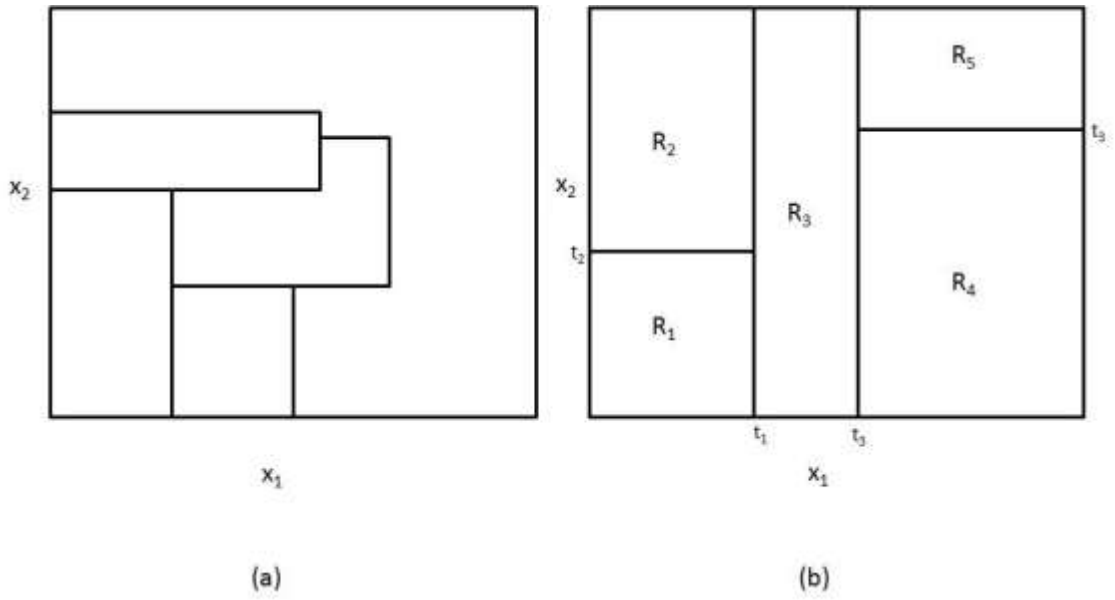


Figura 16. Segmentación del espacio de entrada en una regresión lineal de una respuesta Y y dos entradas x_1 y x_2.

En la Figura 16 primero se dividió $X_1 = t_1$, entonces la región $X_1 \leq t_1$ se dividió en $X_2 = t_2$ y la región $X_1 > t_1$ se dividió en $X_1 = t_3$. Finalmente, la región $X_1 > t_3$ se dividió en $X_2 = t_4$. El resultado de este proceso es una partición en cinco regiones $R_1, R_2, \dots, R_5$ mostradas en las figuras.

El modelo de regresión que resulta para la predicción de Y con una constante c_m en una región R_m es

$$\hat{\boldsymbol{f}}(\boldsymbol{x}) = \sum_{\boldsymbol{m=1}}^{\boldsymbol{M}} \boldsymbol{c_m} \boldsymbol{I}\{(\boldsymbol{X_1}, \boldsymbol{X_2}) \in \boldsymbol{R_m}\} \qquad (12)$$

donde:

$\hat{\boldsymbol{c}}_{\boldsymbol{m}} = \boldsymbol{ave}\ (\boldsymbol{y_i}|\boldsymbol{x_i} \boldsymbol{\epsilon}\ \boldsymbol{\mathcal{R}_m})$ es la constante asignada a cada región (13)

I{·} es la imagen producida por las rectas X_1, X_2 y los ejes coordenados

Este mismo modelo se puede representar por el árbol de la Figura 17, en este ejemplo se pueden observar las mismas particiones representadas en la Figura 16 (b).

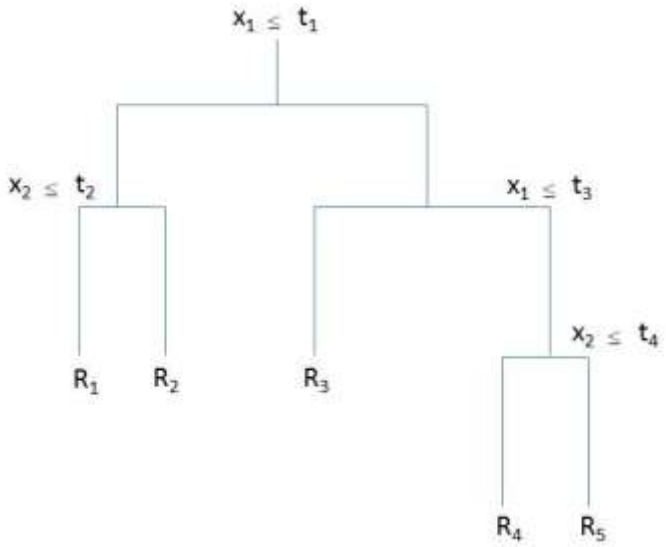


Figura 17. Representación del modelo de predicción en forma de un árbol binario recursivo.

En el ejemplo anterior se presenta un árbol de regresión, cuando se trata de un árbol de clasificación los únicos cambios que se requieren se refieren al criterio para dividir los nodos e injertar el árbol, en el entendido que la acción de injertar es encontrar un cierto subárbol $T_{\propto}$ a ser insertado en el árbol resultante T tal que $T_{\propto} \subseteq T_0$ donde T_0 es el árbol original y que minimice cierto criterio de complejidad del árbol resultante. En el caso de la regresión se utilizó el error cuadrático donde el enfoque es dividir los nodos del árbol solo si eso decrece en la suma de los cuadrados debido a que la división excede cierto umbral, la forma en que para cada α se minimiza este criterio se le considera una medida de pureza del árbol resultante.

Sin embargo, cuando el objetivo del algoritmo es la clasificación, esta medida no es utilizable; si por ejemplo el objetivo obtener una clasificación tomando valores

$1,2,\ldots,k$ es necesario hacer un cambio en los criterios para la división de los nodos y los injertos que se hacen al árbol.

En un nodo m, el cual representa una región R_m con N_m observaciones, la proporción de observaciones de la clase k en el nodo m es:

$$\widehat{\boldsymbol{p}}_{\boldsymbol{mk}} = \frac{\mathbf{1}}{\boldsymbol{N_m}} \sum_{\boldsymbol{x_i} \in \boldsymbol{R_m}} \boldsymbol{I}(\boldsymbol{y_i} = \boldsymbol{k}) \qquad (14)$$

donde i establece el recorrido de todos los pares ordenado (x_i, y_i) en la imagen de la región m.

$$\boldsymbol{k}(\boldsymbol{m}) = \mathbf{arg\,max_k}\, \widehat{\boldsymbol{p}}_{\boldsymbol{km}}, \qquad (15)$$

esto es, la clase mayoritaria en el nodo m.

Una forma de determinar qué tan correctamente están siendo clasificadas las instancias en las diferentes clases es la medida de impureza, como ya se mencionó. Son diferentes las medidas de impureza que se pueden utilizar en el nodo: error de mala clasificación, índice Gini, entropía-cruzada o desviación.

Siendo esta última, la entropía cruzada, la que se utilizó en la presente tesis, y que a continuación se procede a explicarla.

La ecuación que determina la entropía cruzada es

$$\sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \widehat{\boldsymbol{p}}_{\boldsymbol{mk}} \log \widehat{\boldsymbol{p}}_{\boldsymbol{mk}} \qquad (16)$$

para dos clases, si $\boldsymbol{p}$ es la proporción en la segunda clase esta medida es

$$-\boldsymbol{p}\log \boldsymbol{p} - (\mathbf{1}-\boldsymbol{p})\log(\mathbf{1}-\boldsymbol{p}) \qquad (17)$$

la cual es especialmente sensible a los cambios en las probabilidades.

Una ventaja de utilizar particiones binarias donde las posibles resultantes son pertenece o no pertenece (0 y 1), es que el cálculo de las posibles particiones se simplifica muchísimo. Por otro lado si se utilizaran multi-particiones (más de 2) se

tendría el problema que el espacio de entrada se segmentaría demasiado rápido, dejando pocas instancias para el siguiente nivel.

Las consecuencias de una mala clasificación dependen mucho de la naturaleza de la misma. Por ejemplo, no es lo mismo clasificar de forma errónea si una persona está en posibilidades de sufrir un ataque cardiaco, que el clasificar mal un producto al hacer una investigación de mercado.

Para apoyar en la evaluación de cuántas son las entidades mal clasificadas existe la llamada matriz de pérdida o matriz de confusión, la cual se define como una matriz $\boldsymbol{L_{kxk}}$ donde $\boldsymbol{L_{ij}}$, $i \neq j$ son los errores incurridos en la mala clasificación, si no se incurre en algún error de clasificación entonces $\boldsymbol{L_{ij} = 0}$, $\forall\, \boldsymbol{i \neq j}$.

Para problemas de clasificación en medicina, los términos que se utilizan para caracterizar la regla son la sensibilidad y especificidad. Estos se definen de la siguiente forma:

- Sensibilidad: es la probabilidad de predecir una enfermedad dado que es verdad que se padece dicha enfermedad
- Especificidad: la probabilidad de predecir que no se tiene la enfermedad dado que no se tiene la enfermedad.

Las ecuaciones que les definen son las siguientes (Bojarczuk, et al., 2004).

$$\boldsymbol{Se = \frac{VP}{(VP+FN)}} \qquad (18)$$

$$\boldsymbol{Es = \frac{VN}{(VN+FP)}} \qquad (19)$$

donde

$VP(verdadero\ positivo) =$ la regla predice que la instancia tiene una clase determinada y la instancia tiene esa clase.

FP $(falso\ positivo)$ = la regla predice que la instancia tiene una clase dada pero la instancia no tiene esa clase.
VN $(verdadero\ negativo)$ = la regla predice que la instancia no tiene una clase determinada, y en verdad la clase no la tiene.
FN $(falso\ negativo)$ = la regla predice que la instancia no tiene una clase determinada pero la instancia la tiene.

El algoritmo C4.5 es un algoritmo para extraer reglas de clasificación que sigue el paradigma de “divide y vencerás”.
La construcción del árbol binario de decisión se hará de forma recursiva:

1. Se selecciona un atributo y se coloca en el nodo raíz y se hace una rama para cada valor posible (solo hay dos valores posibles). Lo cual divide la muestra en dos subconjuntos.
2. El proceso se repite recursivamente para cada rama, utilizando solo aquellas instancias que corresponden a la rama.
3. Si en un momento todas las instancias de un nodo tienen la misma clasificación, se detiene el desarrollo en esa parte del árbol.
4. El proceso se detiene al terminar de regresar de la recursión de cada una de las ramas que quedaron pendientes (Witten, et al., 2005).

Ahora el problema es determinar cuál es el atributo que se debe tomar para la división lo cual se hace calculando el valor de información promedio que proporciona el atributo y tomando en cuenta el número de instancias que quedaron en cada rama.

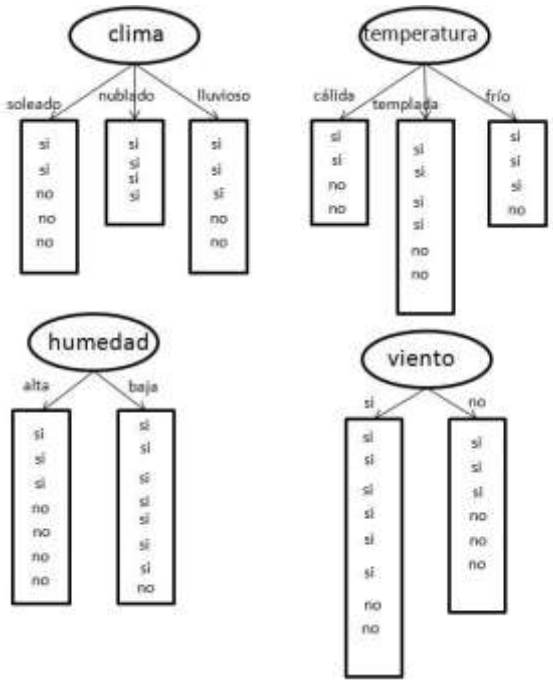


Figura 18. Nodos del árbol de los datos de ejemplo del weather data de weka (conjunto exhaustivo de herramientas para análisis y minería de datos).

En la Figura 18 se presenta un ejemplo explicativo (Witten, et al., 2005). Supóngase que se haya en el momento de segmentar un nodo del árbol, en el cual se encuentran como posibles atributos a ser seleccionados para dividir la rama: clima, temperatura, humedad, viento.

Es necesario determinar cuál es el nivel de información que aporta cada atributo y tomar el máximo de acuerdo a la ecuación (15).
Una forma práctica de resolver este problema es determinar el nivel de información que aporta cada atributo de la siguiente manera.

Supóngase que se quiere determinar el nivel de información del primer atributo “clima”:

1. Se calcula el nivel de información promedio que aporta cada valor del atributo, en este caso:

soleado implica 5 valores con este atributo de 14 posibles, probabilidad de $(\frac{5}{14})$,

nublado implica 4 de 14 probabilidad de $\frac{4}{14}$ y

por último lluvioso implica 5 de 14, probabilidad de $\frac{5}{14}$.

2. Se calcula el valor de la información de la siguiente forma, para el primer atributo clima:

   se toma el número de “si” y “no” que hay con lo que se van a evaluar 3 funciones

   $info([2,3])$ para el valor “soleado”,

   $info([4,0])$ para el valor “nublado”,

   $info([3,2])$ para el valor “lluvioso”.

3. El cálculo de dicho valor se basa en la entropía, de acuerdo a la ecuación (17) cuando se trata n argumentos esta ecuación queda de la siguiente forma

   $entropía(p_1, p_2, \dots, p_n) = -p_1 \log p_1 - p_2 \log p_2 \dots - p_n \log p_n$

   donde los argumentos para la entropía son la información promedio

$$info([2,3]) = entropoía\left(\frac{2}{5}, \frac{3}{5}\right) = -\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5}$$

4. Un ejemplo del cálculo de la información promedio es el siguiente: tomando en cuenta el número de instancias que hay debajo de cada rama en el subárbol clima se tiene el siguiente conteo:

$$info([2,3],[4,0],[3,2]) = \left(\frac{5}{15}\right) \times 0.971 + \left(\frac{4}{14}\right) \times 0 + \left(\frac{5}{14}\right) \times 0.971 = 0.693\ bits$$

5. Es en base a esta ganancia de información que se determina cual argumento es tomado para realizar la división.

Las reglas obtenidas por el método C4.5 adolecen del problema que pueden llegar a ser muy extensas e incluso pueden existir un mínimo de reglas redundantes o inconsistentes, dada la forma en que trabaja el método (Witten, et al., 2005).

Una forma de asegurar la robustez de las reglas encontradas es maximizar el número de muestras con el cual se está trabajando y probar de forma exhaustiva la validez de las mismas. Un método que cuenta con estas características es la programación genética. Esta técnica forma parte de la familia de los algoritmos evolutivos que son considerados algoritmos de optimización.

III.8 Programación genética.

La programación genética (Genetic Programming-GP por sus siglas en inglés) forma parte de la familia de algoritmos evolutivos y se utiliza para encontrar modelos con la máxima adaptación.

La idea subyacente en todos los algoritmos evolutivos es la misma: dada una población de individuos dentro de algún medio ambiente que tiene recursos limitados, la competencia por dichos recursos causa la selección natural, la sobrevivencia del más fuerte (Eiben, et al., 2003).

En otras palabras, la idea principal es evolucionar una serie de individuos (población), donde cada individuo representa una solución candidata a un problema dado. En cada generación los individuos se seleccionan para su reproducción, la selección probabilística que se hace de los individuos dirige al algoritmo a la selección de los mejores individuos de esa generación. Las nuevas generaciones son afectadas por operadores que modifican a los individuos seleccionados (recombinación y mutación). El proceso de selección y generación de nuevos individuos se repite por un cierto número de generaciones, de tal forma que la calidad de los individuos se espera mejoren en cada ocasión, para al final seleccionar al mejor de los individuos (Bojarczuk, et al., 2004).

Los principales componentes de los algoritmos evolutivos son:

- La representación de los individuos,

- La función de evaluación de la adaptabilidad,
- La población,
- El mecanismo de selección de padres,
- Los operadores de variación (recombinación y mutación) y
- El mecanismo de selección de los sobrevivientes.

Según el algoritmo evolutivo que se trate, dichos componentes toman características muy particulares.

Originalmente la GP se concibió para evolucionar programas de computadora. Como en el caso de los demás algoritmos evolutivos la GP transforma estocásticamente a los individuos esperando obtener mejores programas. Dada su naturaleza no se pueden garantizar los resultados, sin embargo su carácter aleatorio proporciona un medio de escape de las trampas en las cuales puede caer un algoritmo determinístico.

Conforme ha pasado el tiempo la GP se ha utilizado para evolucionar otro tipo de individuos y no solamente programas de computadora, incluso es más común utilizarle, por ejemplo, para evolucionar expresiones o algún tipo de configuración.

En el caso de la toma de decisiones médicas se han planteado como la representación más adecuada de conocimiento las reglas del tipo $SI-ENTONCES$ En las secciones anteriores de este Capítulo se planteó el problema de clasificación y su solución hasta llegar a expresiones de este tipo.

Con la programación genética el problema de clasificación es replanteado como un problema de optimización, donde las reglas pueden ser representadas como individuos a ser evolucionados para obtener los mejores adaptados. Esto es aquellas reglas que describan mejor el problema.

Un elemento básico para iniciar con la solución de un problema mediante la GP es determinar la representación que se va a utilizar.
Hay dos representaciones, la primera desde el contexto del propio problema. En este caso la forma $SI-ENTONCES$ que hemos mencionado. Es decir, la representación se hace a nivel de fenotipo.

A la codificación de los fenotipos, esto es ya dentro del algoritmo evolutivo, se les llama Genotipo. En la GP una estructura comúnmente utilizada es mediante árboles binarios completos (Figura 19). Es con esta estructura con las que van a trabajar los operadores genéticos.

Para poder definir el genotipo es necesario haber definido dos diferentes conjuntos de elementos, el conjunto de terminales $\{T\}$ y el conjunto de funciones $\{F\}$.
Se puede observar en la Figura 19 que en las hojas del árbol binario aparecen pares parámetro-valor, los cuales conforman al conjunto T, mientras que en los

nodos intermedios se observan operadores binarios del tipo >=, <, =, AND, OR, los cuales conforman el conjunto F.

Otro elemento importante es la selección de la función de evaluación de adaptabilidad (fitness en inglés), la que permite determinar el nivel de adaptación de un individuo para resolver el problema establecido y que está íntimamente relacionada con la representación seleccionada y con el contexto del problema a resolver.

Como ya se mencionó la representación a nivel del Fenotipo son reglas del tipo $SI-ENTONCES$ que permiten predecir el diagnóstico de una enfermedad.

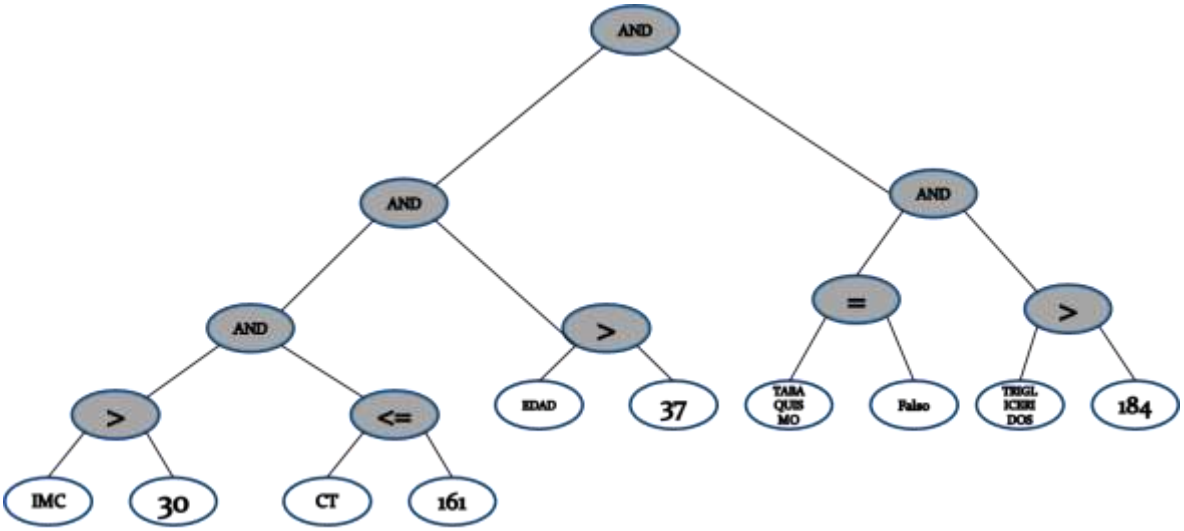


Figura 19. Representación del genotipo de una de las reglas descubiertas.

En la práctica médica normal, particularmente en la epidemiología, para la formulación de este tipo de reglas se utilizan los estudios de cohorte[5].
Un estudio de cohorte o de seguimiento, es aquel procedimiento epidemiológico analítico, no experimental, en donde un grupo de individuos con un factor de

riesgo, cohorte expuesta, se compara con otro sin el factor de riesgo, cohorte no expuesta, con el objetivo de observar en cada uno la aparición y evolución de la enfermedad (Kahl, 1990).

En los estudios de cohorte se utilizan diferentes indicadores para evaluar la exactitud predictiva de la regla, los mismos que van a ser utilizados en la GP como medida del fitness, y que se vieron al explicar el algoritmo C4.5 la sensibilidad de la regla (18) y la especificidad (19), ver sección anterior.

Adicionalmente en el caso de la GP la reglas deben ser simples, por lo que se define el parámetro Simplicidad (Sy) como

[5] Cohorte conjunto, número, serie (Ruiz, et al., 2004)

$$\boldsymbol{Sy} = \frac{\boldsymbol{maxnodos - 0.5 \cdot numnodos - 0.5}}{\boldsymbol{maxnodos - 1}}, \qquad (20)$$

donde maxnodos es el número máximo de nodos permitidos para el árbol, y numnodos es el número de nodos, funciones y terminales, de un árbol (Bojarczuk, et al., 2004).

La exactitud predictiva, es una manera de reunir los dos aspectos anteriores

De acuerdo al modelo planteado por Bojarczuk et al (2004) una forma de calcularle es

$$\boldsymbol{Ep} = \boldsymbol{Se} \cdot \boldsymbol{Sp} \cdot \boldsymbol{Sy} \qquad (21)$$

el cual fomenta que la GP maximice ambos Se y Sp al mismo tiempo.

Se encontró que esta forma de calcular la exactitud predictiva adolece de dar el mismo peso a cada factor de la fórmula, por lo que se agregaron factores de ponderación α para la Se y β para Sp. De acuerdo a Ruiz et al (2004) es poco frecuente que exista una prueba altamente sensible y específica al mismo tiempo, por lo que se decidió por una versión ponderada donde haya un mayor control sobre el fitness de los individuos de la forma

$$\boldsymbol{Ep} = (\boldsymbol{\alpha Se}) + (\boldsymbol{\beta Sp}) + (\mathbf{1} - \boldsymbol{\alpha} - \boldsymbol{\beta})(\boldsymbol{Se} \cdot \boldsymbol{Sp}). \qquad (22)$$

Pasando al manejo de la población un primer elemento que se debe definir es la forma de inicializarle. El método más común es el llamado half-and-half donde la mitad de la población es generada bajo el método *full*, donde se generan árboles completos y todas las hojas están en la misma profundidad, y la otra mitad bajo el método "*grow*", los árboles son creados de diferentes tamaños y formas (Koza, 1992).

En el caso de la selección de padres se utiliza la selección por torneo donde un grupo de individuos es seleccionado al azar de la población. Se les compara entre ellos y entonces los mejores individuos son seleccionados como padres.

En el caso de la recombinación el método comúnmente utilizado es el cruzamiento por subárboles, ejemplificado en la Figura 20, en este método se seleccionan dos puntos de corte diferentes dentro de cada árbol padre. Entonces el padre uno sirve de base para la creación del hijo uno, habiendo cortado el subárbol que está debajo del punto de corte y reemplazándole por el correspondiente subárbol proveniente del corte en el padre dos. El mismo procedimiento se repite con el hijo 2.

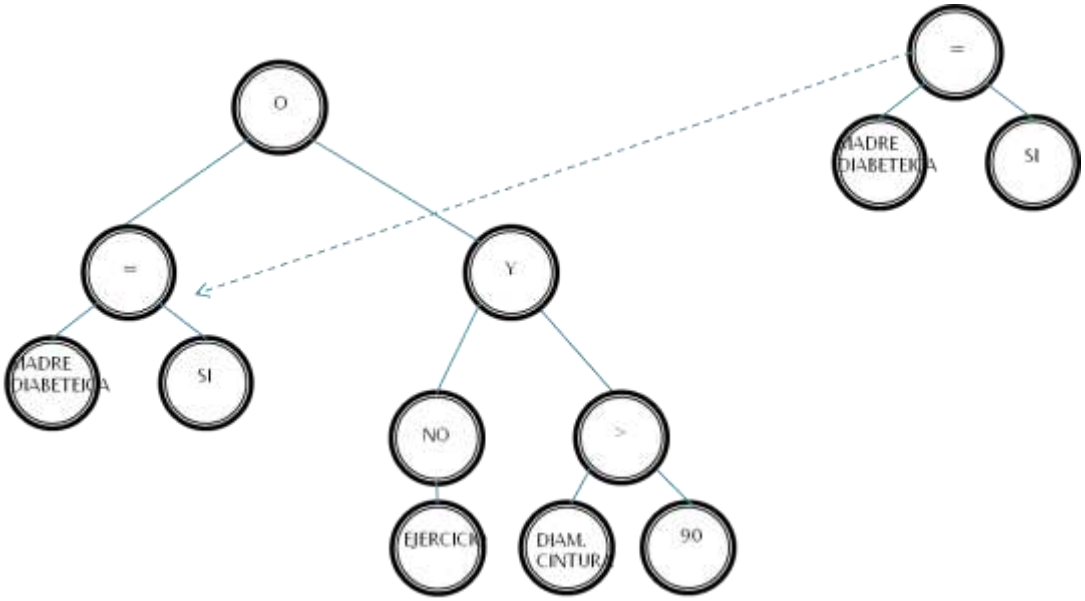


Figura 20. Ejemplificación del cruzamiento por subárbol.

Otro operador genético que asegura cierta variabilidad en los individuos que forman la población es la llamada mutación. En este operador cierto individuo seleccionado aleatoriamente sufre un cambio en su genotipo. El método comúnmente más utilizado es la mutación por subárboles, donde igualmente se selecciona un punto de corte dentro de un individuo seleccionado al azar para la mutación, para entonces reemplazar el subárbol debajo del punto de corte por un

sub-árbol válido generado de forma aleatoria en base al conjunto de funciones y terminales.

Otro operador genético es la selección de sobrevivientes, donde se utilizó el método de la ruleta. En este método cada individuo abarca una parte de la ruleta proporcional a su fitness y a semejanza de lo que sucede en el juego se selecciona un individuo aleatoriamente con una probabilidad de ser seleccionado proporcional al fitness.

III.8.1 Programación Genética con una sintaxis restrictiva.

Una propiedad que debe guardar la programación genética es la propiedad de cerradura, la cual se subdivide en dos propiedades: consistencia de tipo y evaluación segura (Koza, 1992).
La consistencia de tipo implica que se puedan unir y mezclar nodos arbitrariamente. Como resultado de esto es necesario que cualquier sub-árbol pueda ser utilizado en cualquier posición de argumento para cualquier función dentro del conjunto de funciones, ya que es posible que el cruzamiento genere esa combinación.
La evaluación segura se refiere a que algunas funciones generadas pueden llegar a generar problemas en tiempo de ejecución. El ejemplo más simple de esta problemática es la división entre cero.

Ahora bien la propiedad de cerradura es una condición muy limitada cuando se está trabajando con reglas del tipo SI *antecedente* ENTONCES *consecuente1* EN CASO CONTRARIO *consecuente2*.
En el caso de la presente tesis solamente se trabaja con un solo consecuente, ya que lo que se busca es determinar la pertenencia o no a cierto tipo de síndrome metabólico, esto es a cierto conglomerado. Adicionalmente el antecedente tiene ciertas características importantes que se deben preservar:

- En el antecedente de una regla no debe aparecer dos veces un mismo parámetro con dos valores diferentes, por ejemplo:
  SI edad > 23 y CT > 211 y edad > 65
- Un parámetro puede ser categórico o cuantitativo, por ejemplo:
  - Si FUMA es categórico ya que divide a los individuos en dos categorías.
  - CT(colesterol total) es cuantitativo ya que puede tomar una valor real dentro de cierto intervalo
- Al construir nuevos individuos o al modificarles en las hojas los pares parámetro-valor deben ser congruentes de acuerdo al tipo de parámetro, ejemplo: no es válido generar un nodo terminal con Si FUMA > 190.

Una forma de expresar estas restricciones es a través de una gramática. La que debe estar expresada en su forma DNF[6], a continuación se muestra parte de la gramática:

- tipos de argumentos:
  ARGUMENTO=DOBLE|ENTERO|CATEGORICO
- Un ejemplo de una producción para un nodo no terminal
  COMPARADOR_NUMERICO= ”>” | ”<=”
  COMPARACION=atributo+COMPARADOR_NUMÉRICO
  CONJUNCION=AND+COMPARACION|COMPARACION|CONJUNCION
  ANTECEDENTE=CONJUNCION|OR+CONJUNCION|ANTECEDENTE
  nonTerminal = nonTerminal+ANTECEDENTE
- Un ejemplo de una producción para un nodo terminal
  TERMINALS= ”>=” | ”>”
  TerminalNode=atributo|TERMINALS|valor_aleatorio

Mediante las reglas de producción se controlan las diferentes condiciones que se establecen para cada tipo de argumento y para la estructuración de la propia regla. Junto con el fitness la gramática trabaja para asegurar la formación de

[6] Forma Normal Disyuntiva

reglas simples, congruentes y entendibles, en el sentido que no sean demasiado grandes.

III.9 Resumen.

En este Capítulo se partió del planteamiento de que el problema a resolver para apoyar la toma de decisiones es un problema de clasificación y se ha mostrado como el generar una clasificación de riesgo para el síndrome metabólico puede considerarse un problema de este tipo.

Se ha propuesto una serie de métodos de aprendizaje de máquina como una forma de solucionar el problema de clasificación planteado. Dado que no se tiene conocimiento a priori de las clases que conforman el problema, es necesario utilizar un grupo de algoritmos de aprendizaje de máquina conocidos como de aprendizaje no supervisado. El método propuesto es la red neuronal de Kohonen.
Se ha justificado el uso de la red neuronal por lo importante que es preservar las características topológicas de la información a trabajar. Además, por las características propias de la información médica, se propone el uso de una variante de la KNN, una KNN semántica, esto porque el contexto de la información médica con la que se va trabajar es más parecido a un problema de semántica que a un problema geométrico.

Habiendo encontrado los conglomerados que constituyen la clasificación es necesario extraer los patrones que definen la pertenencia a cada conglomerado, expresándolos en la forma de reglas del tipo SI-ENTONCES
Con este propósito se propone el uso del algoritmo C4.5 como uno que precisamente utiliza este tipo de representación y que es similar a la lógica que utilizan los médicos en el proceso de diagnosis.

Hasta este punto de alguna forma se ha ilustrado como se puede resolver el problema de clasificación y su consecuente re expresión en forma de reglas entendibles para el experto.

El siguiente paso es convertir el problema de clasificación en un problema de optimización para que de esta forma se asegure que las reglas resultantes son robustas y reúnen ciertas características que les hagan útiles para el experto médico.

Pensar en realizar este mismo trabajo por métodos manuales no es viable para el especialista médico, por el trabajo que representaría e incluso por la complejidad matemática que implica la alta dimensionalidad de los datos. Cada registro médico tiene varios indicadores (niveles de colesterol, edad, peso, sexo, etc.) que tendrían que correlacionarse.

De acuerdo a la literatura antes de procesar los algoritmos es fundamental entender a fondo el contexto de la información con la cual se va a trabajar, desde el punto de vista médico y de los procesos que involucra la toma de decisiones en el manejo del síndrome metabólico. Para tal efecto en el Capítulo IV se detalla la metodología utilizada para en análisis del proceso de toma de decisiones.

Capítulo IV

Análisis del contexto médico y del manejo del síndrome metabólico

IV.1 Introducción.

Las fases 1,2 y 3 de la metodología de soporte a la toma de decisiones con enfoque a la ingeniería de procesos establecen que el primer paso es estudiar el ámbito del problema antes de poder definirlo. Primeramente se debe determinar que existe un problema de toma de decisiones, así el objetivo de la fase es conocer el(los) proceso(s) relacionado(s) con el problema detectado. Las acciones a realizar son: elaborar una descripción textual del o los proceso(s) y un documento referente al modelado de proceso(s) el cual permitirá ampliar la comprensión del mismo. De forma más específica las acciones propuestas a seguir son:

1. Entender el síndrome metabólico, ubicando en la literatura: como se le define, sus características; de las fuentes médicas con autoridad identificar por qué no se utiliza en la diagnosis; de las herramientas institucionales para el manejo de las enfermedades relacionadas con el síndrome metabólico qué información se requiere para su diagnosis, cómo afecta al cuerpo humano, y los factores que le provocan.
2. De acuerdo a la metodología se deben ubicar los procesos mediante observación, entrevistas y recolección de documentos relacionados. La idea es dar seguimiento a los procesos relacionados con el manejo del síndrome metabólico en medicina familiar, identificar que tanto las personas que trabajan con él lo conocen, sobre todo por la poca difusión que hay del síndrome.

3. Para entender la naturaleza de los datos con los que se trabajaron, se participó en la revisión que el cuerpo médico del hospital de zona 8 realizó durante dos meses, de las nuevas guías médicas de la DM2 y la de la HTA. Donde se identificaron las decisiones importantes que se toman en el manejo del síndrome metabólico y cuáles son las herramientas con que se cuentan para tomar dichas decisiones.
4. Para determinar qué datos quedan registrados en la memoria institucional y cuáles de ellos estaban accesibles para el estudio a realizar, se revisó con el epidemiólogo la información que él genera y que es fuente oficial del almacén de datos del IMSS.

De aquí que para entender el problema primero se presenta el contexto médico del síndrome y una definición amplia del mismo. Igualmente se analiza la forma en que el IMSS, maneja de forma institucional al paciente con este síndrome. Posteriormente para ubicar el proceso de toma de decisiones en el manejo del síndrome metabólico se analizó un caso de estudio real del manejo del síndrome, en medicina de primer nivel (familiar) a través del cual se definió el problema de decisión y de forma profunda se analizó el manejo del síndrome metabólico. De acuerdo a las acciones de la fase 3 de la metodología, se hizo un modelado del caso de estudio y se determinó cuál es la información más relevante en el manejo del síndrome metabólico. Finalmente, como lo establece la fase 4, se establecieron los objetivos de las decisiones.

En la Sección IV.2 se presenta el contexto médico del padecimiento, las enfermedades que aparecen relacionadas con el síndrome metabólico (comorbilidad) y cuáles son los indicadores principales que permiten detectarle.
En las secciones IV.3 a la IV.5 se presenta el caso de estudio práctico llevado a cabo en el Hospital General de Zona número 8 ubicado en la ciudad de Ensenada Baja California, México.

En la Sección IV.6 se presenta el modelo de procesos generado sobre el caso de estudio. Para finalizar en la Sección IV.7 presentando la información relevante que se maneja y el conjunto de datos con los cuales se cuenta para realizar el presenta trabajo, para entonces finalizar con un resumen del Capítulo.

IV.2 Contexto médico del síndrome metabólico.

De acuerdo a Alonso (2008) se trata de una *"entidad clínica controvertida que aparece, con amplias variaciones fenotípicas*[7]*, en personas con una predisposición endógena*[8]*, determinada genéticamente y condicionada por factores ambientales".*

Los factores que propician la aparición del síndrome metabólico son múltiples y de muy diversa naturaleza (en la Figura 21 se les esquematiza) y se puede resumir en 4 grandes grupos:

1. Determinantes de la salud: estilo de vida, sedentarismo, hábitos alimenticios, zona geográfica de residencia.
2. Factores sicológicos: negación del problema, discriminación por padecer sobrepeso, no ser perseverante en los cambios de hábito, entorno familiar, falta de tiempo para preparar sus propios alimentos, etc.
3. Factores genéticos: antecedentes genéticos en parientes de primera generación, características del genotipo de acuerdo a la raza, etc.
4. Factores asociados principalmente con la obesidad-sobrepeso y las enfermedades relacionadas: dietas ricas en grasas saturadas, hidratos de carbono, bajas en fibras, basadas en alimentos chatarra, etc. Puede haber

[7] Fenotípica: perteneciente o relativo al fenotipo.
[8] Endógeno: Que se origina en virtud de causas internas.

otras causales de la DM2 y la HTA pero para efectos de la presente tesis no son relevantes.

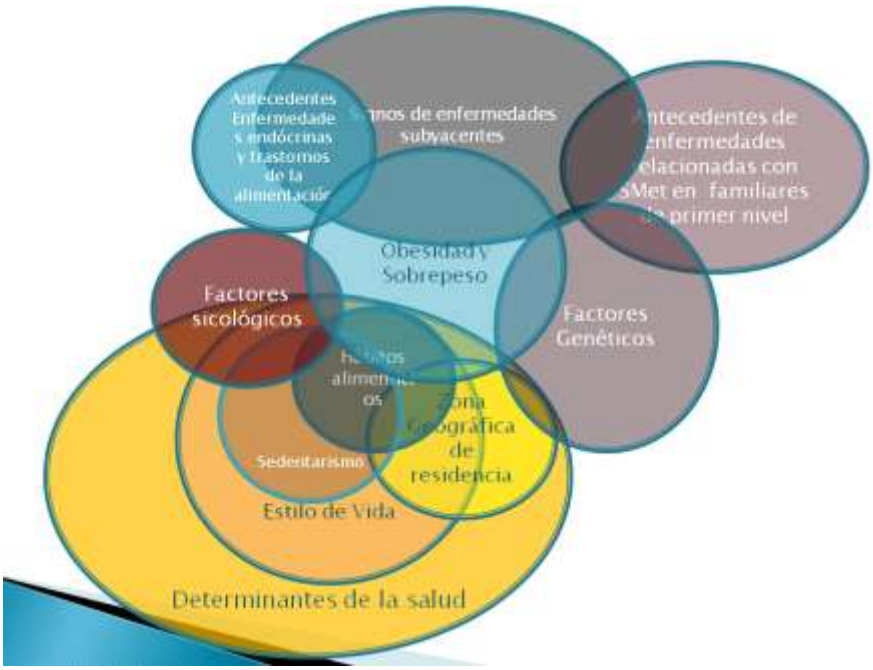


Figura 21. Esquema multifactorial del síndrome metabólico.

Las principales características del síndrome metabólico incluyen resistencia a la insulina, obesidad abdominal, presión sanguínea elevada, y anormalidades en los lípidos (i.e. niveles elevados de triglicéridos, y niveles bajos de lipoproteínas de alta densidad).

Fue inicialmente definido por un panel de expertos de la Organización Mundial de la Salud (OMS) en 1985. En el año 2001 el Departamento de Salud y Servicios Humanos de los Estados Unidos publicó el cuadro que se presenta en la Tabla II. (HEALTH, 2001):

Tabla II. Factores para determinar si se padece o no el síndrome metabólico.

Factor de Riesgo	Límite	
	Hombre	Mujer
Obesidad Abdominal*	Circunferencia de la cintura > 102 cm	Circunferencia de la cintura > 88 cm
Triglicéridos	≥ 150 mg/dl	
HDL colesterol	< 40 mg/dl	< 50 mg/dl
Presión de Sangre	$\geq 130/\geq 85$ mm/dl	
Glucosa en ayunas	≥ 110 mg/dl	

Se identifican clínicamente al síndrome metabólico al presentar al menos tres de los cuadros que aparecen en la tabla anterior.

Otros organismos han incluido como un factor importante la resistencia a la insulina, e incluso la OMS ha incluido como criterio necesario a la DM2.
La resistencia a la insulina se define como la condición en la cual los tejidos dejan de responder a la insulina. Si se tiene resistencia a la insulina el cuerpo producirá más y más insulina, pero como los tejidos no responden a ella no será capaz de usar el azúcar adecuadamente.

De acuerdo a (Ceriello A, 2004) la DM2 y el riesgo cardiovascular *"tienen factores genéticos y ambientales en común. Uno de estos antecedentes es la resistencia a la insulina, un componente importante del síndrome metabólico (obesidad abdominal, disminución de las lipoproteínas de alta densidad [HDL], aumento de triglicéridos e hipertensión) que parece presentarse mucho antes de la diabetes sintomática. Cuando las células beta no pueden sostener el aumento de la producción de insulina, aparece la disminución de la tolerancia a la glucosa."*

Otro elemento importante a considerar es el estrés oxidativo que se define como la condición de producción oxidativa incrementada en las células animales caracterizada por la liberación de radicales libres y la consecuente degeneración celular (Ceriello A, 2004). El estrés oxidativo aumenta la resistencia a la insulina.

Partiendo de todos los antecedentes citados se puede establecer una definición amplia del síndrome metabólico como:

"Una alteración generalizada del organismo, provocada por causas multifactoriales cuya etiología[9] es variada y la cual puede llegar a desembocar en enfermedades crónico degenerativas tales como la obesidad y el sobrepeso, la HTA, la dislipidemia y la DM2. Enfermedades que han adquirido el nivel de pandemia".

IV.3 Manejo institucional de las enfermedades relacionadas con el síndrome metabólico

Primeramente hay que diferenciar el manejo que se hace del síndrome metabólico en medicina de primer nivel y en piso de hospital. El énfasis en medicina de primer nivel es preventivo, mientras que en el hospital está más relacionado con la comorbilidad del síndrome metabólico.

Por esta razón se decidió hacer el estudio en medicina de primer nivel, principalmente centrado en el manejo que se hace de los pacientes con riesgo cardiovascular, DM2 y en general personas con obesidad y sobrepeso.

En el momento de la elaboración de la presente tesis los esfuerzos realizados por el Sector Salud del gobierno mexicano, y particularmente por el IMSS, se han

[9] Estudio de las causas sobre alguna enfermedad

reflejado en programas para la atención de la HTA y la DM2. Programas que han sido prioritarios durante muchos años.

Más recientemente el sobrepeso y la obesidad han tomado una importancia relevante. Un esfuerzo significativo realizado por IMSS en la lucha contra la obesidad y las enfermedades relacionadas es la estrategia SODHI (en apoyo a obesos, diabéticos, e hipertensos) (IMSS, mayo 2008).

El programa SODHI implica una serie de acciones conjuntas entre el área médica y la de prestaciones sociales, con el objetivo de prevenir el llegar a padecer dichas enfermedades, cambiando los hábitos alimenticios, evitando el sedentarismo y sus futuras complicaciones.

Esta estrategia plantea las acciones coordinadas entre el médico familiar, nutriólogo, trabajador social, sicólogo y servicio de los centros de seguridad social. En estos sitios se le brinda al paciente una orientación nutricional, actividad física, educación para la salud y actividades educativas. Sin embargo el problema ha sobrepasado, y con mucho, a los resultados obtenidos.

Tanto las acciones del médico familiar como las de las instancias de apoyo tienden a atacar el problema cuando ya han pasado varios años de evolución del síndrome y la obesidad, o la enfermedad crónico degenerativa, se ha complicado. Se encontró que de acuerdo a diversas fuentes no hay un tratamiento específico para él, y solo se hace referencia a los tratamientos que previenen su comorbilidad. Un ejemplo de esto lo tenemos en la Figura 22 donde (Alonso, 2008) presenta un algoritmo que establece su manejo únicamente basándose en el hecho de padecer o no el síndrome, mediante el control de los indiciadores que previenen su comorbilidad.

De aquí que una de las causas principales por las cuales no se usa el diagnóstico del síndrome metabólico como en la consulta es por esta falta de tratamiento y la poca relevancia que tiene frente a su comorbilidad, al menos en México. Sin embargo para que dichos programas apoyen los tratamientos preventivos de la

comorbilidad es importante la detección temprana del síndrome metabólico. Ya que, como se mencionó anteriormente, las manifestaciones del síndrome se presentan antes que la sintomatología de las enfermedades relacionadas con él.

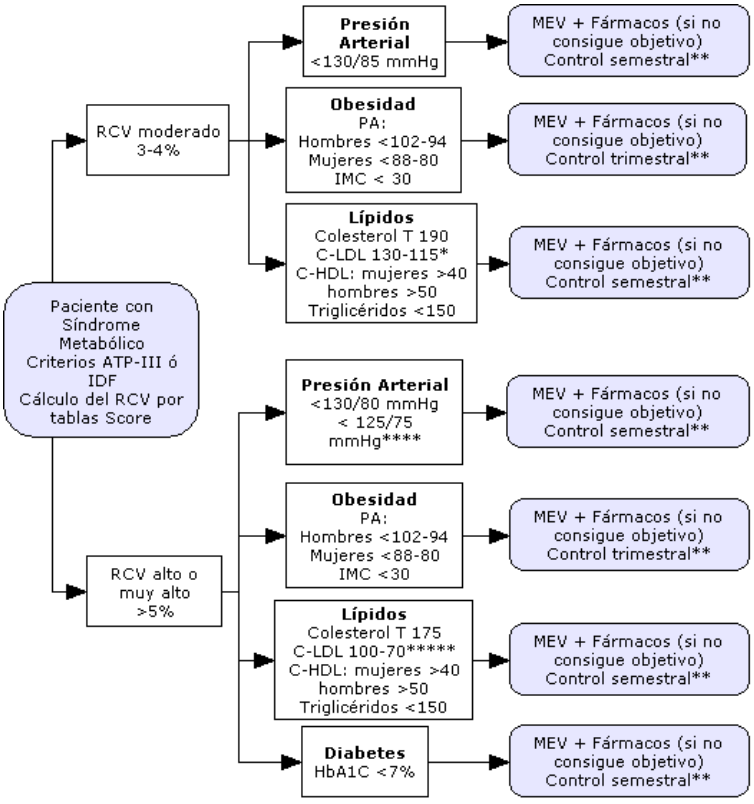


Figura 22. Algoritmo de manejo del síndrome metabólico tomado de (Alonso, 2008).

Llegar a contar con una clasificación de riesgo específica del síndrome metabólico que permita jerarquizarle desde que el síndrome empiece a manifestarse, puede llegar a ser una herramienta útil en este sentido, aun cuando no se pretenda curar el síndrome por sí mismo.

Habiendo establecido el contexto médico de la enfermedad otro aspecto importante es entender el manejo que actualmente hace el médico del síndrome metabólico y cuáles son las incertidumbres a las que se enfrenta al detectarle y querer prevenir sus consecuencias.
Para estudiar a fondo los procesos en el manejo del síndrome metabólico se aplicaron diferentes técnicas de la ingeniería de procesos que a continuación se describen.

IV.4 Estudio del proceso de manejo del síndrome metabólico en medicina de primer nivel.

Para entender el contexto del problema que se va a estudiar y el manejo que los especialistas hacen de la problemática implicada sentido se realizó un estudio de un caso práctico en el hospital general de zona número 8.

En este análisis un primer paso fue identificar los procesos y actores que toman parte en la atención de primer nivel (medicina familiar) lo cual se presenta en la siguiente sección.

IV.5 Proceso del manejo del síndrome metabólico.

Se aplicaron diferentes actividades de la ingeniería de procesos, entre ellas:

- Análisis del manejo del síndrome metabólico y su comorbilidad en medicina familiar. Se analizaron las guías médicas y las normas nacionales para la prevención y el tratamiento de la obesidad, dislipidemias, DM2 y la HTA, revisándolas en forma particular y con el epidemiólogo.
- Aplicación de entrevistas semi estructuradas a 6 médicos familiares, 5 asistentes médicos, 1 trabajador social, 1 nutriólogo, 1 sociólogo y 1 enfermera del programa PREVENIMSS[10].
- Se aplicó una técnica de análisis denominada seguimiento de sombra que consistió en seguir durante un lapso a las asistentes médicas para analizar el trabajo de apoyo que hacen al manejar el síndrome metabólico junto con el médico familiar
- Se participó en las reuniones del cuerpo médico de revisión de las guías médicas 2010 para el manejo de la HTA y la DM2
- Se elaboraron diferentes diagramas de acuerdo a la metodología de análisis de toma de decisiones grupales desarrollada por la M.C. María de Jesús Pacheco (2004).

La técnica de ingeniería de procesos que se utilizó persigue una serie de objetivos que se pueden resumir en la siguiente lista:

- Identificar los actores y los roles que estos juegan dentro del proceso del manejo del síndrome metabólico.
- Ubicar los pasos que se siguen en el manejo del síndrome metabólico
- Ubicar las decisiones, cursos de acción y personas que intervienen la toma de decisiones
- Identificar las preocupaciones e incertidumbres de los actores al tomar decisiones
- Ubicar los datos que se usan, quien los usa y quien los genera

En el caso de la presente tesis todos estos objetivos se refieren al manejo del síndrome metabólico.

[10] Estrategia diseñada e implementada por el IMSS de programas integrados de salud

El Anexo F contiene el protocolo de entrevista que se utilizó, la cual es una entrevista semi estructurada, con preguntas abiertas. Dichas entrevistas se transcribieron y se analizaron obteniendo la siguiente descripción del proceso:

Proceso de toma de decisiones del síndrome metabólico.

- Un paciente puede llegar a consulta con un médico familiar o con un médico especialista
- En ambos casos le recibe un asistente médico, quien se encarga de registrar ciertos datos antropométicos (peso, talla, estatura, presión arterial) antes de pasarle con el médico correspondiente. Ésta información la registra en la hoja de datos antropométricos, la cual se le pasa al médico correspondiente.
- En el caso del médico familiar:
    - Mediante simple observación, determina si el paciente padece de obesidad o cierto sobrepeso. Lo que establece cierto riesgo de padecer el síndrome metabólico. Existe una norma mexicana sobre la obesidad (NOM-174—SSA1-1998) que es la base para tomar la decisión sobre la existencia del riesgo.
    - Como fuentes de información que el médico usa las principales son:
        - la hoja de datos antropométricos donde viene registrada la presión arterial, el peso y la talla, con lo cual se puede calcular el índice de masa corporal (IMC).
        - los antecedentes del paciente asentados en el expediente como la tensión arterial sistémica, análisis de sangre, antecedentes familiares, etc.
    - En base a los datos antes mencionados y a los criterios de la Figura 22 determina la existencia de riesgo cardiovascular. A menos que aparezcan signos o síntomas que indiquen la presencia de DM2 o HTA.

- En el procedimiento que se sigue durante la consulta la mayoría de los médicos familiares identifican la existencia del síndrome aplicando los criterios de la OMS o alguna otra fuente, sin embargo dichos criterios adolecen de no ayudar a determinar la evolución del padecimiento, por lo que generalmente se vuelven repetitivos en las consultas subsecuentes, sobre todo cuando el paciente no ha evolucionado satisfactoriamente. Cabe aclarar que las personas no asisten a consulta con el médico familiar por causas del síndrome metabólico, sin embargo, recientemente, las directrices de atacar la obesidad han dado cierta relevancia a tratar de detectar los problemas íntimamente relacionados con el síndrome, además de que se ha incrementado el número de personas que asisten por colesterol alto. Dentro del procedimiento existente cuando se hizo el análisis ya habían intenciones de hacer cambios significativos integrando las acciones de la enfermera de PREVENIMSS en cada consultorio para una mejor detección de este tipo de problemática (entre otras).

- En el caso del médico especialista:
  - El paciente es dirigido al especialista desde alguna otra instancia
  - El médico especialista igualmente puede establecer a simple vista si el paciente sufre de obesidad o cierto sobrepeso, lo que ya marca un riesgo.
  - De acuerdo a los procedimientos normalmente establecidos si el paciente presenta el síndrome metabólico es muy factible que el médico familiar haya derivado al paciente con el especialista por una enfermedad relacionada con el mismo, en tal caso mucha de la información en que se basó el médico familiar es pasada integra al médico especialista.

- El médico especialista por su parte vuelve a hacer una valoración del riesgo, en este caso más a fondo, debido principalmente a la mayor duración de la consulta. Es posible que producto de esta valoración se determine la existencia del riesgo, en tal caso los procedimientos establecidos permiten trabajar en acciones conjunto entre el médico familiar y el especialista. Lo que se detectó en el análisis es que en la mayoría de los casos en los cuales se toman acciones contra el síndrome metabólico por estas vías ya hay un lapso grande durante se ha dado un fuerte deterioro de la situación del paciente con relación al síndrome metabólico.
- De acuerdo a los procedimientos del IMSS es la norma mexicana sobre la obesidad la base principal de información para la toma de decisión sobre la existencia del riesgo.
- De acuerdo a los procedimientos que rigen la derivación de un paciente con un especialista, éste le atiende solo durante el tiempo que la causa que hizo que el paciente se direccionara a él subsista, por lo que el procedimiento prevé el regreso con el médico familiar quien dará seguimiento a su evolución.

- De acuerdo al riesgo establecido por el médico correspondiente hay 3 posibles acciones a tomar
  - Atender el padecimiento, lo cual lo lleva a cabo el médico correspondiente. En estos casos las enfermedades relacionadas con síndrome metabólico son: hipertensión arterial, dislipidemias (problemas de grasa en el torrente sanguíneo), problemas cardiovasculares, problemas de glucosa en el torrente sanguíneo (principalmente DM2), problemas obesidad y sobrepeso. En tal caso el riesgo de padecer el síndrome metabólico es un hecho.
  - Canalizar al paciente a trabajo social, donde la herramienta que se utiliza es el programa SODHI. Se trata de un esfuerzo conjunto

médico-trabajo social cuyo objetivo principal es cambiar los hábitos alimenticios y evitar el sedentarismo, brindando orientación nutricional, actividad física, educación para la salud, y actividades educativas. En este caso la intensión es reducir el riesgo.

  - Canalizar al paciente con el nutriólogo, su objetivo principal es modificar factores de riesgo y mejorar la actitud alimenticia de las personas, mediante la aplicación de esquemas nutricionales que permitan un cambio en los hábitos alimenticios. Generalmente esto sucede cuando hay un grado de obesidad que representa un alto riesgo cardiovascular.

- De la atención a los pacientes, tanto del médico familiar como del médico especialista se generan datos.
  - En el caso de los datos generados en medicina familiar, estos son recolectados e integrados a una base de datos, para ser procesados por el sistema SIAS que presenta información al personal del IMSS sobre diversos temas; cabe aclarar que no hay información específica sobre el síndrome metabólico, pero hay mucha información relacionada a él.
  - En el caso de los datos generados con los médicos especialistas, estos son recolectados por el sistema SIMO, el cual también presenta y complementa la información producida por el SIAS. En este caso tampoco se produce información específica sobre el síndrome metabólico, pero al igual que en el SIAS hay mucha información relacionada con él.

IV.6 Modelado del caso de estudio.

La ingeniería de procesos es una disciplina que consiste en una colección de técnicas para el análisis, diseño, y evolución de los procesos[11] basados en el uso del modelado[12] de procesos.

Como los procesos son realizados por varias personas, la ingeniería de procesos ayuda a estudiar las necesidades y preocupaciones de esas personas, la cooperación adecuada para llevar a cabo dichos procesos, y a partir de esto, brinda soporte al trabajo cooperativo entre individuos y equipos (Pacheco Soto, 2004).

Muchos de los trabajos previos de extracción de reglas en medicina se centran en el aspecto duro del proceso, esto es principalmente buscando en los datos (aspecto técnico). La presente tesis se apoya en la ingeniería de procesos para incluir el aspecto social del proceso de toma de decisiones en el manejo del síndrome metabólico, en otras palabras toma en cuenta la forma en que un médico enfrenta el problema de tomar decisiones relacionadas con el síndrome metabólico y su comorbilidad, al momento de generar y evaluar las reglas descubiertas.

Dentro del modelado de procesos es posible modelar diferentes aspectos de una organización:

- Funcional – las actividades que se desarrollan
- Comportamiento – cuando se llevan a cabo las actividades
- Organizacional – quienes son responsables de realizarlas
- Informacional (relevante para la regla) – información que se maneja para lograr lo anterior, entre ellas.

[11] Un proceso es un conjunto de roles que colaboran y llevan a cabo actividades parcialmente ordenadas, con el fin de alcanzar algunas metas comunes

[12] Modelo es una representación abstracta de la realidad revela lo que la persona cree que es importante para el entendimiento del mismo modelo.

Se basa en diferentes técnicas diagramáticas para lograr lo anterior, entre ellas, las gráficas ricas, los diagramas Integration Definition for Function Modeling IDEF0, y diagramas RAD.

Las gráficas ricas nos representan el proceso en forma caricaturesca, resaltando actividades o subprocesos que se realizan por medio de nubes (Monk, et al., 1998), la Figura 23 es la gráfica rica de atención a pacientes con síndrome metabólico: a éstas entran y de ellas salen flechas indicando los artefactos (información) que se requiere (entradas) y aquellos que se generan (salidas). También indican por medio de íconos, los actores (personas, sistemas o dispositivos) encargados de realizar las actividades, uniendo estos íconos a las nubes de las actividades por medio de líneas sin flechas; cuando más de una actor participa en la realización de la actividad estos quedan unidos a la misma.
Se pueden apreciar 4 grandes bloques:

- Agendado de la consulta del paciente (donde se da el primer intento de identificar el riesgo, sobre todo cardiovascular) y atención del paciente en medicina familiar, el médico determina el padecimiento y tratamiento. Como parte de un procedimiento ya establecido el médico siempre evalúa, aunque sea visualmente, el posible riesgo cardiovascular y los problemas de obesidad, así como antecedentes familiares de diabetes.
- Control y seguimiento del paciente diabético e hipertenso, donde la toma de decisiones se da principalmente en la evolución del padecimiento
- Apoyo integral al paciente donde se integra el programa SODHI y PREVENIMSS buscando modificar la dinámica de vida hacia tener una vida sana
- Seguimiento y coordinación entre el médico familiar y el médico especialista, coordinando las acciones cuando ya se ha presentado la comorbilidad del síndrome.

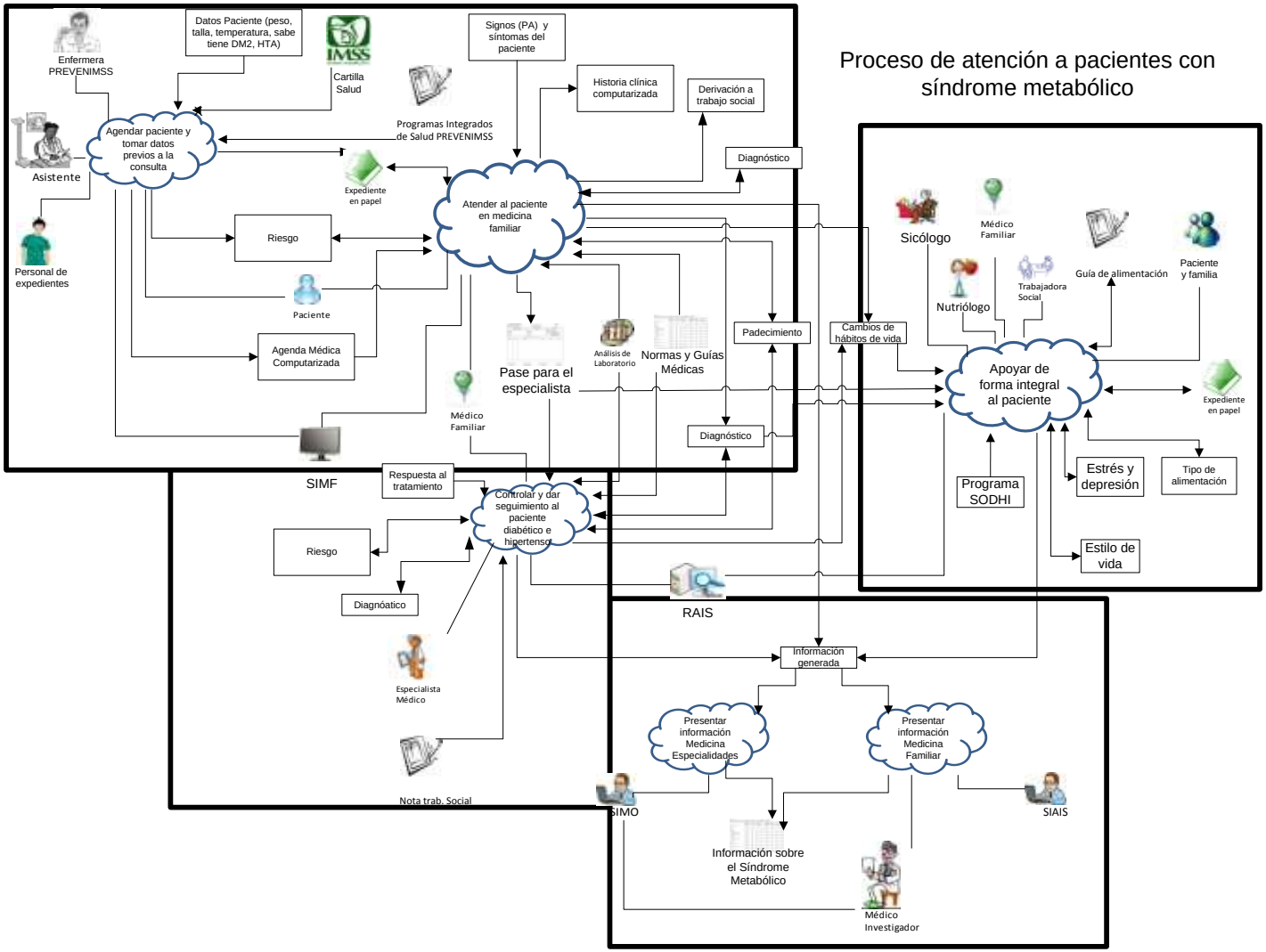


Figura 23. Gráfica rica de manejo del síndrome metabólico en medicina familiar.

En el apéndice F se adjuntan el resto de los diagramas producto del análisis realizado.

La metodología establece que hay que definir primeramente el problema de toma de decisiones, ya que si no existe problema, por ejemplo si las decisiones a tomar con repetitivas y hay incertidumbre muy baja, no tendría sentido brindar apoyo. El elemento primordial para el modelado de la toma de decisiones son los diagramas de influencias, tomar una decisión se define como hacer una estimación respecto a lo que se debería de hacer en cierta situación después de haber obtenido información relevante y deliberando en algunos cursos alternativos de acción; los diagramas de influencia son una representación gráfica del proceso de toma de decisiones. La Figura 24 presenta el diagrama de influencias general del proceso de toma de decisiones sobre el síndrome metabólico. Como se mencionó en el Capítulo II, los elementos que conforman estos diagramas son: las decisiones, representadas por rectángulos de esquinas rectas; las incertidumbres, representadas por los círculos; los resultados finales, representados por los rectángulos con esquinas redondeadas; y las influencias; representadas por las flechas.

En este diagrama se puede apreciar claramente cada uno de los elementos descritos anteriormente, se observa como en los resultados finales están las enfermedades para las cuales el médico cuenta con una guía clínica; cómo las incertidumbres que maneja el médico tienen que ver con casos en los cuáles no hay signos, señales o síntomas evidentes de la enfermedad subyacente relacionada con el síndrome metabólico. Sin embargo el manejo de este tipo de pacientes se basa en valorar el riesgo de padecer el síndrome y prevenir el desarrollo de su comorbilidad. Sin embargo no se usa la diagnosis del síndrome metabólico ya no hay tratamiento. Sin embargo se le toma como un elemento importante de prevención reconocido mundialmente en la prevención de las enfermedades crónico degenerativas (Amy, 2007).

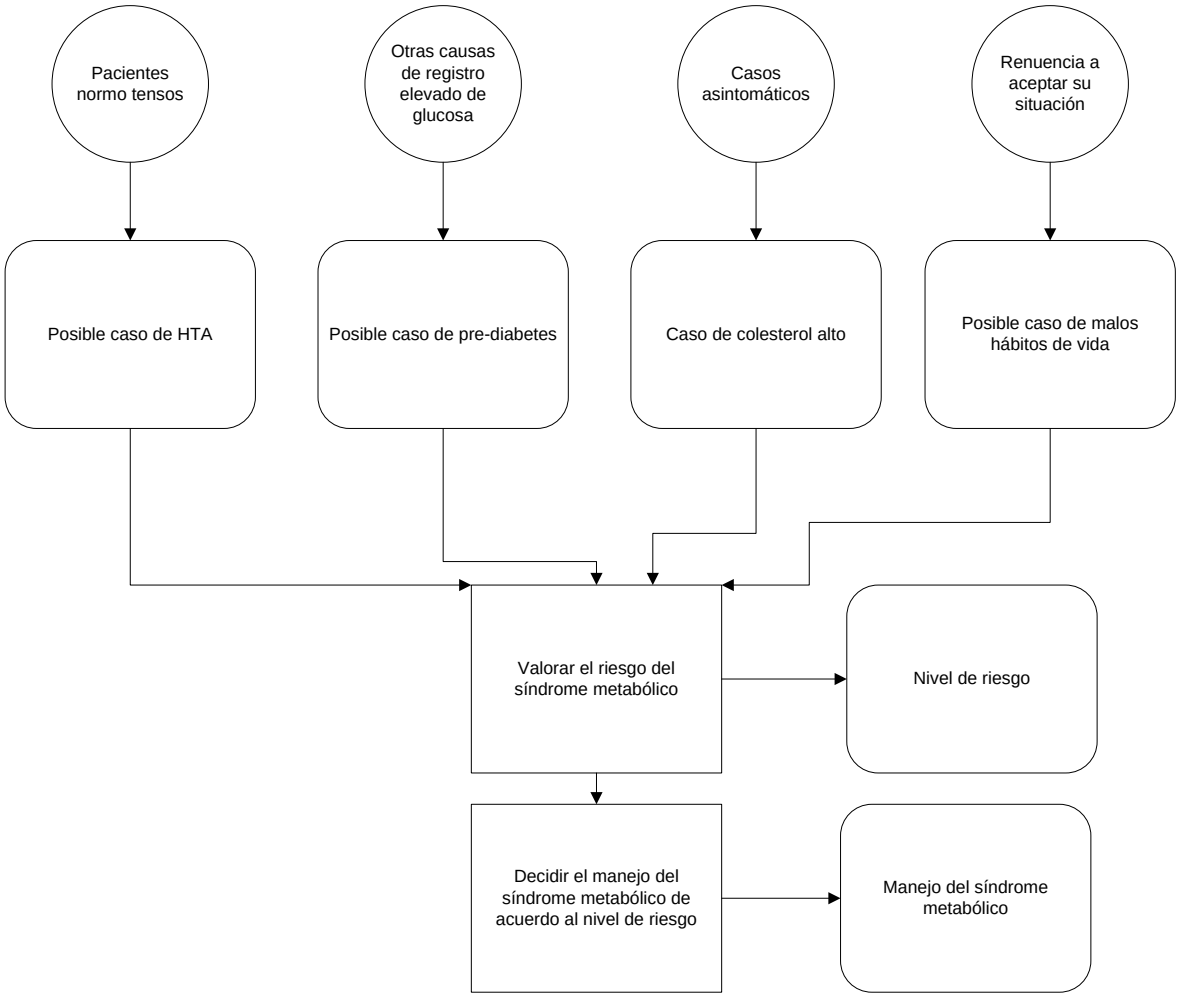


Figura 24. Diagrama de influencias del manejo del síndrome metabólico del médico familiar.

En la Figura 25 se describen 4 grandes bloques de preocupaciones:

1. Al entrevistar por primera vez pacientes que refieren sentirse muy mal y que posiblemente están relacionados con la comorbilidad del síndrome, principalmente HTA, sin señales claras.
2. 1Casos de pre diabetes.
3. Casos asintomáticos con posibles problemas relacionados con dislipidemias

4. Pacientes que son renuentes al tratamiento indicado

Estos bloques representan las principales incertidumbre que hay que apoyar para resolver el problema de toma de decisiones durante el manejo de los pacientes con síndrome metabólico.

Cada uno de los actores, médico familiar, asistente médico familiar, trabajador social, nutriólogo, sicólogo, enfermera PREVENIMSS, que aparecen en el modelado de los subprocesos de toma de decisiones del manejo del síndrome metabólico por medio de las gráficas ricas con preocupaciones y los diagramas de influencia que se detallan a continuación.

Asistente médico:
En la Figura 24 se presentan las incertidumbres que tiene la asistente médico sobre todo en su rol de la coordinación entre el médico, paciente y los demás servicios:

- Falta de indicaciones del médico sobre el nivel de riesgo que tiene el paciente *"somos nosotras quienes estamos pendientes de si un paciente viene o no, si está siguiendo algunas de las indicaciones que les da el doctor y si hay que enviarlos a otros servicios, me gustaría tener más información porque hay veces que es muy poco lo que me indica el doctor"*

  La falta de indicaciones por parte del médico al tener que tomar la decisión de derivar al paciente a alguno de los servicios de apoyo. *"Eso nos toca a nosotras incluso hay ocasiones que sin indicaciones del doctor nosotras vemos que el pacientito está muy mal y lo enviamos a trabajo social o con la enfermera de PREVENIMSS, me gustaría que hubiese más control sobre cómo se debe manejar al paciente".*

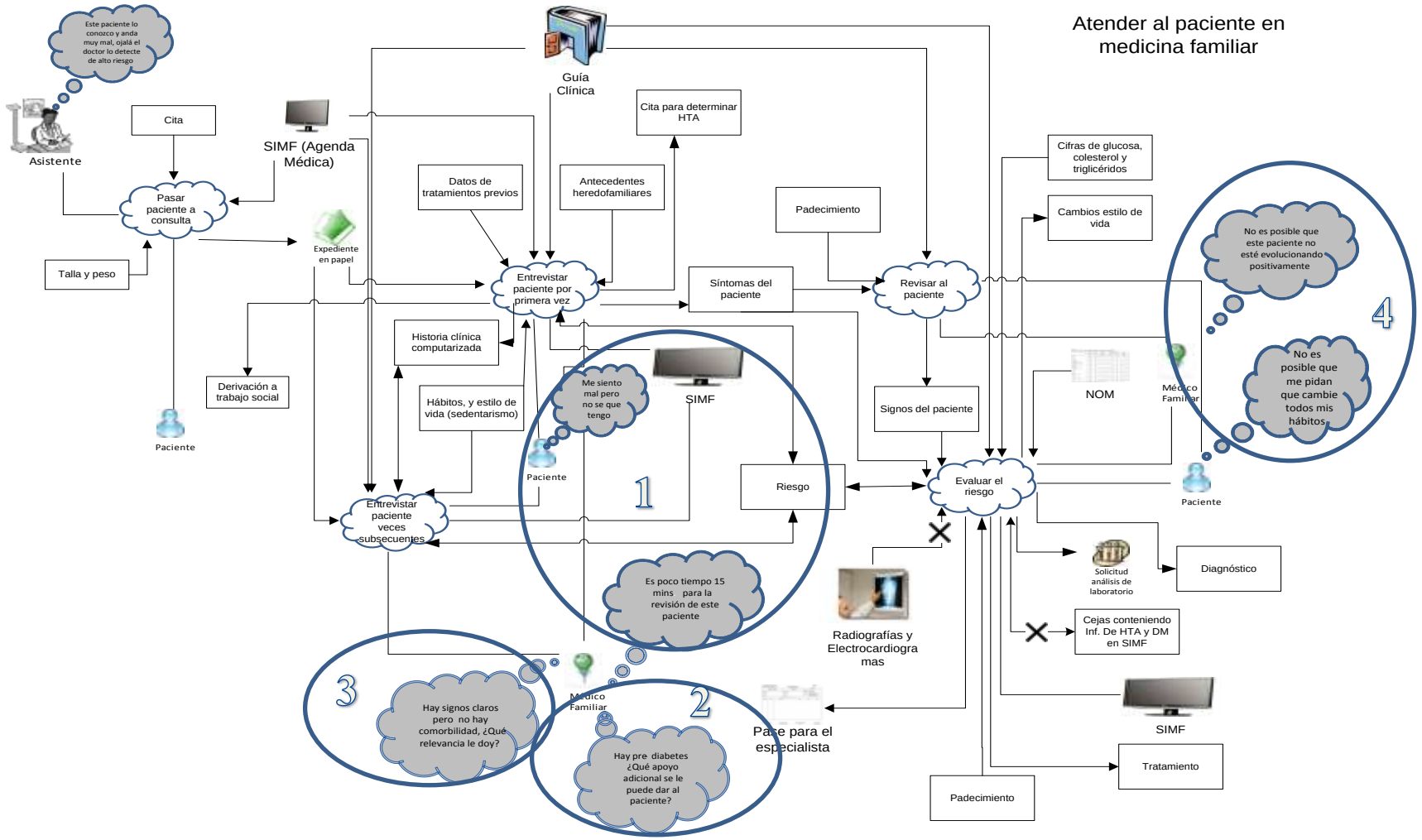


Figura 25. Gráfica rica con preocupaciones del manejo del paciente con síndrome metabólico.

- En opinión de la enfermera de PREVENIMSS el problema radica en que muchas veces el paciente no va y no sabe por qué. Cabe aclarar que en el proceso establecido actualmente se está cambiando para integrar una enfermera PREVENIMSS a cada consultorio.
    - Una sola persona atiende a todos los consultorios *"es mucho el trabajo porque vienen de todos los consultorios, aunque actualmente principalmente se atienden embarazadas y diabéticos, pudiese ser muy útil nuestra participación para detectar este tipo de problemas (el síndrome metabólico) pero no me daría abasto.*
    - La falta de seguimiento, *"nosotras somos enfermeras, no asistentes, aquí aparte de tomarles los signos hacemos la primera evaluación del riesgo, las personas vienen, se les toman los datos, se les envía al consultorio y dejan de venir, no hay seguimiento"*
    - Control del paciente que deja de asistir. *"así como está implementado, son pocos los que vienen por esa razón (el síndrome metabólico), me gustaría que hubiera más control del consultorio ya que prácticamente lo pacientes vienen solitos"*

Por otro lado se puede observar en la Figura 26 como cuando ya se ha declarado la comorbilidad del síndrome la situación cambia radicalmente y las preocupaciones se relacionan con el control del paciente.

En la Figura 26 se puede observar como el médico tiene un nivel alto de incertidumbre al lidiar con:

- pacientes asintomáticos pero que tienen alta probabilidad de padecer el síndrome metabólico. *"Se les manda a hacer estudios cuando el paciente no responde y no sabemos qué está pasando", "principalmente nos fijamos si el perímetro abdominal es grande, estaría bien contar con una*

clasificación que nos ayudara con el riesgo cuando el paciente todavía no presentan síntomas de enfermedad".

Otro elemento importante es el poco tiempo con el que cuenta para realizar la consulta, en ella tiene que revisar al paciente y los resultados de los análisis clínicos (que probablemente puedan estar incompletos) sin contar con alguna herramienta computacional específica que le facilite determinar rápidamente el nivel de riesgo relacionado con el síndrome metabólico. *"institucionalmente contamos con 15 minutos para la consulta" "en ese tiempo todo de ser rápido y hay que revisar muchas cosas, pues estaría bien que en el expediente electrónico hubiese algo que facilitara evaluar el riesgo rápidamente y en vez de repetir muchas cosas durante la entrevista pudiésemos destinar más tiempo a la diagnosis y el tratamiento.*

De análisis realizado con los diagramas de rol y actividad (RAD) del médico familiar en consulta, ver Figura 27, se observa:

- Las fuentes principales de información son los antecedentes del paciente (personales y familiares), signos identificados por simple inspección visual (tipo de obesidad), estudios de laboratorio (de sangre principalmente), síntomas referidos por el paciente (mareos, disnea, dolor de cabeza, etc.), los datos del expediente electrónico (antropométricos, relacionados con DM2, HTA, anotaciones, indicaciones del especialista, antecedentes heredofamiliares).
- Las incertidumbres: cuando es la primera vez que se le atiende y no cuenta con antecedentes, si se le debe enviar al especialista, y si ya se requieren más estudios de laboratorio, si el paciente está respondiendo al tratamiento, como está evolucionando el paciente antes de que aparezcan las

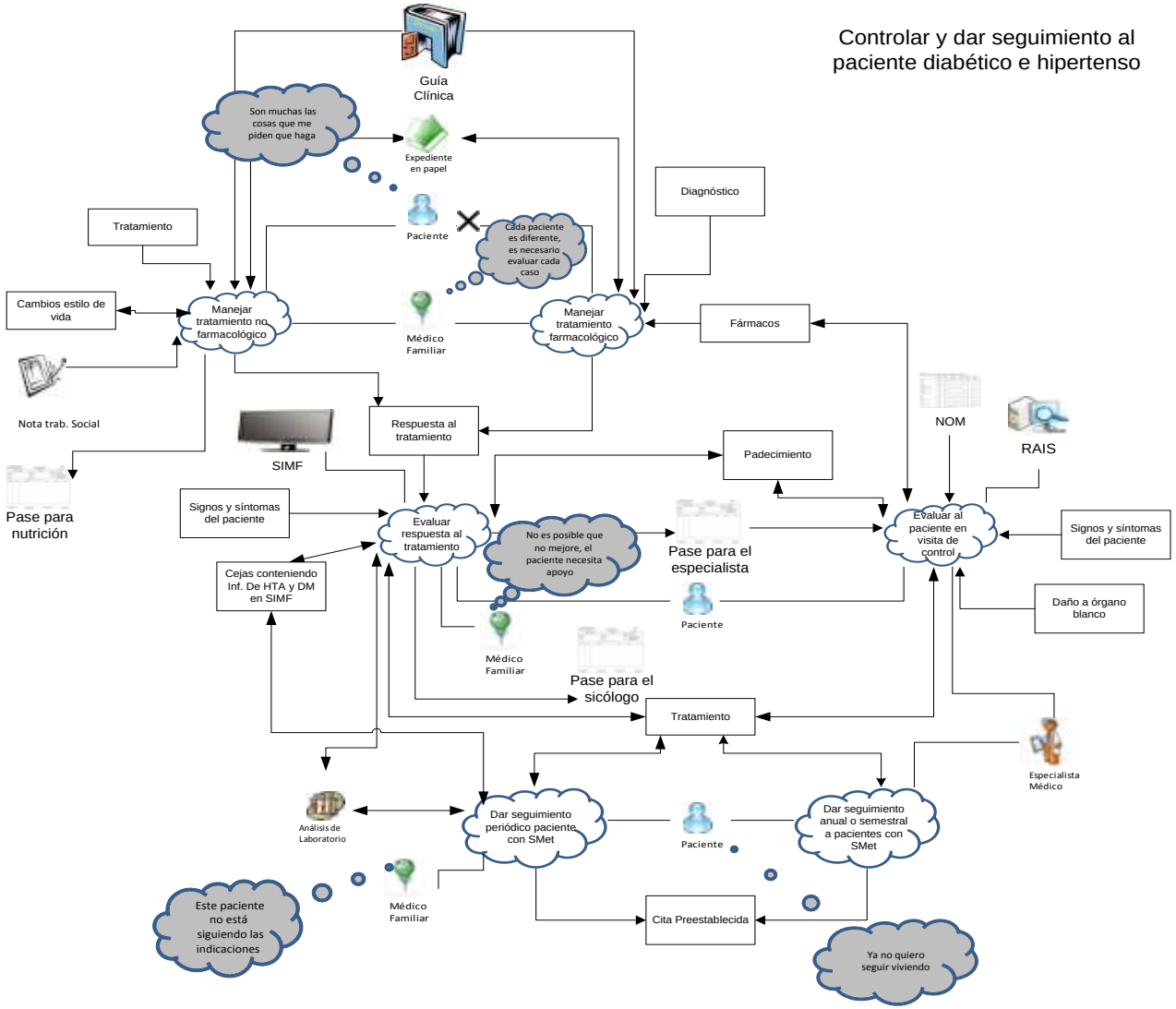


Figura 26. Gráfica rica con preocupaciones del control y seguimiento que se hace del paciente diabético e hipertenso.

enfermedades asociadas con el síndrome metabólico (comorbilidad), sí el paciente requiere otro tipo de apoyos, identificar el riesgo que representa para el paciente el síndrome metabólico (sobre todo el cardiovascular).

- Los cursos de acción alternativos: cuando se tiene un paciente de primera vez, si se requiere un especialista, si se ha identificado la probabilidad de padecer síndrome metabólico antes de que aparezcan la comorbilidad, sí se detectan nuevos padecimientos durante la consulta, si se quiere enviar al paciente a otro tipo de apoyos (nutriólogo, sicólogo, trabajo social, enfermera de PREVENIMSS).
- Las acciones principales: revisar el expediente electrónico y en papel, revisar al paciente, registrar los datos antropométricos del paciente, dar seguimiento al paciente (llamarle, tener claro el nivel de riesgo que presenta el paciente.

En la Figura 28 se observan las incertidumbres que tiene el nutriólogo por:

- Las limitaciones económicas de los pacientes. "*Es muy difícil cambiar los hábitos alimenticios de los pacientes, porqué no solamente se trata de recomendarles se les dan dietas, pero se les hace más fácil seguir comprando la torta o las papitas, justificándose que no hay dinero, aunque estos productos ciertamente no son económicos, estaría bien que en la consulta se les haga conciencia de cómo las dietas y guías que se les dan aquí pueden salvar sus vidas y no se trata de si quiero o no".*

En la Figura 29 se muestran las incertidumbres en el manejo de sicología donde:
Es mínimo el número de pacientes que llegan por causas preventivas. *Aquí nos llegan principalmente los pacientes cuando ya tiene diabetes, ya que es muy difícil que te digan que nunca más vas a volver a comer algo, siendo que toda la vida has estado acostumbrado a comerlo, debe haber un manejo sicológico del*

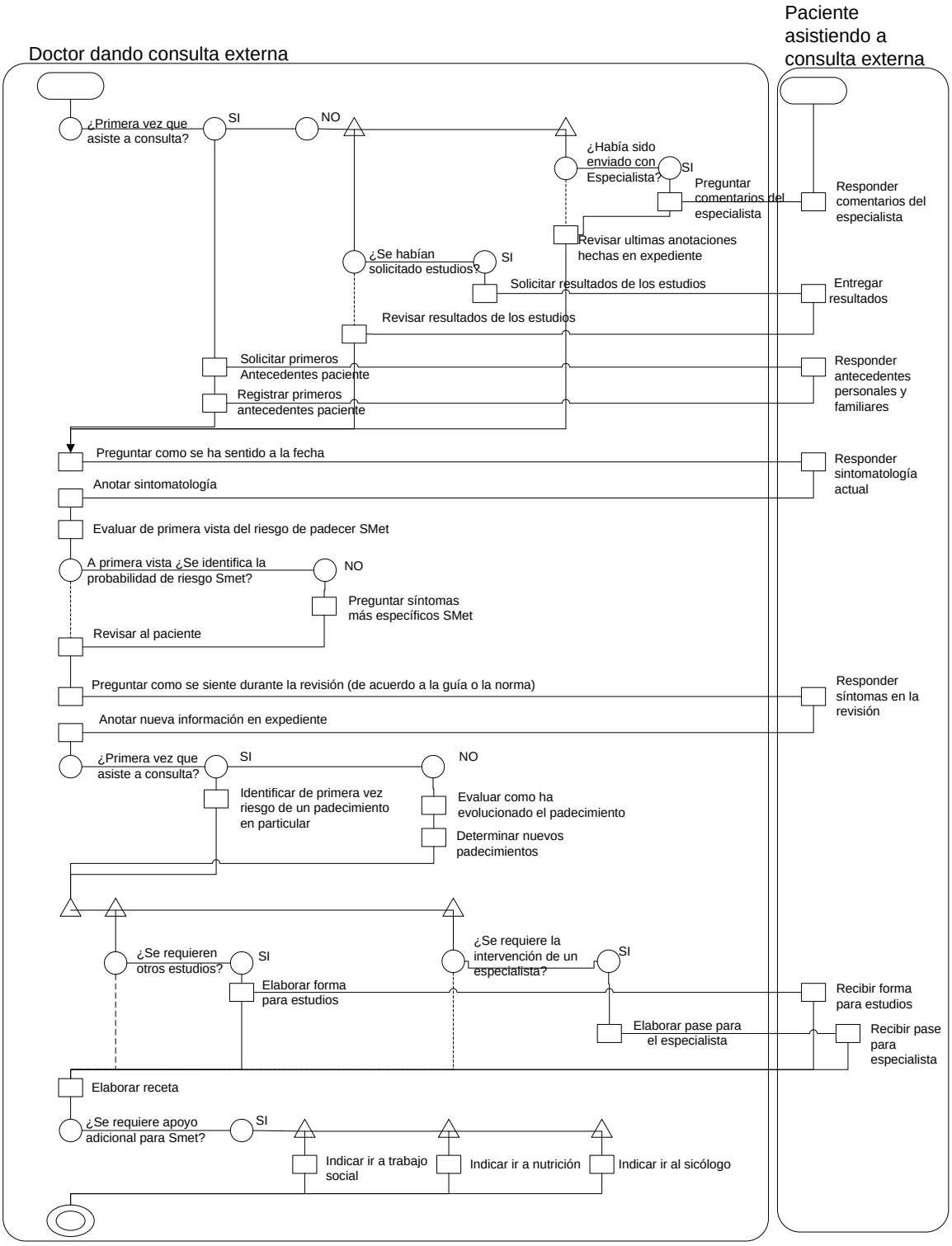


Figura 27. RAD del médico familiar dando consulta.

paciente en el mismo consultorio, no que haya un sicólogo, pero el médico debería manejar estos problemas de forma preventiva para que no lleguen a estos casos"

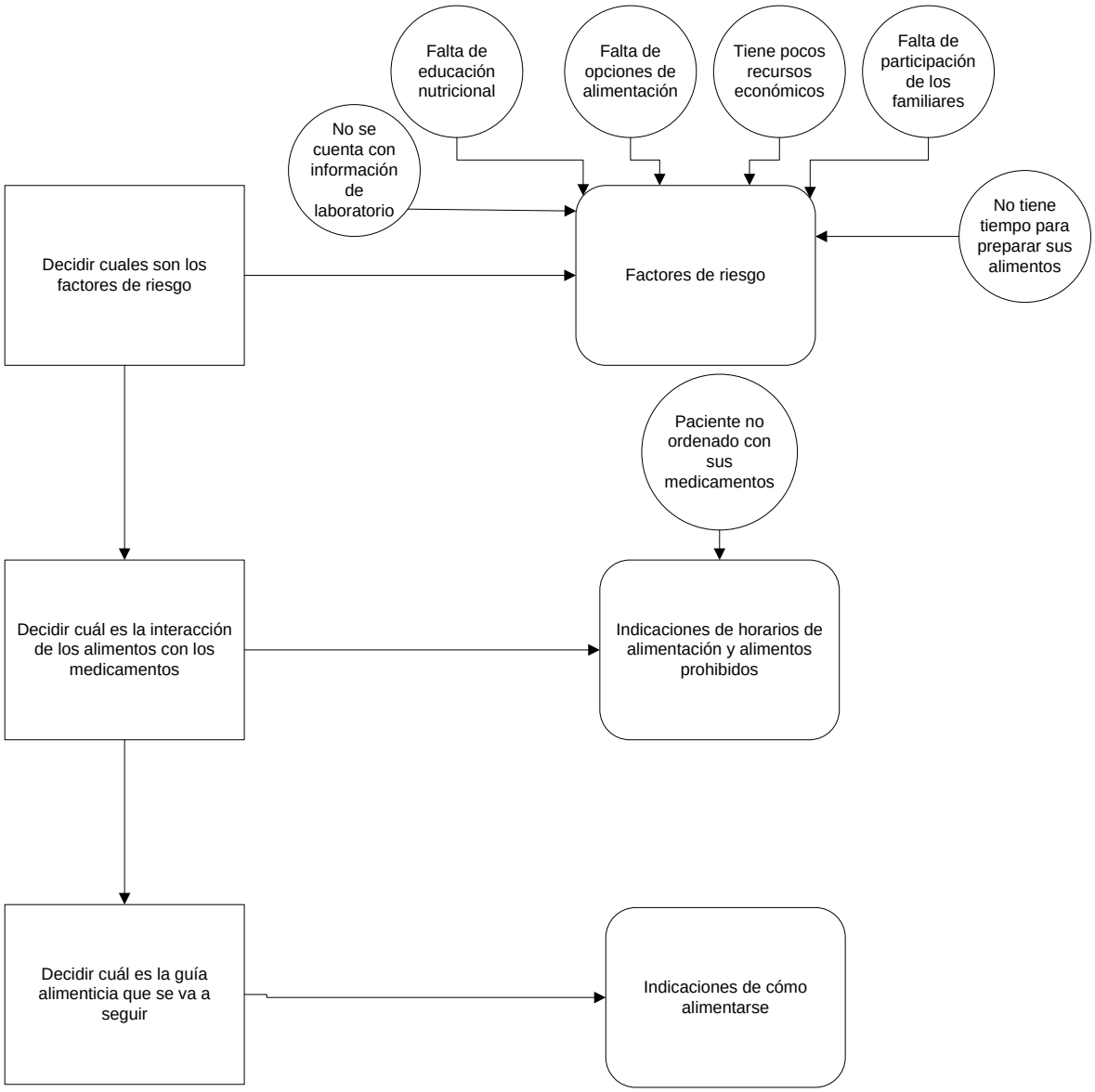


Figura 28. Diagrama de influencias del nutriólogo.

- Principalmente cuando hay complicaciones con otro tipo de padecimientos sicológicos. *"Aquí me llegan pacientes con muchos años de abandono, que ya no quieren vivir, si estaría bien que se apoyase al paciente en la consulta*

haciéndole ver que la situación se puede cambiar, pero sobre todo antes que el caso sea tan grave como te menciono".

Producto de este mismo análisis se encontró que se cuenta con bastante

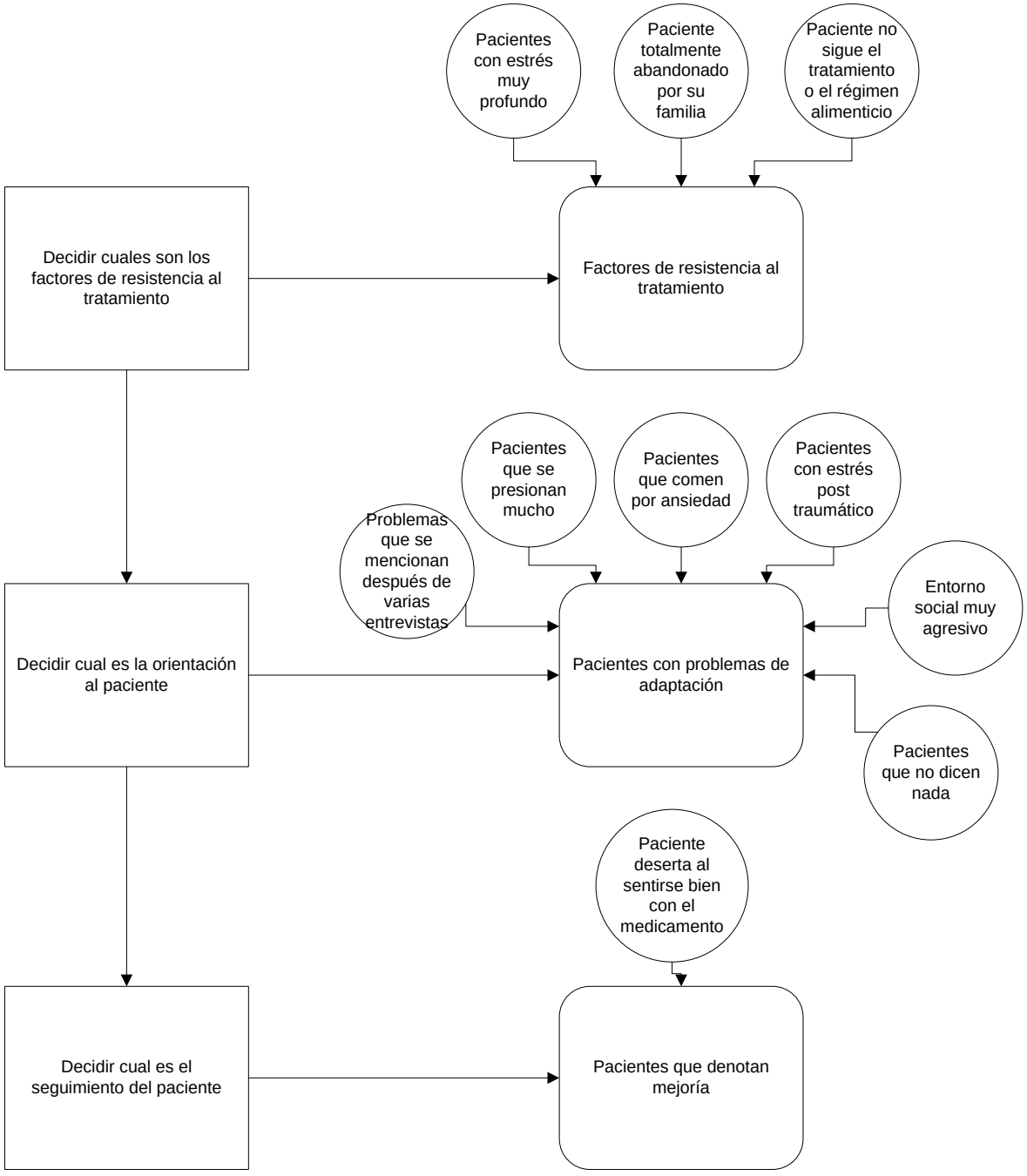


Figura 29. Diagrama de influencias sicólogo.

información para el manejo del síndrome metabólico ya que los sistemas de información del hospital han sido modificados recientemente para forzar a que siempre se registren los datos antropométricos de cada paciente y para proporcionar el IMC del paciente. Adicionalmente los análisis de sangre (biometría hemática) son bastante comunes y quedan registrados en los mismos sistemas, lo que hace que muchos de los datos necesarios para el manejo del síndrome metabólico se encuentren disponibles. En la siguiente sección se analizan cuáles son los datos relevantes y sus principales características.

IV.7 Información relevante en el manejo del síndrome metabólico.

Existen varias herramientas con las que cuenta un médico del IMSS para el manejo del síndrome metabólico, entre ellas las guías de práctica clínica para el diagnóstico y tratamiento de la HTA y DM2 en el primer nivel de atención. Adicionalmente la práctica médica se rige por las normas mexicanas tales como la norma oficial mexicana NOM-037-SSA2-2002 para la prevención y control de las dislipidemias, norma oficial mexicana NOM-015-SSA2-1994 para la prevención, tratamiento y control de la diabetes, y la norma oficial mexicana NOM-30-SSA2-1999 para la prevención, tratamiento y control de la hipertensión arterial.

Estas herramientas están basadas en hechos y evidencias, los cuales presentan recomendaciones en donde aparecen reglas que marcan límites para ciertos indicadores (estos se recolectaron y colocaron en el apéndice B).

Entre los datos más relevantes se encuentran:

- Los índices de glicemia, producto de pruebas clínicas de glucosa en ayuno y prueba de glucosa. Este tipo de estudios sirven para la determinación de prediabetes, diabetes y dislipidemias
- Pruebas de intolerancia a la glucosa, se trata de estudios más específicos que se aplican cuando se ha detectado un riesgo de padecer diabetes
- Análisis de biometría hemática, donde se determinan datos como colesterol total, triglicéridos, niveles de colesterol de baja densidad y de alta densidad
- Tomas de la tensión arterial, de forma rutinaria se hacen mediciones de la presión arterial a los pacientes al acudir a consulta en atención de primer nivel (estas mediciones son puntuales), en el caso de detectar riesgo cardiovascular se deben hacer al menos 2 mediciones en un intervalo semanal, cuando menos, para determinar hipertensión arterial.
- Medición del Índice de masa corporal, de forma rutinaria al registrar los datos de peso y talla en el expediente electrónico el sistema calcula y registra el índice de masa corporal.
- En pacientes hipertensos se practican análisis de fórmula roja, creatina sérica, glucosa sérica y electrolitos. Donde se registran colesterol total, HDL, triglicéridos y ácido úrico, nivel sérico de sodio
- En la entrevista médica otros datos relevantes que se recolectan son el hábito de fumar y beber. Los antecedentes genético hereditarios, o de pacientes que padecen o padecieron enfermedades relacionadas con el síndrome metabólico.
- Otros datos que son básicos es el género y la edad.

Hay una clara relación entre el síndrome metabólico y el desarrollo de las enfermedades crónico degenerativas, y se le reconoce un valor predictivo importante para la detección temprana de su comorbilidad (Deen, 2004). Muchos de los estudios de laboratorio y la información antropométrica tienen una obvia relación con el síndrome. En todo caso el proceso de descubrimiento de conocimiento que se aplicó (KDD) no se puede centrar en un solo dato por el

contrario la información debe ser multidimensional, donde se tomen edad, sexo, colesterol total, colesterol de baja densidad, nivel de azúcar en la sangre, etc. y debe ser el propio proceso el que descubra la forma en que los elementos citados se relacionan para determinar el nivel en que un paciente puede estar afectado por el síndrome metabólico.

Por otro lado, la información registrada se encuentra dentro de la memoria organizacional, mucha de ella está dispersa en diferentes bases de datos y para obtenerle se debe llevar a cabo un proceso de recolección, e incluso algunas de las fuentes, como lo es el almacén de datos del IMSS, son de uso restringido y solo se puede llegar a ella mediante los resúmenes que se le presentan a los médicos, los cuales, para efectos prácticos, son de poca o nula utilidad en apoyo a la tarea a realizar.

Sin embargo, en control epidemiológico se realizan periódicamente estudios de campo donde se hacen varios de los exámenes de laboratorio citados, generando una serie de datos a los que se tiene libre acceso, referidos a entidades bien identificadas (generalmente a una empresa), con perfiles socioeconómicos bien definidos y fácilmente ubicables en el tiempo.

El primer paso del proceso de descubrimiento de conocimiento es la recolección de los datos; en este caso la fuente principal de datos fueron los recolectados por el área de epidemiología. Donde primeramente se nos proporcionó una muestra proveniente de una empresa (CFE), los atributos que contiene se enlistan en la **Tabla III**, se les divide en 3 subgrupos.

Tabla III. Registro de datos de los archivos de la CFE.

Información original de la CFE.	
Campo	Descripción
Edad	Edad del paciente
Sexo	Género del paciente
Peso	Peso del paciente en kilogramos
Talla	Talla del paciente en metros
IMC	Índice de masa corporal $\frac{peso}{talla^2}\ \frac{kg}{m^2}$
Clase de Obesidad	Clasificación del índice de masa corporal por tipo de obesidad
Cintura	Cintura en metros
Riesgo cardiovascular 1	Basado únicamente en los indicadores enlistados anteriormente
Tensión arterial sistólica	Medición puntual de la tensión arterial sistólica del paciente
Tensión arterial diastólica	Medición puntual de la tensión arterial diastólica del paciente
Clasificación de la tensión arterial	Clasificación de la tensión arterial sistémica del paciente basada en las mediciones anteriores
Sobre riesgo	Sobre riesgo en condiciones cardiovasculares.
Td	Dato no identificable
Tabaquismo	Si la persona tiene el hábito de fumar
alcoholismo	Si la persona tiene el hábito de beber
DM2	Si la persona padece diabetes Mellitus tipo 2

Información original de la CFE (continúa)	
Campo	Descripción
Has	Si la persona padece hipertensión arterial
Glicemia	Medición puntual del nivel de glicemia prueba de glicemia en ayunas del paciente
Colesterol	Medición puntual del nivel colesterol total
Triglicéridos	Medición puntual del nivel de triglicéridos del paciente
Vldl	Medición puntual del nivel colesterol de alta densidad del paciente
Recv	Medición del riesgo cardiovascular tomando en cuenta el resto de los indicadores.

1. Generales: edad, sexo, nombre (el cual se omitió desde el principio por confidencialidad), si tiene los hábitos de beber o fumar.
2. Antropométricos: peso, talla, cintura; y lo que de esta información se desprende como el índice de masa corporal IMC y la clase de obesidad.
3. Información de laboratorio: tensión arterial sistólica, tensión arterial diastólica, si padece DM2, si padece de HTA, niveles de glicemia, niveles de colesterol total, triglicéridos, colesterol de alta densidad; y la información que de ella se desprende como riesgo cardiovascular antes de análisis clínicos, riesgo cardiovascular después de los análisis clínicos, sobre riesgo cardiovascular.

Estos son datos puros, sin ninguna modificación por parte del analista, y corresponden a los registros obtenidos en los estudios de campo en dicha empresa. En estas visitas se toman las medidas antropométricas y se hace la

recolección de muestras de sangre, principalmente para realizar análisis de química sanguínea.

El siguiente paso del KDD después de la recolección, es la reducción y proyección de los datos. En este paso se aplica un pre-procesamiento que consiste de varias técnicas:

- Limpieza de datos
- Integración de datos
- Transformación de datos

El objetivo es obtener un conjunto de datos libres de errores, normalizados en las escalas que manejan sus valores, habiendo eliminado datos redundantes o repetidos, eliminando registros con gran cantidad de valores erróneos o perdidos, etc. En otras palabras de tener un conjunto de datos listos para ser minados (atributos) al cual se le denomina vista minable, y que permitirán que los algoritmos de minería de datos puedan trabajar correctamente.

En el siguiente Capítulo se analizarán a fondo los siguientes procesos a los que se somete a la información, de acuerdo al proceso del KDD el análisis y la exploración, la minería de datos, y la evaluación de los resultados.

IV.8 Resumen

En este Capítulo se ha analizado el contexto del problema del manejo del síndrome metabólico desde dos perspectivas diferentes: desde el punto de vista de la medicina y el de la ingeniería de procesos.

Se ha establecido la relación del síndrome con las enfermedades crónico degenerativas. Se ha descrito al síndrome metabólico con un problema multifactorial y se han esquematizado la interrelación de dichos factores. Finalmente se ha elaborado un definición amplia del síndrome metabólico.

Desde el punto de vista de la ingeniería de procesos, se ha elaborado una descripción textual del proceso que se sigue en el manejo del síndrome metabólico en la atención médica de primer nivel. Se presentaron una serie de diagramas donde se pueden identificar claramente el proceso completo de manejo del síndrome metabólico y finalmente se ha establecido cuales son las decisiones y cuáles son las incertidumbres que se tienen durante el manejo del síndrome.
Finalmente se han planteado a muy grandes rasgos cuáles son las fuentes de información con que se cuentan y con cuáles datos se contaron para el análisis que se realizó, así como, partiendo de los datos obtenidos se realiza un pre-procesamiento de la información para generar una vista minable con la cual se va a trabajar.

Hasta este momento se han definido los diferentes algoritmos para la extracción de conocimiento y el contexto de la información. El siguiente paso es, el procesamiento preliminar que se debe hacer de la información con que se cuenta, la aplicación de los algoritmos y la obtención de la correspondiente clasificación de riesgo del síndrome metabólico, lo cual se presenta en el siguiente Capítulo.

Capítulo V

Proceso de clasificación del síndrome metabólico

V.1 Introducción

En la fase 5 de la metodología se establece el análisis del modelo de decisión adecuado, lo que primeramente implica la conformación del conjunto de datos listos para la fase de minería de datos (vista minable). Los datos puros reciben un preprocesamiento que permite limpiarlos y resolver algunos problemas relacionados con la naturaleza de los datos y las escalas en las que están expresados sus valores.

Una vez lista la vista minable se presenta el proceso de definir los parámetros de la red neuronal de Kohonen.

Al final se presenta las reglas de clasificación obtenidas bajo dos métodos diferentes, C4.5 y programación genética, la interpretación que se hizo de los resultados, y el material de apoyo adicional que se generó para acompañar la clasificación.

El IMSS cuenta con un almacén de datos, el cual se compone por información derivada de diferentes áreas, entre ellas la de control epidemiológico. El problema es que el acceso a dicho almacén es restringido y no se tuvo acceso a él.

No obstante si se pudo acceder a la información que genera control epidemiológico del hospital número 8 del IMSS.

Las guías clínicas para la prevención y el tratamiento de la DM2 y la HTA proporcionan gran cantidad de indicadores sobre la detección y el tratamiento de estas enfermedades, de ellas se extrajo un diccionario de atributos compuesto de:

el atributo, la fuente de donde se obtiene, y las notas que contiene la guía clínica en relación a su manejo (el detalle se presenta en el Apéndice B). El problema con esta información es que muchos de los atributos se obtienen en pruebas de laboratorio muy específicas que no son de uso común y generalmente no están disponibles para todos los pacientes.
Entonces el primer paso es reducir este universo de datos acotándolo a solo aquellos datos que se obtienen de los exámenes clínicos rutinarios y las revisiones antropométricas que se hacen previas a la consulta.

El área de control epidemiológico constantemente realiza trabajos de campo con el fin de establecer control sobre la situación que guarda la población con respecto al riesgo de padecer cierta enfermedad. Particularmente aquellos estudios que se realizan para determinar el riesgo de padecer algún tipo de evento cardiovascular cerebral (EVC) contienen información muy relacionada con el síndrome metabólico, en la

Tabla IV presentamos la información obtenida en la Comisión Federal de Electricidad (CFE). Este tipo de estudios tienen características muy deseables para el tipo de análisis al cual se les quiere someter en esta tesis:

1. Son perfectamente identificables, esto es: se sabe a qué grupo pertenecen, bajo qué condiciones se realizaron y en que fechas
2. Al integrar varias muestras generalmente las consideraciones que se deben tomar para el manejo de la información son fácilmente homologables en los diferentes grupos de datos.
3. Son rastreables en el tiempo, generalmente las muestras pertenecen a empresas en las cuales periódicamente se realizan este tipo de pruebas.
4. Al comparar el número de atributos de la
5.
6. Tabla IV con la gran cantidad de atributos que aparecen en el Apéndice B se observa que dichos atributos forman parte del conjunto definido en el

apéndice, pero son muchos menos, lo que favorablemente reduce el espacio de datos.

En la siguiente sección se analiza el proceso de la conformación de la vista minable. Los datos recolectados pueden presentar diferentes problemas: desconocimiento de la naturaleza de los datos, la forma en que están codificados, las escalas pueden ser incompatibles de una fuente a otra, pueden existir valores omitidos, valores erróneos, existir inconsistencias entre las diferentes fuentes, datos repetidos, etc.

En la Sección V.3 se analiza el proceso de pre procesamiento de la información. Un primer paso en este pre procesamiento es corregir las inconsistencias entre fuentes y hacer ajustes preparatorios en los datos, antes del proceso de minería, buscando quitar inconsistencias, de tal forma que los datos después de este pre procesamiento conformen la vista minable.

Contando con la vista minable se procede a resolver el problema de clasificación. La vista minable contiene registros que implica un paciente con ciertas características o atributos. En términos matemáticos, cada uno de estos registros es un vector de información multivariado, y en su conjunto conforman el espacio de datos del problema.

En el problema de clasificación que se quiere resolver a través de la minería de datos es encontrar grupos o conglomerado de pacientes con características que les hacen similares, en base a los datos de la vista minable y algún criterio de similitud establecido por el algoritmo a trabajar.

En la Sección V.4 se presenta el proceso de selección de los valores de parámetros para la red neuronal de Kohonen. La selección de estos parámetros representa un proceso laborioso del cual depende completamente el éxito del

proceso de descubrimiento de conocimiento, y se basa en ciertos criterios pre establecidos en la literatura.

En la Sección V.5 ya habiendo ubicado las condiciones adecuadas se procede a realizar la clasificación de toda la muestra, mediante la red neuronal de Kohonen.

En la Sección V.6 se aplica el algoritmo C4.5 para generar las reglas de clasificación que permitan determinar la pertenencia a cada uno de los subgrupos obtenidos.

En la Sección V.7 se describe el proceso de refinamiento de las reglas mediante la aplicación de la programación genética, cuyo objetivo principal es obtener un segundo conjunto de reglas más robustas y sencillas.

Finalmente en la Sección V.8 se presenta un resumen del Capítulo.

V.2 Conformación de la vista minable.

Adicionalmente a que los datos se deben recolectar, la conformación de la vista minable conlleva ciertos pasos preparatorios antes de poder hacer uso de la información.

Al momento de la recolección de la información se debe determinar el tamaño de la muestra y la diversidad que debe existir entre el tipo de individuos que la conforman. El tamaño es importante no solo por la significancia estadística que al final debe tener el resultado obtenido, sino también por la disponibilidad de datos con los cuales se cuenta para trabajar y por las características que se busca presenten dichos datos para que los algoritmos seleccionados funcionen correctamente. En la siguiente sección se detallan estas características.

Una vez obtenida la información primero se debe comprender a fondo el contexto médico de la información con la que se va a trabajar. Se deben revisar las escalas con las cuales se están manejando los valores de cada atributo, entender el significado médico de cada uno de ellos, las posibles consideraciones especiales que se tomaron al momento de la recolección de los datos, y finalmente detectar datos con problemas.

Generalmente, entendido el contexto de la información y asegurado que la información ya se encuentra en condiciones adecuadas, se tiene que llevar a cabo algún tipo de transformación en los datos de la vista minable. El tipo de transformaciones van desde la re expresión de las escalas de valores en las que viene la información hasta la creación de algún tipo de atributo intermedio para facilitar el procesamiento. En la sección V.3 se explica a detalle este proceso.

Lista la vista minable, se procede a correr los algoritmos seleccionados utilizando parte de los datos de prueba. Al inicio se debe tomar el tiempo suficiente para establecer los parámetros adecuados con los que corre el algoritmo, varios de los algoritmos seleccionados no son determinísticos, así que la validez de los resultados se determina a través de muchos experimentos guiados por condiciones que en la misma literatura dicta deben cumplirse para que el algoritmo trabaje correctamente. En las secciones V.4 a V.6 se presenta en forma detallada este proceso.

Establecidos los parámetros idóneos para cada algoritmo se procede a correr los algoritmos y a obtener los resultados, ya con toda la muestra de datos. Los resultados obtenidos en este proceso deben ser revisados y analizados en conjunto con el especialista epidemiólogo. Buscando validarles contra la experiencia previa del especialista y las fuentes de información existentes relacionadas con la clasificación que se está obteniendo.

Tabla IV. Información generada en control epidemiológico para el riesgo de padecer un EVC.

Información de control epidemiológico	
Campo	Descripción
Edad	Edad del paciente
Sexo	Género del paciente
Peso	Peso del paciente en kilogramos
Talla	Talla del paciente en metros
IMC	Índice de masa corporal $\frac{peso}{talla^2}$
Clase de Obesidad	Clasificación del índice de masa corporal por tipo de obesidad
Cintura	Cintura en metros
Riesgo cardiovascular 1	Basado únicamente en los indicadores enlistados anteriormente
Tensión arterial sistólica	Medición puntual de la tensión arterial sistólica del paciente
Tensión arterial diastólica	Medición puntual de la tensión arterial diastólica del paciente
Clasificación de la tensión arterial	Clasificación de la tensión arterial sistémica del paciente basada en las mediciones anteriores
Sobre riesgo	Sobre riesgo cardiovascular.
Td	Dato no identificable
Tabaquismo	Si la persona tiene el hábito de fumar
Alcoholismo	Si la persona tiene el hábito de beber
DM2	Si la persona padece diabetes Mellitus tipo 2
HTA	Si la persona padece hipertensión arterial

Información original de la CFE (continuación)	
Glicemia	Medición puntual del nivel de glicemia prueba de glicemia en ayunas del paciente
Colesterol	Medición puntual del nivel colesterol total
Triglicéridos	Medición puntual del nivel de triglicéridos del paciente
Vldl	Medición puntual del nivel colesterol de alta densidad del paciente
Recv	Medición del riesgo cardiovascular tomando en cuenta el resto de los indicadores.

V.2.1 Análisis del contexto médico de la información.

Toda la información presentada en la

Tabla **IV**, de acuerdo a la literatura, está claramente relacionada con el síndrome metabólico.

Lo primero que se observa al analizar la

Tabla **IV** es que muchos de los datos están relacionados entre sí:

- El índice de masa corporal (IMC) se obtiene mediante el peso y la talla, así que es un dato calculado, no recolectado.
- La clase de obesidad se obtiene mediante el IMC

- El riesgo cardiovascular 1 es una clave que indica la existencia de dicho riesgo basada en el análisis de edad, sexo, peso y talla. Hay otro atributo (Recv) que reexpresa este valor.
- Los datos sobre tensión arterial están muy relacionados entre sí, y son mediciones puntuales de estos signos tomadas en el estudio de campo realizado.
- El sobreriesgo es un dato que determina el epidemiólogo, pero que en este caso el epidemiólogo prefirió omitir.
- El dato TD no tiene uso y se decidió omitir.
- Los datos tabaquismo y alcoholismo son claves que indican el hábito de consumir alguno de estos productos en forma regular, de acuerdo a un criterio de consumo generalmente aceptado para este tipo de estudios.
- Los datos DM2, HTA y glicemia son indicadores de si el paciente presenta alguna de estas enfermedades.
- Colesterol, triglicéridos y VLDL son mediciones puntuales de estos signos tomadas en el estudio de campo realizado, todas ellas son mediciones de diferentes tipos de grasas que se encuentran en la sangre.
- Recv es una clave que remplaza al dato riesgo cardiovascular, solo que ahora tomando toda la información y no solo los datos que toma el primer dato. El problema con este atributo es que en algunos casos no se cuenta con la información de laboratorio o de toma de presión arterial, en tales casos se omite.

En concordancia con lo que establece Han (2006) los ajustes preparatorios que realizaron en los datos fueron:

- Se creó un atributo nuevo donde se complementó el manejo del atributo recv y el de riesgo cardiovascular, de tal forma que se tenga un solo atributo.
- El IMC ya que este dato se puede calcular en base a la talla y el peso.

- Se re expresó el dato clase de obesidad como Clase de IMC ajustando y estandarizando su uso al que normalmente se encuentra en la literatura al respecto.
- Se eliminaron de los datos los atributos que no tienen uso o que no son claros: riesgo cardiovascular, recv, td, HAS (hipertensión arterial sistémica)[13],

Habiendo hecho esta primera validación de los datos el siguiente paso consiste en determinar el tamaño y las características que debe tener la información que conforma a la vista minable.

V.2.2 Determinación de la muestra.

Por las características de los procesos a realizar se establecieron los siguientes requisitos para la muestra de datos:

1. Cumplir con un mínimo de registros que le haga una muestra representativa,
2. Estar conformada por al menos tres subgrupos diferentes de individuos. De acuerdo a Kahl (1990) la selección del grupo se pude realizar de diferentes maneras, una buena opción es la de grupos especiales de personas trabajando en entidades o compañías estables, lo que les hace de fácil localización.
3. Los subgrupos deben estar conformados por individuos relacionados de alguna forma y debe haber una diversidad de los perfiles que representan cada uno de ellos.

En relación al punto uno se tomó como base la mecánica presentada para el cálculo de la muestra de la encuesta nacional de salud (Secretaria de Salud Pública, 2006) de acuerdo a la siguiente fórmula:

[13] Este dato se omitió por la baja confiabilidad que refirió de él el epidemiólogo.

$$\boldsymbol{n = \frac{Z^2_{\frac{\alpha}{2}}(1-P)}{r^2 P}} \quad (23)$$

donde

P es la proporción a estimar (proporción de menor importancia)

$Z^2_{\alpha/2}$ cuantil de una distribución asociada a un nivel de confianza deseado

r es el error relativo máximo a aceptar

n es el número de registros que debe contener la muestra.

Los valores determinados para cada uno de estos parámetros fueron: $P = 8.1\%$, $Z^2_{\alpha/2} = 1.97$ con una $N = 29287$, $r = 25\%$ y un nivel de confianza del 95%. Lo que arrojó un tamaño de la muestra de 1071 en total.

Independientemente a la importancia que reviste el tamaño de la muestra, igual o más importante es la diversidad que debe haber en los datos a trabajar. Tanto la KNN como la GP, son algoritmos que no están basados en una inferencias estadística pura, funcionan más como una heurística donde la variedad de individuos que conforman la muestra tiene suma relevancia. Para la KNN si no hubiese esa variedad la formación de conglomerados sería muy pobre y en el caso de la GP la muestra funciona como una población de prueba para las reglas encontradas, si esta población no contiene una variedad rica de ejemplos muy probablemente el resultado del algoritmo tienda a un máximo local, lo que no asegura la optimalidad del resultado.

En este sentido para la generalización que se haga de los resultados obtenidos sí tiene importancia el tamaño de la muestra pero no tanto como lo tendría si los algoritmos seleccionados se basasen en una generalización estadística.

Los criterios utilizados para determinar si la muestra seleccionada tiene la variabilidad deseada fueron: los subgrupos deben ser representativos de diferentes niveles socioeconómicos, principalmente porqué el nivel socioeconómico determina el tipo de alimentación; y de diferentes perfiles laborales, lo que marca mucho las costumbres sociales del grupo.

Tabla V. Porcentajes como se distribuye la población por género y peso en los archivos de la muestra.

Concepto	*Archivo*		
	Hutchingson	CFE	CEMEX
Género			
hombres	51%	69%	86%
mujeres	49%	31%	14%
Peso			
Bajo	0.76%	1.31%	0.00%
Normal	31.22%	18.34%	26.92%
Sobrepeso	40.36%	45.85%	42.31%
Obesidad I	25.63%	22.27%	19.23%
Obesidad II	1.27%	8.73%	11.54%
Obesidad III	0.51%	1.75%	0.00%
Obesidad IV	0.25%	1.75%	0.00%

De acuerdo a la disponibilidad de información reciente con que cuenta en el área de control epidemiológico se estableció que la muestra se conformara de tres archivos tomados de diferentes empresas en diferentes momentos (todos dentro de los últimos 3 meses).

Las empresas seleccionadas para la muestra fueron:

1. Comisión Federal de Electricidad (CFE), con un perfil socioeconómico de clase media baja
2. Maquiladora de ropa Hutchinson, con perfil socioeconómico bajo
3. Cementos de México (CEMEX), con perfil económico medio alto.

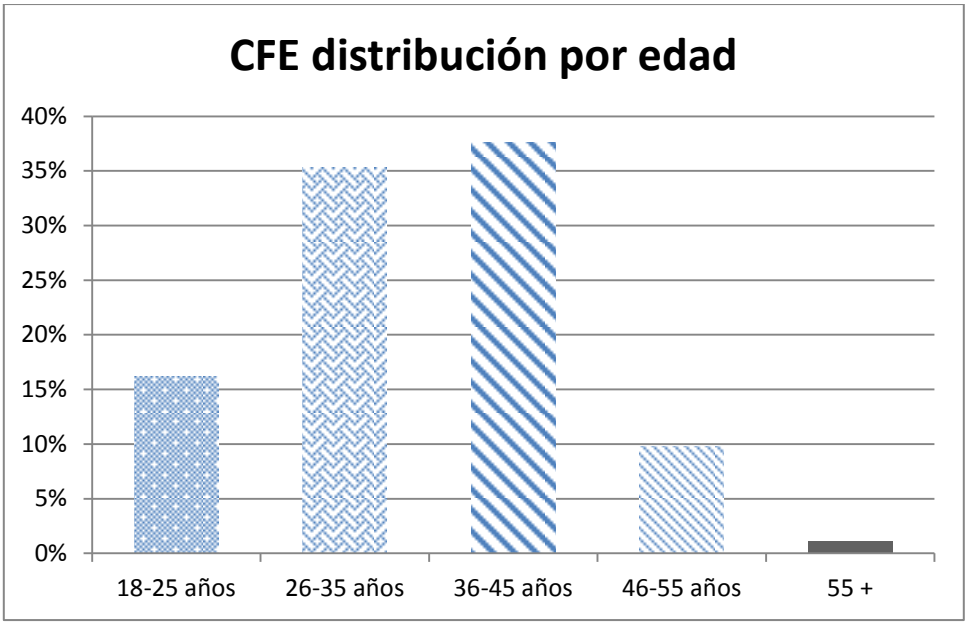
En relación al punto 3, se tomó para evaluarlo tres datos que se usan en todo estudio médico relacionado con la obesidad: sexo, edad y el índice de masa corporal.

La Tabla V presenta las distribuciones por sexo y tipo de obesidad en cada una de las empresas seleccionadas.
En términos del género se puede observar hay una mezcla variada entre el número de mujeres y hombres que componen las tres empresas, con una marcada tendencia a haber más mujeres en CEMEX, sin embargo la proporción total entre hombres y mujeres no se ve afectada grandemente porque Hutchingson es el archivo más grande y CEMEX el más pequeño de la muestra, nivelándose la situación en el consolidado.
En el caso de los niveles de obesidad era de esperar que el número de personas con sobrepeso y obesidad fuese mucho mayor que el de personas en peso normal. Sin embargo, se puede apreciar en la gráficas que existen personas con peso normal e incluso bajo, las cuales enriquecen el análisis a practicar evaluando personas que no forzosamente padecen el síndrome metabólico.

Finalmente se estableció tomar el archivo de la CFE como datos de entrenamiento el cual tiene la mezcla más variada en el tipo obesidad y corresponde a un perfil socioeconómico intermedio que se consideró más representativo.
A continuación se presentan algunos estadísticas sobre este archivo.

En la Figura 30 se observa que la mayoría de los individuos se hayan en el rango de edad 26-45 años, edades entre las cuales se desarrollan la mayoría de los padecimientos relacionados con el síndrome metabólico (Secretaria de Salud Pública, 2006).

Por otro lado en la Figura 31 se presenta esta misma muestra por género, donde se observa que hay una mezcla muy similar entre hombres y mujeres, lo que asegura que no habrá un sesgo. Se puede observar Figura que la obesidad y sobrepeso son problemas similares en ambos géneros, con una pequeña tendencia a ser más marcado el problema del sobrepeso en los hombres.

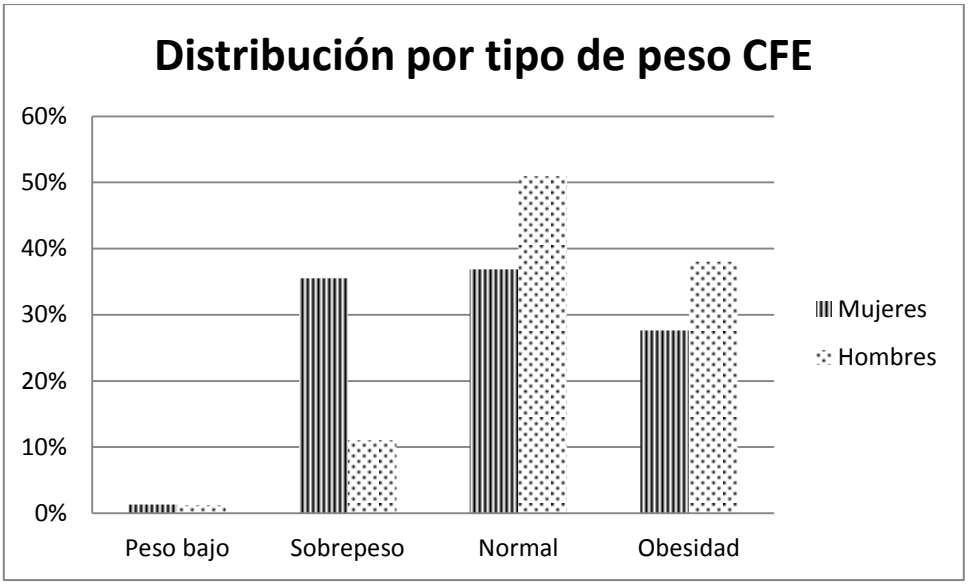


Edad	Total	Registros
18-25 años	14%	38
26-35 años	32%	84
36-45 años	33%	89
46-55 años	9%	23
55 +	2%	4
Total		238

Figura 30. Distribución por edad de la muestra para pruebas de la CFE.

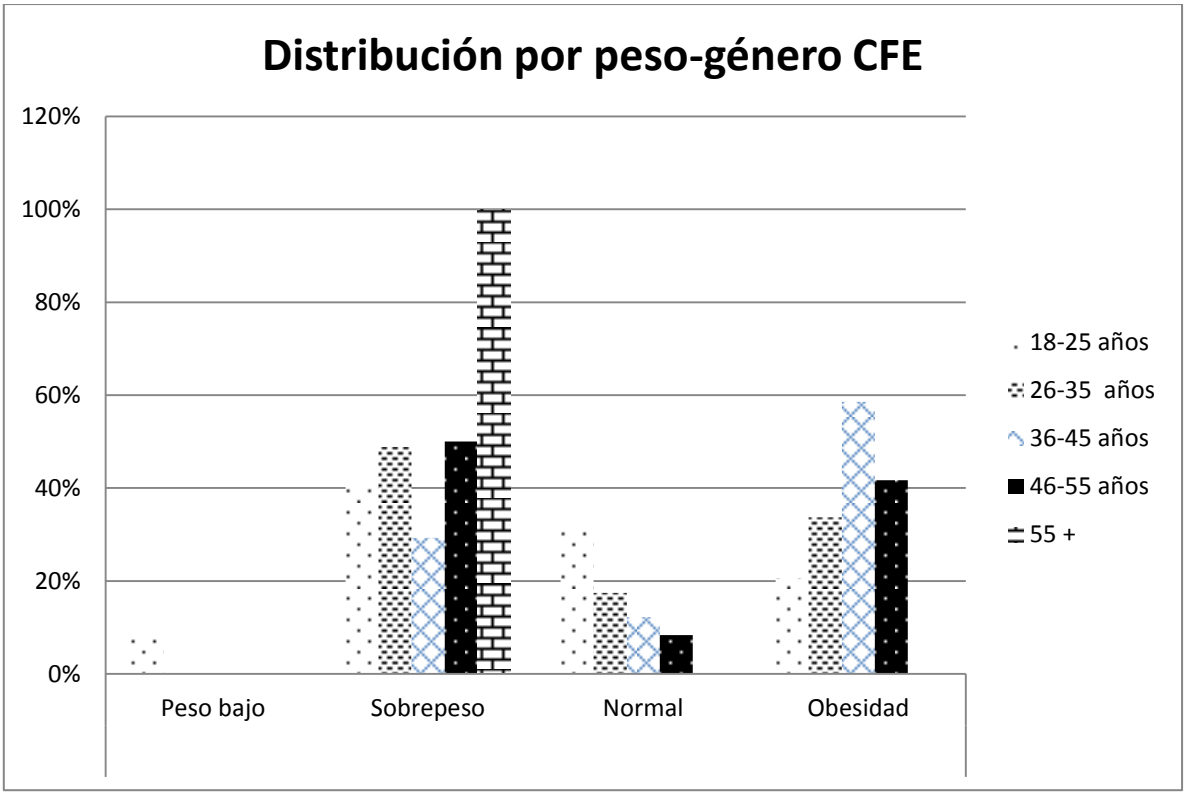
Finalmente la Figura 31 se presenta la misma muestra ahora haciendo además cortes por el índice de obesidad (calculado como lo establece el Apéndice B). Podemos observar una marcada tendencia a la obesidad y sobrepeso en todos los grupos de edades. Sin embargo, las mezclas de índices de obesidad son más variadas y muy similares, lo que da una gran variedad de condiciones entre los individuos que conforman la muestra.

Habiendo determinado que el archivo de la CFE será tomado como datos de entrenamiento, primero se debe realizar una limpieza y transformación de los datos.



género	peso	Porcentaje del subgrupo
Mujeres	Bajo Peso	1%
	Normal	36%
	Sobrepeso	37%
	Obesidad	28%
Hombres	Bajo Peso	1%
	Normal	11%
	Sobrepeso	51%
	Obesidad	38%

Figura 31. Distribución por género e índice de obesidad de la muestra de prueba de la CFE.



Edad	Peso	% del subgrupo
18-25 años	Peso bajo	8%
	Sobrepeso	41%
	Normal	31%
	Obesidad	21%
26-35 años	Peso bajo	0%
	Sobrepeso	49%
	Normal	17%
	Obesidad	34%
36-45 años	Peso bajo	0%
	Sobrepeso	29%
	Normal	12%
	Obesidad	59%
46-55 años	Peso bajo	0%
	Sobrepeso	50%
	Normal	8%
	Obesidad	42%
55 +	Peso bajo	0%
	Sobrepeso	100%
	Normal	0%
	Obesidad	0%

Figura 32. Distribución de peso por rangos de edad.

V.3 Pre procesamiento de la información.

El registro de la CFE (ver Apéndice A) tiene 18 componentes que deben ser limpiados antes de procesarles, cada uno de ellos son de naturaleza muy distinta

con características muy particulares y hay que buscar homogeneizarles para poderles procesar.

V.3.1 Limpieza de los datos.

De acuerdo a Witten (2005) el primer paso es la limpieza de los datos y son 4 los aspectos que se revisan:

1. Esparcimiento de los datos, consiste en revisar si en la matriz de información no hay muchas filas en la cual la mayoría de sus elementos contengan ceros y solo de forma muy esporádica contengan valores diferentes de cero.
2. Tipo de atributos, dependiendo del método de minería de datos a trabajar los atributos numéricos, como escalas ordinales, solamente se utilizan en comparaciones del tipo menor-qué y mayor-qué; sin embargo existen otros métodos que los utilizan como escalas radiales y utilizan cálculos de distancia. De acuerdo al método es muy factible que se tenga que realizar algún tipo de normalización de los valores numéricos.
3. Valores perdidos, en la práctica los archivos pueden contener registros en los cuales hayan valores faltantes, esto posiblemente se deba a algún tipo de error, omisión o consideración especial sobre la fuente de la información. Son diferentes las acciones que se pueden tomar antes de correr los métodos para manejar la información faltante, este tipo de acciones van desde omitirles, hasta tomar valores promedio o representativos. En todo caso la mayoría de los métodos suponen implícitamente que los valores perdidos simplemente se desconocen.

4. Inexactitud de los valores, se debe verificar cuidadosamente buscando atributos con valores engañosos, repetidos, u obsoletos.

De acuerdo a los pasos anteriores del archivo de la CFE se revisó (apoyados por Excel): por inspección visual si existían filas con puros ceros, o vacías, y aquellas donde ciertos atributos no tenían información (valores perdidos); en base al conocimiento del contexto médico de la información se revisaron los valores que presentaban los datos y se identificaron las escalas que usaron; se hicieron comparaciones entre los valores de atributos que estaban altamente relacionados, o aquellos que eran productos de un cálculo.

Producto del proceso de limpieza se encontró que:

- No había ningún atributo que presentara problemas particulares de esparcimiento de la información,
- Se encontraron seis atributos (IMC y clase de obesidad, tensión arterial y clasificación de la tensión arterial, riesgo cardiovascular y Recv) que eran los mismos datos con diferentes representaciones.
- Se encontraron 41 registros con una gran cantidad de valores perdidos en los atributos relacionados con las mediciones de grasas (lípidos) y glicemia, la razón para la falta de información es que algunas de las personas implicadas no se presentaron a los análisis de sangre.
- Se encontraron dos pares de atributos duplicados y un atributo no identificable (IMC y clase de obesidad, tensión arterial y clasificación de la tensión arterial), el atributo no identificable es TD).

La Figura 33 esquematiza el preprocesamiento de la información que se utilizó para el entrenamiento de la red neuronal.

Los primeros tres pasos de la Figura 33 tienen como propósito identificar la distribución de los valores que pueden tomar los datos y como los sesgos en dichas distribuciones pueden afectar en el proceso de formación de conglomerados. Producto de este análisis fue el tener argumentos para decidir el

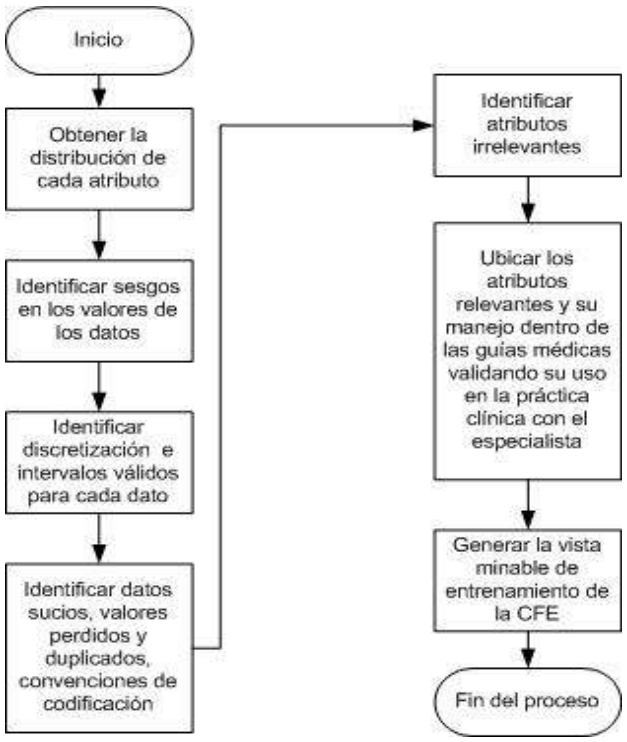


Figura 33. Pre-procesamiento de información

modelo de discretización, que en este caso fue por clases (en los datos categóricos).

V.3.2 Transformación de los datos.

Los últimos cuatro pasos de la Figura 33 tienen como propósito principal la transformación de los datos, de acuerdo a Han (2006). A continuación se hace una descripción de lo que se hizo en estos pasos:

- Identificar:

- datos sucios, se encontró que el atributo has no tiene mucha validez ya que en la mayoría de los casos no concordó con el atributo clasificación de la tensión arterial, el cual si avaló el especialista.
- valores perdidos y duplicados, se encontró que 78 registros (lo que representa 28% de todo el archivo) presentaban, en los datos relacionados con grasas en la sangre, omisiones debido a que las personas no quisieron practicarse el análisis de sangre. Este porcentaje es aceptable y se manejaron a través de valores perdidos (lo cual sucede frecuentemente con la información). El algoritmo KDD en weka (conjunto exhaustivo de herramientas para análisis y minería de datos) ofrece facilidades para manejar este tipo de valores en los atributos, en este caso se optó por la facilidad de omitirlos al cálculo, lo que implica que el algoritmo no los considera al establecer el criterio de similitud, pero sí los clasifica en base a todos los demás atributos.
- convenciones de codificación: la clave del sexo está codificada como F o M, todos los demás datos categóricos son binarios y se manejan con un 0 o 1, solo en el caso de tipo de obesidad y tipo de tensión arterial se manejan valores tipo clase descritos textualmente.

- Identificar atributos irrelevantes, ya se describió este proceso en la limpieza.
- Ubicar el manejo de los atributos relevantes, se ubicaron identificándoles en las guías clínicas encontrando que todos los atributos en el vector de entrada son relevantes.
- El paso de generar la vista minable implica la aceptación por parte de todos los implicados en que los datos están listos para ser procesados.

La transformación de datos se refiere a transformar o consolidar los datos de la forma adecuada para que el algoritmo trabaje. En los datos de la CFE se encontraron varios tipos de codificaciones:

- Claves numéricas binarias (riesgo, tabaquismo, alcoholismo, triglicéridos, DM2),
- Claves en forma de clases con valores codificados como cadenas de caracteres (sexo, clase de obesidad, clasificación de la tensión arterial),
- Datos continuos codificados con valores discretizados (peso, cintura, tensión arterial sistólica, diastólica, glicemia, colesterol, triglicéridos)
- Datos numéricos de tipo real (talla, IMC, VLDL)

La KNN para formar conglomerados con los vectores de datos de entrada necesita que las escalas sean similares, ya que si no lo son, el criterio de similitud (distancia euclidiana) se ve afectado por el peso que tiene una escala muy grande frente a una escala muy pequeña. Por ejemplo si una variable tiene valores entre [0,…,1000] y otra en el intervalo de [0…10] el primero dominará casi completamente al segundo ya que tendrá gran impacto en la medición de la distancia (Vensato, et al., 1999)

Para resolver el problema se deben escalar los valores utilizando una normalización de tipo clasificación usando la metodología de análisis de histogramas (Han, et al., 2006).

Este método consiste en analizar el histograma valor-frecuencia de cada atributo y encontrar la forma de juntar intervalos de tal forma que se reduzca el número de ellos. Una vez definida la nueva agrupación se asigna un valor de clase para cada nuevo intervalo producido. En la Tabla VI se presenta el resulta de aplicar el método descrito al atributo peso.

Como se describió en el Capítulo III antes de ingresar los datos a la red neuronal se les debe integrar una parte semántica, con este propósito se construyó una base de datos en Access 2007 para poder facilitar el proceso (ver sección III.6).

Tabla VI. Transformación del dato peso usando clases.

Normalización	Rango de valores
1	(...,58)
2	[58,67)
3	[67,77)
4	[77,87)
5	[87,96)
6	[96,106)
7	[106,115)
8	[115,…)

Al finalizar estos procesos se obtuvo la vista minable que aparece en la **Tabla VII** lista para ser trabajada por la red neuronal de Kohonen.

Tabla VII. Registro de la vista minable.

Campo	Descripción
Edad	Edad del paciente
Sexo	Género del paciente
Peso	Peso del paciente en kilogramos
Talla	Talla del paciente en metros
Clase de obesidad	Clasificación del índice de masa corporal por tipo de obesidad
Cintura	Cintura en metros
Tensión arterial sistólica	Medición puntual de la tensión arterial sistólica del paciente

Vista Minable (continúa)	
Tensión arterial diastólica	Medición puntual de la tensión arterial diastólica del paciente
Clasificación de la tensión arterial	Clasificación de la tensión arterial sistémica del paciente basada en las mediciones anteriores
Sobre riesgo	Sobre riesgo en condiciones cardiovasculares.
Tabaquismo	Si la persona tiene el hábito de fumar
alcoholismo	Si la persona tiene el hábito de beber
Dm2	Si la persona padece diabetes Mellitus tipo 2
Glicemia	Medición puntual del nivel de glicemia prueba de glicemia en ayunas del paciente
Colesterol	Medición puntual del nivel colesterol total
Triglicéridos	Medición puntual del nivel de triglicéridos del paciente
Vldl	Medición puntual del nivel colesterol de alta densidad del paciente
Riesgo detectado	Medición del riesgo cardiovascular tomando en cuenta el resto de los indicadores.

V.4 Selección de los valores de los parámetros de la red neuronal.

Se ubicaron diferentes alternativas para trabajar las redes neuronales de Kohonen, y se determinó utilizar el *toolbox* para Matlab, creado en la Universidad de Helsinki en Finlandia (Vensato, et al., 1999), porque además de proporcionar un software robusto, cuenta con una serie de herramientas complementarias para apoyar en el análisis de resultados.

De acuerdo a Choppin (1998) antes de iniciar con el entrenamiento de la red, hay que seleccionar el tipo de entrenamiento que va a tener la red, hay dos posibles opciones: secuencial o por lotes.

- En el entrenamiento secuencial en cada paso de iteración, se selecciona de forma aleatoria un vector x de entrada y se calculan todas las distancias entre él y todos los vectores de pesos utilizando la medida de distancia.
- En el caso del entrenamiento por lotes (por lotes) el entrenamiento también es iterativo pero en vez de utilizar un solo vector x de entrada a la vez, se usa el conjunto de datos completo presentándolo al mapa antes de que los ajustes sean hechos. En cada paso de entrenamiento el archivo es particionado de acuerdo a las regiones de Voronoi del mapa, siendo este proceso más parecido al manejo del algoritmo de K-Medias y no tanto un entrenamiento en el sentido estricto de una red neuronal

Por las consideraciones anteriores se decidió el tipo de entrenamiento secuencial.

Los parámetros a seleccionar fueron: el tipo de vecindario, que en este caso se seleccionó el tipo hexagonal sobre el rectangular, debido a que da un mayor número de vecinos (6 contra 4) y mejora la exactitud de la red; número de iteraciones (épocas) en entrenamiento rudo y fino de la red, donde se seleccionó el máximo posible (1000 y 10000 respectivamente) de acuerdo a como lo que

establece la literatura, el tipo de inicialización (lineal o aleatoria), se seleccionó lineal ya que asegura el cardado de toda la red; y el tipo de vecindario, que en esta ocasión de seleccionó gaussiano, ya que aseguraba una mejor actualización de las neuronas de la red.

El siguiente paso fue el establecimiento de los parámetros del algoritmo, los cuales se dividen en características estructurales de la red y sintonía de los parámetros del algoritmo.

V.4.1 Características estructurales de la red.

De acuerdo a (Choppin, 1998) son cuatro las características estructurales a definir: el tamaño de la malla, su forma, el tipo de superficie, y el tipo de inicialización.

De acuerdo a lo establecido en el Capítulo III, el valor de una red neuronal de Kohonen se basa en la reducción de la dimensionalidad de los datos a través de una cuantización de vectores y en la preservación de la topología intrínseca dentro de los datos.
La calidad de la cuantización se mide por el error de cuantización definido por

$$\boldsymbol{Eq} = \frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{k=1}}^{\boldsymbol{n}}\|\boldsymbol{\xi_k} - \boldsymbol{x_i}\| \qquad (24)$$

Donde E_q es la distancia promedio actual entre cada punto de la distribución inicial y su vector representativo, x_i representa cada punto de la distribución inicial y ξ_k representa una neurona (vector representativo) de la malla.

La calidad topológica se mide mediante el llamado error topológico, de acuerdo a (Choppin, 1998) este error corresponde a una representación $dy - dx$ donde la idea es construir una gráfica con un punto para cada par de unidades de todo el mapa, cada punto en esta gráfica (ψ) se obtiene de comparar un par de unidades en el mapa $i_1(x_{i1}, y_{i1})$ e $i_2(x_{i2}, y_{i2})$, de tal forma que cada punto ψ de la gráfica

tiene un par ordenado (dy, dx), donde dy es la distancia entre las coordenadas de la malla de esas dos unidades

$$\boldsymbol{dy(i_1, i_2) = \left\| y_{i_1} - y_{i_2} \right\|} \quad (25)$$

Y dx es la distancia euclidiana (calculada con los componentes de los vectores) entre vectores ($p - dimensionales$) representativos asociados con cada nodo

$$\boldsymbol{dx(i_1, i_2) = \left\| x_{i_1} - x_{i_2} \right\|} \quad (26)$$

El proceso de construcción se describe en la
Figura 34, donde se puede apreciar una línea en donde (a) (b) o (c) pasa por (0,0) y (1, d_m), donde d_m es la distancia media entre el vector representativo de dos unidades vecinas, esto es dos unidades donde la distancia topológica es uno; a la cual se le llama línea de auto organizado (SO). Una gráfica dy-dx (
Figura 34) donde todos los puntos que están cercanos indican que hay una buena correlación entre las distancias topológicas y los vectores representativos. Ahora si tenemos que el espacio del mapa (d) es de menor dimensionalidad que el espacio de entrada(p) $d < p$ entonces la mayoría de los puntos estarán por debajo de SO. Lo que básicamente revela que, conforme la distancia topológica entre unidades se incrementa, la distancia correspondiente entre el espacio de vectores representativos no se incrementa tan rápidamente. Generalmente este es el caso ya que se busca representar un espacio de alta dimensionalidad (como de los vectores de la CFE) en un espacio bidimensional. El error topológico entonces es esta distancia a SO.
Ahora bien tanto en el caso de la cuantización de vectores, como en el de la reducción de la dimensionalidad de la representación siempre va a haber un error.

Si ponemos tantos nodos (vectores representativos) como datos de entrada tendremos que el error de cuantización tiende a hacerse cero, y en el caso del tamaño del espacio de la malla, si fuese $d = p$, el error topológico tendería a desaparecer.

En la Tabla VIII se resumen los primeros intentos para establecer las características estructurales de la red. La principal característica que se trabajó fue el tamaño de la malla y su efecto en el error de cuantización.

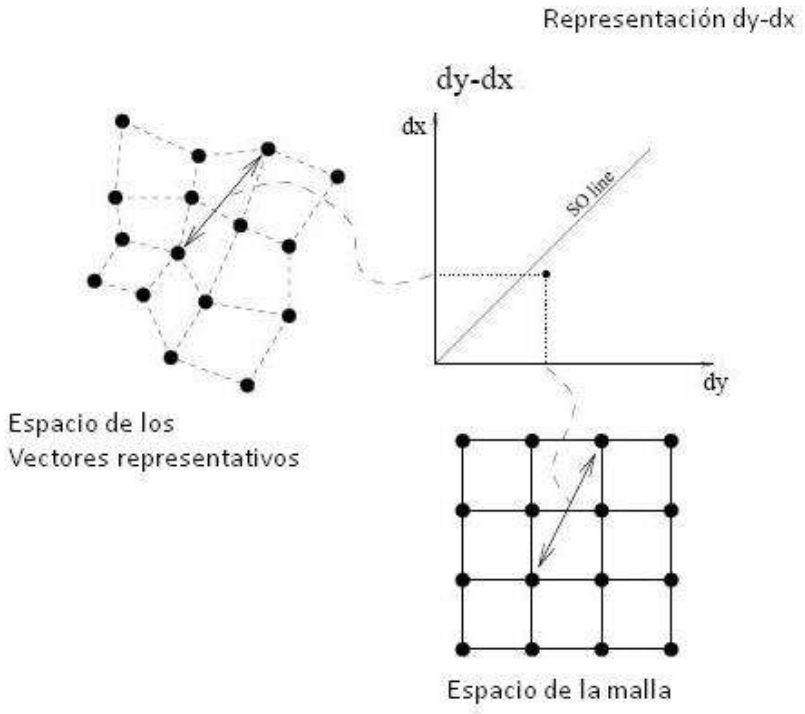


Figura 34. La representación dy-dx de un mapa de Kohonen dando información acerca de la conservación de la topología, tomado de (Choppin, 1998)**.**

De acuerdo a (Choppin, 1998), el tamaño de la malla está restringido por dos factores: la reducción de la dimensión de los datos (sin llegar a la pérdida de información); y que dicho tamaño tiene un límite superior natural, el número de datos de entrenamiento.

Este último factor tiene implicaciones tanto en el costo computacional que conlleva el procesar la red neuronal como, y más importante aún, por la pérdida de generalización que se da cuando una red demasiado grande tiende a tener un error de cuantización muy pequeño por estar altamente ajustada a los datos de

entrenamiento. Este efecto lo podemos observar en la Tabla VIII, donde se presentan los resultados de los experimentos realizados para establecer las características estructurales de la red.

Tabla VIII. Resumen de los experimentos realizados para establecer las características estructurales de la red.

# Pba.	Tamaño mapa	Iteraciones de entrenamiento	Iteraciones de ajuste	Tipo de superficie	Inicialización	vecindario	Qerror promedio
1	14x6	1000	10000	Rect	Lininit	Gaussiano	3.533
2	14x10	1000	10000	Rect	Lininit	Gaussiano	3.326
3	14x14	1000	10000	Rect	Lininit	Gaussiano	3.121
4	17x14	1000	10000	Rect	Lininit	Gaussiano	3.040
5	20x20	1000	10000	Rect	Lininit	Gaussiano	2.659
7	40x27	Pequeño	Pequeño	Rect	Lininit	Gaussiano	1.281
10	40x37	Pequeño	Pequeño	Rect	Lininit	Gaussiano	0.962
11	12x7	1000	10000	Hexa	Randinit	Gaussiano	3.518

*nota en todos los casos el error topológico fue muy similar y para la prueba 10 se obtuvo 0, para la prueba 11 el error topográfico es 0.004

Otra herramienta fundamental para determinar el buen funcionamiento de los parámetros seleccionados son las llamadas matrices U.

De acuerdo a Ultsch (2003) las matrices U entregan un paisaje de la relación de distancia de los datos de entrada en el espacio de datos, que visto en un diagrama tridimensional, Figura 35, estas distancias representan alturas donde los vectores con alturas muy grandes (cimas) son muy distantes con relación a otros vectores en el espacio de datos y las distancias muy pequeñas (valles) representan vectores que son muy cercanos a otros vectores en el espacio de datos, de tal forma que las “montañas” representan los límites de los conglomerados.

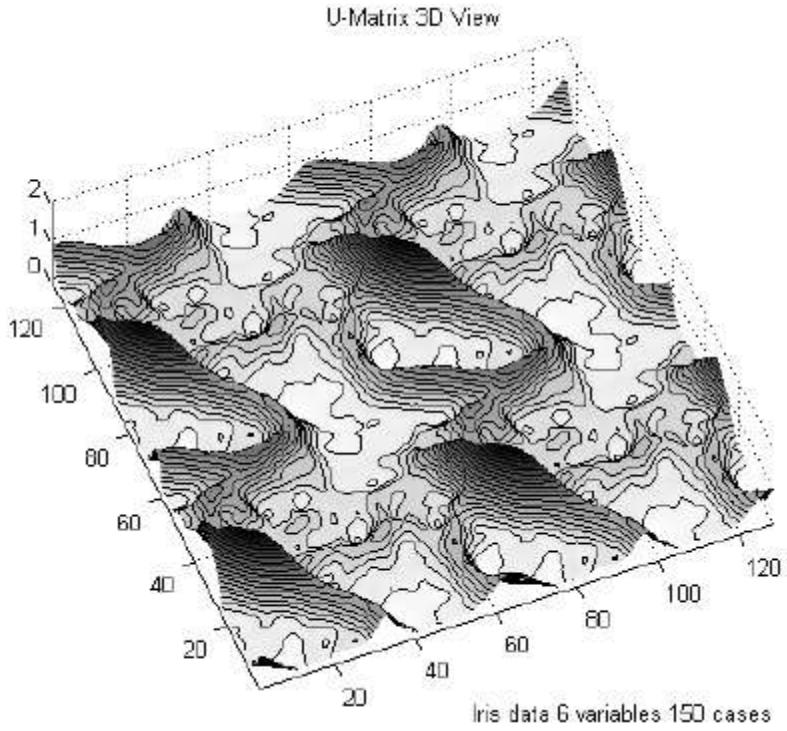


Figura 35. Ejemplo de una gráfica tridimensional de una matriz U (Ultsch, 2003)**.**

En otras palabras una matriz U es una representación gráfica que despliega la estructura de la distancia local de la malla. Sea una neurona (vector representativo) n de la malla, $NN(n)$ el conjunto de vecinos cercanos en la malla, y $w(n)$ el vector de pesos asociados con la neurona n, entonces

$$\boldsymbol{U - altura(n) = \sum_{m \in NN(n)} d(w(n) - w(m))} \qquad \textbf{(27)}$$

donde $d(x, y)$ es la distancia euclidiana.

En una representación bidimensional la altura calculada se representa mediante un código de color donde el azul representa individuos muy similares y los colores rojos individuos muy diferentes.

En la Figura 36 se presenta la matriz U del experimento 10 con un error de cuantización de 0.962 y un error topológico de 0. En la Figura se observa una gran

cantidad de conglomerados estos son las áreas más oscuras, donde el gran número de ellos hace que la clasificación pierda todo valor de generalización, ya que está demasiado ajustada al espacio de entrada, efecto que se le denomina overfitting.

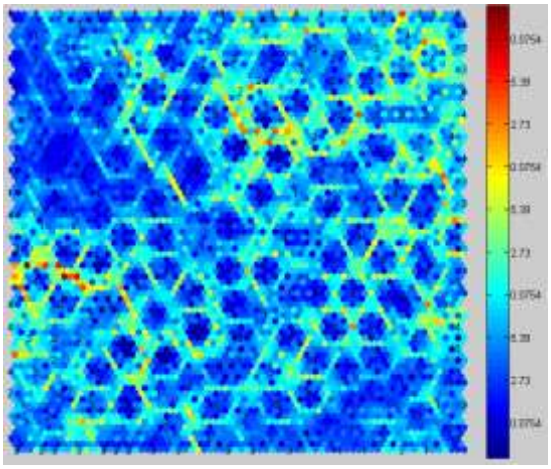
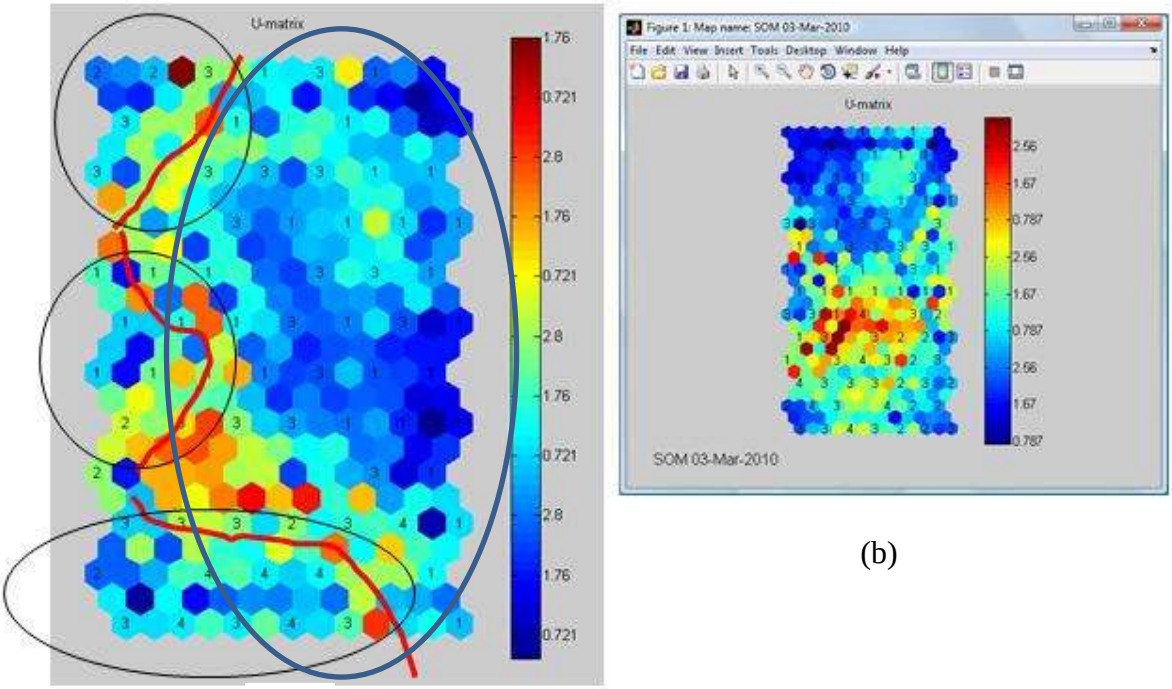


Figura 36. Matriz U de una malla de 40x37, en ella se observa el efecto del overfitting.

Al tomar una matriz de salida de 12x7 se incrementó el error de cuantización a 3.518 y el error topológico a 0.004, que comparado con la matriz de 14x6 da un error de cuantización muy similar. Sin embargo al analizar la Figura 37(a) se observa que la matriz de 12x7 da una mejor definición de los conglomerados que la matriz de 14x6 (b), al mostrar alturas que forman muros (marcados con las líneas) que delimitan más claramente cuatro conglomerados.

Habiendo establecido estas primeras conclusiones se realizaron más experimentos con tamaños de malla más semejantes al descubierto, buscando identificar ahora como se comportaba la contraparte del error topológico. Los resultados de estos experimentos se presentan en la Figura 38. Incipientemente se empiezan a delimitar dos grandes conglomerados el de la derecha de la línea, que al revisarle se encontró que correspondía a las persona obesas y de forma todavía muy dispersa el de la izquierda de la línea, el cuál hasta este momento no queda muy clara su conformación.



(a) (b)

Figura 37. (a) matriz U de una malla 12x7, se puede observar la formación de 4 conglomerados; (b) matriz U de malla 14x6 donde se diluyen algo los conglomerados.

Como resultado de estas pruebas se seleccionó el tamaño de la matriz de 12x7, el cual aun cuando tiene un error de cuantización relativamente alto 3.518, permite un nivel de generalización bastante aceptable y mantiene el error de topología en el mismo nivel que las demás pruebas, además tiene una mejor definición de los conglomerados a trabajar. Al final del entrenamiento de la red se obtuvo que el número ideal de iteraciones (épocas) para la sintonía ruda de la red fue de 1000, mientras que en la sintonía fina fue de 10000, lo cual concuerda en forma general con lo descrito por (Choppin, 1998).

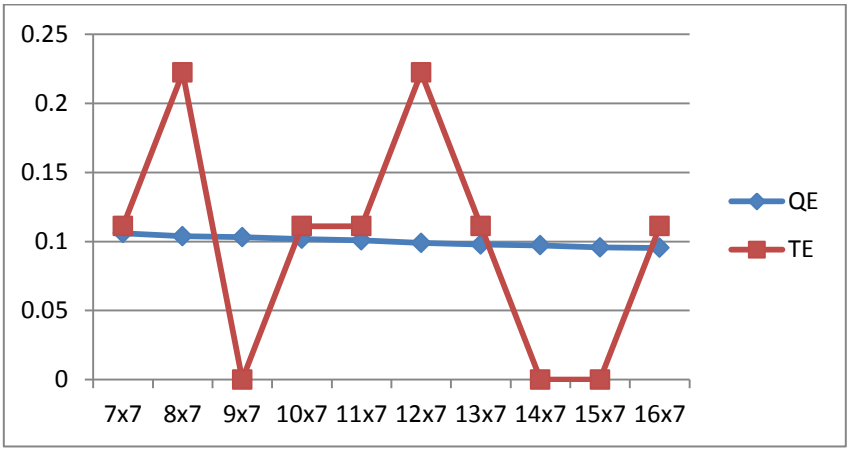


Figura 38. Comportamiento entre el error de cuantización (QE) y el error topológico (TE) en las diferentes pruebas con los datos de la CFE.

Al analizar los conglomerados obtenidos se encontró que el gran conglomerado de la derecha (ovalo grande) en la Figura 37 (a) se conformaba de pacientes con algún tipo de obesidad, pero en los otros tres conglomerados no había una clara diferenciación de los patrones. De acuerdo a Choppin (1998) se podía estar presentando el problema de dentro de los componente hubiese una muy alta correlación. Cuando se tienen componentes dentro de los vectores altamente correlacionados se afecta el cálculo de la medida de similitud y por lo tanto perjudica la discriminación entre subgrupos no son tan claramente diferenciables.

Para valorar este efecto se obtuvieron las matrices U de todos los componentes por separado y se les compararon Figura 39. Si dos matrices U son muy similares implica que ambos componentes se comportan de la misma forma al determinar el criterio de similitud.

Se puede observar claramente como las matrices U de peso y cintura, y las de triglicéridos y VLDL son prácticamente iguales, lo que implica una alta correlación entre datos.

Analizando la información con el especialista epidemiólogo se encontró una muy alta correlación entre los siguientes datos:

- Tensión arterial sistólica, tensión arterial diastólica, tensión arterial sistémica y la hipertensión arterial
- Cintura y peso
- VLDL y triglicéridos
- Riesgo cardiovascular y peso

De acuerdo a Choppin (1998) es necesario descartar uno de los atributos ya que finalmente el comportamiento de uno se ve reflejado en el otro. Después seleccionar los atributos a descartar junto con el epidemiólogo la vista minable quedando como sigue:

Edad:	constante aleatoria entre [18-75]
Genero:	valor booleano aleatorio [1,2]
IMC:	constante aleatoria entre [0-7]
CT:	constante aleatoria entre [200-240]
TG:	constante aleatoria entre [150-1000]
GLICEMIA:	constante aleatoria entre [67-200]
ALCOHOLISMO:	valor booleano aleatorio [1-2]
TAS:	constante aleatoria entre [1-4]
TABAQUISMO:	valor booleano aleatorio [1-2]
DM2:	valor booleano aleatorio [1-2]

Donde la información en corchetes dicta el dominio que puede tomar el atributo.

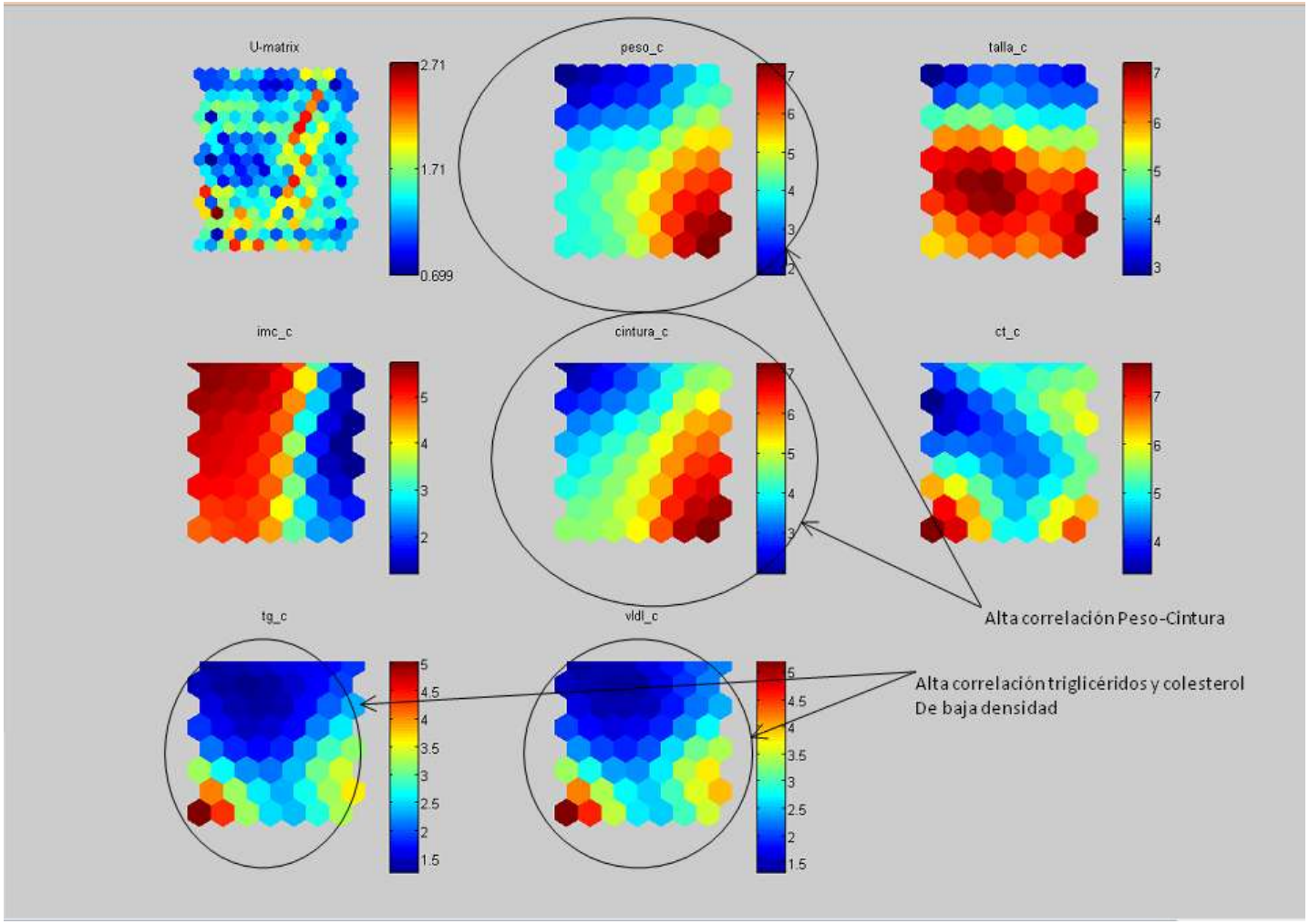


Figura 39. Representación en forma de Matriz U de la agrupación realiza por la red neuronal de Kohonen en varios de los atributos de los datos de entrada.

V.4.2 Resultados red neuronal de Kohonen.

Con la nueva vista minable se volvió a entrenar la red y se obtuvo un primer conjunto de conglomerados, ver Figura 40, donde al analizarles se empiezan a percibir dos grandes conglomerados separados por una gran pared de neuronas con muy poca similitud entre los individuos asignados a ellas.

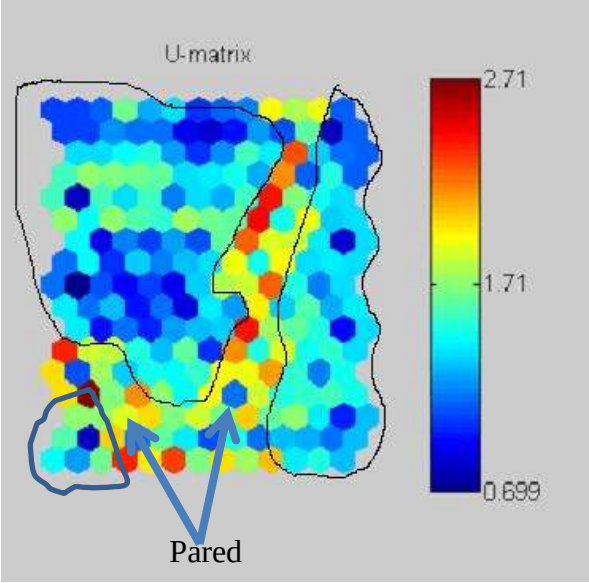


Figura 40. Primeros conglomerados usando solamente los datos de la CFE.

Con estos resultados se procedió a integrar el resto de los datos de la vista minable (Hutchinson y CEMEX), nuevamente se inició un proceso de entrenamiento bajo el criterio de validación cruzada.

Durante este nuevo entrenamiento se utilizaron procedimientos adicionales descritos por Alhoniemi (1999) para dejar lista la red neuronal. De acuerdo a este autor, es muy útil el uso de otras herramientas adicionales con las cuales se puede determinar si la red no está sub o sobre entrenada (mapas de Sammon) y también el mapeo de los componentes principales, el cual ejecuta un proceso de clasificación de los componentes de entrada de alguna forma similar a lo que se

realiza con la red neuronal pero directamente sobre los vectores de datos de entrada.

Los mapas de Sammon tienen como objetivo proyectar en un punto, en un espacio bidimensional, los vectores en el $L - espacio$, donde L es el número de componentes de cada vector y $L \gg 2$. Este proceso es similar al realizado por la red neuronal con la diferencia significativa de que en este caso se trabaja directamente en el espacio de los datos de entrada, lo cual permite hacer una comparación entre ambos resultados. El mapeo se realiza de forma similar a como lo hace la red neuronal, pero para establecer los vectores principales se calculan los 2 eigenvectores más grandes de la matriz estimada de covarianza (Sammon, 1969).

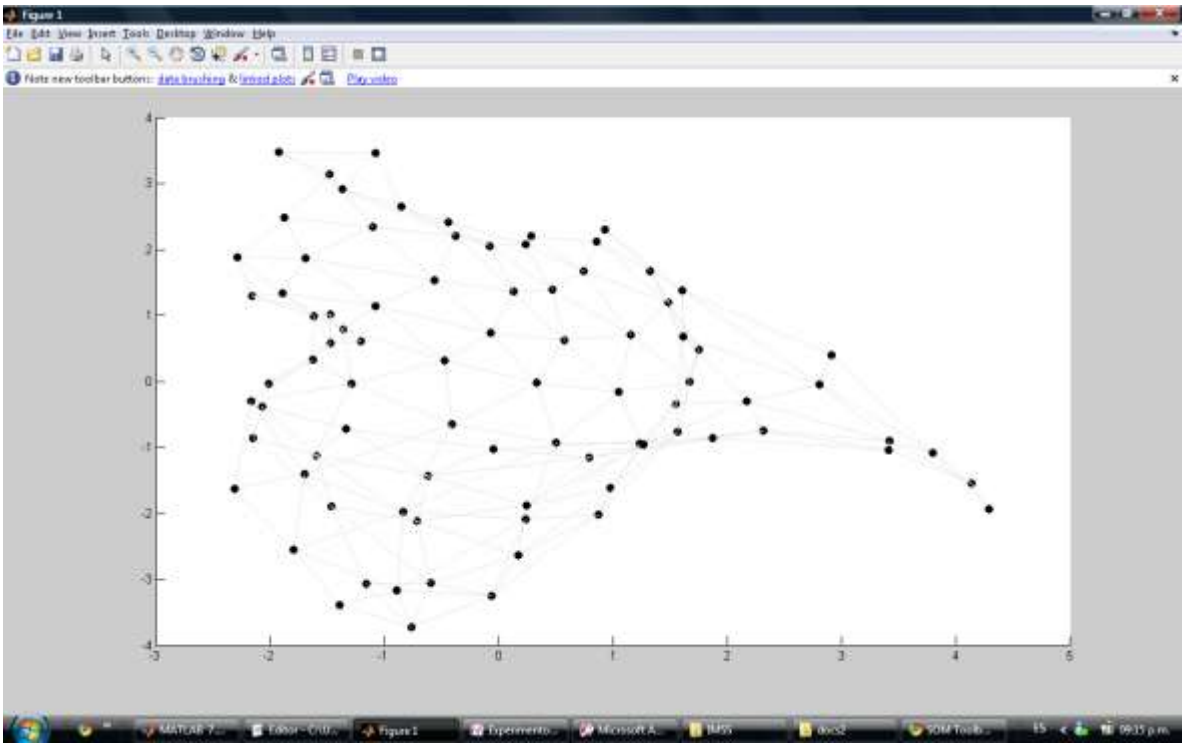


Figura 41. Muestra de un mapa de Sammon plegado hacia la izquierda, efecto que se da por un bajo entrenamiento, 100 iteraciones entrenamiento rudo y 1000 entrenamiento fino

En este caso el *toolbox* para KNN tiene la capacidad de tomar como base la red generada por la red neuronal de Kohonen para elaborar el mapeo de Sammon, lo que permite ubicar ciertas características que permiten validar:

- Si el mapa de Sammon está sub entrenado aparecerán en el mapa, en la Figura 41 se puede observar dicha torsión sobre sí mismo.

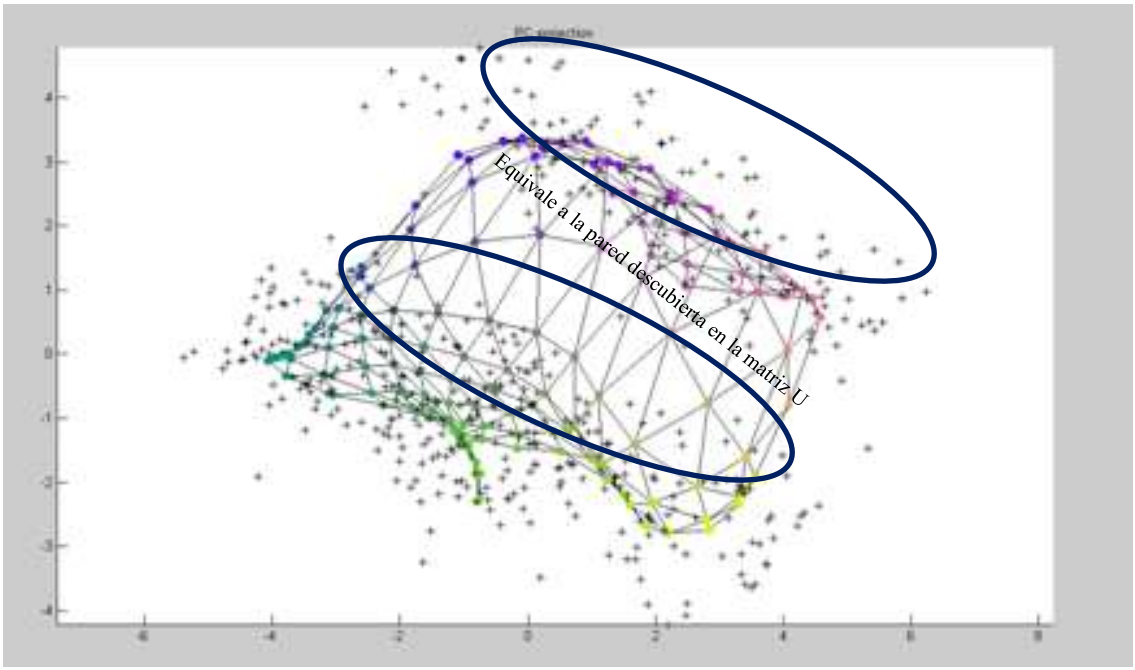


Figura 42. Mapeo de Sammon tomando como base los datos consolidados.

- Mientras que el mapa de la Figura 42 no hay pliegue y por simple inspección visual se puedan diferenciar los principales conglomerados (ovalos), claramente hay una diferenciación entre los dos grandes conglomerados antes referidos precisamente en la parte media del mapa.
- Los datos de entrada original se agrupan alrededor de dichos conglomerados, igualmente en la Figura 42 se puede observar como las nubes de datos se forman alrededor de ambos conglomerados.

En el caso de mapeo de componentes principales el *toolbox* permite observar la forma en que los eigenvectores son mapeados a la malla generada por la red neuronal, en la Figura 43 se pueden observar los resultados.

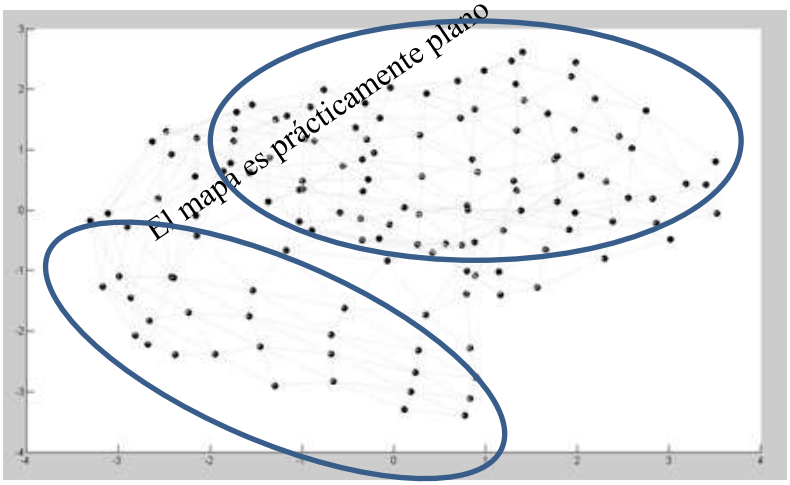


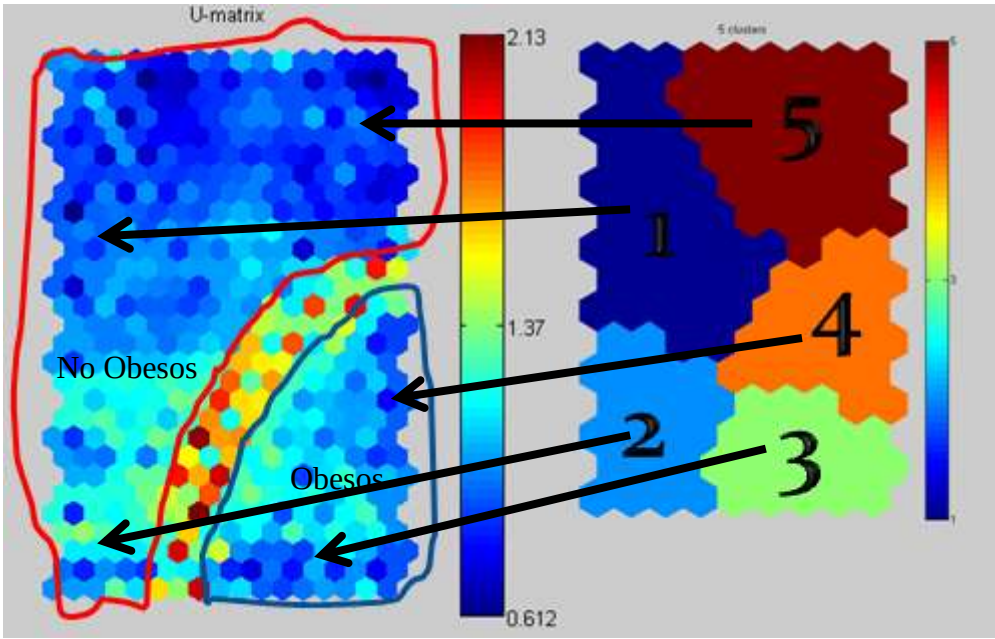
Figura 43. Mapeo de componentes principales de la red neuronal obtenida.

En este caso también, si el mapa está sobre o sub entrenado hay torsiones en el mapeo y de la misma forma se puede observar como sus componente se agrupan en dos grandes conglomerados, el más poblado en la parte superior de la Figura y el menos poblado en la inferior, con una separación entre ellos que indica la delimitación de ambos conglomerados.

Habiendo validado la red neuronal de acuerdo al procedimiento antes descrito, el mismo *toolbox* proporciona herramientas para la generación de conglomerados del mapa utilizando K-Medias basado en la malla generada.

Producto de este proceso se obtuvieron los 5 conglomerados presentados en la Figura 44. En la Figura (a), se puede observar cómo se preservan los dos grandes conglomerados antes descritos separados por una gran pared, y como

adicionalmente en (b), el algoritmo K-Medias, descubre otros 3 sub conglomerados, que corresponden a sub agrupaciones entre personas obesas y no obesas, todos ellos señalados con las flechas.



(a) (b)

Figura 44. (a) matriz U de los dos grandes conglomerados Obesos y No obesos y (b) formación de conglomerados usando K-Medias con la malla resultante red KNN.

V.5 Extracción de reglas de clasificación usando el algoritmo C4.5.

Una vez determinados los conglomerados en los cuales se dividió la información de entrada, el siguiente paso fue establecer las reglas que definen a cada conglomerado en otras palabras a cada clase. Para este propósito se seleccionó el algoritmo C4.5.

A dicho algoritmo se le proporcionaron los mismos datos de entrada que a la red neuronal más el indicador de clase que generó la propia red para cada sujeto, o

sea la clase a la cual pertenece. La razón fue para que el algoritmo C4.5 determinara las reglas para cada clase de acuerdo al procedimiento descrito en el Capítulo III.

Dado que se contaba con una cantidad limitada de información de entrada, se nuevamente se utilizó la técnica de validación cruzada para el entrenamiento (cross validation), con 10 separaciones (folds).

Producto de este análisis se obtuvo un conjunto de 57 reglas, que ya presentan la jerarquización de los diferentes subgrupos.

Al analizar las reglas con el especialista se determinó que cada clase agrupaba cierto tipo de pacientes que delineaba la regla, aunque algunas de ellas no eran muy claras (cortas), lo que concuerda con lo que se había previsto.

Sin embargo por la forma en que trabaja el algoritmo C4.5 existe cierta problemática sobre todo en las reglas del Tipo 5. Como se explicó en el Capítulo III el algoritmo C4.5 establece la ramificación mediante injertos al árbol base (prunning en inglés) de acuerdo a la cantidad de información que un atributo en particular ofrece. Este procedimiento puede llegar a causar la sobre adaptación del árbol a los datos de entrada (overfitting en inglés) (Witten, et al., 2005).

Al revisar el conglomerado del tipo 5, se encontró que de acuerdo al mapa semántico existe mucha cercanía entre éste y el del tipo 4 donde se ubica a las personas con obesidad mórbida. El grupo de personas con obesidad IV dentro de los datos es muy pequeño, apenas 5 personas, de las cuales la mitad son jóvenes con niveles normales en los indicadores de colesterol, triglicéridos y glicemia, mientras que la otra mitad son personas mayores con problemas fuertes de dislipidemia, lo que hace este un subgrupo demasiado disperso.

Una forma de evaluar a las reglas es la exactitud predictiva, a continuación se presentan las reglas generadas con C4.5 y las estadísticas generadas por weka:

Instancias clasificadas correctamente	536	80.9668 %
Instancias clasificadas incorrectamente	126	19.0332 %
Estadística Kappa	0.7553	
Error absoluto medio	0.1173	
Error cuadrático medio en raíz	0.2474	
Error relativo absoluto	37.5991 %	
Error cuadrático relativo en raíz	62.6507 %	
Número total de instancias	662	

Exactitud detallada por clase (conglomerado)

	VP	FP	Precisión	Clase
	0.817	0.08	0.813	tipo1
	0.752	0.046	0.779	tipo2
	0.897	0.058	0.836	tipo5
	0.781	0.027	0.833	tipo3
	0.736	0.035	0.762	tipo4
Peso				
Promedio	0.81	0.055	0.809	

donde VP son los verdaderos positivos, FP son los falsos positivos y Precisión es la exactitud predictiva.

Tabla IX. Matriz de confusión.

Frecuencia					
a	b	c	d	e	Clase
161	21	14	0	1	a=tipo1
21	88	7	1	0	b=tipo 2
9	3	148	3	2	c=tipo 3
1	1	2	75	17	d=tipo 4
6	0	6	11	64	e=tipo 5

En la matriz de confusión (Tabla IX) la diagonal principal establece el número de individuos clasificados correctamente, todos los individuos contabilizados fuera de esta diagonal no fueron correctamente clasificados y por consiguiente representan un error en la exactitud de la regla. Al analizar el detalle por clase se observa las reglas que determinan los tipos 2 y 4 tiene relativamente baja su precisión lo que implica que posiblemente existan problemas para estas dos reglas. Esto aunado al hecho que la precisión general está por debajo del 90% implica que es necesario optimizar las reglas generadas usando la GP encontrar reglas más robustas.

V.6 Extracción de reglas de clasificación usando Programación Genética.

Tanto las redes neuronales como la programación genética pueden ser vistas como técnicas alternativas para las mismas tareas, e.g., clasificación y problemas de aproximación. En particular la programación genética basada en gramáticas restrictivas es utilizada para la extracción de conocimiento partiendo de las bases de datos médicas (Brameir, 2004).

Una forma común de trabajar con GP es mediante la utilización de un "*framework*" que ayude al investigador a poder generar rápidamente un software robusto, y poder entonces concentrarse en la parte medular de la investigación (Poli, 2008). El *framework* que se decidió utilizar es JCLEC (Ventura, et al., 2007) , el cual es un sistema que implementa una gran variedad de algoritmos evolutivos y tiene la gran ventaja de que es fácil de configurar, además de que su código es abierto y permite la inclusión o modificación de componentes.

El usuario de un framework debe conocer principalmente qué tipo de programación genética va a utilizar, el tipo de representación a utilizar, los conjuntos de terminales y funciones a utilizar, cuáles son los parámetros que se deben establecer, la forma en que se van a evaluar los individuos (función fitness), y en su caso la gramática que va a restringir la formación de los individuos. De acuerdo a (Poli, 2008) una forma común de controlar estos parámetros durante los diferentes experimentos es mediante un control de experimentos como el de la Tabla X.

Tabla X. Control de experimentos con el JCLEC para el síndrome metabólico.

Objetivo	Encontrar una serie de reglas que mejor detecten el síndrome metabólico de acuerdo a la clasificación que tiene los casos de fitness
Conjunto Funciones	AND, OR, =, <,≥
Conjunto Terminales	Edad: constante aleatoria entre [18-75] Genero: valor booleano aleatorio [1,2] IMC: constante aleatoria entre 0-7 CT: constante aleatoria entre 200-240 TG: constante aleatoria entre 150-1000 GLICEMIA: constante aleatoria entre 67-200 ALCOHOLISMO: valor booleano aleatorio TAS: constante aleatoria entre 1-4 TABAQUISMO: valor booleano aleatorio DM2: valor booleano aleatorio

En el caso de la presente tesis se seleccionó GP basada en árboles con una gramática restrictiva, de acuerdo a la implementación de (Bojarczuk, et al., 2004); Enseguida establecemos función de evaluación de adaptabilidad (fitness). (Bojarczuk, et al., 2004) plantea una función de fitness donde intervienen 3 elementos: sensibilidad de la regla (Se), especificidad de la regla (Sp) y la simplicidad de la regla (Sy), ver ecuaciones (18), (19) y (22), respectivamente, en la Sección III.8. Estas ecuaciones son bien conocidas por los investigadores médicos que evalúan una nueva prueba de diagnóstico. Esta actividad califica el examen determinando los conceptos de sensibilidad y especificidad.

En el caso normal de una prueba de cohorte, se desarrolla una gran cantidad de trabajo discriminando la importancia que tiene cada prueba clínica que se hace hasta determinar aquella que diagnostique la condición X (Ruiz, et al., 2004).

De acuerdo a (Bojarczuk, et al., 2004) se considera que la comprensibilidad de la regla es mayor mientras más simple (pequeña) sea la regla, por lo que se busca mediante un factor reducir lo más posible su magnitud (profundidad del árbol).

La regla de evaluación de adaptación (fitness) se utiliza un criterio muy simple expresado en la ecuación (23) de la Sección III.8 y reexpresada como función de fitness de la siguiente forma

$$\boldsymbol{fitness} = \boldsymbol{Se} \cdot \boldsymbol{Sp} \cdot \boldsymbol{Sy} \qquad (28)$$

esta ecuación tiene como propósito maximizar tanto la sensibilidad como la especificidad y al mismo tiempo impulsar la reducción de la regla.
Sin embargo de acuerdo a (Ruiz, et al., 2004) buscar una prueba que sea altamente sensible y específica es poco frecuente y puede ser un tanto idealista pretenderla. En el enfoque de la presente tesis, la aplicación previa del mapa semántico a través de la red neuronal ayudó a preestablecer un conjunto

relativamente pequeño de atributos fuertemente relacionados entre ellos y el síndrome metabólico.

Producto de este análisis se halló un eje central que funciona como patrón de oro[14] sobre el síndrome metabólico: los análisis sobre colesterol total, triglicéridos y glicemia.

Al obtener los primeros resultados aplicando la ecuación (33) se observó que la restricción en el tamaño era un parámetro que pesaba mucho más que los otros dos al grado de provocar la pérdida de diversidad en la población En la mayoría de los casos la sensibilidad y la especificidad no están relacionadas; sin embargo, en el caso específico de los atributos donde los resultados no son dicótomos, los que generan puntos de corte dentro de las reglas (como es el caso de los resultados de los estudios arriba mencionados de sensibilidad y especificidad) están altamente relacionados, al subir uno, baja el otro. Este efecto produjo que al buscar subir ambos parámetros pesaran demasiado poco, mientras la simplicidad de la regla pesaba demasiado produciendo individuos extremadamente aptos pero de poco valor clínico por lo pequeño de la regla, que era demasiado simple, razón por la cual se optó por un esquema ponderado en cuanto a la función valuación de adaptación (fitness)

$$\boldsymbol{fitness} = (\propto \cdot \boldsymbol{Se}) + (\boldsymbol{\beta} \cdot \boldsymbol{Sp}) + (\mathbf{1} - \boldsymbol{\alpha} - \boldsymbol{\beta}) \cdot (\boldsymbol{Se} \cdot \boldsymbol{Sp}) \qquad (29)$$

en la cual se omite la simplicidad de la regla.

De acuerdo a (Ruiz, et al., 2004) cuando haya grandes costos o riesgos (incluidos los riesgos emocionales) en un resultado falso positivo, deben buscarse pruebas con una alta especificidad. Por otro lado, cuando los costos están más relacionados con los riesgos que es no detectar la enfermedad se requieren pruebas con alta sensibilidad. En el caso del síndrome metabólico la mayoría de las reglas descubiertas tienen una alta sensibilidad, por lo que, habiendo revisado

[14] De acuerdo a (Ruiz, et al., 2004) el patrón de oro es el examen que da certeza de la condición

con el especialista médico, se decidió ponderar más alta la especificidad de la regla.

La forma normal en la cual se presentan estos resultados es la tabla tetracórica como la presentada en la Figura 45, un resumen de las tablas correspondientes a cada una de las reglas descubiertas se presenta en el apéndice C.

Regla 1		**Concordó tipo de Síndrome Metabólico**	
Resultado de la prueba		Positivo	Negativo
	Positivo	183 (a)	105(b)
	Negativo	14 (c)	360 (d)
		197 (a+c)	465 (b+d)
Sensibilidad		92.9% (a/(a+c))	
Especificidad		77.4% (b/(b+d))	
Exactitud		82.0% ((a+c)/(a+b+c+d))	

Figura 45. Ejemplo de tabla tetracórica Regla 1 del síndrome metabólico obtenida por GP.

En la Tabla XI se presenta un resumen de la exactitud predictiva de las diferentes reglas generadas por GP. Se puede observar que a excepción de la regla 1 todo el conjunto de reglas tiene una exactitud predictiva por arriba del 90% que es el valor establecido como aceptable cuando se trata de una clasificación de riesgo totalmente inédita, como es el caso de la clasificación de riesgo del síndrome metabólico.

Tabla XI. Resumen de las exactitud de las reglas generadas por GP.

	Sensibilidad	Especificidad	Exactitud
Regla 1	93%	77%	82%
Regla 2	56%	97%	90%
Regla 3	75%	97%	93%
Regla 4	90%	96%	95%
Regla 5	75%	98%	92%

A continuación se enlista las reglas para determinar las 5 clases del síndrome metabólico:

- **REGLA 1**: Personas con peso por debajo de lo normal y, normal y sobrepeso con colesterol entre 165-226 mg/dl
- **REGLA 2**: : Personas con peso por debajo de lo normal y, normal y sobrepeso con Edad < 65 años y colesterol mayor a 180 mg/dl y triglicéridos mayor a 177 mg/dl
- **REGLA 3:** Personas con obesidad tipo 1 y 2 con colesterol total mayor a 195 mg/dl y triglicéridos mayor a 177 mg/dl
- **REGLA 4:** Personas con obesidad tipo 1, 2, 3 ,4 con colesterol mayor a 211 mg/dl o mujeres con sobrepeso y colesterol mayor a 242 mg/dl que fuman
- **REGLA 5:** Edad < 49 años , colesterol total menor a 165 mg/dl, triglicéridos menor a 218 mg/dl

Si se compara este conjunto de reglas con las obtenidas mediante el algoritmo C4.5 es muy clara la simplicidad de las reglas descritas por la GP en comparación con la complejidad de las descubiertas mediante C4.5.

Entonces de acuerdo al nivel de predicción alcanzado y la simplicidad del conjunto de reglas, de acuerdo a los criterios establecidos en los objetivos de la tesis, la GP cumplió con la expectativa de generar reglas más robustas desde el sentido de su simplicidad y exactitud.

Para evaluar el tercer criterio de los objetivos, la utilidad, de acuerdo a la fase 7 de la metodología se deben estructurar los elementos de coordinación y colaboración que van a apoyar a la toma de decisiones. Esto implica cuestionar el conocimiento descubierto con los criterios y conocimientos previos del especialista epidemiólogo.

V.7 Clasificación del síndrome metabólico de acuerdo a la visión socio-técnica de la metodología.

Al presentar al epidemiólogo los resultados obtenidos se compararon ambos conjuntos de reglas, cuestionándoseles como si los resultados fuesen producto de un estudio de cohorte normal. Los nuevos criterios que surgieron fueron los siguientes:

1. En el caso del síndrome metabólico es mejor una regla más específica, ya que mientras más lo sea podrá ayudar a determinar los casos que no son tan claros. En cambio sí es altamente sensible, será muy buena dentro de su predicción, pero probablemente no tanto (si tiene baja especificidad) para identificar a todos los afectados.
2. Adicionalmente el número de verdaderos negativos es especialmente significativo dentro de la especificidad, ya que este indicador nos dice realmente cuantos individuos fueron identificados como que no pertenecen a un clase dada y en realidad no pertenecen. Siendo una clasificación inédita no hay forma de comprobar que se está clasificando mal a un paciente por lo que es importante que este indicador sea alto.

Al analizar los casos de GP se encontró que hay dos reglas 1 y la 4 de la (GP) que de acuerdo al criterio del médico pueden ser cuestionadas, aun cuando su fitness

es alto: la regla 1 de GP tiene 61 verdaderos negativos, mientras que la de C4.5 tiene 105 un 59% más; en el caso de la regla 4 de GP tiene un 80.5% de especificidad mientras la de C4.5 tiene 89.7.

En ambos casos al comparar el contenido de la regla se encontró que la regla de GP contiene a la regla de C4.5, en otras palabras la regla de GP es una generalización de la de C4.5.
Por lo cual el especialista decidió tomar las reglas 1 y 4 de C4.5, y las otras 3 de la programación genética. Es importante recalcar que la GP cumplió con los objetivos previstos generando reglas más simples y con una alta exactitud predictiva, sin embargo en el caso de las reglas 1 y 4 fueron criterios muy particulares que se dieron por el enfoque socio-técnico con el cual se trabajó.

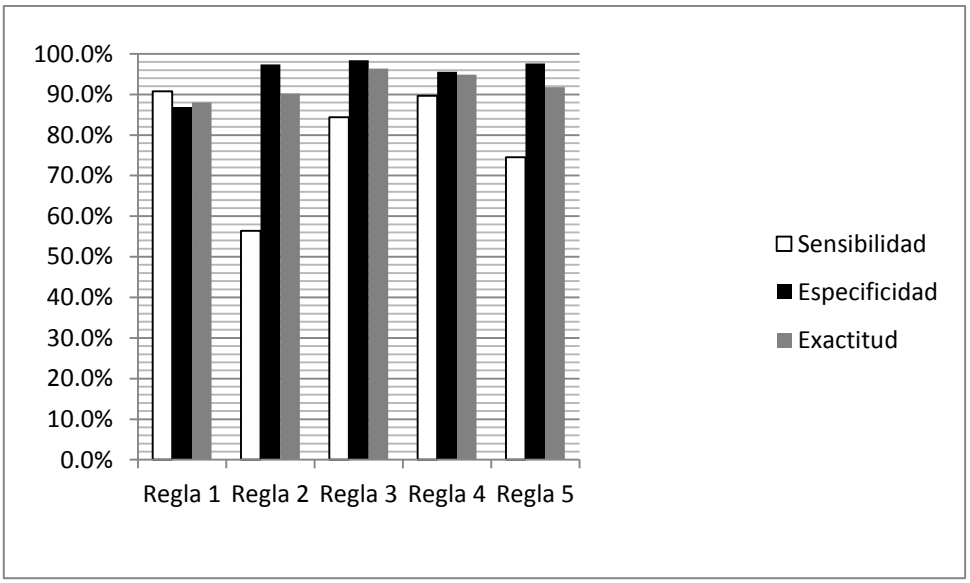
En la Figura 46 se presenta un resumen de los resultados de las diferentes reglas habiendo hecho los ajustes antes mencionados. En esta Figura se puede observar que únicamente en el caso de la regla 2 se ve castigada la sensibilidad de la regla, sin embargo la exactitud predictiva es relativamente buena. También se puede observar como en general todas las reglas tienen una exactitud por arriba del 90%, que en un estudio de cohorte normal es el mínimo requerido para que un clasificación totalmente inédita tenga una relevancia, lo que nuevamente es un criterio socio-técnico.

Habiendo cubierto toda esta serie de requisitos para dar por útil la clasificación obtenida, se procedió, junto con el epidemiólogo, a analizar qué tipo de pacientes ubica cada regla. Producto de este análisis se identificó la siguiente tipificación:

- que hay 3 clases[15] de síndrome metabólico que refieren a personas sin problemas de obesidad:

[15] Las clases no se reenumeraron, sino que les asignó el número de conglomerado de la Figura 44.

- Clase 5: Personas que presentan problemas leves de dislipidemia principalmente con alteraciones en triglicéridos
- Clase 1: Personas que pueden llegar a presentar problemas de sobrepeso con un nivel de dislipidemia moderado y que incluso pueden presentar alteraciones en azúcar en la sangre (glicemia).



	Sensibilidad	Especificidad	Exactitud
Regla 1	91%	87%	88%
Regla 2	56%	97%	90%
Regla 3	84%	98%	96%
Regla 4	90%	96%	95%
Regla 5	75%	98%	92%

Figura 46. Resumen de la exactitud predictiva de las reglas.

- Clase 2: Personas que pueden llegar a presentar problemas de sobrepeso y que ya presentan problemas fuertes de dislipidemia, ya sea afectados en colesterol y/o triglicéridos.

- Las siguientes 2 clases de síndrome metabólico se refieren a personas con problemas de obesidad en diferentes niveles:
    - Clase 3: Personas con obesidad tipo I y II con diferentes niveles de dislipidemias
    - Clase 4: Personas con obesidad tipo I y II de más de 47 años y personas con obesidad mórbida.

Al asignar un tipo a cada clase se pudo identificar una relación jerárquica entre las diferentes clases, que se representa en la Figura 47.

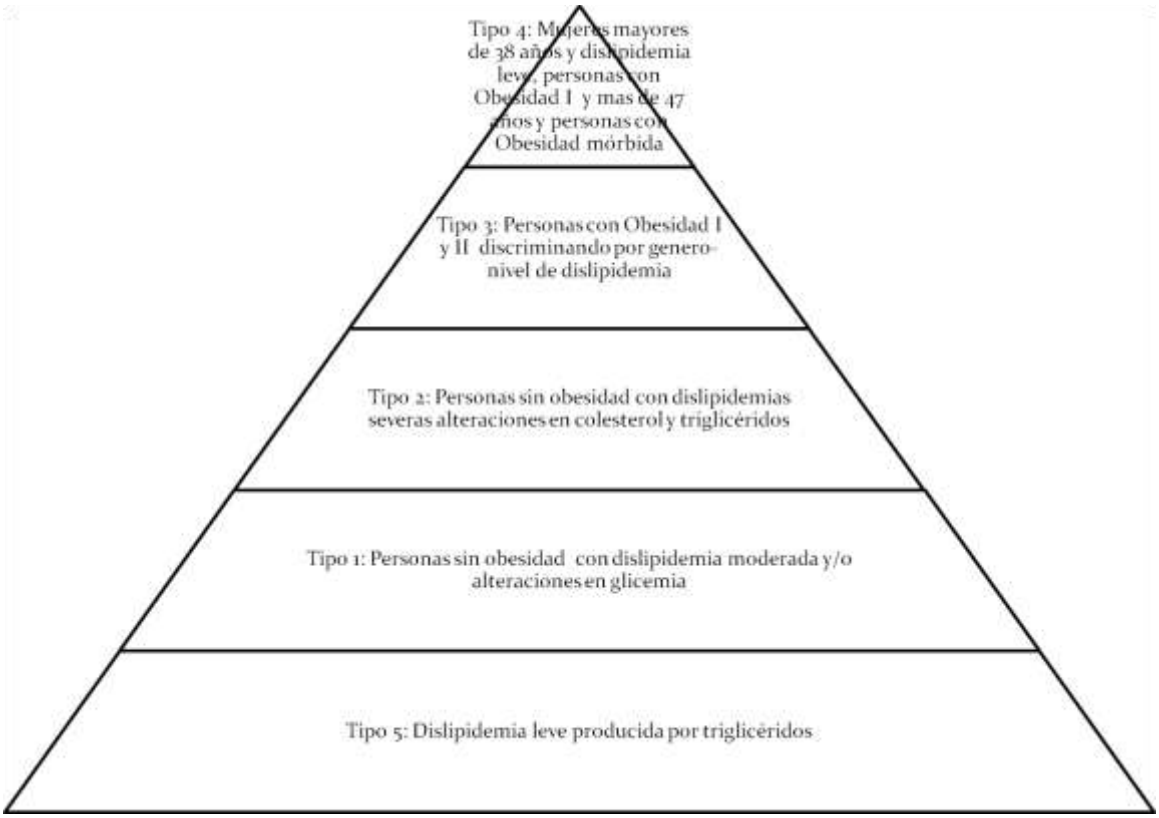


Figura 47. Clasificación del síndrome metabólico.

La aportación que representa esta clasificación jerárquica, es que al resolver el problema de clasificación que implicaba el apoyo que requiere el médico, se han podido identificar una serie de etapas que sigue la evolución del síndrome desde sus manifestaciones más tempranas hasta el establecimiento de la comorbilidad. Esta aportación permite establecer una traza de cómo antes de la comorbilidad se establezca hay una evolución identificable a través de indicadores que normalmente se cuentan dentro de los estudios que de forma común se le hacen a un paciente. Desde el punto de vista del riesgo cardiovascular, agrega una perspectiva nueva, la evaluación del nivel de riesgo desde el punto de vista del síndrome metabólico.

En el Apéndice C se presenta un detalle del conjunto de reglas descubiertas.
En la Figura 48 se presentan dos ejemplos de reglas de la Clasificación de Riesgo, en la Figura 48(b) se presenta la regla que determina la pertenencia a la clase 3 y la (b) para la clase 2. Se puede observar la diferencia en lo específico de ambas reglas la clase 2 es más general y abarca a personas sin obesidad con problemas de dislipidemias, indicado por los niveles arriba de los límites de colesterol y triglicéridos. La clase 3 corresponde a un subgrupo de personas con obesidad tipo 1 y 2, con edad mayor a los 38 años y algunas consideraciones por género donde se ve claramente que en el caso de las mujeres los problemas de dislipidemias les coloca en este grupo sin mayores consideraciones.

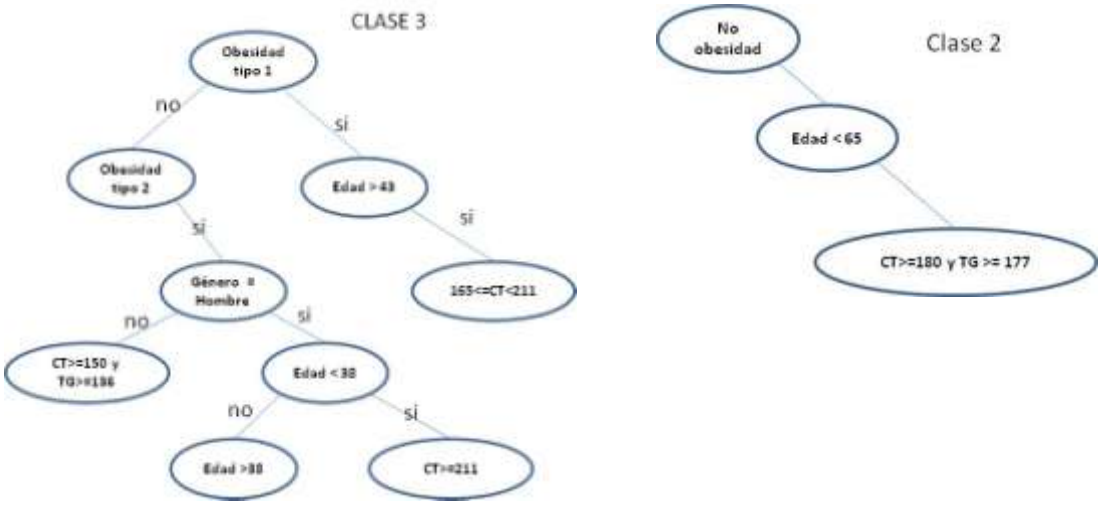


Figura 48. Muestras de reglas para determinar (a) clase 3 y (b) clase 2 del síndrome metabólico.

En el caso de las reglas que determinan la clase 3 tienen una sensibilidad del 84.4%, una especificidad de 98.4% y una exactitud predictiva del 96.4%. Y la regla está expresada de la siguiente forma:

Obesidad tipo 1.

- Mayores de 43 años y Ct entre 165-211 mg/dl
- Ct mayor-igual a 211

Obesidad tipo 2

- Hombres menores de 38 años y Ct mayor-igual a 211 mg/dl
- Hombres mayores de 38 años
- Mujeres con Ct mayor-igual a 150 mg/dl y Tg mayor-igual a 136 mg/dl

En el caso de las reglas que determinan la clase 2 tienen una sensibilidad del 56.4%, una especificidad de 97.4% y una exactitud predictiva del 90.4%. Y la regla está expresada de la siguiente forma:

- Personas sin obesidad con edad menor a 65 y Ct mayor-igual a 180 mg/dl y Tg mayor-igual a 177 mg/dl.

V.7.1 Relaciones topológicas entre las diferentes clases del síndrome metabólico.

Una contribución importante de los resultados obtenidos habiendo aplicado los mapas semánticos, es la preservación de las características topológicas de la información. Se puede observar en la Figura 49 la colindancia que existe entre los diferentes tipos de síndrome metabólico. Por ejemplo el tipo 5 tiene colindancia con el tipo 1 y el tipo 4, estas colindancias son significativas ya que de acuerdo a los presentado en el Capítulo III son producto del parecido que hay entre los 3 tipos. Lo cual implica que un sujeto del tipo 3 evoluciona hacia un tipo 1 y tipo 4 pero no a un tipo 2 y 3.

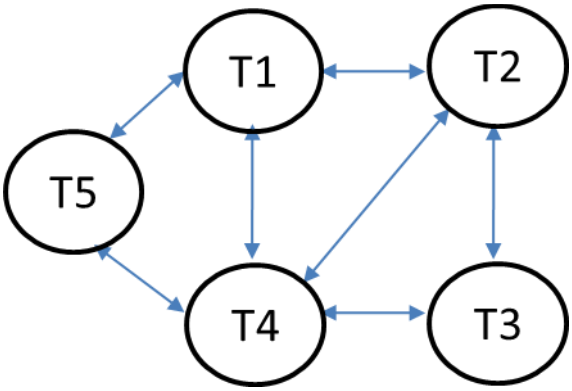


Figura 49. Grafo de las relaciones entre las diferentes reglas descubiertas.

Una forma de analizar las relaciones topológicas obtenidas es mediante un grafo el cual permite observar las diferentes transiciones entre las clases. La Figura 49 presenta dichas transiciones donde T_n corresponde a la $Regla_n$. Producto de este análisis es la forma en que un sujeto que tiene cierto tipo de síndrome

V.7.2 Apoyo a las decisiones clínicas, análisis de transiciones.

Tabla XII. Cuadro resumen de apoyo a la toma de decisiones clínicas con respecto al síndrome metabólico y su progresión o regresión según sea el caso.

Origen	Evoluciona a
Tipo 5: Personas con dislipidemias leves causadas por desajustes en los niveles de triglicéridos	Tipo 1, Tipo 4 (sin incluir obesidad mórbida).
Tipo 1: Personas sin sufrir obesidad con problemas de dislipidemias declarados	Tipo 2, Tipo 4 (sin incluir obesidad mórbida), Tipo 5
Tipo 2: Personas sin obesidad con dislipidemias severas con alteraciones en colesterol y triglicéridos	Tipo 1, Tipo 3, Tipo 4 (sin incluir obesidad mórbida)
Tipo 3: Personas con obesidad I y II discriminando por género-nivel de dislipidemia	Tipo 2 y Tipo 4 (sin incluir mujeres de más de 38 años)
Tipo 4: Mujeres de más de 38 años con problemas de dislipidemias o personas con obesidad I y edad > 47 años o personas con Obesidad Mórbida	Tipo 1, Tipo 2, Tipo 3, Tipo 5

De acuerdo a la Figura 49 se obtuvo la Tabla XII de apoyo a la toma de decisiones clínicas relacionadas con el síndrome metabólico, donde se resume la progresión que puede tener el síndrome metabólico y que implica el incremento en el riesgo de padecer enfermedades relacionadas con él.

Otra parte importante de los resultados obtenidos es la relevancia de los atributos seleccionados, existen atributos que de acuerdo a los análisis practicados tienen suma importancia para determinar si un individuo pertenece o no a un conglomerado de datos, de igual forma hay atributos que no son relevantes en este aspecto.

V.8. Resumen.

En el presente Capítulo se ha demostrado que tanto las KNN y su producto los mapas auto organizados SOM, así como el descubrimiento de conocimiento a través de la GP son herramientas muy útiles para un estudio de cohorte sobre el síndrome metabólico.

Las reglas descubiertas vienen a confirmar las ideas a priori que los expertos tienen sobre las causas que provocan la problemática relacionada con el síndrome metabólico y a la vez representan un aporte en el sentido que se generó la primera clasificación formal del síndrome metabólico.

Adicionalmente el manejo de ambas técnicas conjuntamente, red neuronal y GP, han permitido obtener una serie de reglas robustas con un nivel aceptable de exactitud predictiva.

Producto de las técnicas utilizadas se obtuvo una serie de transiciones entre las diferentes reglas que también son un aporte en el manejo del síndrome metabólico. Las reglas y las transiciones obtenidas son la base para la elaboración

de un programa de apoyo a la toma de decisiones clínicas relacionadas con el síndrome metabólico, que se presenta en el siguiente capítulo.

Capítulo VI

Diseño e implementación del CDSS

VI.1 Introducción.

La fase 8 de la metodología establece que es en esta fase que se desarrolla el soporte adecuado para el problema de toma de decisiones que se está resolviendo.

De acuerdo a (Tan, et al., 1998) entre las diversas formas que puede llegar a tomar un HDSS, están las *"herramientas computacionales para la consulta de un paciente-específico – por ejemplo, sistemas de diagnóstico experto diseñados para proveer consejos o sugerencias en el diagnóstico diferencial".*

Un elemento fundamental en el diseño de este tipo de sistemas es que no reemplazan al especialista, ni tampoco toman la forma oracular de consulta donde el especialista es un espectador pasivo.

A este tipo de sistemas se les define como CDSS. Tienen las características de proveer modelos teóricos de las decisiones clínicas y métodos estadísticos de reconocimiento de patrones, entre otras más. Además deben tomar en cuenta los diferentes niveles de conocimiento, que pueden presentar los usuarios sobre el problema en particular.

Para el diseño del sistema usaremos el Lenguaje Unificado de Modelado (Unified Modeling Language-UML por sus siglas en inglés) usando un proceso de desarrollo iterativo.

En la Sección VI.2 se presentan los requerimientos de usuario en forma de los casos de uso que definen la funcionalidad del sistema.
En la Sección VI.3 se presentan se presenta el diseño arquitectónico del SATDSmet.
Las secciones VI.4 y VI.5 presentan el diseño del SATDSmet; la primera sección presenta el diagrama de clases del SATDSmet; y en la otra se presenta el comportamiento del sistema mediante los diagramas de secuencia.
En Sección VI.5 se presenta la implementación del sistema para entonces concluir en la Sección VI.6 con un resumen del capítulo.

VI.2 Requerimientos del Sistema de Apoyo a la Toma de Decisiones del manejo del síndrome metabólico.

De acuerdo a la metodología en las fases 3 y 4 se identificaron los componentes del problema de decisión y los objetivos de la decisión. Los modelos y la información obtenida durante estos análisis permiten el registro y la validación de una serie de requerimientos funcionales y no funcionales establecidos y validados por el usuario.
En este análisis se encontró que son dos los procesos de toma de decisiones fundamentales en el manejo del síndrome metabólico en medicina familiar:

1. En el manejo que hace el médico familiar al determinar el nivel de riesgo del paciente con síndrome metabólico.
2. Cuando se establece un “tratamiento” que se relaciona con el síndrome metabólico”.

Todos los demás manejos que realizan los otros actores involucrados en el manejo del síndrome metabólico dependen de estos dos momentos de toma de decisión.

Los principales requerimientos funcionales detectados en estos dos procesos de toma de decisiones son los siguientes:

1. Cuando se está realizando la valoración del nivel de riesgo se requiere una herramienta (clasificación) que ayude a evaluar el nivel de riesgo del paciente que lo lleve a desarrollar enfermedades como DM2 y los eventos cardiovasculares cerebrales (EVC).
2. Cuando un paciente no presenta signos que por la valoración a simple vista se puedan detectar (obesidad, acantosis nigricams[16], falta de aliento, etc.) el sistema debe presentar información relevante de acuerdo a la condición previa del paciente.
3. La información debe ser útil para el diagnóstico (conocimiento nuevo) y en base a ese conocimiento indicar acciones a tomar.
4. El sistema debe guardar la historia del paciente y ayudar a determinar que cursos de acción a tomar para que la enfermedad no evolucione empeorando la condición del paciente.
5. En el sistema deben aparecer todas las herramientas que el médico familiar necesite al momento de tomar una decisión además de apoyarle con conocimiento específico de cómo se debe manejar a cada paciente individualmente.
6. Debe ser muy simple en la forma en que se presente la información.
7. El sistema debe permitir personalizar la información.
8. El sistema debe presentar la información de cómo ha evolucionado el paciente de tal forma que si no mejora su condición, se puedan tomar decisiones de enviarlo a algún tipo de tratamiento no farmacológico.
9. El sistema debe incluir una sección especializada para el epidemiólogo donde se valúe el riesgo en grupos de pacientes y que presente indicadores de la forma en que el riesgo está distribuido en la población.

[16] Una condición de la piel caracterizada por la aparición de zonas oscuras aterciopeladas principalmente en la piel de los pliegues de las axilas (Amy, 2007)

Primeramente el software se construye para satisfacer las necesidades del cliente, en ellos se definen los problemas y necesidades del usuario. Un requerimiento es *"cualquier función, restricción o propiedad que debe ser provista, cumplida o satisfecha para cubrir las necesidades del sistema propuesto por el usuario"* (Abbott, 1983).

En UML los requerimientos del sistema se establecen a través de casos de uso, entendiendo como un caso de uso es *"una descripción de un conjunto de secuencia de acciones, incluyendo variaciones, que un sistema lleva a cabo y que conduce a un resultado observable de interés para un actor determinado"* (Jacobson, et al., 2000).

La Figura 50 muestra los casos de uso del CDSS creado al cuál se le nombró Sistema de apoyo a la Toma de Decisiones del Síndrome Metabólico (SATDSMet). Se puede observar que hay dos actores, los cuales corresponden a los dos diferentes niveles de especialización identificados entre usuarios (el médico epidemiólogo y el médico familiar).

Tanto el médico familiar como el médico epidemiólogo pueden atender a un paciente en particular, en este sentido ambos médicos juegan el rol de tomadores de decisiones, sin embargo el especialista epidemiólogo tiene otros casos de uso (representados por los óvalos) específicos de su rol cómo epidemiólogo.
El caso de uso *evaluar riesgo* es el más importante para el tomador de decisiones al momento de la consulta, ya que en él se tratan la mayor parte de los requerimientos funcionales descritos por el usuario. Para poder evaluar el riesgo el sistema debe proporcionar conocimiento nuevo que apoye en el manejo de la toma de decisiones, el caso de uso *clasificar un paciente* integra la nueva clasificación de riesgo obtenida en el proceso KDD, de tal forma que este caso de uso se incluye dentro del caso de uso *evaluación de riesgo*. El caso *evaluación de riesgo* presenta el conocimiento nuevo en forma de información personalizada

sobre la situación del paciente, su posible evolución y las acciones que se pueden seguir para apoyarle. El caso de uso *consultar guías clínicas* tiene como propósito aportar toda la información necesaria para la toma de decisiones.

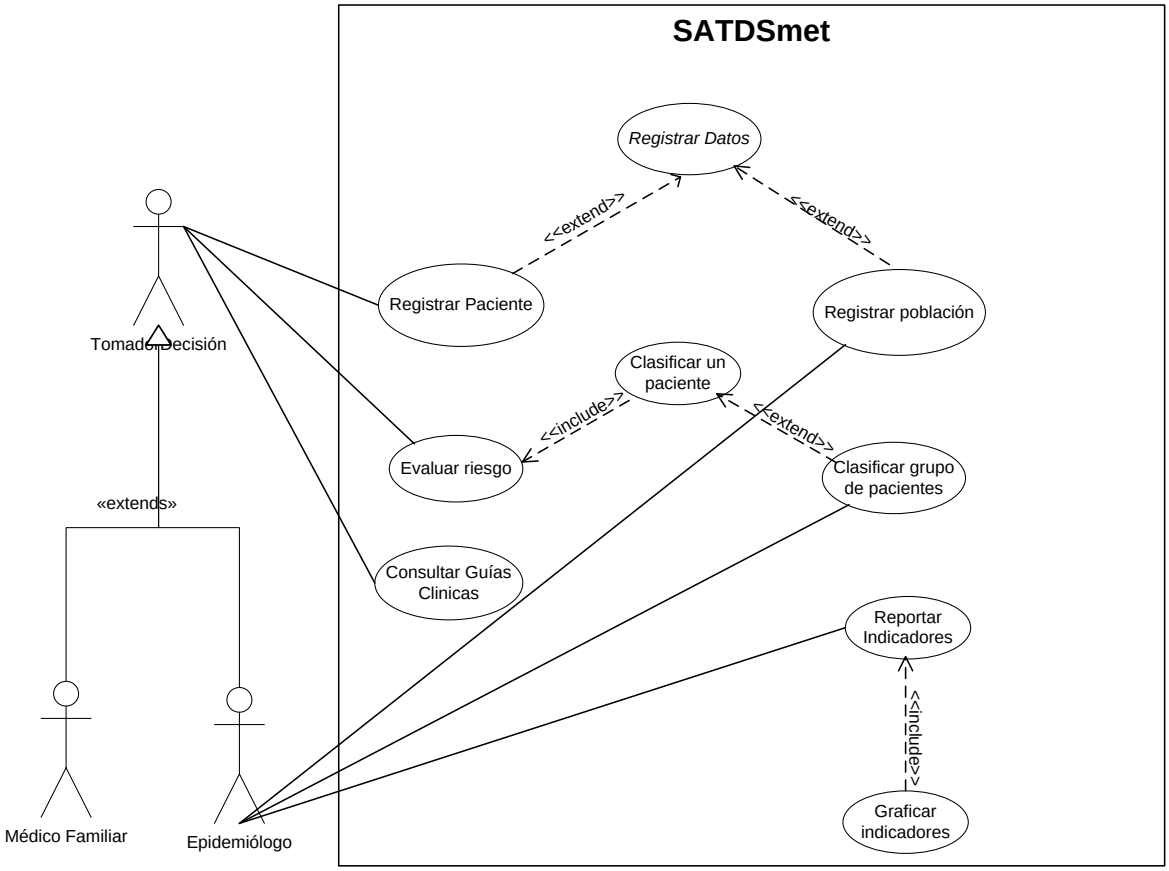


Figura 50. Diagrama de casos de uso del SATDSmet.

Independiente del manejo antes descrito el especialista epidemiólogo tiene otros requerimientos, tanto el caso de uso *registrar población* como clasificar grupo de pacientes, son casos especiales del caso *registro de datos*, como ya se mencionó el epidemiólogo tiene acceso a todos los casos de uso antes descritos en su rol de tomador de decisiones, y adicionalmente en el caso de uso *reportar indicadores* tiene acceso a otro tipo de herramientas estadísticas para evaluar poblaciones de pacientes.

Enseguida se presenta la descripción de los casos de uso principales (en el Apéndice F se encuentre el resto).

Caso de uso: *Evaluar riesgo*

Actores: Médico Familiar, Médico Epidemiólogo

Propósito: Establecer el nivel de riesgo para cierto nivel de Síndrome Metabólico

Descripción:

El caso de uso extiende al caso de uso *Registrar Datos Actualizados del Paciente* e inicia cuando al capturar la información del paciente el médico solicita se evalúe el riesgo de padecer el Síndrome Metabólico. El SATDSmet primero clasifica el nivel en que el paciente padece el síndrome, por lo que el caso incluye el caso de uso *Clasificar un Paciente*. Habiendo obtenido la clasificación el SATDSmet permite 4 acciones, las cuales se pueden presentar todas juntas o por separado:

1. Presentar la clasificación del paciente. De acuerdo a la clasificación de riesgo del síndrome metabólico y en base a los datos capturados el sistema deberá clasificar al paciente y los datos más relevantes que le llevaron a dicha clasificación (edad, sexo, IMC, colesterol total, triglicéridos, glicemia, fuma, bebe, tiene diabetes tipo 2).
2. Presentar el nivel de riesgo que representa la clasificación obtenida de acuerdo a la clasificación jerárquica de los diferentes tipos de síndrome metabólico.
3. Presentar las posibles acciones a seguir según el grupo que le corresponde dentro de la clasificación de riesgo. El sistema recomienda una serie de acciones para detener o revertir los efectos del síndrome.
4. Presentar la evolución que puede seguir el paciente al pasar de un nivel a otro nivel dentro de la clasificación, de acuerdo al grafo de colindancia topográfica obtenido durante la clasificación.

El caso de uso finaliza en el momento que el usuario cierra el sistema.

Caso de uso: *Clasificar al paciente***.**

Actores: Médico Familiar, Médico Epidemiólogo.

Propósito: Establecer el nivel de Síndrome Metabólico que tiene el paciente de acuerdo a la clasificación de riesgo obtenida.

Descripción: El caso de uso inicia cuando al evaluar el riesgo se debe clasificar al paciente de acuerdo a la clasificación de riesgo del síndrome metabólico para un paciente en particular. De acuerdo a los datos del paciente (género, peso, talla, IMC, colesterol total, triglicéridos, glicemia, si fuma o no, si bebe o no, si padece DM2).

Otra opción es cuando el médico epidemiólogo evalúa un grupo de personas, y se requiere evaluar uno a uno cada paciente. El SATDSmet clasifica uno a uno todos los pacientes determinando la clasificación del síndrome metabólico de cada paciente, por lo que el caso de uso *clasificar grupo de pacientes* es una extensión de este caso de uso.

Caso de uso: *Clasificar grupo de pacientes*.

Actores: Médico Epidemiólogo.

Propósito: Evaluar el nivel de riesgo de un grupo de pacientes.

Descripción:

El caso de uso inicia cuando el médico epidemiólogo ha finalizado de capturar o registrar un grupo de pacientes y quiere que el SATDSmet le ayude a clasificar el tipo de síndrome metabólico. El sistema debe evaluar individualmente cada paciente, por lo que el caso extiende al caso de uso *clasificar un paciente*.

Caso de Uso: *Consultar Guías Clínicas.*

Actores: Médico Familiar, Medico Epidemiólogo.

Propósito: Permitir la consulta de las guías clínicas sobre HTA y DM2.

Descripción: El caso de uso inicia cuando el médico desea consultar las guías clínicas vigentes por alguna duda que le surge. El sistema deberá presentar una opción donde el médico pueda escoger entre la guía de HTA y la DM2. El

SATDSmet abrirá en línea la guía seleccionada para que el médico la pueda consultar.

En la siguiente sección se trata el comportamiento del SATDSmet desde dos diferentes vistas, la dinámica (diagrama de secuencia) y la estática (diagrama de clases).

VI.3 Arquitectura del SATDSmet.

De forma análoga a como la arquitectura en construcción define una serie de elementos que guiarán las construcción física de un edificio la arquitectura de software preestablece los elementos básicos de cómo estará constituido un sistema.

La arquitectura de un sistema como *"conjunto de decisiones significativas acerca de la organización de un sistema software, la selección de los elementos estructurales a partir de los cuales se compone el sistema, y las interfaces entre ellos, junto con su comportamiento, tal y como se especifica en las colaboraciones entre esos elementos, la composición de estos elementos estructurales y de comportamiento en subsistemas progresivamente mayores…"* (Jacobson, et al., 2000).

En la Figura 51 se presenta la arquitectura del SATDSmet. Los rectángulos sólidos representan los nodos, representan fuentes de procesado o de cómputo; rectángulos con copete, mecanismo de agrupamiento; rectángulos representan componentes; parte física y reemplazable del sistema; círculos y semicírculos atados a un componente, los cuales representan las interfaces de los componentes; y algunos otros íconos de uso común, pantallas y archivos.

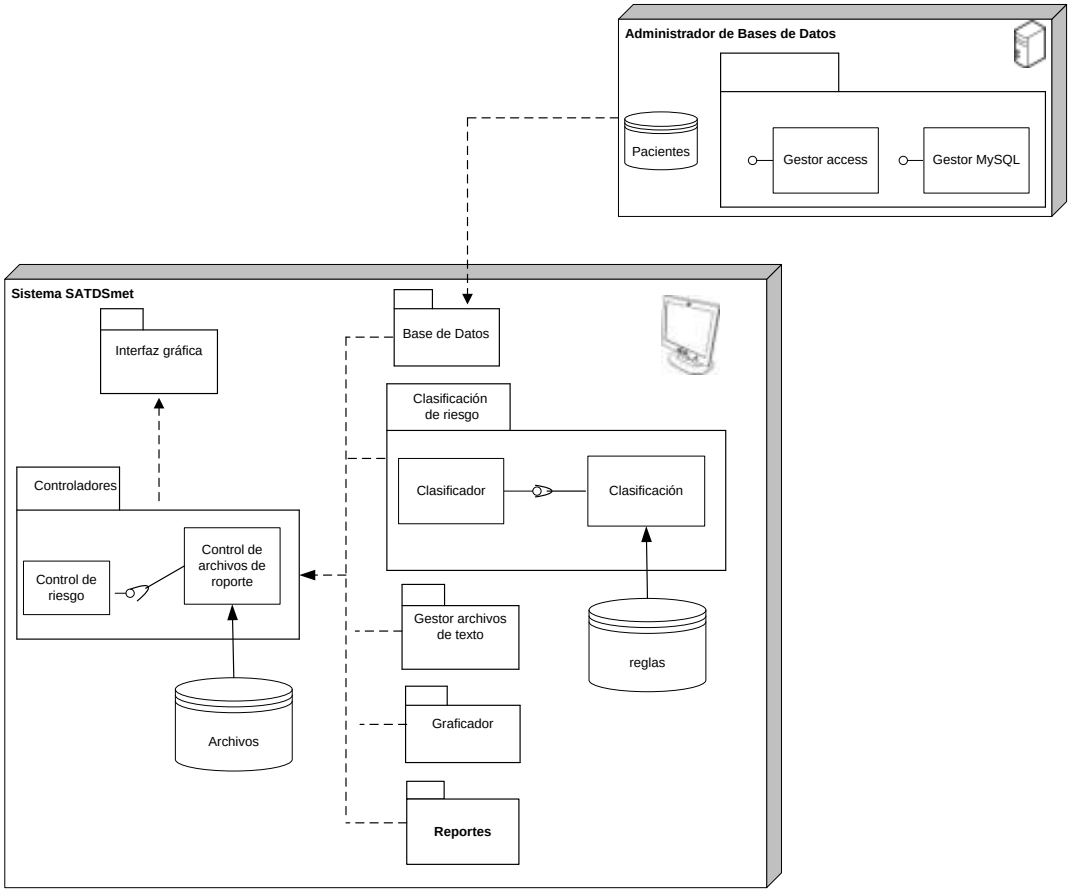


Figura 51. Arquitectura del SATDSMet.

Estea diseño corresponde a la separación natural de las diferentes partes que componen el SATDSmet.

- **Sistema SATDSmet** contiene una serie de paquetes que se dividen en vistas (interfaz de usuario), controladores, y paquetes correspondientes al modelo del sistema. En este último grupo se encuentran los de manejo de la base de datos, archivos de soporte, la clasificación de riesgo, graficadores y reportes.
- **Administrador de Bases de Datos** El componente principal de este nodo es la base de datos de pacientes del SATDSMet. El nodo se compone de

dos gestores de bases de datos (Access y MySQL). Access es compatible con el paquete SPSS que utiliza el epidemiólogo. Mientras que MySQL permite la instalación en cualquier equipo, ya que es de libre distribución.

Se seleccionó manejar una base de datos principalmente para el registro histórico de la información de los pacientes. En este sentido las preocupaciones principales que se tuvieron al seleccionar las herramientas a trabajar, fueron los recursos de cómputo con los que cuentan los diferentes actores detectados (médico familiar y epidemiólogo).

VI.4 Diseño (Diagramas de secuencia).

La vista dinámica de un sistema presenta las tareas que son realizadas por medio de objetos, los que interactúan entre ellos pasándose mensajes. Los diagramas de secuencia son los que principalmente se utilizan para modelar esta parte, en ellos principalmente se presentan:

- El comportamiento entre los objetos, las entidades y los objetos de control.
- Muestran las interacciones que ocurren a través del tiempo entre los objetos asociados con cada una de sus clases.
- Establece la distribución de las operaciones entre clases.
- Y se realiza un diagrama de secuencia por caso de uso.

Un diagrama de secuencia consiste de una serie de objetos, los rectángulos que aparecen en secuencia en la parte superior; una línea de vida del objeto, los rectángulos largos debajo de cada objeto; una serie de mensajes por los cuales se comunican los objetos, las flechas que aparecen en una secuencia de tiempo que se lee de izquierda a derecha y de arriba hacia abajo.

En la Figura 52 se presenta el diagrama de secuencia del caso de uso Evaluar Riesgo, donde el rol tomador de decisiones, ya sea médico familiar o médico epidemiólogo, evalúa el riesgo que representa para un paciente el síndrome de acuerdo a la clasificación de riesgo del síndrome metabólico establecida.

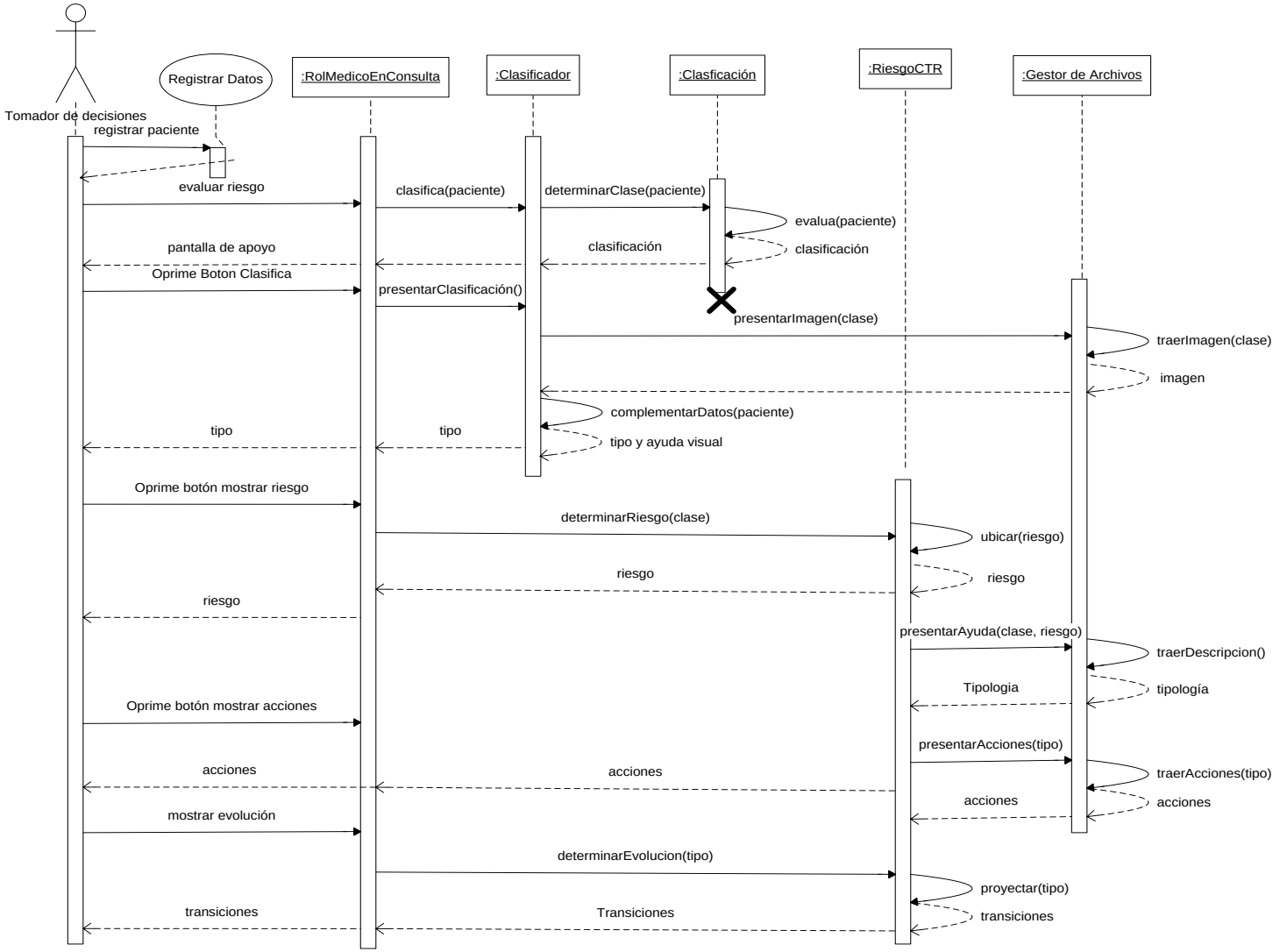


Figura 52. Diagrama de secuencia del caso de uso Evaluar Riesgo.

En el diagrama de secuencia del caso de uso *evaluar riesgo* (Figura 52) el médico solicita al sistema el apoyo para clasificar al paciente de acuerdo al nivel que presenta del síndrome metabólico, y al mismo obtener una serie de apoyos que le permitan decidir entre diferentes tipos de acciones a seguir.

1. El Tomador de Decisiones inicia la interacción con el sistema al seleccionar oprimir el botón de evaluar el riesgo.
2. La RolMedicoEnConsulta interactúa con el clasificador solicitando se determine la clasificación a la cual pertenece el paciente.
3. El Clasificador accede a la clasificación, obteniendo el indicador de a qué clase pertenece el paciente.
4. El Tomador de Decisiones selecciona oprimir el botón Clasifica.
5. La RolMedicoEnConsulta solicita al clasificador se presente la clasificación del paciente.
6. El Clasificador solicita al gestor de archivos la imagen que le permitan explicar al paciente cuál es su estado.

El gestor de archivos de apoyo accede a la colección de imágenes con que cuenta el sistema de acuerdo al género, tipo y clase de síndrome metabólico detectado y devuelve la imagen correspondiente al clasificador.

1. Habiendo obtenido la imagen el Clasificador le adjunta los datos más relevantes del paciente que le fueron pasados por parámetro
2. El clasificador integra toda esta información y la entrega a la RolMedicoEnConsulta quien se encarga de presentarla al tomador de decisiones.
3. El Tomador de Decisiones selecciona oprimir el botón mostrar riesgo.
4. El RolMedicoEnConsulta solicita al controlador RiesgoCTR determine el riesgo

5. El controlador RiesgoCTR solicita al Gestor de Archivos determina la tipología y la devuelve el RolMedicoEnConsulta quien se encarga de presentar la información al tomador de decisiones.
6. El Tomador de Decisiones oprime el botón mostrar acciones.
7. El RolMedicoEnConsulta solicita al controlador RiesgoCTR determine las acciones a seguir con el paciente.
8. El controlador RiesgoCTR solicita al Gestor de Archivos obtenga las acciones a seguir y las devuelve al RolMedicoEnConsulta quien se encarga de presentar la información al tomador de decisiones..
9. El Tomador de Decisiones oprime el botón de determinar la posible evolución del paciente.
10. El RolMedicoEnConsulta solicita al controlador RiesgoCTR se determinen la posible evolución del paciente.
11. El controlador RiesgoCTR proyecta la evolución del paciente mediante el tipo y la devuelve al RolMedicoEnConsulta quien se encarga de presentar la posible evolución del paciente al tomador de decisiones.
12. En el caso de que el Tomador de Decisiones seleccione ver toda la información junta, el RolMedicoEnConsulta realizan los pasos 10 a 18 de forma automática sin esperar se opriman los botones

En los pasos 10 y 18 el sistema determina cuales son los caminos a seguir de acuerdo a la clasificación y se le presentan diversas alternativas al médico, el conforme a su conocimiento y experiencia determina el curso de acción a seguir.

En la Figura 53 se presenta diagrama de secuencia del caso de uso evaluar grupo de pacientes, que a continuación se detalla:

1. El médico epidemiólogo después de acceder a un archivo de pacientes selecciona la opción de determinar el tipo de síndrome metabólico que tiene cada uno de los pacientes.

2. El RolMedEpi solicita al clasificador se clasifiquen todos los individuos del archivo que se seleccionó
3. El clasificador controla el determinar uno a uno la clasificación de cada uno de los pacientes de la población. El clasificador pasa a la clasificación cada paciente.
4. La Clasificación evalúa al paciente y establece la clasificación correspondiente.
5. El médico epidemiólogo solicita los reportes
6. El RolMedEpi solicita al reporteador genere los diferente reportes.
7. El reporteador genera los reportes de exactitud predictiva y las tablas tetracóricas
8. La RolMedEpi se encarga de presenta los archivos generados.
9. El médico epidemiólogo solicita una gráfica de la clasificación de lapoblación
10. El RolMedEpi solicita al Clasificador genere la gráfica por tipo de síndrome metabólico.
11. El graficar genera la gráfica correspondiente

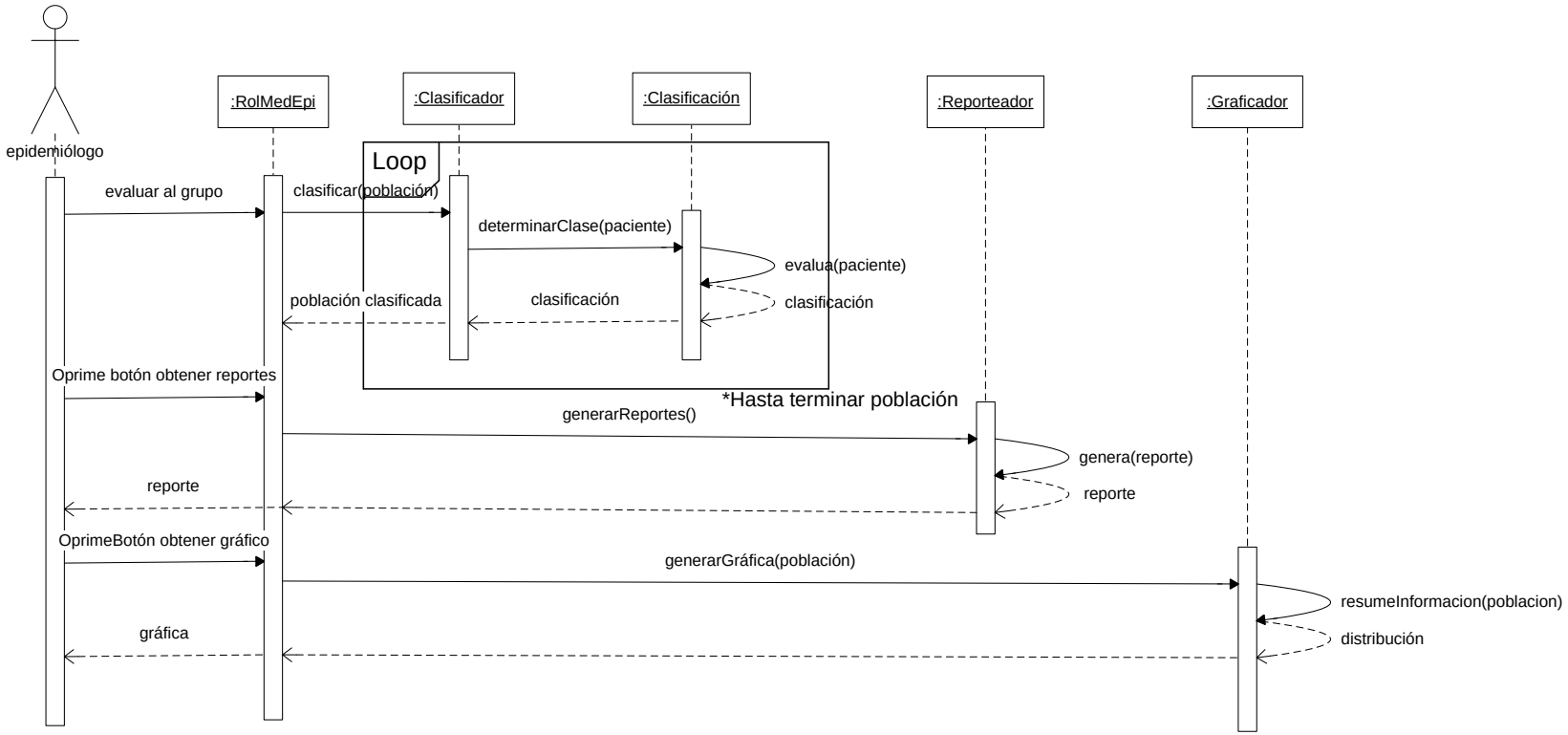


Figura 53. Diagrama de secuencia del caso de uso Evaluar Grupo de Pacientes.

VI.5 Diseño (Diagramas de clases).

Los diagramas de clase presentan los boques de construcción básicos con los cuales se construye cualquier sistema orientado a objetos: las clases.

En la Figura 54 se presenta al diagrama de clases del SATDSmet, en él que primeramente se observan dos clases RolMedEnConsulta y RolMedEpi ambas diferencian la funcionalidad que el sistema proporciona a los dos actores principales que se identificaron en los casos de uso.

De acuerdo a esta diferenciación de las funciones por rol, la clase RolMedEnConsulta establece asociaciones con clases que cubren la funcionalidad del caso de uso *registrar paciente*. La clase RiesgoCtr se encarga de proporcionar la información de riesgo para un paciente en particular, la clase Paciente contiene los datos del paciente que han sido registrados en la Base de Datos, la clase Conexión se establece el diálogo con la base de datos. Por lo que la funcionalidad que representa esta clase corresponde a la que se requiere para la toma de decisiones cuando se está atendiendo un paciente en particular. Hay que recordar que de acuerdo a lo presentado en la Figura 50 el actor epidemiólogo es una especialización del rol Tomador de Decisiones y que en este sentido ambos médicos (familiar y epidemiólogo) hacen uso de la funcionalidad de la clase RolMedEnConsulta.

Por su parte la clase RolMedEpi establece relaciones con las clases Reporte y Gráfica las cuales contienen la funcionalidad que requiere el epidemiólogo al estar evaluando una población, y esta información queda registrada precisamente en la clase Población.

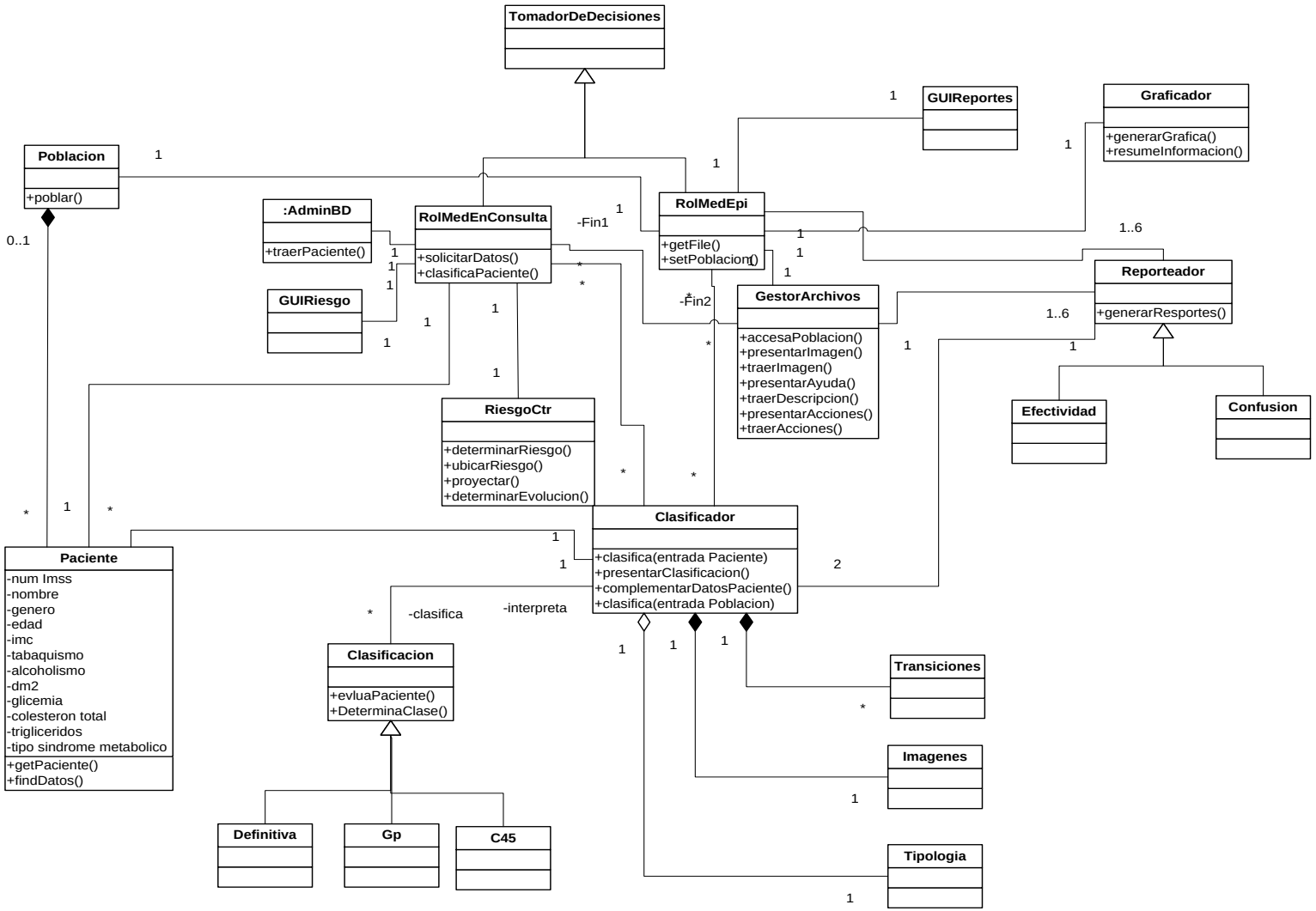


Figura 54. Diagrama de clases del SATDSmet.

En ambos casos las clases RolMedEnConsulta y RolMedEpi establecen relación con la clase Clasificador. La clase Clasificador establece una relación de colaboración con la clase Clasificación la cual clasifica de acuerdo a la clasificación de riesgo encontrada para el síndrome metabólico, y el Clasificador se encarga de interpretar dicha clasificación mediante tres acciones:

- Clasificar, pide a la clasificación clasifique a un(os) pacientes, en el caso de que sean varios pacientes se asegura de clasificar a toda la población.
- Tipificar, en el caso de que sea un solo paciente el que se está manejando se encarga de armar las clases Transiciones (evolución del paciente), Imagen(imagen de apoya según la clasificación dada) y Tipología (descripción de la clase establecida)
- Comparar, dentro de la funcionalidad definida para el médico epidemiólogo permite comparar entre las 3 formas de clasificación que maneja la clasificación: bajo reglas de C4.5, GP y Definitiva.

Después de haber presentado los modelos de análisis y diseño elaborados a continuación se presentan la forma en que se implementó el sistema propuesto.

VI.6 Implementación.

Tanto el médico familiar como el epidemiólogo cuentan con computadoras para trabajar y tienen acceso al sistema de expediente médico electrónico del IMSS.
En el caso del médico familiar el uso de este equipo durante la consulta se restringe a la captura de información y la consulta de datos que el propio expediente proporciona. Por lo que la estrategia de implementación consistió en emular el funcionamiento del expediente electrónico (principalmente en la captura y la presentación de datos de apoyo que ya presenta) y agrega en el emulador el acceso a la pantalla de soporte a la toma de decisiones muy similar a las pantallas de consulta que tiene acceso dentro del sitio del IMSS. En este sentido se diseñó

una aplicación sin conexión a la red que emulara el comportamiento que se tendría al integrar el apoyo al expediente médico electrónico, corriendo en una PC. En el caso del epidemiólogo hace un uso más abierto de su equipo de cómputo usando paquetes como EXCEL, ACCESS, SPSS para evaluar grupos de pacientes y la presentación de cuadros estadísticos sobre los indicadores evaluados.

De acuerdo a la Figura 51 el diseño arquitectónico del SATDSmet se constituye de dos componentes:

Administradores de Bases de Datos se implementaron accesos a dos diferentes bases de datos, en el caso de MySQL se seleccionó por ser software libre lo que permite se instale el sistema en cualquier computadora, en el caso de Access se busca aprovechar que se encuentra instalado en las máquinas de epidemiología y dar la facilidad de acceso directo a los datos de los pacientes de consulta para poder extraer información y realizar análisis de poblaciones con ellos.

Sistema SATDSmet, tiene una arquitectura cliente-servidor trabajando en forma local en cada computadora, en otras palabras no hace uso de internet. En la programación del sistema se seleccionó el lenguaje de programación Java por su orientación a objetos y por ser software libre, además de contar con una gran cantidad de librerías que permiten el manejo de gráficas y reportes. Otra ventaja importante de java es su portabilidad que le permite instalarle en cualquier sistema operativo. El SATDSmet ocupa el patrón de diseño Modelo-Vista-Controlador (Model View Controler-MVC por sus siglas en inglés) y en la implementación se construyeron 6 paquetes:

- La interfaz de usuario, que soporta los dos manejos definidos en los casos de uso para los actores médico familiar y médico epidemiólogo. Y que corresponden a las vistas del sistema.
- Controladores, que se integra de dos componentes: el control de riesgo, cuya función es integrar todas las acciones relacionadas con el manejo del

síndrome ligado a la clasificación; y el control de archivos de soporte, que tramita todos los recursos que le son necesarios al controlador de riesgo para integrar los servicios dados a las interfaces de usuario.

- Base de datos, que corresponde a la parte del modelo relacionada con el manejo de la base de datos.
- Clasificación de riesgo, integrada por la parte del modelo donde reside la clasificación de riesgo, contiene dos componentes la clasificación que funciona como la base de reglas. Y el clasificador que integra las reglas y las aplica para un paciente o para una población de pacientes.
- Gestor de archivos de texto, forma parte del modelo integrando al controlador un objeto relacionado con alguna funcionalidad (imagen, cuadro, resumen, presentación, etc.)
- Graficador, contiene los componentes que dentro del modelo proporcionan la funcionalidad de graficar.
- Reporte, contiene los componentes que dentro del modelo proporcionan la funcionalidad de general un reporte.

Independientemente al SATDSmet se utilizaron diferentes subsistemas para llevar a cabo el proceso KDD:

- Matlab (Matrix Laboratory): es un lenguaje de alto nivel de programación y medios ambientes interactivos que posibilita procesar tareas intensas rápidamente para el cómputo técnico. Este lenguaje contiene herramientas que posibilitan el análisis de diferentes tipos de redes neuronales y herramientas de soporte para el análisis de resultados.
- Weka (Weikato Environment for Knowledge Analysis): software de aprendizaje automático y minería de datos elaborado en java, que contiene diferentes modelos de clasificación, entre ellos los algoritmos de expectativa y maximización y el algoritmo C4.5 para generar reglas de clasificación.

- JCLEC: Es un framework elaborado en java que permite la implementación de diferentes tipos de algoritmos genéticos entre ellos la programación genética.
- EXCEL: la mayoría de la información recolectada en el almacén de datos son archivos de Excel recolectados por el epidemiólogo. Se aprovechó la facilidad de conexión entre Excel y Access para realizar el pre procesamiento de los datos que se describe en el Capítulo III.

En el caso del proceso KDD llevado a cabo en la presente tesis el objetivo es la extracción de reglas que constituyen una clasificación de riesgo del síndrome metabólico. Este proceso se lleva a cabo una vez y producto de él es una clasificación (con 3 diferentes conjuntos de reglas) que en el diseño arquitectónico del sistema se queda registrado en el objeto del mismo nombre, y este objeto es la base para el CDSS elaborado.

A continuación se presenta la implementación de la funcionalidad determinada en el análisis anteriormente presentado.

VI.5.1 Implementación del diseño de la funcionalidad propuesto.

En este Capítulo se presenta la implementación que se hizo de la arquitectura.
El sistema del expediente electrónico es la principal herramienta con la cual cuenta el médico tomador de decisiones al momento de realizar su consulta. En el quedan registrados todos los datos necesarios para clasificar al paciente. La forma ideal de implementar el apoyo para el manejo del síndrome metabólico, es incluyéndolo como una liga de tal forma que el médico pueda acceder a él en cualquier momento de la consulta.
Sin embargo el expediente médico electrónico es un sistema institucional que no puede ser modificado sin la autorización de las autoridades del IMSS. Al evaluar

esta situación con el especialista se decidió emular el funcionamiento expediente e incluir la liga en dicho emulador, para evitar que llegase a haber cierto desinterés por parte del médico familiar hacia el sistema al presentársele como algo por separado del expediente médico electrónico.

En la Figura 55 se presenta la interfaz de captura para el rol tomador de decisiones, adicionalmente se representa parte del funcionamiento del sistema. Al momento del registro el tomador de decisiones teclea el número de identificación del derechohabiente, si el paciente ya existe el sistema recupera automáticamente los datos de él, en el caso de no ser así, el sistema permite la captura de todos los datos que aparecen en la interfaz.

Al capturar el peso y la talla el sistema automáticamente presenta el IMC (índice de masa corporal) y el tipo de obesidad que presenta el paciente (ver apéndice B). Ambos indicadores son fundamentales para la toma de decisiones relacionadas con el síndrome metabólico, es al momento de evaluarles que probablemente se requiera del apoyo para determinar las diferentes alternativas de acción.

En la Figura 56 se presentan segmentos de la gráfica rica donde se observan 3 momentos dentro del proceso en que el tomador de decisiones puede llegar a necesitar apoyo para manejar la incertidumbre al momento de la toma de decisiones al evaluar la respuesta al tratamiento, al momento de dar el seguimiento periódico que se hace de pacientes que pueden padecer el síndrome metabólico, y cuando se evalúa el tratamiento farmacológico. En cualquiera de ellos el tomador de decisiones puede solicitar el apoyo.

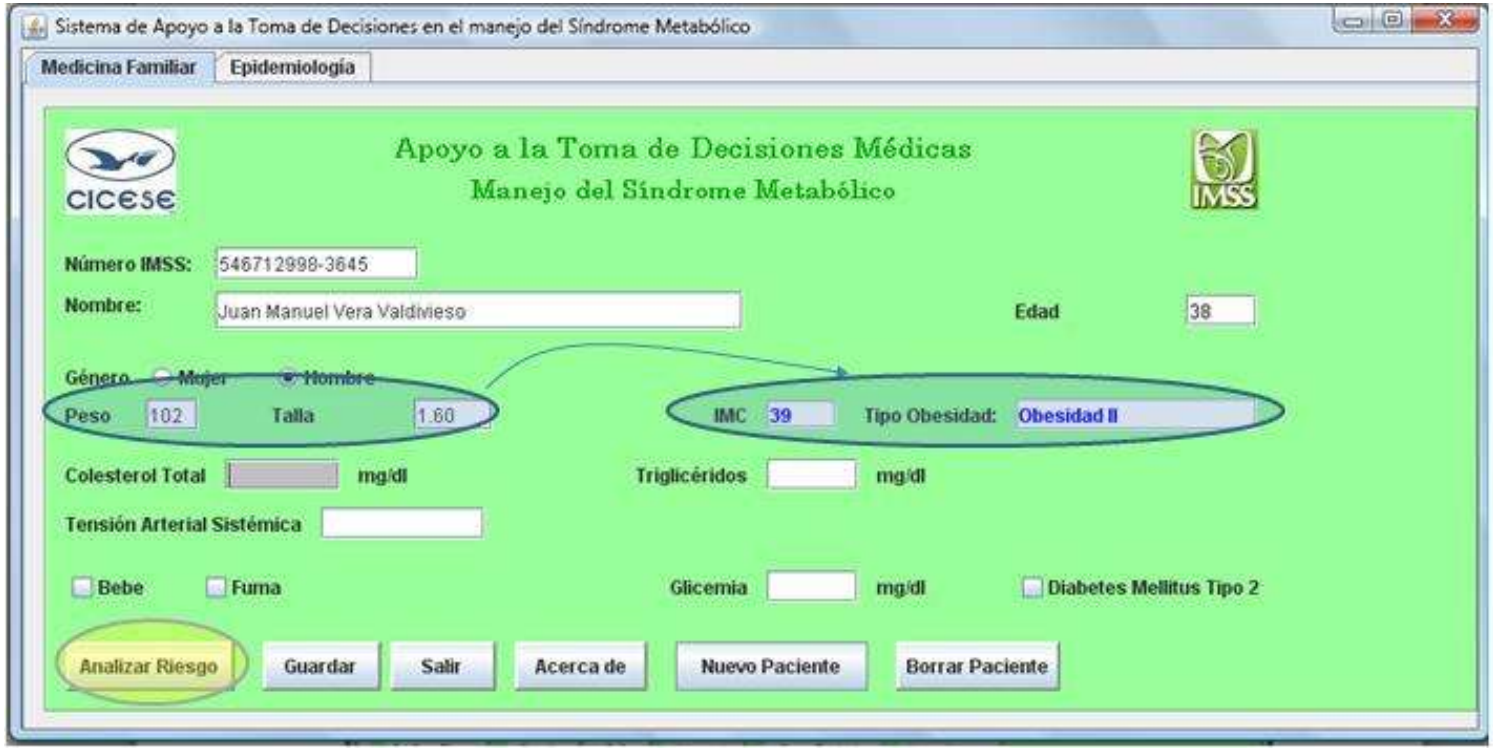


Figura 55. Interfaz de captura para el tomador de decisiones.

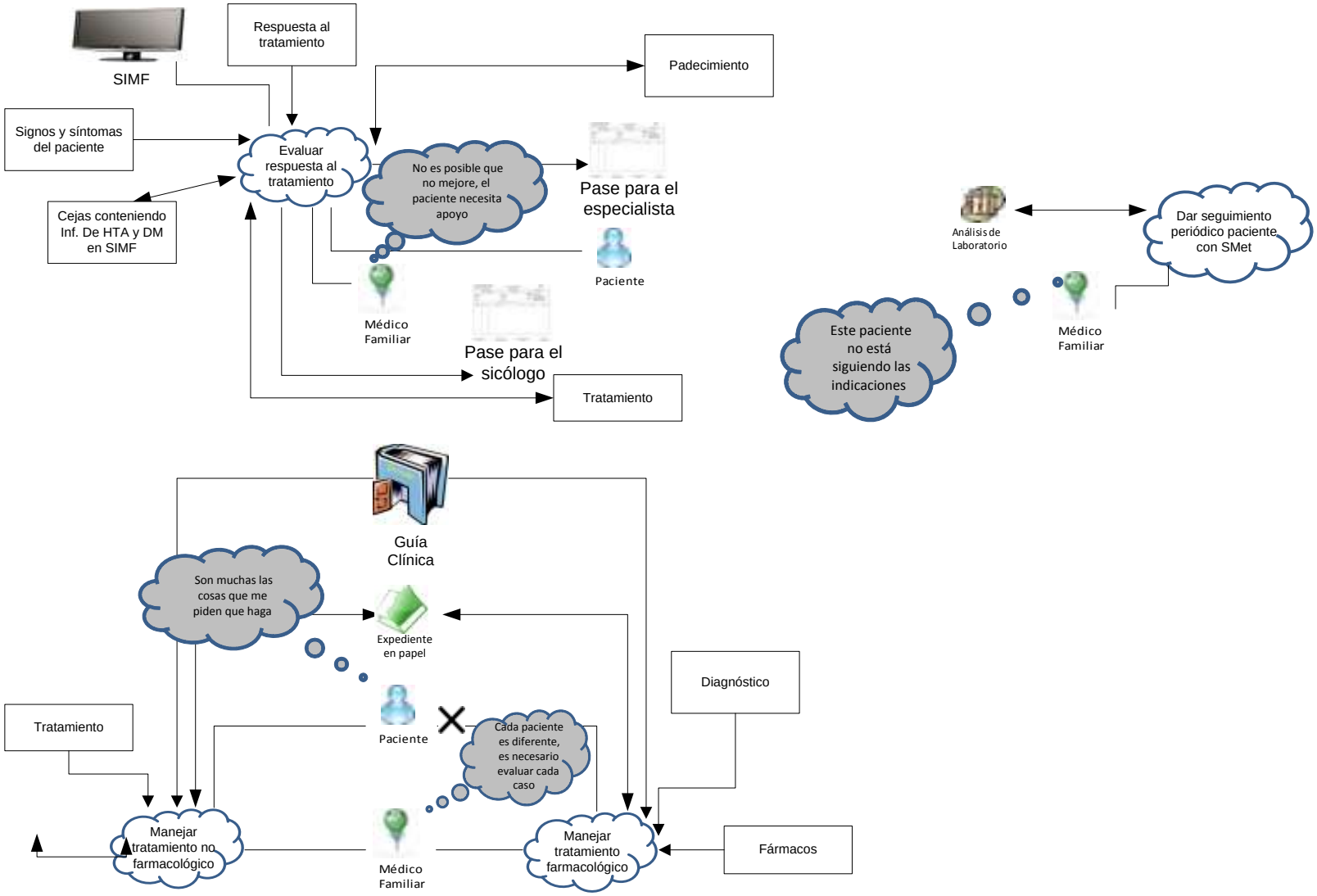


Figura 56. Segmentos de la gráfica rica donde se maneja la incertidumbre al tomar decisiones sobre el síndrome metabólico

En la Figura 55 también se observa un botón que activa la ayuda que proporciona el sistema.

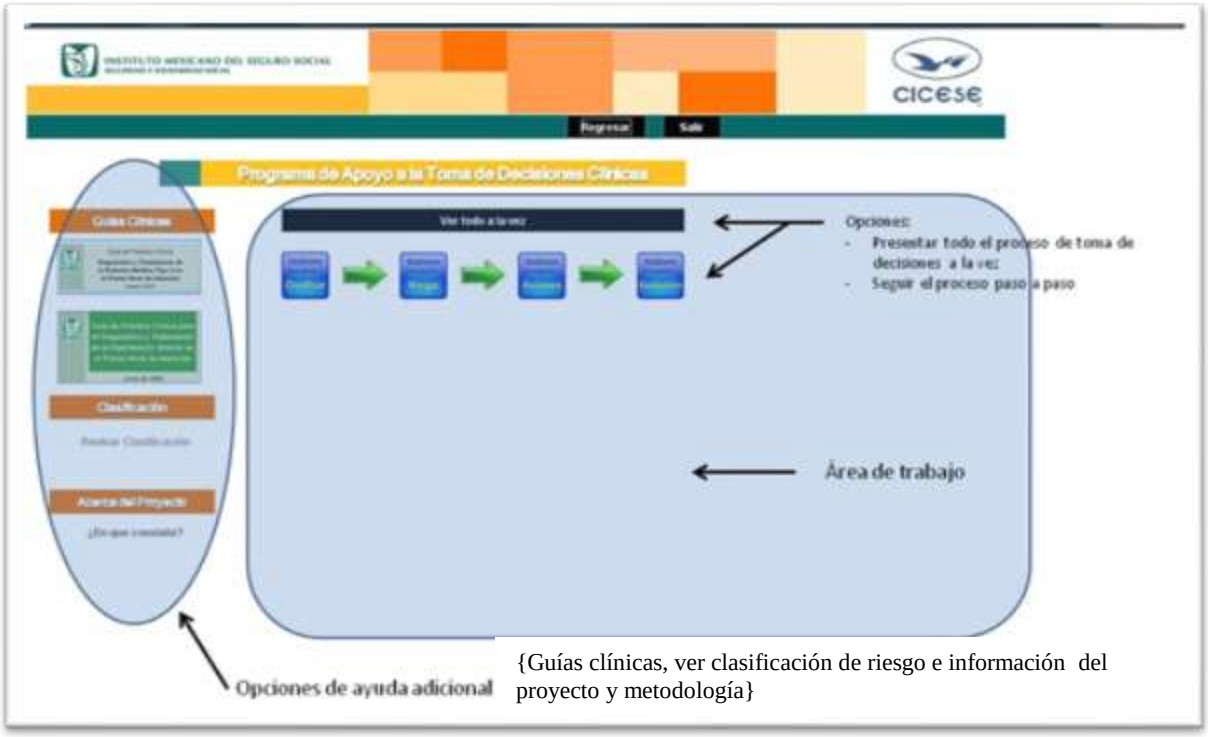


Figura 57. Pantalla del sistema que da el apoyo a la toma de decisiones en la consulta de un paciente.

Al seleccionar esta opción el emulador presenta la pantalla que se muestra en la Figura 57, esta pantalla emula una página de internet del instituto. Como se puede observar en la Figura 57 la página se divide en dos secciones, la central en la que hay dos opciones de manejo: ver todo el material de apoyo a la vez o seguir paso a paso el proceso de análisis en la toma de decisiones del síndrome metabólico (representada por los botones azules en secuencia).

La otra sección es la de opciones de ayuda adicional donde existen 3 opciones: consultar las guías clínicas, ver datos sobre la clasificación de riesgo del síndrome metabólico y por último información acerca del proyecto y su metodología.

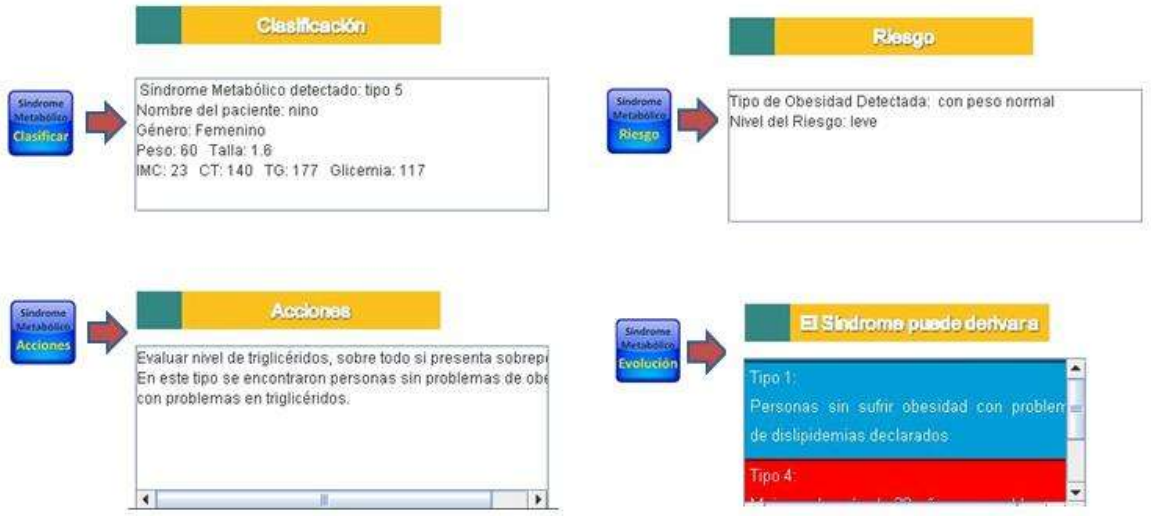


Figura 58. Esquematización de los datos de apoyo que se obtienen durante el proceso.

En la Figura 58 se muestra el tipo de apoyo que el sistema presenta para un paciente en particular, la cual presenta la información según el caso de uso Evaluar Riesgo (ver Sección VI.3).

En la Figura 59 se presenta el acceso a esta parte del sistema, de acuerdo a lo definido en los casos de uso, está restringido solo para el uso del médico epidemiólogo. Una vez que se proporciona el password se tienen que acceder al archivo que contiene la población o en su defecto capturar hasta 20 pacientes en los renglones que presenta el sistema.

Una vez que se ha tenido acceso a la información de la población la primera acción a la cual se tiene acceso es clasificar a cada uno de los miembros de la población. Y habiéndose clasificado se tiene acceso a todos los elementos de análisis que proporciona el sistema y que se esquematizan en la Figura 60.

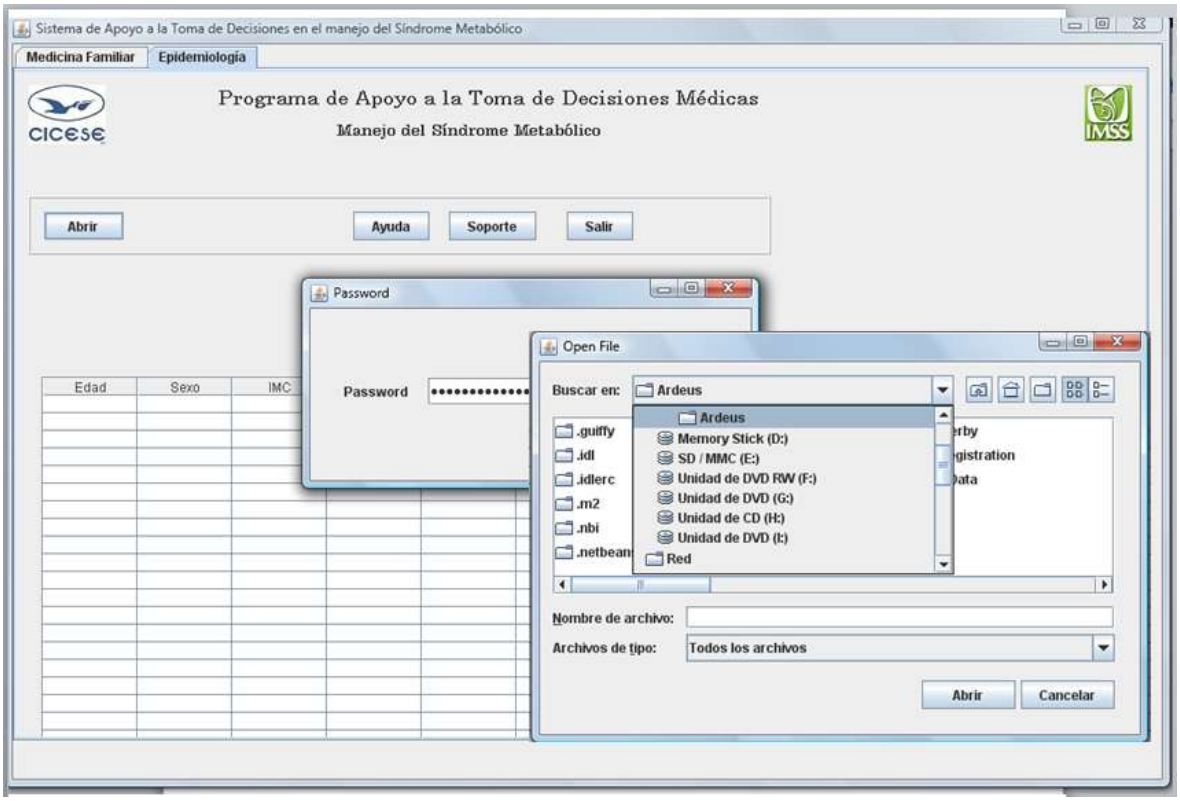


Figura 59. Acceso a archivos secuenciales conteniendo la población a trabajar.

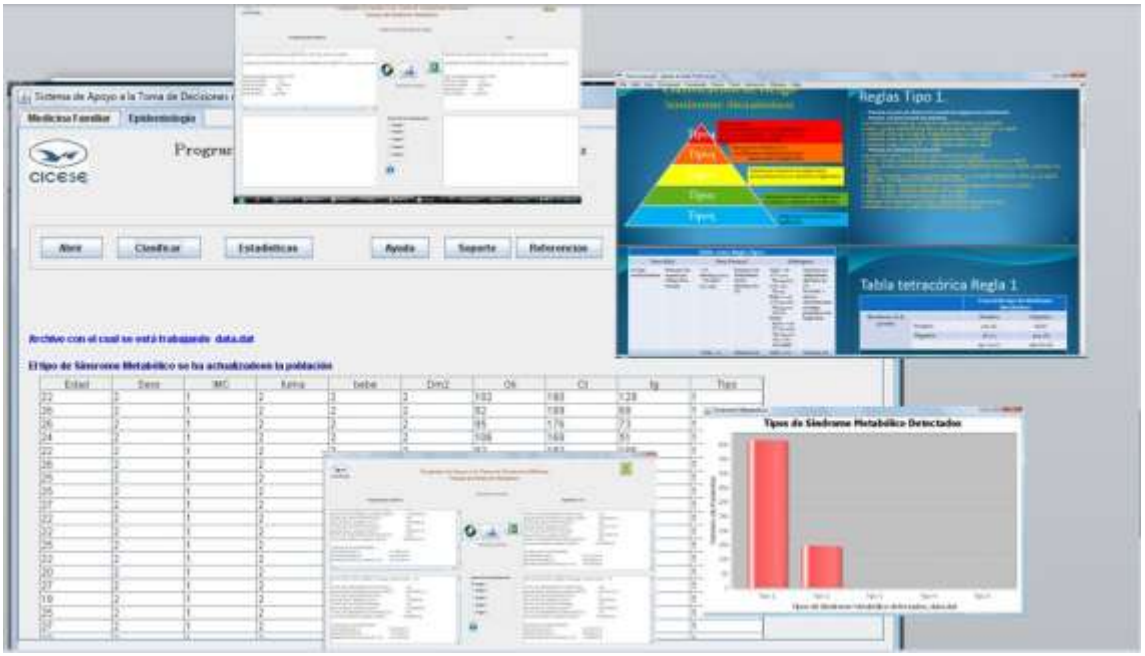


Figura 60. Herramientas estadísticas que presentan el sistema de apoyo a la toma de decisiones.

VI.6 Resumen.

En el presente Capítulo se ha presentado el análisis, diseño e implementación del sistema de apoyo a la toma de decisiones clínicas que apoya la toma de decisiones en el manejo del síndrome metabólico basado en la clasificación de riesgo que se obtuvo mediante el proceso KDD implementado como una máquina de conocimiento.

Se siguió la traza del caso de uso Evaluar Paciente para explicar la forma en que el médico tomador de decisiones interactúa con el sistema y cómo le ayuda a dilucidar el curso de acción en el momento en que se presenta cierta incertidumbre.

Se presentó la arquitectura del sistema y cómo partiendo de ella y guiado por los casos de uso se desarrolló un sistema de software orientado a objetos. Y finalmente se presentó el diseño de la aplicación que se realizó.
Se puede concluir que el producto del proceso KDD sí funciona como base para la construcción del sistema descrito.
En el siguiente Capítulo se presentan los resultados de exactitud predictiva de las reglas obtenidas y de utilidad percibida.

Capítulo VII

Evaluación del modelo predictivo y el CDSS

VII.1 Introducción.

De acuerdo a la metodología en la fase 8 se debe realizar una evaluación del modelo de soporte a la toma de decisiones. En el caso de la presente tesis son dos los elementos a evaluar: la clasificación de riesgo del síndrome metabólico y el CDSS elaborado, el SATDSmet.

En el primer caso la clasificación obtenida es inédita, por lo cual no es posible compararla con la experiencia pasada del médico clasificando el nivel en que un paciente sufre el síndrome metabólico.
Sin embargo, las reglas que definen a los grupos (conglomerados) encontrados en la fase de minería de datos se pueden comparar con los puntos de corte encontrados en las guías médicas del manejo de la DM2, HTA y dislipidemias. Producto de este análisis, realizado en conjunto con el médico epidemiólogo, se encontró que las reglas que describen los subgrupos son congruentes, en términos generales, con lo descrito en las guías clínicas, donde un elemento central de la clasificación es el nivel de grasa (colesterol y triglicéridos) en la sangre. En base a la experiencia del epidemiólogo se hizo una tipología del perfil de pacientes en cada grupo, la cual sirvió para validar la congruencia de la clasificación realizada. Habiendo validado la congruencia de la clasificación se procedió a hacer una prueba de la exactitud predictiva de las reglas obtenidas utilizando la metodología descrita en Ruiz et al (2004), los datos del Almacén de datos, y la clasificación asignada a dichos datos en el algoritmo KDD. La evaluación comprende los tres indicadores descritos en la Sección V.6,

sensibilidad, especificidad, y exactitud predictiva, y se resumen en la siguiente sección.

En el caso del CDSS se realizó una evaluación cualitativa de la utilidad y la facilidad de uso percibida. Para realizar esta prueba se integró un grupo de 8 doctores (2 epidemiólogos y 6 médicos familiares) del hospital número 8 del IMSS. Se habilitó un consultorio donde cada médico evaluó el riesgo del paciente de padecer un EVC o DM2 con un paciente seleccionado al azar usando la clasificación y el CDSS, los experimentos se explican a fondo en la Sección VII.3.

VII.2 Evaluación de la clasificación de riesgo del síndrome metabólico.

De acuerdo a lo descrito en la sección V.6 para la evaluación de resultados se utilizó la metodología descrita por Ruiz et al (2004), todas las tablas generadas en este proceso se presentan en el Anexo D. A continuación se presenta una serie de resúmenes de los resultados obtenidos.

Se define a la sensibilidad como *la capacidad de la prueba para clasificar correctamente a un enfermo como enfermo* (Ruiz, et al., 2004).

Por su parte la especificidad se define como *la capacidad que tiene la prueba de clasificar a un sujeto sano como sano* (Ruiz, et al., 2004).

De acuerdo a estos autores cuando hay grandes costos o riesgos, incluidos los riesgos emocionales, en un resultado falso positivo, se deben emplear pruebas específicas, en el caso del síndrome metabólico ya se mencionó en la sección IV.3 que una razón por lo que no es común usar la detección del síndrome metabólico es que no hay un tratamiento específico para el síndrome. Sin embargo, sí es importante detectarle de forma temprana puesto que permite evaluar el riesgo de evolucionar hacia otro tipo de enfermedades. Es en este sentido que no es tan

relevante la sensibilidad de la regla como la especificidad de la misma. De acuerdo a los mismos autores es poco frecuente que existan pruebas altamente sensibles y específicas al mismo tiempo, de hecho no hay una relación entre sensibilidad y especificidad así que ambas pueden llegar a ser muy bajas o muy altas. Sin embargo, cuando la regla establece un punto de corte, como es el caso de las reglas obtenidas, un cambio en el punto de corte siempre producirá una mejoría en uno de los indicadores a costa del otro.

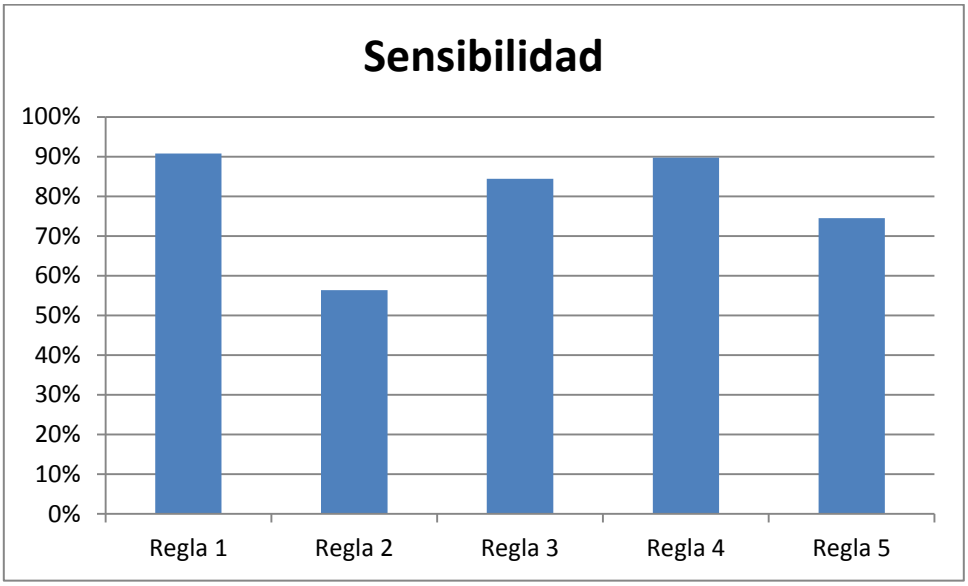


Figura 61. Resultados de la sensibilidad de las reglas descubiertas.

En la Figura 61 se puede observar que 3 de las reglas tiene una sensibilidad por arriba del 80%. Las reglas 2 y 5 tienen valores significativamente más bajos que los de las otras 3 reglas.

En la Figura 62 se presentan los resultados de la especificidad de las reglas obtenidas. En la Figura se puede observar cómo se cumple lo establecido en el párrafo anterior, ya que siendo la regla 1 la que tiene la sensibilidad tiene una sensibilidad por debajo del 90% y es la más baja de los resultados obtenidos. También se puede observar el efecto de la ponderación hecha al trabajar las

reglas, en el sentido que la mayoría de las reglas tienen una especificidad por arriba del 90%.

De acuerdo a la Figura 44 las clases 2 y 5 corresponden a conglomerados que están en los extremos del súper conjunto de las personas que no padecen obesidad pero que si tienen problemas relacionados con el síndrome metabólico. Estos subgrupos tiene una gran diversidad en los individuos que los componen, por lo que el proceso red neuronal de Kohonen semántica arrojó bastante información en relación a la tipología que define a estos dos grupos. En la clase 2 se tienen personas con dislipidemias leves y alteraciones en los índices de glicemia, mientras que en la clase 5 están personas con problemas de dislipidemias incipientes identificadas principalmente mediante los niveles de triglicéridos. Esto hace que en la clasificación en ambos subgrupos sea especialmente importante la especificidad arrojada por las reglas, mediante el proceso realizado se logró que ambas clases tuviesen una especificidad por arriba del 96%

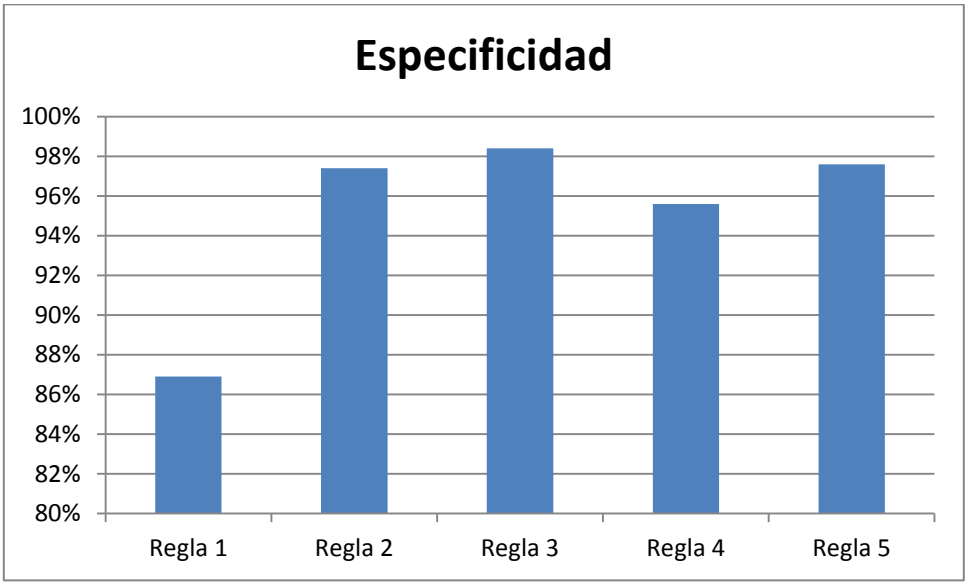


Figura 62. Especificidad de las reglas obtenidas.

Una forma de reunir los dos aspectos, sensibilidad y especificidad, es mediante una tercera característica denominada exactitud predictiva que se define por la probabilidad que corresponde al porcentaje del total que el examen de sujetos correctamente clasificados – tanto verdaderos positivos como verdaderos negativos (Ruiz, et al., 2004).

Un resultado relevante de la evaluación de la clasificación de riesgo del síndrome metabólico de acuerdo a los parámetros establecidos por el especialista epidemiólogo, es que al ser una clasificación totalmente nueva su exactitud predictiva debe ser en promedio por arriba del 90%, el promedio de la exactitud predictiva de la prueba realizada arroja un 92%.

A continuación se presentan los resultados obtenidos al ser evaluado el sistema de soporte a la toma de decisiones por un grupo de médicos.

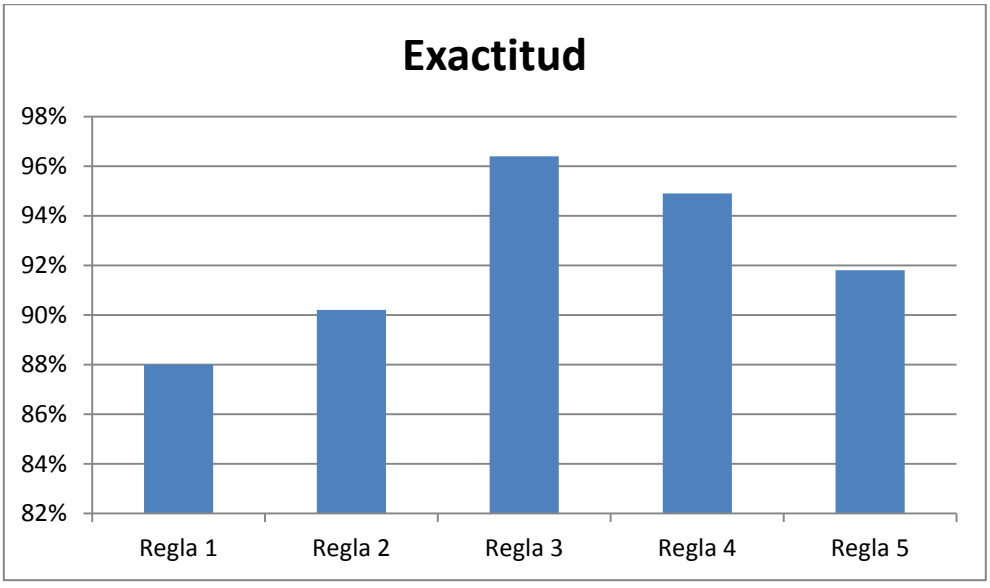


Figura 63. Exactitud predictiva de las reglas encontradas.

VII.3 Evaluación con usuarios.

Después de evaluar la exactitud de las reglas obtenidas, en esta sección se presentan los resultados de la evaluación de la utilidad y la funcionalidad percibida del sistema de soporte a la toma de decisiones desarrollado. Antes de presentar los resultados de la evaluación realizada es necesario definir el problema de evaluación.

VII.3.1 Definición del problema de evaluación.

El objetivo principal de la presente tesis es la extracción de reglas utilizando la programación genética como base para la elaboración de un CDSS para el manejo del síndrome metabólico. Por lo que la finalidad principal de la evaluación se puede dividir en dos partes:

1. Ubicar que tan exactas son las reglas extraídas para predecir que una persona que padece el síndrome sí lo padece y para predecir si una persona que no padece el síndrome en realidad no lo padece.
2. Ubicar que tanta utilidad le encuentran, tanto el médico familiar como el epidemiólogo, al CDSS elaborado.

Por lo que se quiere es evaluar la percepción de utilidad y facilidad de uso del sistema SATDSMet al momento de la consulta y al mismo tiempo evaluar, esas mismas percepciones, sobre la clasificación de riesgo obtenida.

Se formaron dos grupos de prueba: tomador de decisiones en consulta y epidemiólogo.

El primer grupo se conformó de 6 médicos familiares de la Clínica 8 del IMSS. En la Tabla XIII se puede observar la composición de dicho grupo (muestra).

Tabla XIII. Configuración del grupo de médicos familiares seleccionados para la prueba.

No. Médico	Área a la que pertenece	Puesto en la organización	Años de ejercer	Experiencia en consulta	Conoce al Síndrome
M1	Epidemiología	Doctora Investigadora	20	SI	SI
M2	Internista	Estudiante	1	NO	SI
M3	Medicina Familiar	Médico	6	SI	SI
M4	Medicina Familiar	Médico	3	SI	SI
M5	Medicina Familiar	Médico	1	SI	SI
M6	Medicina Familiar	Médico	5	SI	SI

En el caso del grupo de epidemiología se contó con la participación de dos especialistas (el jefe del departamento y la médico epidemióloga). Dado el tamaño de la muestra los resultados son evaluados cualitativamente. En la siguiente sección se describe la evaluación.

VII.3.2 Diseño de la evaluación del CDSS.

Las preguntas de investigación que guiaron la evaluación fueron las siguientes:

- ¿Cuál es la utilidad de contar con una clasificación de riesgo del síndrome metabólico?

- ¿Qué tanto un sistema de apoyo a la toma de decisiones ayudará en el manejo del síndrome metabólico?

En la siguiente sección se presentan las actividades que se llevaron a cabo por los tomadores de decisiones y por los médicos epidemiólogos para evaluar la facilidad de uso y la utilidad de la herramienta proporcionada.

VII.3.3 Actividades realizadas.

Las tareas se dividen en dos grupos: aquellas que tienen que ser realizadas por el médico, tanto familiar como epidemiólogo, en su rol de tomador de decisiones al momento de la consulta de un paciente.

La forma en que se estructuraron las actividades fueron las siguientes:

- Seleccionaron seis médicos familiares y dos epidemiólogos
- Selección de tres pacientes al azar
- Tarea 1: familiarizarse con la clasificación
- Tarea 2: determinar el riesgo con 10 mins. de tiempo usando el sistema de apoyo

Las actividades que se realizaron fueron: introducción del proyecto, seleccionar un paciente al azar, estudiar los expedientes seleccionados (el paciente no era conocido por el médico), familiarizarse con la clasificación de riesgo del síndrome metabólico haciendo una breve pre evaluación de la situación del paciente, explicación del funcionamiento del sistema, realización de las tareas, aplicación de la evaluación y recolecta de los resultados.

En la introducción del proyecto se explicó a los médicos el propósito que se buscaba y la metodología que se iba a seguir. Además se explicó brevemente el proceso de minería de datos y su relación con la investigación en medicina, así

como la forma en que se obtuvo la clasificación de riesgo, y como se utilizó como base para desarrollar el SATDSMet.

En la capacitación del uso del sistema se vio primero un ejemplo determinando el riesgo de un paciente usando la clasificación a mano para que el médico se familiarizara con ella. Habiéndose familiarizado se le presentó al médico un ejemplo del uso del SATDSMet para ilustrar los apoyos que proporciona.

Se pidió al médico:

1. Ingresar la información del paciente en el sistema SATDSMet
2. Hacer una primera evaluación del paciente y el posible manejo del síndrome con la información con que hasta ese momento contaba.
3. Seleccionar la opción de analizar riesgo
4. Seguir el proceso de valoración del riesgo
5. Determinar el riesgo y los posibles cursos de acción a seguir con el paciente.

Al finalizar las tareas se aplicó una herramienta para evaluar una tendencia de la opinión sobre la utilidad y la facilidad de uso percibida.
Adicionalmente se realizó una pequeña entrevista para evaluar sus percepciones tanto de la clasificación como del sistema computacional, y finalizando se tuvieron pequeñas charlas para discutir los resultados obtenidos, grabándose todas estas conversaciones para su posterior análisis.

Para realizar las tareas se utilizó un consultorio libre con una computadora con el sistema instalado. El tiempo de duración de prueba y evaluación tomó de 40 a 45 minutos con cada médico.

En el caso del epidemiólogo las actividades que se realizaron fueron: introducción del proyecto, seleccionar un archivo de población al azar, familiarizarse con la

clasificación de riesgo del síndrome metabólico, clasificar a toda la población y evaluar tendencias en los reportes y la gráfica presentada.

Los epidemiólogos recibieron la introducción y la explicación de la clasificación de riesgo al mismo tiempo que los otros médicos.

En la capacitación del uso del sistema con el epidemiólogo para la funcionalidad se tomó un archivo de ejemplo y se realizaron las siguientes tareas:

1. Ingresar al sistema y seleccionar un archivo de población.
2. Clasificar toda la población
3. Obtener los reportes y evaluar los resultados obtenidos
4. Obtener la gráfica de los resultados obtenidos
5. Evaluar las tendencias mostradas

Al igual que en el caso anterior, con el epidemiólogo también se utilizó la herramienta para evaluar la tendencia, así como una entrevista y la discusión sobre la percepción que se tuvo de las herramientas.

La evaluación se realizó en las mismas condiciones que las de los tomadores de decisiones con un tiempo promedio de 35 minutos.
A continuación se presentan los resultados cualitativos obtenidos.

VII.3.4 Análisis cualitativo de los resultados encontrados en la evaluación.

En este Capítulo se presentan los comentarios obtenidos en las entrevistas y el análisis cualitativo que se hizo de ellos.

Para tener una idea de la tendencia de la utilidad y la facilidad de uso percibida se utilizó una herramienta denominada TAM (Davis, 1989), aun cuando se trata de una herramienta estructurada por el número de experimentos que se pudieron realizar los resultados solo dan un indicio de la tendencia que siguen dichas percepciones.

A continuación se presentan los resultados obtenidos de la evaluación.

VII.3.4.1 Utilidad de la clasificación.

En relación a la utilidad de la clasificación de riesgo los 8 médicos marcaron estar muy de acuerdo (de acuerdo al TAM), que pudiese llegar a ayudarles a mejorar el manejo de las enfermedades relacionadas con el síndrome metabólico y con relación a mejorar el manejo del propio síndrome metabólico, se utilizó una escala Likert de 7 valores (Cañadas, et al., 1998), las opiniones variaron de ligeramente de acuerdo hasta extremadamente de acuerdo con un promedio de 6 que indica bastante de acuerdo. Esto también se ve reflejado en los siguientes comentarios realizados en las entrevistas.

Los comentarios que se hicieron son los siguientes:

- *"Al usar la información de riesgo presentada nos ayudará a evitar tener que manejar pacientes con enfermedades crónico degenerativas" [M6]*
- *"La información de riesgo permitirá coordinar mejor las acciones del médico con los programas integrales del IMSS" [M4]*
- *"Permitirá contar, oportunamente, con información del pacientes que empiezan con el síndrome metabólico" [M1]*
- *"Permite difundir el mismo criterio sobre el síndrome metabólico, el paciente siempre busca una segunda opinión, si los conceptos del síndrome metabólico se unifican aumentará la confianza del paciente en el diagnóstico." [M3]*

- *“La clasificación permite integrar el concepto del síndrome metabólico como elemento de diagnóstico” [M5]*
- *“Al no haber diferentes criterios en el manejo del síndrome mejora la confianza del paciente en el diagnóstico y las medidas tomadas.” [M6]*
- *“Permite la ubicación instantánea de pacientes de alto riesgo.” [M2]*
- *“Permite el seguimiento de la evolución de un paciente con síndrome metabólico.” [M1]*
- *“Con la clasificación se pueden tomar diferentes medidas preventivas, no es lo mismo tener un paciente con un nivel de bajo riesgo que uno con muchas alteraciones.” [M4]*
- *“Al pasar el paciente de un tipo de alto riesgo a uno de bajo riesgo nos da una idea de que tanto está funcionando el tratamiento integral.” [M2]*
- *“Actualmente en la práctica médica no le damos gran importancia a este tipo de síndrome, y podría ser bastante útil para la detección temprana de riesgo.” [M5]*

En 5 de los 11 comentarios presentados, se menciona la importancia que la clasificación de riesgo tiene para prevenir, de forma temprana las enfermedades crónico degenerativas relacionadas con el síndrome metabólico. Hay 3 comentarios que establecen la utilidad de incluir el síndrome en la consulta y como la inclusión de la clasificación puede ayudar a unificar criterios y al mismo tiempo ayudar a mejorar la actitud del paciente hacia el manejo preventivo de la comorbilidad del síndrome metabólico. Por último hay otros 3 comentarios que mencionan la utilidad que tiene la clasificación para evaluar la evolución del paciente y la eficacia del tratamiento integral.

Por lo que se puede establecer que los médicos entrevistados encontraron útil la clasificación obtenida tanto como una herramienta para la prevención de la comorbilidad del síndrome metabólico como un apoyo al manejo que se hace de la enfermedad (tratamiento integral).

VII.3.4.2 Utilidad del sistema de apoyo a la toma de decisiones.

En relación a si el sistema de apoyo a la toma de decisiones es útil, los 8 médicos marcaron estar muy de acuerdo (según escala Likert) que el contar con un sistema de apoyo a la toma de decisiones sobre el síndrome metabólico puede ayudar a evaluar el riesgo que representa dicho síndrome de una forma más efectiva y que la información que presenta el sistema es útil en su trabajo, con un promedio de 6 que indica bastante de acuerdo. Esto también se ve reflejado en los siguientes comentarios realizados en las entrevistas.

Los comentarios hechos fueron los siguientes:

- *"Con esto se integran las acciones a seguir al manejo del propio expediente electrónico lo que me parece muy bien." [M4]*
- *"Este sistema puede ayudar a la comunicación médico-paciente." [M1]*
- *"Mira uno de los principales problemas es la renuencia del paciente a aceptar su situación, como que el sistema nos ayudaría a asustarle un poquito y ver si con eso mejora su participación." [M3]*
- *"Me gusta que la información ya está de acuerdo al paciente-riesgo y no es algo general." [M3]*
- *"La información que se presenta permite al médico incluir rápidamente el manejo del síndrome metabólico en la consulta." [M6]*
- *"Nos da información oportuna." [M5]*
- *"Nos puede ayudar a dirigir un tratamiento más específico, a dar un enfoque más personalizado." [M2]*

En los comentarios se establece que los médicos entrevistados consideran útil el CDSS, tanto para apoyo en la consulta, toma de decisiones, como una herramienta de apoyo para el manejo del paciente (buscando resolver uno de los

problemas principales que es la renuencia del paciente a seguir las instrucciones del médico).

VII.3.4.3 Facilidad de uso del Sistema de Apoyo a la Toma de Decisiones.

En relación a la facilidad de uso de la sola clasificación de acuerdo a la escala Likert la mayor parte de los médicos calificaron entre ligeramente de acuerdo y muy de acuerdo, hubo una calificación de extremadamente de acuerdo y otra de para nada de acuerdo. Al final el promedio fue de 6 que indica muy de acuerdo en que fue muy simple su manejo. Esto también se ve reflejado en los siguientes comentarios realizados en las entrevistas.

Los comentarios realizados fueron los siguientes:

- *"El manejo del sistema es muy sencillo solo es picar un botón y seguir la secuencia que aparece en la pantalla." [M1]*
- *"Pues como proporciona conocimiento que fácilmente se integra al del médico no es complicado." [M3]*
- *"Es muy concreto sobre los niveles que puede llegar a presentar un paciente." [M5]*
- *"De entrada establece la posible evolución del paciente, eso facilita el manejo." [M6]*
- *"De forma instantánea permite determinar si el paciente padece o no el síndrome." [M2]*
- *"Es una herramienta totalmente automatizada no le veo problema." [M4]*

En términos generales los médicos entrevistados perciben al CDSS como fácil de utilizar.

A continuación se discuten los resultados obtenidos.

VII.3.5 Discusión de los resultados.

De acuerdo a los comentarios realizados por los médicos en su papel de tomadores de decisiones, hay una marcada tendencia a considerar útil la clasificación de riesgo obtenida. Ninguno de ellos comentó que no sirviera tomar en cuenta al síndrome metabólico como apoyo preventivo de las enfermedades con las cuales se relaciona.

Además todos estuvieron de acuerdo en que esta nueva clasificación facilitaría incluir en la consulta la evaluación del síndrome metabólico. Es más clara la utilidad percibida por los médicos como una herramienta de apoyo para potenciar la medicina preventiva al detectar de forma temprana a quienes tiene alto riesgo, y así evitar que un paciente evolucione hacia una enfermedad crónica, que como herramienta para el tratamiento del propio síndrome. Lo cual concuerda con el concepto de que al síndrome metabólico se le da poca importancia como padecimiento, por carecer de tratamiento específico.

Otro aspecto importante es la posible utilidad de la herramienta para apoyar la comunicación y la coordinación entre las diferentes áreas que participan en el tratamiento del síndrome; lo cual, de acuerdo al análisis de procesos realizado, corresponde a una de las incertidumbres más grandes de los otros actores participantes en el manejo del síndrome metabólico.

Los comentarios indican que el sistema sería fácilmente aceptado como una herramienta de apoyo al manejo del síndrome metabólico.

El sistema se consideró muy fácil de usar, el uso manual de la clasificación se consideró un poco más complejo aunque las opiniones siguen siendo favorables. A la mayoría de los médicos les costó un poco de trabajo recordar los cortes exactos que hacen las reglas, pero apoyándose en la tipología que acompaña a la clasificación cuando evaluaron al paciente ya la usaban y hacían referencia a ella sin ningún problema.

En el caso del médico epidemiólogo consideraron las herramientas de evaluación de poblaciones muy fáciles de usar. Los comentarios relacionados a la utilidad de la clasificación se centraron en la importancia que tiene el poder contar con un indicador de riesgo sobre el síndrome metabólico que establece el riesgo cardiovascular en una etapa más temprana. Adicionalmente se comentó que los resultados del proceso de minería de datos, vienen a confirmar cuáles son los principales indicadores para determinar la evolución del síndrome y que están plasmados en la clasificación de riesgo obtenida.

VII.4 Resumen.

En este Capítulo se presentaron los resultados sobre la exactitud predictiva de las reglas encontradas y un análisis cualitativo de los resultados obtenidos en una serie de experimentos donde se aplicó la clasificación de riesgo del síndrome metabólico y el sistema SATDSmet, tanto con el médico en su rol de tomador de decisiones como el del médico epidemiólogo.

De acuerdo a los resultados obtenidos se ha establecido que la clasificación de riesgo puede permitir la inclusión del manejo del síndrome metabólico como un elemento para evaluar el riesgo de la posible evolución hacia una enfermedad crónico degenerativa. Y que en términos del epidemiólogo representa una evaluación que pudiese promover la detección temprana del riesgo cardiovascular.

Otro elemento importante a considerar, es la posibilidad de que la inclusión del sistema apoye tanto a la comunicación de los diferentes actores dentro del manejo del síndrome como en la toma de decisiones al seleccionar los diferentes cursos de acción dentro del manejo del síndrome. En este sentido, un concepto importante fue que la clasificación permitirá unificar los criterios de manejo del paciente, cuando éste reciba retroalimentación de otras fuentes que tengan

acceso a la clasificación. Por lo que tanto la clasificación como el sistema sí pueden llegar a servir de apoyo en el manejo del síndrome metabólico.

En el siguiente Capítulo se presentan una serie de conclusiones en general y las aportaciones de la presente tesis.

Capítulo VIII

Conclusiones

VIII.1 Conclusiones

Hemos presentado el proceso de análisis de toma de decisiones apoyados en una metodología que permitió desarrollar un CDSS, el cual tiene como base un conjunto de reglas extraídas mediante el proceso KDD aplicado a una base de datos médica que contiene información recolectada por el área de epidemiología. La metodología aplicada permitió integrar una visión socio-técnica, que guió todo el proceso de extracción de reglas y de diseño del CDSS, lo que implica una diferencia significativa con relación a los trabajos similares de minería de datos en medicina.

Los resultados obtenidos muestran que las técnicas descritas han sido apropiadas para descubrir conocimiento nuevo y útil en la forma de una clasificación de riesgo de padecer el síndrome metabólico inédita. Se usó un KNN semántica para resolver el problema de clasificación mediante la formación de conglomerados bajo un criterio de similitud que toma 18 atributos. Este proceso ha mostrado que el eje principal sobre el cual gira la clasificación se centra en 6 atributos principalmente: edad, género, colesterol, triglicéridos, DM2 y si la persona tiene la costumbre de beber alcohol. Esto es una aportación en el sentido que sustenta con datos una percepción que ya se tenía del papel principal que juegan las dislipidemias en el desarrollo de las enfermedades crónico degenerativas ligadas con el síndrome metabólico.

Igualmente se ha establecido, por dos métodos distintos, una serie de reglas del tipo Si-Entonces que permiten determinar la pertenencia a los cinco diferentes

conglomerados establecidos. En el primer método se utilizó el algoritmo C4.5 con el objetivo de generar un primer conjunto de reglas, para posteriormente aplicar un algoritmo de programación genética con una gramática restrictiva con el cual se obtuvieron reglas más robustas (cortas, entendibles y con una alta eficiencia).

En este sentido el obtener los dos conjuntos de reglas mediante algoritmos distintos, para después analizar ambos conjuntos mediante un método que se utiliza de forma normal en el área de epidemiología cuando se realizan estudios de cohorte. Este proceso junto con el análisis de sus resultados con el especialista médico permitió incluir una visión socio-técnica que fue más allá de la sola evaluación de la función fitness y produjo finalmente un conjunto de reglas mixto.

Finalmente la metodología permitió el desarrollo de un CDSS que es de utilidad en el momento de la consulta para determinar el riesgo de desarrollar la comorbilidad del síndrome y al tomar decisiones sobre el manejo integral de la obesidad, laDM2 y la HTA.

VIII.2 Aportaciones al conocimiento.

Las aportaciones principales de la presente tesis giran en torno a dos conceptos principales: se obtuvo una clasificación de riesgo del síndrome metabólico para potenciar la detección temprana de la comorbilidad de dicho síndrome, HTA, DM2, dislipidemias y la obesidad y el sobrepeso; se incluyó una visión socio-técnica como guía del proceso de descubrimiento de conocimiento en bases de datos y de la programación genética. A continuación de hace una breve descripción de las aportaciones obtenidas.

- La aportación de una clasificación de riesgo del síndrome metabólico totalmente inédita, que permite evaluar el nivel del padecimiento desde sus etapas tempranas antes de declararse la comorbilidad ligada al síndrome.

- Se estableció que el uso de la programación genética en conjunto con las redes neuronales semánticas de Kohonen es una herramienta muy útil para el análisis de un padecimiento como el síndrome metabólico.
- La inclusión de una visión socio-técnica conjuntando el análisis del proceso de toma de decisiones y el proceso de minería de datos, su importancia al obtener el conjunto de reglas mediante la programación genética y al evaluar cuál debería ser el conjunto de reglas final
- La aportación que representa dicha clasificación para la evaluación del riesgo cardiovascular que realiza el área de epidemiología. Y como las reglas producto de la programación genética se complementaron con otros elementos generados por diferentes algoritmos, sin demérito de la calidad de la reglas encontradas
- La aportación de datos sustentados en un proceso KDD en base a la experiencia de una región de México especialmente afectada por los problemas de obesidad y su comorbilidad, que sustenta la expectativa que se tenía sobre la importancia que tienen las dislipidemias en el desarrollo de la comorbilidad del síndrome metabólico.
- La evaluación de la utilidad percibida por parte de médicos familiares y médicos epidemiólogos.
- La obtención de un conjunto de reglas con márgenes de aceptabilidad buenas.

VIII.3 Trabajo futuro.

Ya se ha establecido la importancia que tienen las enfermedades relacionadas con la obesidad y el sobrepeso en la sociedad mexicana, los grandes esfuerzos que se están realizando para combatirle y el costo social tan grande que las enfermedades con el síndrome tienen. A continuación se enumera el trabajo a futuro a realizar.

- El IMSS cuenta con un almacén de datos con el cual se puede realizar el mismo estudio con un tamaño de muestra mucho mayor abarcando otras regiones de México, dando un mayor sustento estadístico a los resultados obtenidos.
- La clasificación de riesgo del síndrome metabólico aporta los datos suficientes para que en base a ella el IMSS u otra entidad del Sector Salud pueda desarrollar una herramienta específica para el manejo del síndrome, esto es una guía médica.
- Se cuanta con acceso a otras unidades del IMSS donde se pueden realizar estudios de utilidad del CDSS con un número mayor de participantes para evaluar cuantitativamente su utilidad como un apoyo en la práctica clínica.

Literatura Citada.

Abbott, R. 1983. *Program Design by Informal English Descriptions*. Communication of the ACM. 26(11): 882-894 p.

Alhoniemi, E., Holmén, J, Simula, O., y Vesanto, J. 1999. *Process Monitoring and Modeling using the Self-Organizing Map*. Integrated Computer-Aided Engineering 6: 3-14 p.

Amy, F. 2007. *Etiology of the Metabolic Syndrome*. Current Cardiology Reviews 3: 232-239 p.

Bojarczuk, C., Lopes, H., Freitas, A. Michalkiewicz, E. 2004. *A Constrained-Syntax Genetic Programming System for Discovering Classification Rules: Application to Medical Data Sets*. Atificial Intelligence in Medicine 30(1): 27-48 p.

Bovick, A. 2005. *Handbook of Image & Video Processing.* Elsevier Academic Press.1355 pp.

Brameir, M. 2004. *On Linnear Genetic Programming.* Disertación doctoral..(Dortmund, Germany): Universität Dortmund am Fachbereich Informatik. 278 pp.

Cañadas, I., Sánchez A. 1998. *Categorías de respuesta en esclas tipo Likert.* Psicothema. 10(3): 623-631 p.

Ceriello, A. M. 2004. *Is oxidative stress the pathogenic mechanism underlying insulin resistance, diabetes and cardiovascular disease? The common soil hypothesis revisited*. Arterioscler Thromb Vasc Biol 24: 1-8 p.

Choppin, A. 1998. *Unsupervised Classificationof High Dimensional Data by means of Self-Organizing Neural Network*. Tesis de Ingeniería en Informática. (Louvain, Francia): Univeritè Catholique de Louvain. 103 pp.

Cios, K., Pedrycz, W., Swiniarski, R., Kurgan, L. 2007. *Data Mining a knowledge discovery approach.* Springer. 606 pp.

Davis, F. 1989. *Perceived usefulness perceived ease of use, and user acceptance of information technology*. MIS Quarterly: 318-339 p.

Deen, D. 2004. *Metabolic Syndrome: Time for action.* American Family Physician: A peer reviewed journal of the American of Family Physicians: 2875-2882 p.

Duda, R. O., Hart, P. E., Strok D. G. 2001. *Pattern Classification*. Wiley-Interscince. 654 pp.

Eiben, A. E., Smith J.E. 2003. *Introduction to Evolutionary Computing*. Springer. 300 pp.

Fayyad, U., Piatetsky G., Smyth, P. 1996. *From data mining to knowledge discovery in Databases.* American Association for artifitial intelligence: 37-54 p.

Han, J., Kamber M. 2006. *Data Minning: Concepts and Techniques.* Morgan Kaufmann.743 pp.

Hastie, T., Tibshirani R., Friedman J. 2001. *The Elements of Stadistical Learning Data Mining, Inference, and Prediction.* Springer-Verlag. 745 pp.

HEALTH, US. 2001. *ATPIII Guidelines At-A-Glance quick desk reference.* NIH publications. 6 pp.

Hernández, O., Ramírez, M.. J., Ferri, R. 2004. *Introducción a la Minería de Datos.* Person Prentice Hall. 656 pp.

IDF. 2006. *IDF consensus worldwide definition of the metabolic syndrom.* International Diabetes Federation. 24 pp.

IMSS. 2008. *El IMSS desarrolla programa en apoyo a obesos, diabéticos e hipertensos*. Comunicado oficial de la coordinación de comunicación social. 2 pp.

Jacobson, I., Booch G., Rumbaugh J. 2000. *El Proceso Unificado de Desarrollo de Software.* Addison Wesley. 438 pp.

Kahl, M. 1990. *Fundamentos de Epidemiología.* Ediciones Díaz de Santos, S.A. 368 pp.

Kohonen, T. 1990. *The Self-Organizing Maps*. Proceedings of the IEEE. 1464-1480 p.

Koza, J. 1992. Genetic Programming on the programming of Computers by means of natural selection. Bradford Dooks. 840 pp.

Leung, W., Lam, W., Sak, K., Cheng, J. 1999. *Applying Evolutionary Algorithms to Discover Knowledge form Medical Databases.* IEEE Eng. Med. Biol. Mag. 615-619 p.

Marakas, G. 2003. *Decision Support Systems in the 21st century.* Prentice Hall. 611 pp.

Monk, A., Howard S. 1998. *The rich picture: a tool for reasoning about work context*. Interactions 5(2): 21-29 p.

Ofstad, H. *An Inquiry into the Freedom of Decision.* 1961. Norwegian Universities Press. 15 pp.

Pacheco, M. J. 2004. *Soporte a la toma de decisiones con el enfoque a la Ingeniería de Procesos.* Tesis de Maestría en el Centro de Investigación Científica y de Eduación Superior de Ensenada. Departamento de Ciencias de la Computación. Ensenada Baja California México. 277 pp.

Poli, R. 2008. *A Field Guide to Genetic Programming.* Creative Commons. 250 pp.

Ruiz, A., Morrillo, L. 2004. Epidemiología Clínica investigación clínica aplicada. Editorial Médica Panamericana. 576 pp.

Sammon, J. 1969. *A non linear mapping for Data Structure Analysis.* IEEE Transactions on computers. C-18(5): 401-409 p.

Secretaria de Salud Pública**.** 2006. *Encuesta Nacional de Salud y Nutrición 2006.* Gobierno de México. 131 pp.

Skinner, D.C. 19999. *Introduction to Decision Analysis.* Probabilistic Publishing, segunda edición. 369 pp

.

Ta-Cheng, C., Toung-Chou, H. 2006. *A GA´s based approach for minning breast cancer pattern.* Expert System with Applications. 30(4): 674-681 pp.

Tan, J., Sheps S. 1998. *Health Decision Support Systems.* Aspen publisher.408 pp.

Tan, K.C., Teoh, E.J., Yu, Q., Goh, K.C. 2009. *A hybrid evolutionary algorithm for attribute selection in data minning.* Expert System with Applications. 36(4): 8616-8630 p.

Ultsch, A. 2003. *U*-Matrix: a Tool to visualize Cluster in high dimensional Data*. Reporte técnico (Marburg, germany). Technical Report NR. 36: 1-12 pp.

Ventura, S., Romero, C., Zafra, A., Delgado, J., Hervás C. 2007.*JCLEC: A Java framework for evolutionary algorithms.* Soft Computing – A fusion of Foundations, Methodologist and Applications – Special issue. 14(4): 315-357 p.

Witten, I., Frank, E. 2005. *Data Mining Pratical Machine Learning Tools and Techniques*.: Elsevier. 524 pp.

Zupan, J., Gasteiger, J. 1993. *Neural Networks for Chemists An Introduction*. Weinheim : VCH, Verlagsgesellschaft. 305 pp.

Zurada, J.M. 1992. *Introduction to Artificial Neural Systems*. West Publishing Company. 683 pp.

Ligas de internet citadas

Alonso. 2008. Sitio de atención primaria en la red guías clínicas España. URL:http://www.fisterra.com/guias2/smetabolico.asp.

Ricardo, R. 2010. *México gordo, problema de estado.* Sitio oficial del programa Reporte 13. URL: http://www.tvazteca.com/capitulos/reporte-13/26624/mexico-gordo,-problema-de-estado

Téllez, Cecilia. 2010. *México, primer lugar mundial en obesidad infantil; modificará el DIF dieta de desayunos; CCE se compromete a poner en letras grandes contenido nutrimental.* Sitio oficial del diario La crónica de hoy.(Cd. de México). URL: http://www.cronica.com.mx/nota.php?id_nota=483571

Universal. 2010. *75% de las camas de hospital, a obesidad.* Sitio oficial de el diario El Universal, el gran diario de México (Cd. de México). URL: http://www.eluniversal.com.mx/notas/695262.html

Vensato, J. 1999. SOM Toolbox homepage. Página oficial Laboratory of Computer and Information Science of adaptative informatics research centre (Helsinki, Finlandia). URL: http://www.cis.hut.fi/somtoolbox/

Apéndice A. Formatos de archivos del sistema

En este Apéndice se presenta los formatos que maneja el SATDSmet y los criterios utilizados para la transformación de datos.

En la Tabla IV de la Sección V.2 presenta el formato original que el área de epidemiología proporcionó para el trabajo de minaría de datos. Producto del preprocesamiento que se hace de la información previo a iniciar la fase de minería de datos se obtuvo un nuevo formato donde se eliminaron algunos atributos redundantes el cual se presenta en la Tabla XIV.

Tabla XIV. Formato de la CFE después de haber eliminado algunos datos redundantes, inservibles o irrelevantes.

Información de la CFE después de limpieza de datos.	
Campo	Descripción
Edad	Edad del paciente
Sexo	Género del paciente
Peso	Peso del paciente en kilogramos
Talla	Talla del paciente en metros
Clase de IMC	Clasificación del índice de masa corporal expresado como tipo de peso (obesidad y no obesidad)
Cintura	Cintura en metros
Tensión arterial sistólica	Medición puntual de la tensión arterial sistólica del paciente
Tensión arterial diastólica	Medición puntual de la tensión arterial diastólica del paciente
Clasificación de la tensión arterial	Clasificación de la tensión arterial sistémica del paciente basada en las mediciones anteriores
Sobre riesgo	Sobre riesgo en condiciones cardiovasculares.

Información de la CFE después de limpieza de datos (continúa).	
Campo	Descripción
Tabaquismo	Si la persona tiene el hábito de fumar
alcoholismo	Si la persona tiene el hábito de beber
Dm2	Si la persona padece diabetes Mellitus tipo 2
Glicemia	Medición puntual del nivel de glicemia prueba de glicemia en ayunas del paciente
Colesterol	Medición puntual del nivel colesterol total
Triglicéridos	Medición puntual del nivel de triglicéridos del paciente
Vldl	Medición puntual del nivel colesterol de alta densidad del paciente
Riesgo detectado	Medición del riesgo cardiovascular tomando en cuenta el resto de los indicadores.

Los cambios principales se resumen a continuación:

1. Los atributos IMC y Clase de obesidad se reemplazan por el atributo clase de IMC. El atributo IMC es un valor continuo en base al cual el epidemiólogo estableció la clase de obesidad. Sin embargo este último atributo está incompleto ya que solo toma en cuenta cuando existe un sobrepeso o cierto nivel de obesidad. Por lo cual aplicando el mismo criterio usado en la clase de obesidad se amplió para incluir personas con peso normal o por debajo de lo normal cubriendo todas las posibilidades.
2. Riesgo cardiovascular uno se eliminó porqué es redundante con el atributo riesgo detectado.
3. Td se eliminó por no tener utilidad
4. HTA por ser redundante con el atributo clasificación de la presión arterial

A continuación se presenta la normalización de varios de los atributos de la Tabla XIV.

Normalización de atributo edad	
# CLASE	INTERVALO
1	[18-23)
2	[23-28)
3	[28-33)
4	[33-38)
5	[38-43)
6	[43-47)
7	[47,52)
8	[52,...)

Normalización de atributo peso	
# CLASE	INTERVALO
1	(...,58)
2	[58,67)
3	[67,77)
4	[77,87)
5	[87,96)
6	[96,106)
7	[106,115)
8	[115,…)

Normalización de atributo talla	
# CLASE	INTERVALO
1	(...,1.52)
2	[1.52,1.57)
3	[1.57,1.62)
4	[1.62,1.67)
5	[1.67,1.71)
6	[1.71,1.76)
7	[1.76,1.81)
8	[1.81,…)

Normalización de atributo cintura	
# CLASE	INTERVALO
1	(...,76)
2	[76,83)
3	[83,90)
4	[90,97)
5	[97,104)
6	[104,110)
7	[110,117)
8	[117,…)

Normalización de atributo ta_sist (Tensión arterial sistólica)	
# CLASE	INTERVALO
1	(...,91)
2	[91,101)
3	[101,112)
4	[112,123)
5	[123,133)
6	[133,144)
7	[144,155)
8	[155,…)

Normalización de atributo ta_diast (Tensión arterial diastólica)	
# CLASE	# CLASE
1	(...,68)
2	[68,77)
3	[77,85)
4	[85,94)
5	[94,102)
6	[102,...)

Normalización de atributo glicemia	
# CLASE	# CLASE
1	(...,75)
2	[75,90)
3	[90,104)
4	[104,119)
5	[119,…)

Normalización de atributo colesterol	
# CLASE	INTERVALO
1	(...,140)
2	[140,159)
3	[159,177)
4	[177,196)
5	[196,215)
6	[215,234)
7	[234,253)
8	[253,271)
9	[271,290)
10	[290,…)

Normalización de atributo triglicéridos	
# CLASE	# CLASE
1	(…,126)
2	[126,210)
3	[210,293)
4	[293,376)
5	[376,460)
6	[460,…)

Normalización de atributo VLDL	
# CLASE	# CLASE
1	(…,23)
2	[23,37)
3	[37,52)
4	[52,66)
5	[66,80)
6	[80,…)

El SATDSmet tiene integrada una base de datos que registra una tabla denominada PACIENTE que tiene como propósito guardar un registro de los pacientes evaluados durante la consulta. El médico familiar tiene acceso a esta información un paciente a la vez para consultar los antecedentes registrados en el SATDSmet con el siguiente formato:

Tabla XV. Registro de la base de datos del SATDSmet.

Registro de la base de datos del SATDSmet		
Nombre del campo	Descripción	Tipo
nombre	Nombre del paciente	texto
genero	sexo del paciente	número
edad	edad del paciente	número
imc	IMC de paciente	número
peso	peso del paciente	número
talla	talla del paciente	número
ct	colesterol total	número
tg	triglicéridos	número
tas	tensión arterial sistémica	número

Registro de la base de datos del SATDSmet (continúa)		
Nombre del campo	Descripción	Tipo
gli	glicemia	número
bebe	indicador si el paciente tiene la costumbre de beber	número
fuma	indicador si el paciente tiene la costumbre de fumar	número
DM2	indicador si el paciente tiene diabetes mellitus tipo 2	número
class	clase de síndrome metabólico detectado	memo

A continuación se presenta el formato que tiene el SATDSmet para recibir una población de pacientes.

Tabla XVI. Formato del archivo secuencial para cargar poblaciones del epidemiólogo del SATDSmet.

Registro archivo de texto para el SATDSmet		
Num.	Descripción	Tipo
1	Edad del paciente	Número entero2 cuyo valor debe ser mayor a 18 años.
2	Genero del paciente	Solo tiene dos valores posibles: F: Femenino M: Masculino
3	IMC	Número con decimales
4	tabaquismo	Solo tiene dos valores posibles: 0: NO fuma 1: SI fuma
5	alcoholismo	Solo tiene dos valores posibles: 0: NO bebe 1: SI bebe

Registro archivo de texto para el SATDSmet (continúa)		
Num.	Descripción	Tipo
6	DM2	Solo tiene dos valores posibles: 0: NO la padece 1: SI la padece
7	glicemia	Número entero (si lleva decimales será truncado)
8	Colesterol total	Número entero (si lleva decimales será truncado)
Num.	Descripción	Tipo
9	triglicéridos	Número entero (si lleva decimales será truncado)
10	Tipo síndrome metabólico	Valor entero entre 1 y 5. Si se proporciona este dato el sistema infiere que se quiere validar su valor contra la propia clasificación que el sistema hace. Si no se utiliza poner en cero.

Apéndice B. Base de reglas para determinar la comorbilidad del síndrome metabólico.

En este apéndice se presenta una recopilación de diferentes indicadores relacionados con el síndrome metabólico y su comorbilidad recolectados de las guías clínicas de DM2 y HTA del IMSS. Esta recopilación sirvió como una base de reglas existentes y como un punto de referencia para entender el manejo clínico del síndrome metabólico. Principalmente se usó en la tipificación que se hizo de las diferentes clases de síndrome metabólico obtenidas, permitiendo relacionar las reglas de la nueva clasificación con los puntos de corte que se manejan en las reglas que determinan la existencia de la comorbilidad del síndrome metabólico.

Tabla XVII. Reglas que determinan la comorbilidad del síndrome metabólico.

Reglas para determinar la comorbilidad del síndrome metabólico			
Concepto	Prueba de laboratorio	Puntos de corte	Notas
Diabetes	Glucosa en ayuno y Prueba de glucosa	$glucosa \geq 126\, mg/dl$ $glucosa \geq 200\, mg/dl$	Factores de riesgo: • Prediabetes • Edad • Sobrepeso • Dislipidemias
Prediabetes	Glucosa en ayuno	$100 \leq glucosa \leq 126\ mg/dl$	• Prevalencia de DM2 se eleva con la edad • el riesgo de DM2 se eleva conforme se eleva el IMC

<table>
<tr><th colspan="4">Reglas para determinar la comorbilidad del síndrome metabólico (continúa)</th></tr>
<tr><th>Concepto</th><th>Prueba de laboratorio</th><th>Puntos de corte</th><th>Notas</th></tr>
<tr><td>Intolerancia a la glucosa</td><td>Prueba de glucosa</td><td>$140 \leq glucosa \leq 200\ mg/dl$</td><td></td></tr>
<tr><td>Dislipidemias</td><td colspan="2">Análisis de sangre</td><td rowspan="4">Factores de riesgo personales:
Obesidad
Tabaquismo
Sedentarismo
HAS
Niveles bajos de colesterol de alta densidad
Hombres > 45 años
Mujeres > 55 años</td></tr>
<tr><td>Niveles de Colesterol Total (CT)</td><td colspan="2"><table><tr><td>Recomendable</td><td>$< 200\ mg/dl$</td></tr><tr><td>Limítrofe</td><td>$200 - 239\ mg/dl$</td></tr><tr><td>Alto Riesgo</td><td>$ct > 200\ mg/dl$</td></tr></table></td></tr>
<tr><td>Niveles de baja densidad</td><td colspan="2"><table><tr><td>Recomendable</td><td>$< 130\ mg/dl$</td></tr><tr><td>Limítrofe</td><td>$130 - 159\ mg/dl$</td></tr><tr><td>Alto riesgo</td><td>$160\ mg/dl$</td></tr><tr><td>Muy alto riesgo</td><td>$190\ mg/dl$</td></tr></table></td></tr>
<tr><td>Triglicéridos</td><td colspan="2"><table><tr><td>Recomendable</td><td>$< 150\ mg/dl$</td></tr><tr><td>Limítrofe</td><td>$150 - 200\ mg/dl$</td></tr><tr><td>Alto Riesgo</td><td>$> 200\ mg/dl$</td></tr><tr><td>Muy alto riesgo</td><td>$> 1000\ mg/dl$</td></tr></table></td></tr>
<tr><td>Tensión Arterial</td><td colspan="2">Toma periódica (al menos dos tomas) con el esfigmomanómetro.</td><td></td></tr>
<tr><td>Tensión arterial sistémica</td><td colspan="2"><table><tr><td>Óptima</td><td>$< 120/80\ mm\ de\ Hg$</td></tr><tr><td>Normal</td><td>$120 - 129/80 - 84\ mm\ de$</td></tr><tr><td>Normal Alta</td><td>$130 - 139/85 - 89\ mm\ de$</td></tr><tr><td>Hipertensión Arterial</td><td><table><tr><td>Etapa1</td><td>$140 - 159/90 - 99\ mm\ de\ Hg$</td></tr><tr><td>Etapa2</td><td>$160 - 179/100 - 109\ mm\ de\ Hg$</td></tr><tr><td>Etapa 3</td><td>$\geq 180/\geq 110\ mm\ de\ Hg$</td></tr></table></td></tr></table></td><td>La conjunción de la HAS y tabaquismo con implica alto riesgo cardio vascular</td></tr>
</table>

Reglas para determinar la comorbilidad del síndrome metabólico (continúa)			
Concepto	Prueba de laboratorio	Puntos de corte	Notas
Índice de obesidad	Medición de talla y peso		
	Peso insuficiente	< 18.5	
	Normopeso	$18.5 - 24.9$	
	Sobrepeso grado I	$25 - 26.9$	
	Sobrepeso grado II (pre obesidad)	$27 - 29.9$	
	Obesidad Tipo I	$30 - 34.9$	
	Obesidad Tipo II	$35 - 39.9$	
	Obesidad Tipo III (mórbida)	$40 - 49.9$	
	Obesidad Tipo IV	> 50	

Apéndice C. Reglas de la clasificación del síndrome metabólico.

Tabla XVIII. Reglas de clasificación obtenidas por algoritmo C4.5

Clasificación por C4.5 (Reglas para la clase 1)					
Peso Bajo		Peso normal		Sobrepeso	
N.A.	Personas con su peso por debajo de lo normal	CT [195,211) TG [177,259)	Personas con dislipidemias severas afectados en TG.	Edad <28 CT>= 211 TG [94,177) CT>=211 TG<94 Edad >= 28 CT [211,226) TG[94,177) GL<87 Mujer Edad >= 28 CT [211,226) TG [94,177) GL >= 87 NO BEBE	Personas con dislipidemias, afectados en CT. Por edad y sexo se discrimina que no tenga problemas con la glicemia
		Edad < 33 CT [165, 180) TG < 177 CT [195, 211) TG [177,259) CT [165,195) TG >= 177 CT >= 180 TG < 177	Personas con TG al límite o con dislipidemia afectados en TG. Personas con CT al límite	Edad >=33 CT [195,211) TG<177 Edad < 33 CT [180,211) Edad < 23 CT [165,180) Mujer Edad [33,38) CT [165,195) Hombre Edad>=33 CT[165,195)	Personas con sobre peso y sin dislipidemia, los cortes por edad restringen los rangos de colesterol

Clasificación por C4.5 (Reglas para la clase 2)			
Peso normal		Sobrepeso	
CT >= 211 TG [177,259) CT >= 195 TG >=259	Personas con peso normal y dislipidemia severa con afectación de CT y TG	Edad >= 28 CT>=226 TG [94,177) Edad >= 33 CT[195,211) TG>=177 Edad >= 28 CT [211,226) TG [94,177) GL>=87 Bebe=SI Hombre Edad >= 28 CT [211,226) TG[94,177) GL>=87 Bebe=NO	Personas con sobrepeso con dislipidemia afectados en colesterol y/o triglicéridos

Clasificación por C4.5 (Regla 3)			
Peso normal		Sobrepeso	
CT>211 Edad >= 43 CT [165,211)	Personas con Obesidad I con dislipidemia afectados en colesterol	Mujeres CT>=150 TG>=136	Mujeres con Obesidad II
		Hombres Edad < 38 CT>211 Hombres Edad >=38	Hombres haciendo distinción por edad jóvenes con dislipidemia, edad avanzada solo con obesidad II

Clasificación por C4.5 (Regla 4)			
Peso normal		Sobrepeso	
Mujeres Edad >= 38 CT [165,195)	Mujeres en peso normal de edad avanzada y colesterol limítrofe	Obesidad I CT < 165 TG >=218 Obesidad I Edad>=47 CT<165 TG< 218 Obesidad I Edad < 43 CT [165,211)	Personas con Obesidad I con colesterol y/o triglicéridos al límite
		Obesidad II, III, IV Mujeres CT>=150 TG<136 Obesidad II, III, IV Hombres Edad < 38 CT< 211	Personas con Obesidad II , III, IV haciendo discriminación por género en los niveles de colesterol

Clasificación por C4.5 (Reglas para la clase 5)					
Peso normal		Sobrepeso		Obesidad	
Edad >= 33 CT [165,180) TG < 177	Personas con peso normal y triglicéridos al límite	Edad > 47 CT<180	Personas con sobrepeso y edad avanzada	Obesidad I Edad > 47 CT>165 TG< 218	Personas con Obesidad I y dislipidemia, de edad avanzada
				Obesidad II Mujer CT<150	Mujeres con Obesidad II sin problemas de dislipidemia

Detalle del segundo conjunto de reglas obtenidas por medio de programación genética

Tabla XIX. Reglas de clasificación obtenidas por programación genética.

Clasificación por GP (reglas para la clase 1)	
Reglas	Interpretación
Peso bajo o Normal con CT [165, 226) No DM2	Personas con peso bajo o normal que no padecen diabetes mellitus tipo 2 y con los niveles de colesterol al límite

Clasificación por GP (reglas para la clase 2)	
Reglas	Interpretación
Peso Bajo, Normal y Sobrepeso Edad <= 65 CT >=180 TG >=177	Personas que no sufren obesidad con dislipidemia

Clasificación por GP (reglas para la clase 3)	
Reglas	Interpretación
Obesidad cualquier tipo CT>=195 TG>=94	Personas cualquier tipo de obesidad con dislipidemia leve

Clasificación por GP (reglas para la clase 4)	
Reglas	Interpretación
Obesidad cualquier tipo CT<211 DM2=no Peso debajo de lo normal Hombres CT>=241 Fuma=SI DM2=NO	Personas con cualquier tipo de obesidad con colesterol limítrofe sin padecer Diabetes Mellitus 2, ó Hombres de peso por debajo de lo normal con dislipidemia afectada por colesterol

Clasificación por GP (reglas para la clase 5)	
Reglas	Interpretación
Edad <= 49 CT < 165 TG < 218	Personas con dislipidemia severa afectadas en triglicéridos

Comparativo de exactitud predictiva reglas obtenidas por C4.5 contra reglas obtenidas por programación genética, de acuerdo al criterio del fitness.

Tabla XX. Comparativo de exactitud predictiva C4.5 vs GP (criterio del fitness).

Comparativo Exactitud Predictiva							
Reglas para clase 1							
	VP	VN	FP	FN	Se	Es	EP
GP	92	451	14	105	0.467	0.970	0.820
C4.5	179	404	61	18	0.909	0.869	0.881
Reglas para clase 2							
	VP	VN	FP	FN	Se	Es	EP
GP	66	531	14	51	0.564	0.974	0.902
C4.5	57	541	4	60	0.487	0.993	0.909
Reglas para clase 3							
	VP	VN	FP	FN	Se	Es	EP
GP	69	549	17	27	0.719	0.970	0.934
C4.5	80	562	4	16	0.833	0.993	0.970
Reglas para clase 4							
	VP	VN	FP	FN	Se	Es	EP
GP	70	521	54	17	0.805	0.906	0.893
C4.5	77	554	21	10	0.885	0.963	0.953
Reglas para clase 5							
	VP	VN	FP	FN	Se	Es	EP
GP	120	491	6	45	0.727	0.988	0.923
C4.5	16	398	99	149	0.097	0.801	0.625

Tabla XXI. Clasificación del síndrome metabólico.

Clasificación de Riesgo del Síndrome Metabólico		
Tipo 5. Nivel de Riesgo Leve.		
Peso	Tipología	Reglas a aplicar
No Importa	Personas con dislipidemia severa afectadas en triglicéridos	Si Edad <= 49 y CT < 165 mg/dl y TG < 218 mg/dl
Tipo 1. Nivel de Riesgo Bajo.		
Peso	Tipología	Reglas a aplicar
Personas con peso por debajo de lo normal	Sin otro indicador	Si IMC ≤ 18.5
Personas con peso normal	Dislipidemia severa, afectados en triglicéridos.	Si $18.5 \leq IMC < 25$ con • $195 \leq CT < 211$ y $177 \leq TG < 259$
	TG al límite o con dislipidemia afectados en TG. Y Personas con CT al límite	Si $18.5 \leq IMC < 25$ con • $Edad < 33\ y\ 165 \leq CT < 180\ TG < 177$ • $195 \leq CT < 211\ y\ 177 \leq TG < 259$ • $165 \leq CT < 195\ y\ TG \geq 177$ • $CT >= 180\ y\ TG < 177$
Personas con sobrepeso	Personas con dislipidemia, afectadas en CT. Por edad y sexo se discrimina que no tenga problemas con la glicemia	Si $25 \leq IMC < 30$ con • Edad <28 y CT>= 211 y 94≤ TG<177 • CT>=211 y TG<94 • Edad >= 28 y 211≤ CT<226 y 94 ≤TG<177 y GL<87 • Mujer y Edad >= 28 y 211≤ CT<226 y 94 ≤TG<177 y GL >= 87 y NO BEBE

Clasificación de Riesgo del Síndrome Metabólico (continúa)		
Personas con sobrepeso (continúa)	sin dislipidemia, los cortes por edad restringen los rangos de colesterol	Si $25 \leq IMC < 30$ con • Edad >=33 y 195≤ CT<211 y TG<177 • Edad < 33 y 180≤ CT<211 • Edad < 23 y 165≤ CT<180 • Mujer y 33≤ Edad<38 y 165≤ CT<195 • Hombre y Edad>=33 y 165≤ CT<195
Tipo 2. Nivel de Riesgo Medio.		
Peso	Tipología	Reglas a aplicar
Peso por debajo de lo normal, normal y sobrepeso	Personas que no sufren obesidad con dislipidemia	Si IMC < 30 con • Edad <= 65 y CT >=180 y TG >=177
Tipo 3. Nivel de Riesgo Alto.		
Peso	Tipología	Reglas a aplicar
Obesidad I	Con dislipidemia afectados en colesterol.	Si $30 \leq IMC < 35$ • CT>211 • Edad >= 43 y 165≤ CT<211
Obesidad II	Mujeres con Obesidad II	Si $35 \leq IMC < 40$ • Mujeres y CT>=150 y TG>=136
	Hombres haciendo distinción por edad jóvenes con dislipidemia, y edad avanzada aún sin padecer dislipidemia	Si $35 \leq IMC < 40$ • Hombres y Edad < 38 y CT>211 • Hombres y Edad >=38

247

Clasificación de Riesgo del Síndrome Metabólico (continúa)		
Tipo 4. Nivel de Riesgo Muy Alto.		
Peso	Tipología	Reglas a aplicar
Peso Normal	Mujeres en peso normal de edad avanzada y colesterol limítrofe	Si $18.5 \leq IMC < 25$ con • Mujeres y Edad >= 38 y 165≤CT<195
Obesidad I,II,III,IV	Personas con Obesidad II , III, IV haciendo discriminación por género en los niveles de colesterol	Si IMC > 35 • Mujeres y CT>=150 y TG<136 • Hombres y Edad < 38 y CT< 211

Tabla XXII. Tablas tetracóricas reglas de la clasificación del síndrome metabólico.

Reglas para clase 1		**Concordó tipo de Síndrome Metabólico**	
Resultado de la prueba		Positivo	Negativo
	Positivo	179 (a)	61(b)
	Negativo	18 (c)	404 (d)
		197 (a+c)	465 (b+d)
Sensibilidad		90.8% (a/(a+c))	
Especificidad		86.9% (b/(b+d))	
Exactitud		88.0% ((a+c)/(a+b+c+d))	

Reglas para la clase 2		**Concordó tipo de Síndrome Metabólico**	
Resultado de la prueba		Positivo	Negativo
	Positivo	66	14
	Negativo	51	531
		117	545
Sensibilidad		56.4%	
Especificidad		97.4%	
Exactitud		90.2%	

Reglas para la clase 3		**Concordó tipo de Síndrome Metabólico**	
Resultado de la prueba		Positivo	Negativo
	Positivo	81	9
	Negativo	15	557
		96	566
Sensibilidad		84.4%	
Especificidad		98.4%	
Exactitud		96.4%	

Reglas para la clase 4		**Concordó tipo de Síndrome Metabólico**	
Resultado de la prueba		Positivo	Negativo
	Positivo	78	25
	Negativo	9	550
		87	575
Sensibilidad		89.7%	
Especificidad		95.6%	
Exactitud		94.9%	

Reglas para la clase 5		**Concordó tipo de Síndrome Metabólico**	
Resultado de la prueba		Positivo	Negativo
	Positivo	123	12
	Negativo	42	485
		165	497
Sensibilidad		74.5%	
Especificidad		97.6%	
Exactitud		91.8%	

Anexo D. Formato de entrevistas.

PROTOCOLO DE LA ENTREVISTA.

Escenario:
Fecha
Lugar
Hora
Entrevistador

Preguntas de Investigación.
¿Cuál es el proceso del manejo del síndrome metabólico y cuáles son las decisiones que toman para manejarlo?
¿Qué información manejan las personas al tomar decisiones sobre el síndrome metabólico?

PREJUICIOS.

1. Las normas y guías no se siguen al pie de la letra y mucho se utiliza el criterio de la persona al realizar la toma de decisiones del síndrome metabólico.
2. Habiendo forma de detectar tempranamente el síndrome metabólico, no se hace.
3. Aun habiendo diferentes programas institucionales para el manejo de la comorbilidad del síndrome metabólico, la detección en sí del propio síndrome no existe.
4. Mucha de la información útil solo se genera a petición de la persona involucrada y los criterios para solicitarlos se rigen por el sentido común más que por normas o procedimientos. Esto provoca que información útil no exista en la mayoría de los casos.
5. No es claro el curso de acción que sigue el tratamiento del síndrome metabólico, se le relaciona más con su comorbilidad.

Guía.

Buenos días. Estoy realizando una serie de entrevistas con relación al síndrome metabólico para mi trabajo de tesis. La intención es identificar cual es el proceso del manejo del síndrome metabólico que personas participan en la toma de decisiones, que decisiones se toman y la información que se utiliza.

INFORMACIÓN PERSONAL

Nombre del Entrevistado:
Sexo:
Edad:
Puesto:
Años trabajando:
Tiempo de manejar el síndrome metabólico:

1. ¿Conoce el síndrome metabólico?

Sí lo conoce

1.1. Describa Ud. cuál es su rol en el manejo del síndrome metabólico.
1.2. Describa el(los) proceso(s) mediante el(|los) cual(es) maneja Ud. el síndrome metabólico al tratar a un paciente.
1.3. ¿Cuáles son las variantes en el proceso, cuando un paciente viene por primera vez y cuando ya es una visita subsecuente?

NO SE HABÍA DETECTADO

1.4. Describa las decisiones a las que se enfrenta para determinar si una persona corre el riesgo de padecer el síndrome metabólico
1.5. Describa Ud. los diferentes cursos de acción que se puede seguir con un paciente que tiene el riesgo de padecer el síndrome metabólico.
1.6. ¿Cuáles son las preocupaciones e incertidumbres a las cuales se enfrenta para determinar si el paciente padece el síndrome metabólico?
1.7. Puntualice ¿Qué otras personas intervienen en dichos procesos y cuáles son los roles que ellos juegan?
1.8. ¿Cuáles son los datos que permiten determinar el riesgo de padecer el síndrome metabólico? ¿De quién recibe dichos datos?
1.9. ¿Qué datos genera Ud. que están relacionados con el tratamiento del síndrome metabólico?

YA SE HABÍA DETECTADO

1.10. En este caso, describa cuales son las diferencias con el proceso descrito anteriormente, puntualizando la relación con los servicios de enlace, nutriólogo y sicólogo.
1.11. En este caso, describa las decisiones a las que se enfrenta en el manejo del síndrome metabólico y su comorbilidad.
1.12. ¿Cuáles son las preocupaciones e incertidumbres a las cuales se enfrenta en el manejo del síndrome metabólico y su comorbilidad?

1.13. Puntualice ¿Qué otras personas intervienen en dichos procesos y cuáles son los roles que ellos juegan?

1.14. ¿Qué datos adicionales (de acuerdo a la evolución del caso) posiblemente necesite o genere Ud. para el manejo del síndrome metabólico?

Si no lo conoce

1.15. ¿Podría describir cuál es el proceso que Ud. sigue en padecimientos como la obesidad, DM2, dislipidemias, ECV,HTA?.

1.16. ¿Cuáles son las principales preocupaciones e incertidumbres que Ud. tiene al manejar las enfermedades del punto 1.15?

1.17. ¿Qué información necesita Ud. para tratar las enfermedades del punto 1.15

2. ¿Cuáles son las circunstancias que hacen que Ud. decida solicitar estudios adicionales para un paciente?
3. Indique cuales de los siguientes programas e instrumentos utiliza Ud. en su práctica médica
   3.1. Norma Mexicana para el manejo integral de la obesidad
   3.2. Programa SODHI
   3.3. Norma Mexicana, Control de la nutrición, crecimiento y desarrollo del niño y del adolescente.
   3.4. PREVENIMSS
   3.5. Guía Clínica IMSS para el manejo de la diabetes mellitus
   3.6. Norma Mexicana para la prevención, tratamiento, control de la diabetes
   3.7. Guía Médica IMSS para el manejo de la hipertensión arterial
   3.8. Norma Mexicana para la prevención, tratamiento y control de la hipertensión arterial.
4. ¿Podría describir la forma en que utiliza Ud., los elementos enlistados en el punto anterior, para la toma de decisiones relacionadas con el síndrome metabólico?
5. ¿Conoce Ud. los sistemas SIAIS y SIMO?

Si los conoce

5.1. ¿Los utiliza? ¿Cuál de los dos?

5.2. ¿Qué información consulta en dichos sistemas?

5.3. ¿Qué datos de los que Ud. registra sobre el síndrome metabólico se ven reflejados en dichos sistemas?

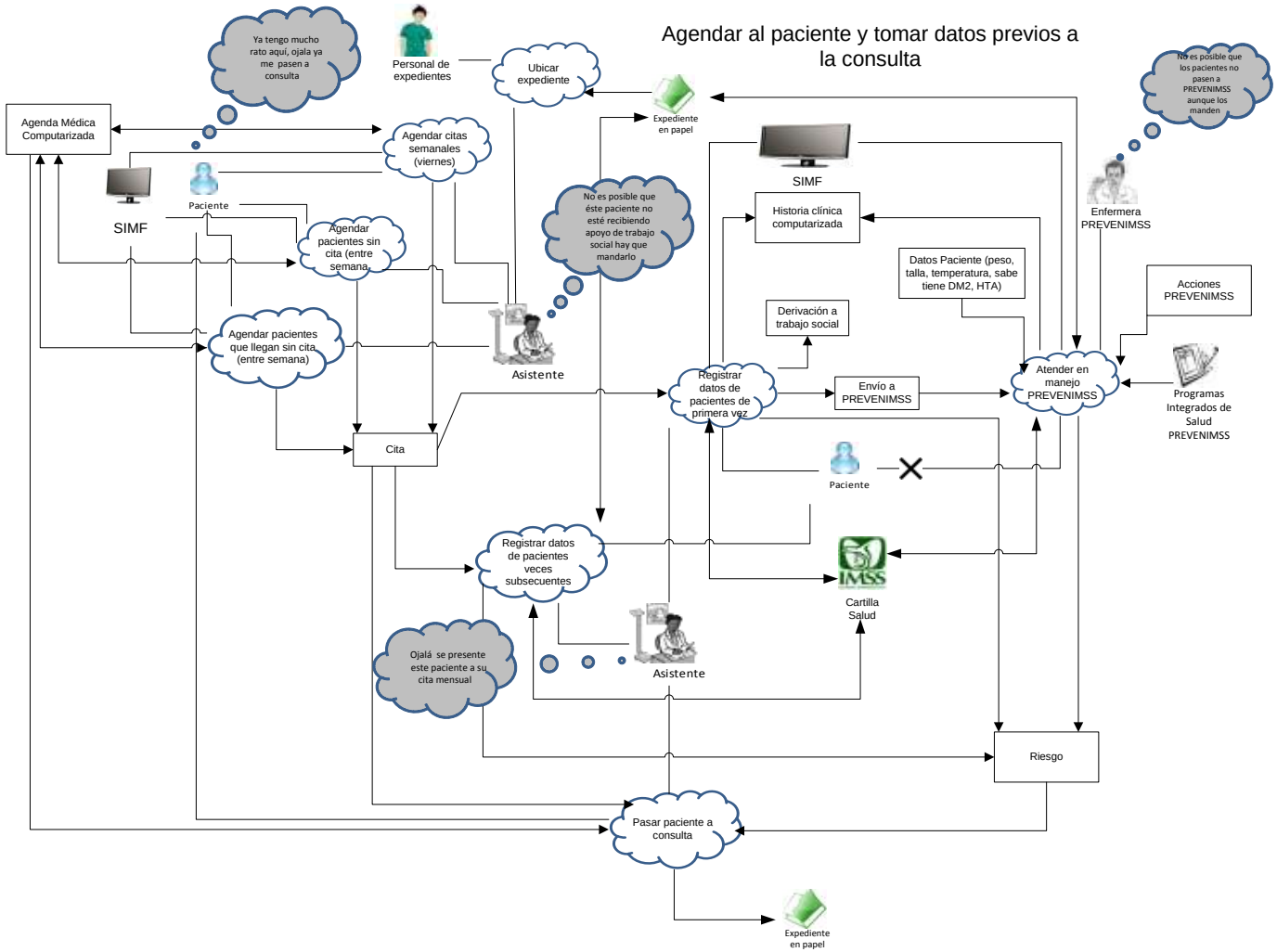


Figura 64. Gráfica rica agendado de pacientes.

Anexo E. Gráficas del modelado de procesos.

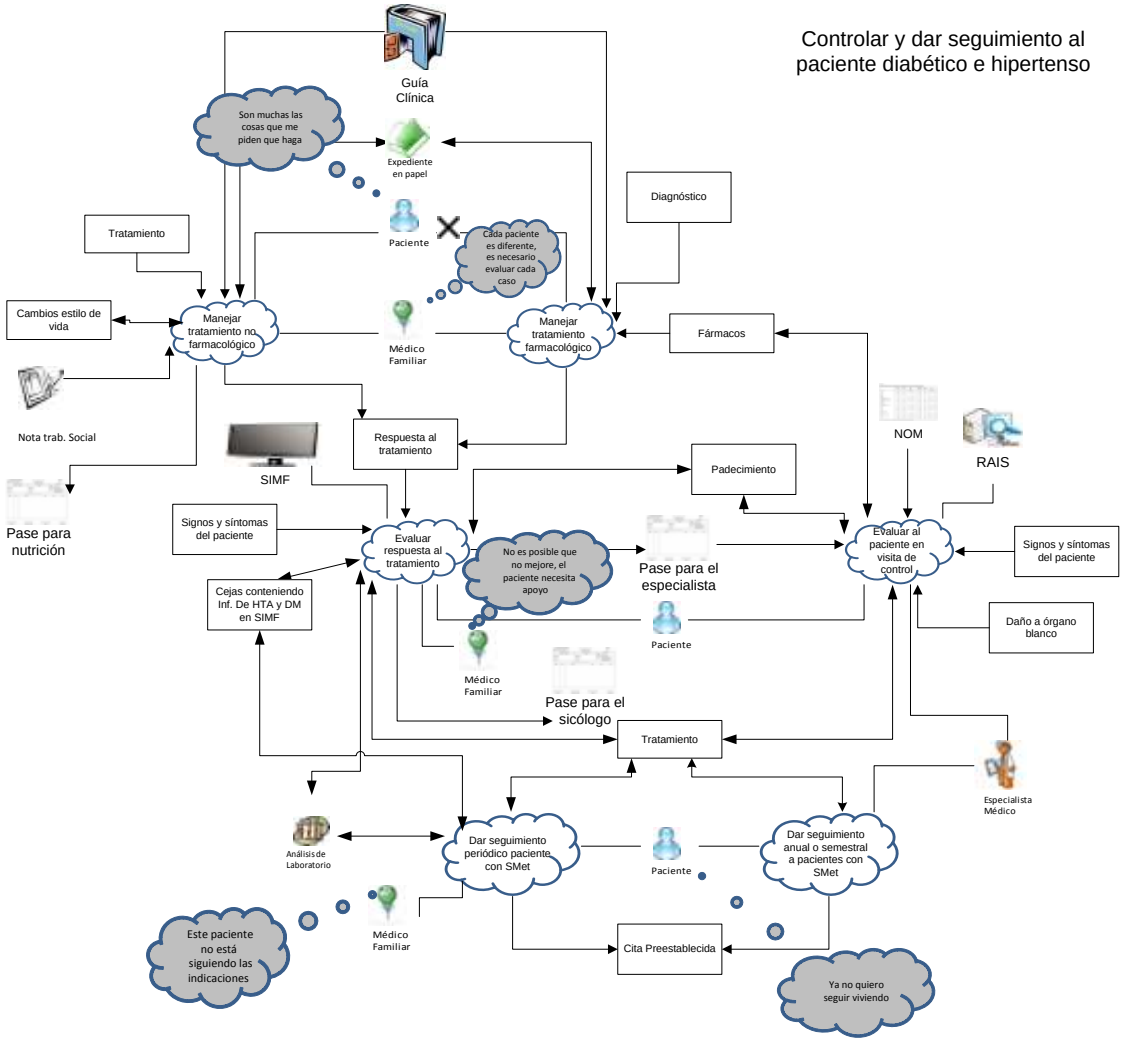


Figura 65. Gráfica rica control de pacientes con DM2 y HTA.

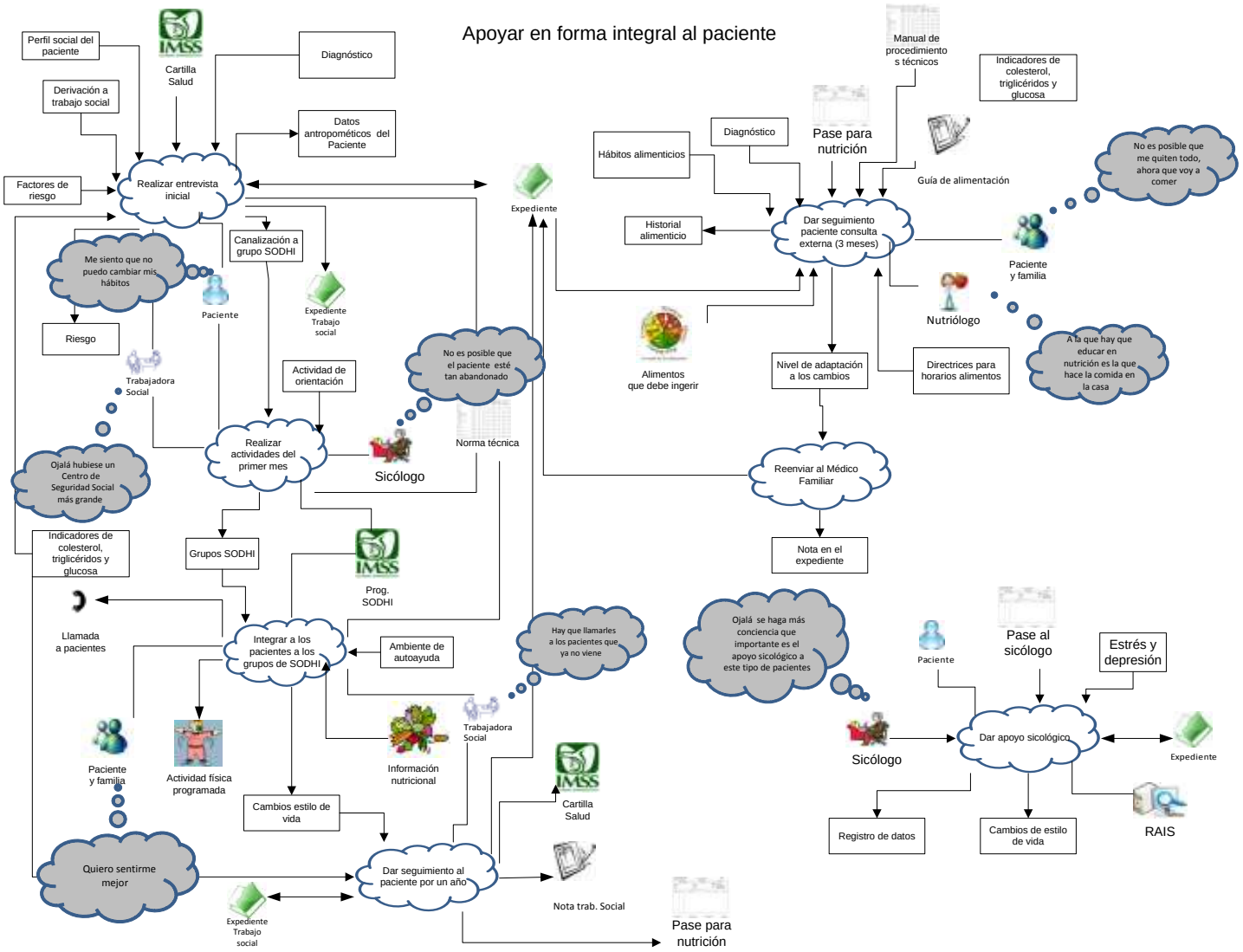


Figura 66. Gráfica rica de control integral del paciente

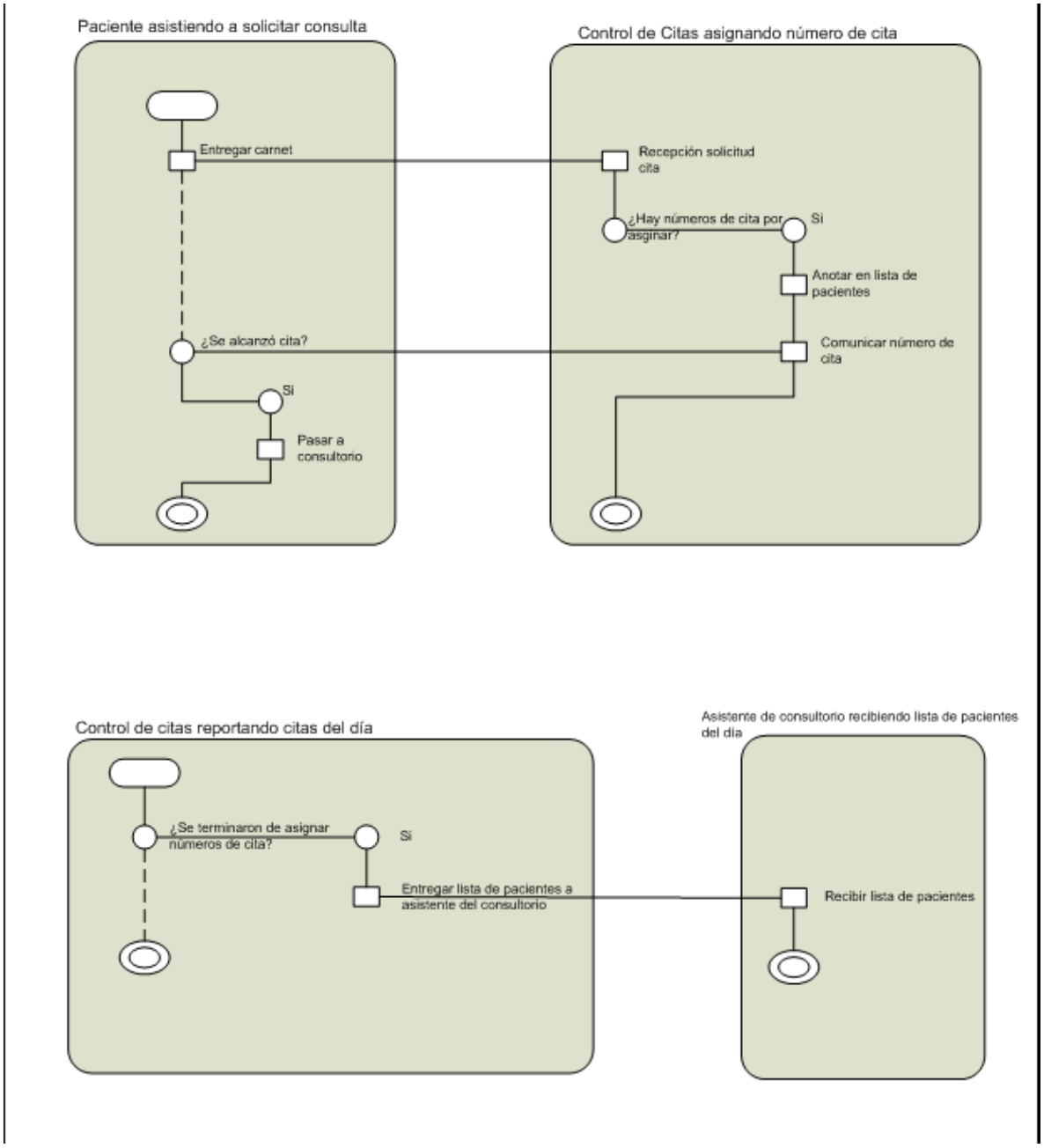


Figura 67. Diagrama IDEF0 para agendado del paciente.

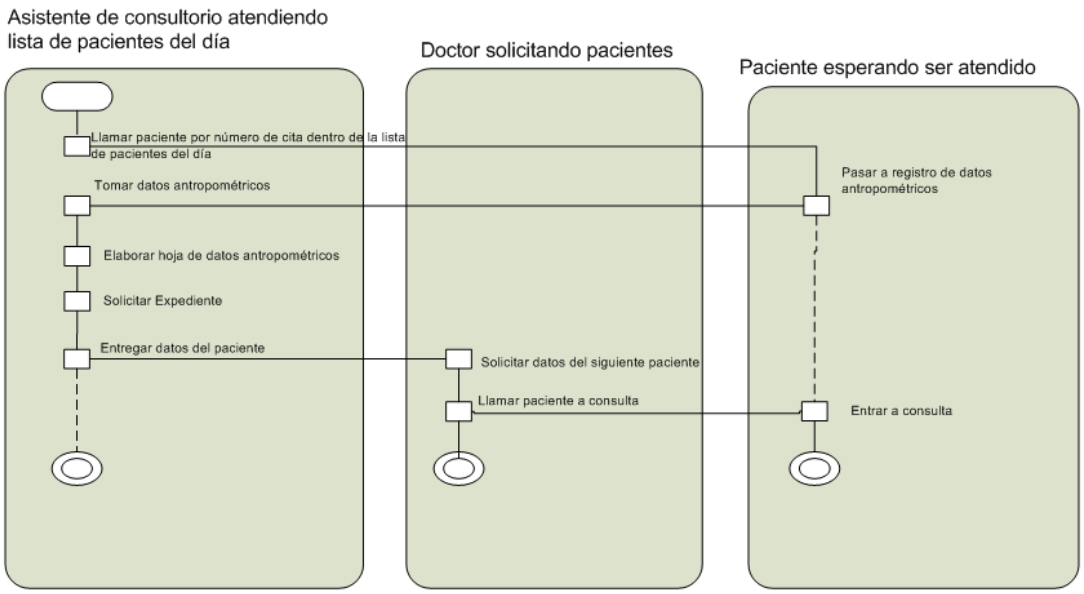


Figura 68. Diagrama IDEF0 para control de consultas medicina familiar.

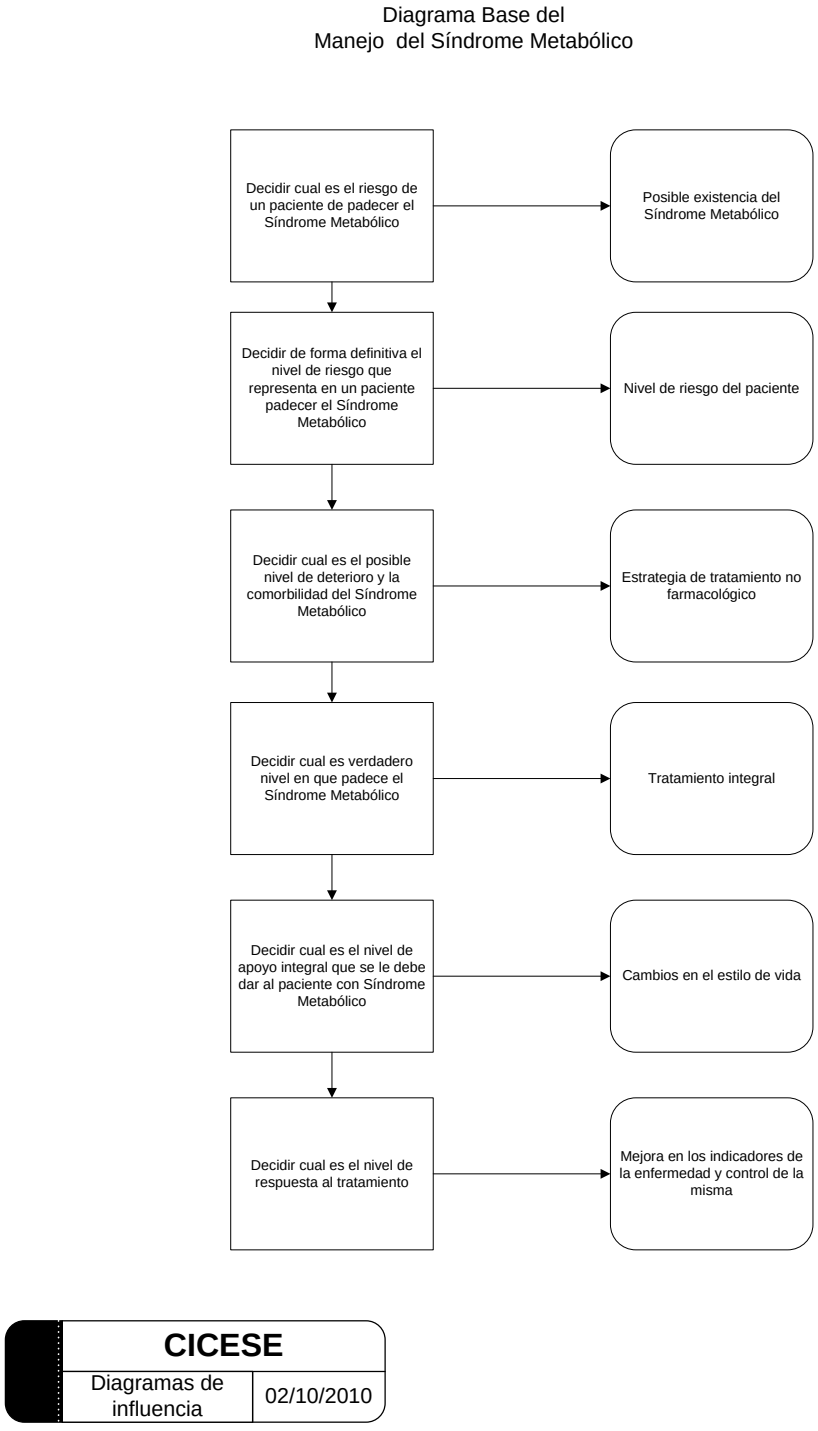
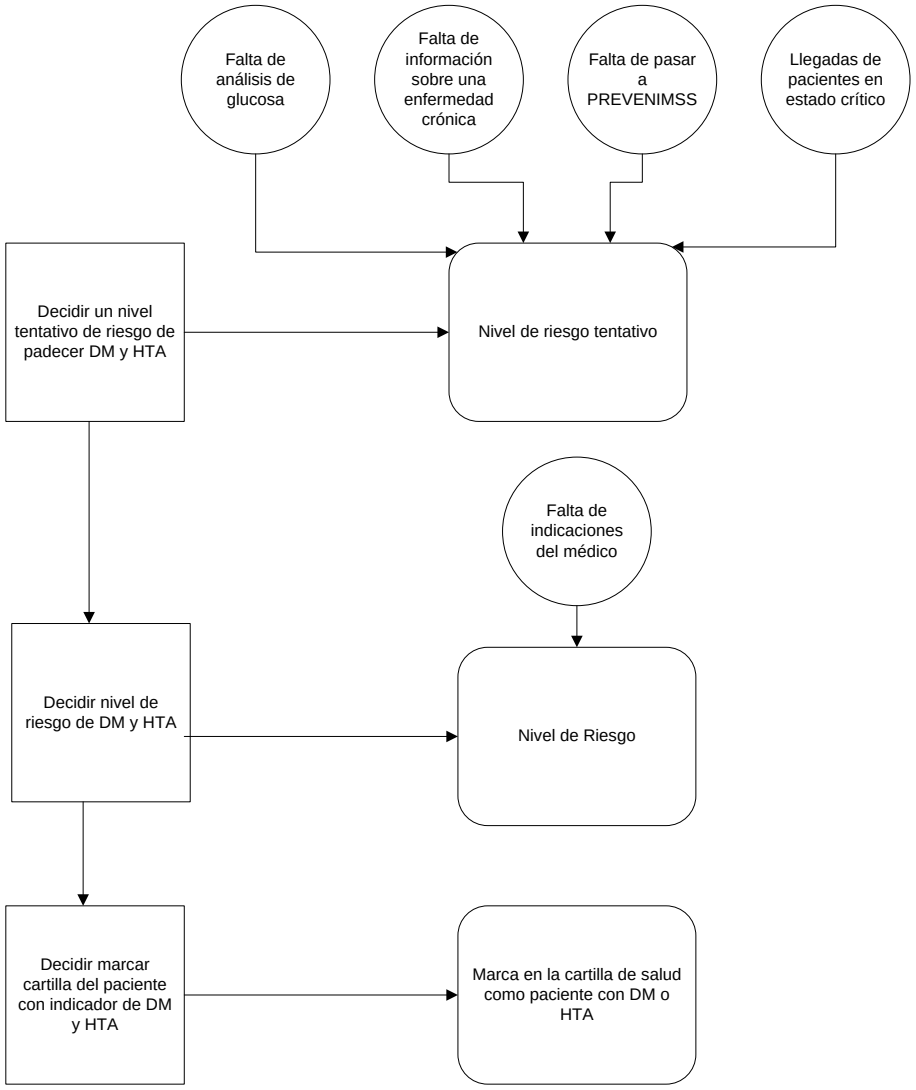


Figura 69. Diagrama de influencia general del manejo del síndrome metabólico.

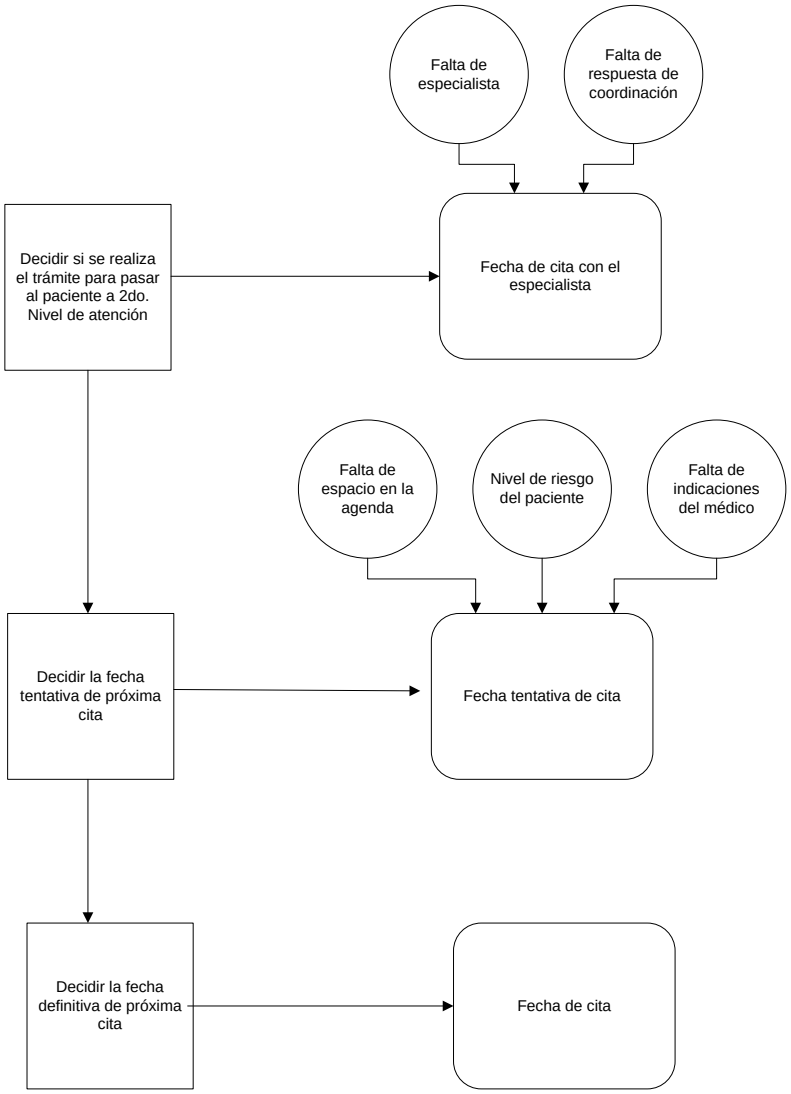
ASISTENTE MÉDICO
Manejo de identificación del paciente hipertenso o diabético



CICESE	
Diagramas de influencia	02/10/2010

Figura 70. Diagrama de influencia manejo de pacientes con DM2 y HTA.

ASISTENTE MÉDICO
Manejo de control de citas para el paciente hipertenso o diabético



CICESE	
Diagramas de influencia	02/10/2010

Figura 71. Diagrama de influencias control de citas para el paciente hipertenso y diabético.

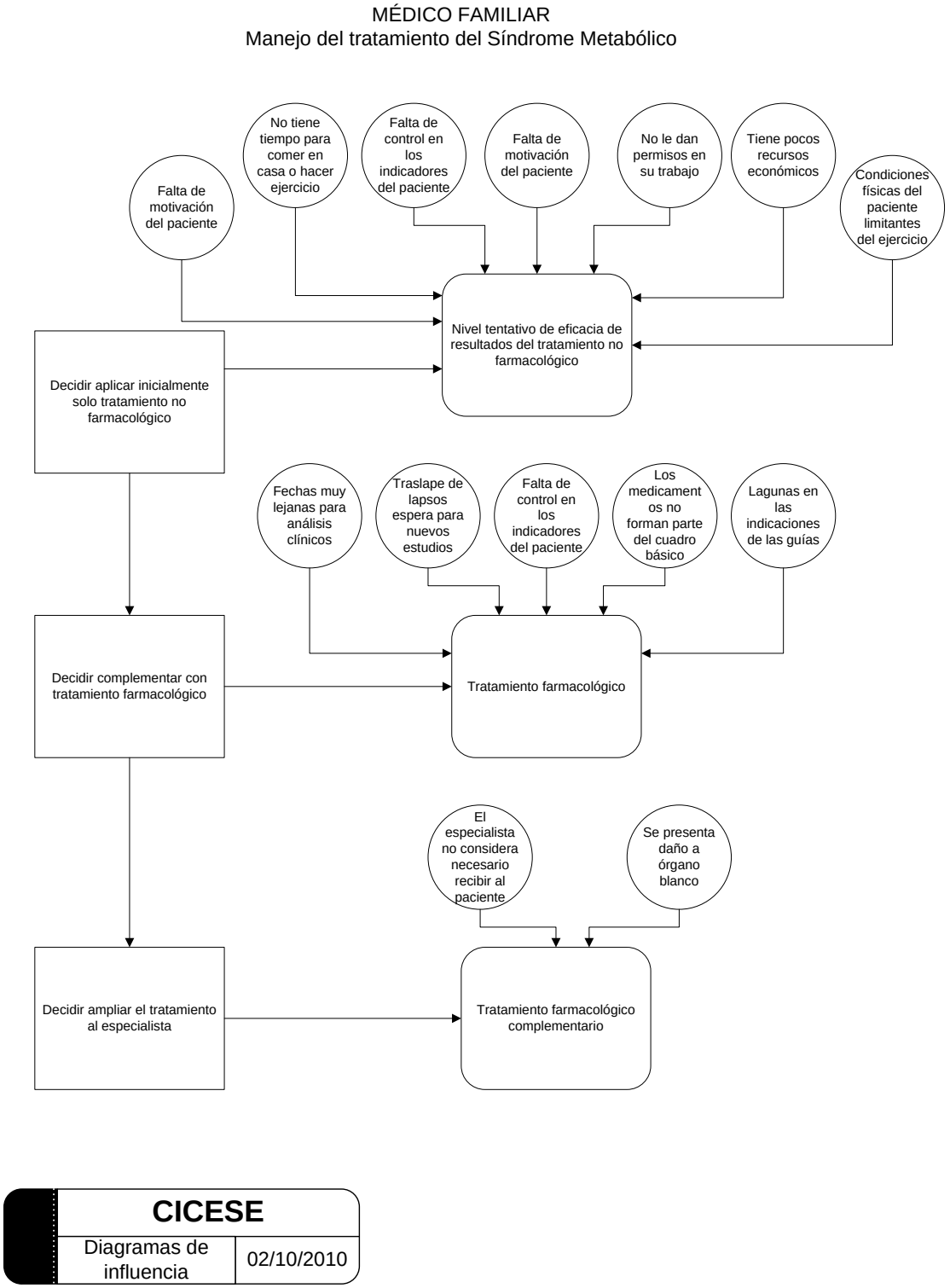
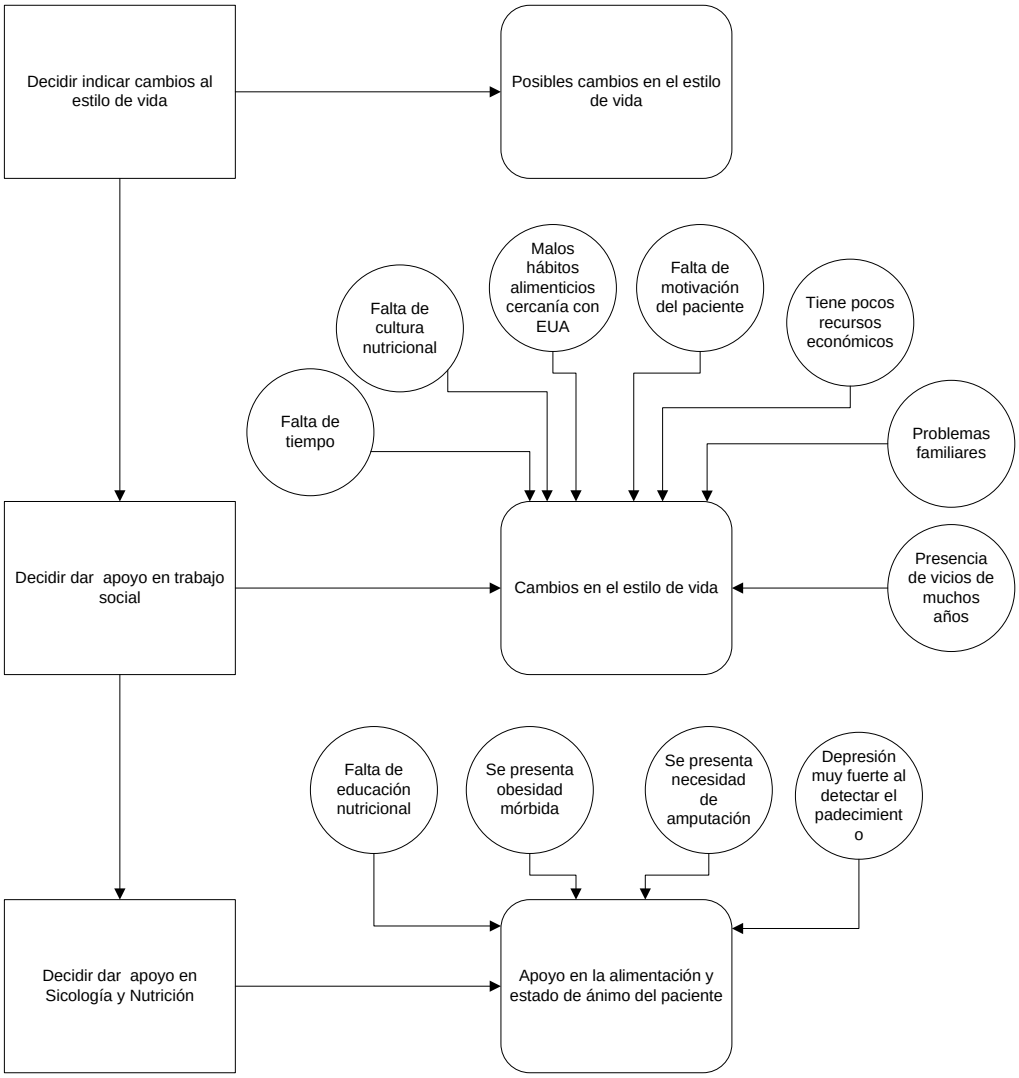


Figura 72. Diagrama de influencias médico familiar determinación del tratamiento en el manejo del síndrome metabólico.

MÉDICO FAMILIAR
Manejo integral del tratamiento del Síndrome Metabólico



CICESE	
Diagramas de influencia	02/10/2010

Figura 73. Diagrama de influencias manejo integral del paciente con síndrome metabólico.

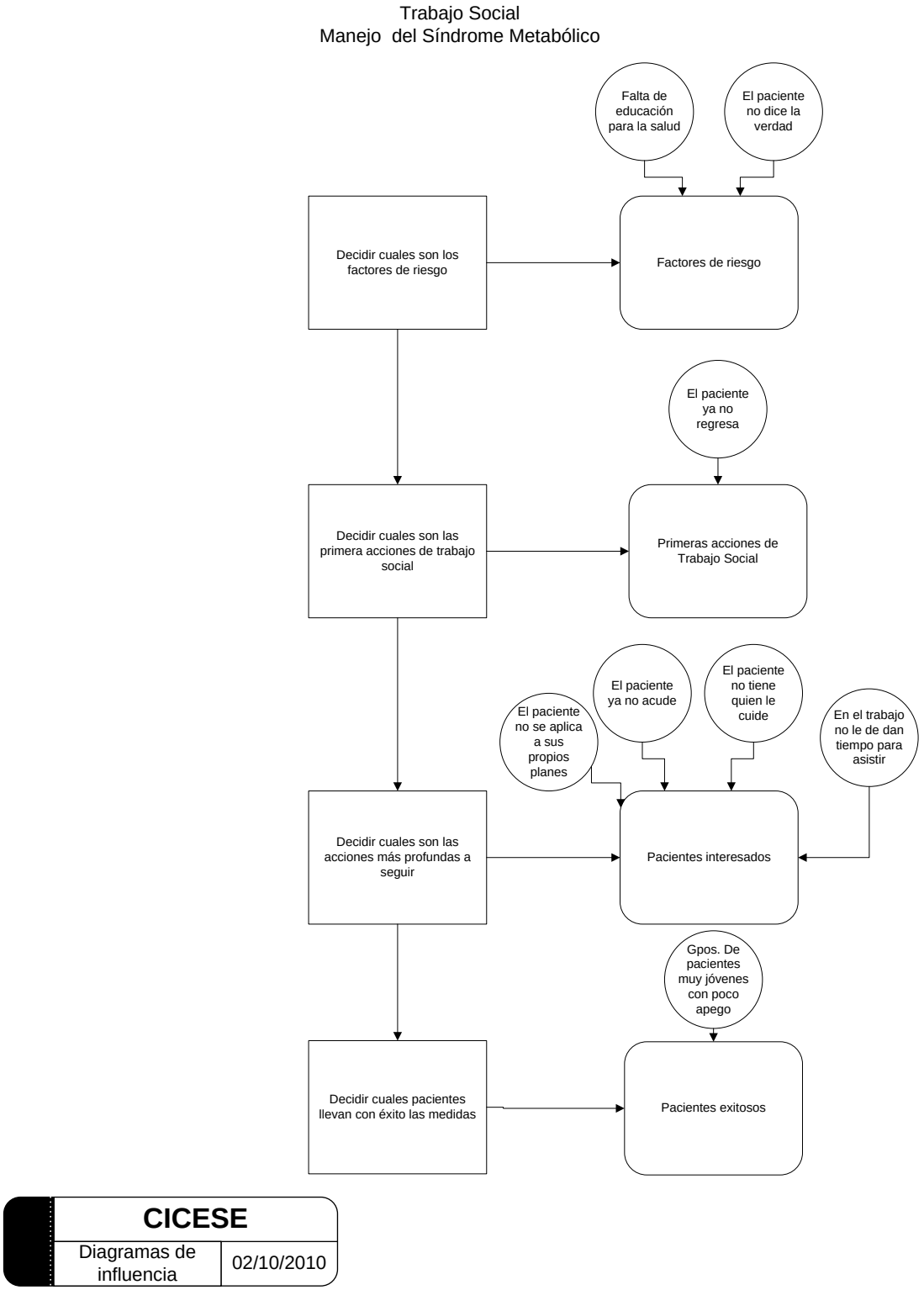


Figura 74. Diagrama de influencias manejo del síndrome metabólico en trabajo social.

Apéndice F. Casos de uso.

Caso de Uso: Registrar datos actualizados del paciente.

Actores: Médico Familiar, Médico Epidemiólogo.

Propósito: Registrar o actualizar los datos actuales del paciente dentro de la base de datos.

Descripción:

El caso de uso inicia cuando el médico recibe a un paciente en consulta (ya sea de primera vez o subsecuente) e introduce en el sistema el número del IMSS del paciente. Si el paciente existe dentro de la base de datos el sistema recupera la información del paciente, si no existe el sistema debe permitir se capturen los datos del paciente (número del IMSS, nombre, edad, género, peso, talla, colesterol total, triglicéridos, tensión arterial sistémica, indicador de si bebe o fuma, índice de glicemia, indicador si se tiene Diabetes Mellitus 2).

Los datos que el médico familiar puede capturar o modificar se dividen en 3 grupos:

- Información personal: número del IMSS, nombre, edad y género.
- Información antropométrica: peso, talla. Con esta información el sistema debe calcular y presentar inmediatamente el IMC y tipo de obesidad detectada (ver Apéndice A).
- Análisis de laboratorio: (colesterol total, triglicéridos, tensión arterial sistémica, indicador de si el paciente bebe, indicador de si el paciente fuma, índice de glicemia e indicador de si el paciente padece diabetes mellitus tipo 2.

Una variante del caso de uso es cuando el médico solicita apoyo para determina el nivel de riesgo, el sistema debe tener una opción de apoyo a la toma de decisiones por lo que el caso se extiende en el caso de uso evaluar riesgo.

Caso de uso: Registrar población.

Actores: Médico Epidemiólogo.

Propósito: Permitir registrar (o leer) los datos de un grupo de pacientes a los que se les quiere evaluar el síndrome metabólico.

Descripción:

El caso de uso inicia cuando el médico epidemiólogo está evaluando algún tipo de riesgo relacionado con el síndrome en un grupo de pacientes y quiere identificar qué clase de síndrome metabólico tienen. El sistema debe tener la funcionalidad de leer los datos de un archivo de texto que el médico epidemiólogo podrá seleccionar libremente de algún

directorio, siempre y cuando los datos estén de acuerdo al formato que aparece en el apéndice A. Adicionalmente el sistema debe presentar una pantalla donde se puedan capturar hasta 20 registros para clasificarles en conjunto.

Caso de uso: Graficar indicadores.
Actores: Médico Epidemiólogo.
Propósito: Presentar una gráfica en línea de la clasificación obtenida del grupo de pacientes**.**
Descripción:
El caso de uso inicia cuando después de evaluar un grupo de pacientes el médico epidemiólogo quiere obtener una gráfica sobre la clasificación obtenida, el sistema de presentar en línea una gráfica de cuantos pacientes clasificaron en cada una de las diferentes clasificaciones del síndrome metabólico.