Modelo para Análisis de Riesgo de la Diabetes Mellitus 2 usando Inteligencia de Negocios y Minería de Datos

Angela María Franco Pérez
Ingeniera de Sistemas, Msc(c)
Código: 300498

Universidad Nacional de Colombia
Facultad de Ingeniería
Departamento de Sistemas e Industrial
Bogotá, D.C.
Enero de 2014

Modelo para Análisis de Riesgo de la Diabetes Mellitus 2 usando Inteligencia de Negocios y Minería de Datos

Angela María Franco Pérez
Ingeniera de Sistemas, Msc(c)
Código: 300498

Trabajo de tesis para optar al título de
Magister en Ingeniería de Sistemas y Computación

Director
Elizabeth León Guzman, Ph.D.
Ingeniera de Sistemas, Ph.D.

Universidad Nacional de Colombia
Facultad de Ingeniería
Departamento de Sistemas e Industrial
Bogotá, D.C.
Enero de 2014

Título en español

Modelo para Análisis de Riesgo de la Diabetes Mellitus 2 usando Inteligencia de Negocios y Minería de Datos.

Title in English

Risk analysis model of Diabetes Mellitus Type 2 using business intelligence and data mining

Resumen: La diabetes mellitus tipo 2 (DM2) es una enfermedad crónica caracterizada por una hiperglucemia y trastornos en el metabolismo de las grasas, hidratos de carbono y proteínas de forma tal que genera defectos en la producción y acción de la insulina en el cuerpo. Esta enfermedad presenta complicaciones crónicas que deterioran la calidad de vida de los pacientes y aumentan significativamente el riesgo de muerte. Para Colombia, es claro que se debe tener una prioridad en la detección temprana y aseguramiento de intervenciones para la diabetes mellitus 2. Este documento de tesis presenta un modelo de análisis de riesgo de la DM2 basado en inteligencia de negocios y minería de datos, el cual permite integrar y transformar datos clínicos, caracterizar pacientes y describirlos teniendo en cuenta sus diagnósticos, consumos y entorno social. Adicionalmente, el modelo permite predecir si un paciente puede o no sufrir comorbilidad asociada a DM2 y de qué tipo puede ser. La caracterización de pacientes se realiza a través de un algoritmo de agrupación y la descripción se hace mediante el uso de reglas de asociación. El modelo de predicción por su parte, utiliza árboles de decisión y redes bayesianas. El caso de estudio consistió de una cohorte de 14162 pacientes reales enfermos de DM2 proporcionados por la empresa Processum LTDA, con registros de diagnósticos, procedimientos clínicos y variables socio culturales desde el año 2009 hasta el 2012.

Abstract: Type 2 Diabetes mellitus (T2DM) is a chronic disease characterized by hyperglycemia and disorders in the metabolism of fat, carbohydrate and protein in a manner that produces defects in the production and action of insulin in the body. This disease presents chronic complications that deteriorate patients life quality and significantly increase the risk of death. In Colombia, it is clear that ensuring early detection and intervention for type 2 diabetes mellitus should be a priority. This thesis paper presents a model for risk analysis of DM2 based on business intelligence and data mining. This model allows to characterize and describe patients by using their diagnoses, clinical procedures and social environment records. Additionally, the model can predict whether a patient may or may not suffer T2DM related comorbidity and which type of comorbidity it can be. Patients characterization is performed through a clustering algorithm, and the description is made by using the association rules. The prediction model uses decision trees and Bayesian networks. The case study consisted of a cohort of 14162 patients with DM2 real patients provided by the company processum LTDA, with records of diagnoses, clinical procedures and sociocultural variables from 2009 to 2012.

Palabras clave: Inteligencia de Negocios, Agrupación, Reglas de Asociación, Clasificación, Perfiles de pacientes, Diabetes Mellitus Tipo 2, Minería de Datos

Keywords: Business intelligence, Clustering, Association Rules, Classification, Patients Profiles, Type 2 Diabetes Mellitus, Data Mining

Nota de aceptación

Trabajo de tesis
Aprobado

"Mención Meritoria"

Jurado
Emiliano Barreto

Jurado
Eduardo Romero

Director
Elizabeth León

Asesor
Fernando Ruíz

Bogotá, D.C., Enero de 2014

Dedicado a

Dios, mis papás Julio y Dora, mi abuelita Flor y mi novio Javier.

Agradecimientos

Quiero expresar mi agradecimiento a Dios por darme fortaleza y permitirme aprender más cada día. A mis papás, Julio y Dora, por su compañía, apoyo, comprensión y motivación a lo largo de este proyecto. A mi abuelita Flor por sus palabras de ánimo y consejos. A mi novio Javier por su paciencia, apoyo incondicional y por ser un motivo de felicidad.

A mi directora Elizabeth por su compromiso, ayuda, paciencia y su enseñanza.

A Liz Garavito por sus consejos, enseñanzas, comprensión y su apoyo en este proyecto.

También quiero extender mi agradecimiento a Processum LTDA, a Fernando Ruiz y a Alexandra Castillo por su ayuda y contribución para que este trabajo llegara a feliz término.

Índice general

Índice de tablas

Índice de figuras

Introducción

La diabetes mellitus tipo 2 (DM2) es una enfermedad crónica caracterizada por una hiperglucemia y trastornos en el metabolismo de las grasas, hidratos de carbono y proteínas de forma tal que genera defectos en la producción y acción de la insulina en el cuerpo. Esta enfermedad presenta complicaciones crónicas que deterioran la calidad de vida de los pacientes y aumentan significativamente el riesgo de muerte [15]. Los pacientes de diabetes utilizan con mucha más frecuencia servicios especializados como odontología, optometría y nutricionistas, los cuales incrementan el costo de la diabetes. Adicionalmente, el costo derivado del tratamiento de la diabetes supone casi el doble de un paciente no diabético y además consumen entre 2 y 6 veces más recursos que los pacientes de edad y sexo similares con otras enfermedades crónicas [2]. Para Colombia se estimó la prevalencia de la DM2 en 4.8 % para los grupos de edad de 20 a 79 años, y un valor de 5.2 % ajustada por edad. Esto quiere decir que aproximadamente un millón y medio de personas sufren de DM2 en el país[13].

Las comorbilidades asociadas a la DM2 pueden llevar a sus pacientes a la muerte o altos grados de discapacidad, particularmente las comorbilidades de tipo micro y macrovascular. El tratamiento de estas complicaciones representa un alto impacto en el costo global de un paciente con DM2. En Colombia, el costo de las comorbilidades de DM2 abarcan un 86 % del costo directo de la enfermedad y un 95 % del costo indirecto [25] y enfermedades como la cardiopatia isquémica o cardiopatía hipertensiva, ambas relacionadas a comorbilidades de la DM2, se encuentran entre las enfermedades con más índice de mortalidad en el país [1].

La correcta identificación de factores de riesgo asociados a una enfermedad crónica en Colombia como lo es la Diabetes Mellitus Tipo 2 (DM2), puede ayudar a su oportuna prevención. Mediante la creación de niveles de riesgo usando estos factores, las Entidades de Salud pueden saber qué población requiere planes específicos de prevención. También al tener un control sobre la población, es posible disminuir la incidencia de la enfermedad o condición crónica y por ende se reducirían también los costos derivados del tratamiento de la enfermedad.

A nivel internacional, la principal aproximación a la gestión en salud es el uso de los sistemas de clasificación de pacientes como los ACG (Adjusted Clinical Groups), desarrollados por la Universidad Johns Hopkins, los cuales permiten clasificar grupos pacientes teniendo en cuenta sus consumos y sus características [3]. En paises como México, Colombia, Perú y Argentina se estan implementando en sus fases iniciales y aún no se han evaluado sus resultados [4].

Adicionalmente, algoritmos de minería de datos con enfoque de aprendizaje maquinal se han utilizado sobre datos clínicos en otros países para encontrar factores de riesgo asociados condiciones clínicas de interés como la gestación [46]. Particularmente se ha trabajado con árboles de decisión y redes bayesianas como clasificadores de pacientes con enfermedades asociadas a distintos factores de riesgo [31, 21], y reglas de asociación para encontrar las posibles relaciones entre dichos factores [9].

Actualmente, en Colombia se trabaja con la aplicación de técnicas estadísticas sobre datos de pacientes para encontrar los factores de riesgo de enfermedades de interés en el país como lo son las cardiovasculares [37]. Cabe resaltar que en estudios relacionados con el tema se tratan áreas como la estadística descriptiva y métodos de regresión lineal y logística, de modo que se puede descubrir conocimiento sobre una población. Sin embargo, la búsqueda de relaciones interesantes entre los datos médicos puede mejorarse con la ayuda de técnicas computacionales de aprendizaje maquinal para descubrir patrones ocultos entre ellos.

La información de DM2 en Colombia se encuentra en bases de datos de las EPS (Entidad Prestadora de Servicios de Salud) por lo que es necesario centralizar los datos médicos y realizar las transformaciones necesarias sobre éstos para que no solo cumplan con una labor administrativa sino que la información que poseen pueda ser utilizada en análisis de riesgo en salud. De esta forma, es posible aplicar fácil y efectivamente algoritmos de minería de datos ya sean técnicas estadísticas o de aprendizaje maquinal que permitan identificar factores de riesgo en poblaciones y niveles de severidad de las enfermedades en los pacientes.

En esta tesis se desarrolló un modelo de análisis de riesgo para la Diabetes Mellitus 2 para la población colombiana, el cual se compone de dos submodelos: un modelo de datos que centraliza toda la información de un conjunto de pacientes colombianos a través de una bodega de datos, y un modelo de minería de datos el cual se basa en la caracterización de los pacientes de DM2 y la predicción de comorbilidades de DM2.

Este trabajo de investigación se realizó con datos reales que pertenecen a una cohorte de 14162 pacientes colombianos residentes en zona urbana del centro del país comprendida entre los años 2009 y 2012, suministrados generosamente por la empresa Processum LTDA[1].

Objetivos

General

Diseñar un modelo de análisis de riesgo para la Diabetes Mellitus 2, contextualizado en Colombia, basado en inteligencia de negocios y minería de datos.

[1]Empresa privada con fines de investigación aplicada para la generación de soluciones innovadoras que buscan favorecer el diseño e implementación de políticas, toma de decisión y optimización de la gestión instituciones de sectores sociales, en particular en el sector salud. Tomado de: http://processum.org/

Específicos

1. Integrar y caracterizar los datos clínicos relacionados a la Diabetes Mellitus 2 en Colombia por medio de una Bodega de Datos, teniendo en cuenta la consistencia, completitud y oportunidades de recolección, extensión o modificación de los mismos.
2. Diseñar un modelo de caracterización personas enfermas de Diabetes Mellitus 2 mediante una técnica de agrupación de computación suave.
3. Describir los perfiles de pacientes generados a partir del uso de reglas de asociación.
4. Diseñar un modelo de predicción de comorbilidades de DM2 basado en un meta-clasificador.
5. Validar los resultados obtenidos por el modelo de clasificación y el de caracterización en términos computacionales y adicionalmente realizar una validación con expertos[2] [3].

Alcances y Limitaciones

Los siguientes son algunos aspectos que restringen los resultados de este proyecto:

1. Los datos reales de pacientes de Diabetes Mellitus tipo 2 presentan problemas de completitud real, por lo que para ciertos análisis no siempre se contará con el 100 % de los datos de los pacientes.
2. No se cuenta con todos los datos de la población colombiana, sino con un conjunto de 14162 pacientes de la zona urbana del centro del país. Esto corresponde a aproximadamente el 1 % de la población de DM2, sin embargo por ser población de área urbana es representativa debido a que la prevalencia de la enfermedad es mayor en zonas urbanas. Adicionalmente, se presenta un problema de identificación de pacientes por medio sistemático lo que reduce considerablemente el número de pacientes con información médica requerida para realizar análisis como el propuesto en este trabajo.
3. Se utilizarán herramientas de software libre para la ejecución de algoritmos y preprocesamiento de los datos.

Contribuciones

Los aportes de este trabajo investigativo son:

1. Diseño e implementación de una bodega de datos que almacena los principales factores que hacen parte de un análisis de riesgo de la Diabetes Mellitus tipo 2 en Colombia. Estos son diagnósticos, procedimientos clínicos y variables socio culturales de pacientes. El diseño se muestra en la sección 2.2.3.

[2] Liz Garavito, Directora ejecutiva de Processum LTDA e investigadora en el área de seguridad social y economía de la salud.

[3] Alexandra Castillo, Enfermera y Epidemióloga.

2. Conjunto de datos reales médicos de pacientes de DM2 preprocesado y listo para aplicar diversas técnicas de minería de datos.

3. Reportes de interés médico generados con OLAP sobre la bodega de datos médicos.

4. Modelo de caracterización de pacientes de DM2 basado en agrupación y reglas de asociación. El diseño se muestra en la sección 2.3.3.1.

5. Modelo de predicción de comorbilidades basado en un meta-clasificador. El diseño se encuentra en la sección 2.3.3.2.

Publicaciones

Las publicaciones obtenidas con este trabajo investigativo fueron:

1. Perez, A. & Guzman, E. An approach to the risk analysis of diabetes mellitus type 2 in a health care provider entity of Colombia using business intelligence Telematics and Information Systems (EATIS), 2012 6th Euro American Conference on, 2012, 1-8. *Artículo publicado.* Valencia, España.

2. Franco, A. León E. Meta-classifier for Type 2 Diabetes Mellitus comorbidities in Colombia. 15th IEEE International Conference on e-Health Networking, Application & Services (IEEE Healthcom 2013), 2013. *Artículo Publicado* Lisboa, Portugal.

Estructura del Documento de Tesis

Este documento de tesis se encuentra organizado de la siguiente manera:

- El capítulo 1 contiene el estado del arte sobre la inteligencia de negocios utilizada para analizar riesgo de enfermedades en el sector salud. Se presenta la introducción a la enfermedad que se quiere trabajar en esta tesis desde el punto de vista médico, se describen las herramientas que presenta la metodología KDD para extraer conocimiento valioso de los datos y su aplicación en el sector salud describiendo los métodos estadísticos y computacionales más usados.

- El capítulo 2 presenta el modelo de análisis de riesgo de la DM2 basado en inteligencia de negocios y minería de datos. Se describe el modelo general y se definen los modelos que lo integran: el modelo de datos, comprendido por la bodega de datos clínicos, y el modelo de minería de datos, el cual comprende la caracterización de pacientes de DM2 y la predicción de comorbilidades asociadas a la DM2.

- El capítulo 3 contiene los experimentos realizados en el modelo de minería de datos. Para la parte de caracterización de pacientes, se describen los experimentos hechos con algoritmos de agrupación, así como sus resultados y para la parte de predicción de comorbilidades, se muestran los experimentos y resultados de los clasificadores seleccionados. En este capítulo también se describe la validación hecha por las analistas de riesgo en salud para los resultados de ambos modelos.

- El capítulo 4 presenta las conclusiones obtenidas del trabajo desarrollado.

- Finalmente, el capítulo 5 muestra algunas ideas y perspectivas de trabajos futuros.

CAPÍTULO 1

Estado del Arte

En este capítulo se presentará de manera conjunta el marco teórico y los trabajos previos que se han realizado en el campo de análisis de riesgo en salud desde el punto de vista estadístico, el cual es el más utilizado en estudios epidemiológicos, y el campo computacional mediante el uso de inteligencia de negocios y técnicas de minería de datos.

Se describirán los conceptos de la enfermedad Diabetes Mellitus 2, factores de riesgo, inteligencia de negocios, bodegas de datos y minería de datos, así como las técnicas más utilizadas en esta rama. También se presentarán de manera general el uso de las técnicas estadísticas y computacionales sobre datos de salud.

1.1. Importancia del análisis de datos médicos y la gestión del riesgo en salud

Hoy en día, se están realizando grandes esfuerzos para que el sector de la Salud tenga una mayor cobertura y realice una buena gestión de sus recursos. Así mismo, se pretende que la población pueda gozar de una mejor calidad de vida mediante la implantación de programas para la prevención de enfermedades de alto riesgo. Con el fin de asegurar el cumplimiento de esta meta, varias organizaciones médicas, han venido desarrollando programas de análisis de riesgo en Salud para encontrar poblaciones expuestas a los factores de riesgo de dichas enfermedades y ofrecerles de forma oportuna programas de prevención.

A nivel internacional, la principal aproximación a la gestión en salud es el uso de los sistemas de clasificación de pacientes como los ACG (Adjusted Clinical Groups), desarrollados por la Universidad Johns Hopkins, los cuales permiten clasificar grupos pacientes teniendo en cuenta sus consumos y sus características [3]. En paises como México, Colombia, Perú y Argentina se estan implementando en sus fases iniciales y aún no se han evaluado sus resultados [4].

Las ventajas que permiten los ACG radican en su unidad de análisis la cual es el paciente. Adicionalmente, solo requiere de las variables: edad, sexo y los diagnósticos que han presentados los pacientes. Sin embargo, entre sus limitantes se encuentran la dependencia de un nivel de informatización de los datos de salud, estandarización de los

datos suministrados y las dificultades de interpretación de resultados por parte de los profesionales clínicos [4].

En Colombia el Ministerio de Protección Social mediante la Resolución 1445 de 2006, plantea que las entidades prestadoras de salud EPS del país deben realizar planes de atención a sus usuarios basándose en la información que puedan recopilar de éstos. Ésto incluye la detección de factores de riesgo, grupos de riesgo y perfiles de pacientes los cuales permitan orientar los planes de prevención, promoción y educación de las EPS. La información de los pacientes debe ayudar también a identificar necesidades de los pacientes y orientar el manejo del riesgo a enfermar de sus afiliados.

Adicionalmente, en Colombia se trabaja con la aplicación de técnicas estadísticas sobre datos de pacientes para encontrar los factores de riesgo de enfermedades de interés [37]. Cabe resaltar que en estudios relacionados con la búsqueda de factores de riesgo se tratan áreas como la estadística descriptiva y los modelos basados en regresiones, los cuales han sido ampliamente usados para descubrir conocimiento sobre una población. Sin embargo, la búsqueda de relaciones interesantes entre los datos médicos puede mejorarse con la ayuda de otras técnicas computacionales para descubrir patrones ocultos entre ellos. La minería de datos ofrece un conjunto de diversas técnicas de análisis como lo son asociación, agrupación y clasificación, las cuales ya han sido utilizadas sobre datos clínicos de otros países para encontrar factores de riesgo asociados a enfermedades de interés [46].

1.2. Diabetes Mellitus Tipo 2

1.2.1. Descripción

La diabetes mellitus tipo 2 (DM2) es una enfermedad crónica caracterizada por una hiperglucemia y trastornos en el metabolismo de las grasas, hidratos de carbono y proteínas de forma tal que genera defectos en la producción y acción de la insulina en el cuerpo. Esta enfermedad presenta complicaciones crónicas que deterioran la calidad de vida de los pacientes y aumentan significativamente el riesgo de muerte [15]. Los pacientes de diabetes utilizan con mucha más frecuencia servicios especializados como odontología, optometría y nutricionistas, los cuales incrementan el costo de la diabetes. Adicionalmente, el costo derivado del tratamiento de la diabetes supone casi el doble de un paciente no diabético y además consumen entre 2 y 6 veces más recursos que los pacientes de edad y sexo similares con otras enfermedades crónicas [2].

1.2.2. Factores de Riesgo

Dentro del contexto médico, un factor de riesgo se define como:

"Un aspecto del comportamiento o del estilo de vida, exposición medioambiental o característica innata o heredada que, sobre la base de evidencia epidemiológica, se conoce que está asociado con una condición de salud relacionada considerada importante para prevenir".

Los factores de riesgo se destacan principalmente por el hecho de ser una evidencia epidemiológica y además, su conocimiento permite implementar acciones de prevención en salud [19]. Como ejemplos de factores de riesgo se encuentran el ser fumador activo y mayor de 40 años para la enfermedad de EPOC (Enfermedad Pulmonar Obstructiva

Crónica), tener un índice de masa corporal superior a 25 para Diabetes Mellitus tipo 2, padecer de una enfermedad marcadora de VIH/SIDA tipo C para VIH/SIDA y haber padecido de mutaciones germinales para Cáncer de Mama.

Las bases de datos médicas contienen un gran número de atributos relacionados entre sí, los cuales se encuentran a su vez ligados al estado del paciente. Se puede decir entonces que los factores de riesgo se describen como conjuntos de atributos relacionados significativamente con un estado particular del paciente. Sin embargo, cuando se trabaja con datos médicos es posible ver un gran desbalance entre el estado de los pacientes, "enfermo" y "sano", puesto que la mayoría de la población se clasifica como "sano". Si se toma como ejemplo tabla 1.1, en la cual se relacionan diferentes características o atributos de los pacientes, se puede observar que el par de atributos Sexo = F y Edad = 40-60 conforman un patrón para la clase o estado "enfermo".

Tabla 1.1. Conjunto de Datos Médicos[34]

Sexo	Edad	Presión Arterial	...	Estado
F	40-60	Alto	...	Enfermo
M	1-14	Bajo	...	Sano
M	14-40	Bajo	...	Sano
F	1-14	Bajo	...	Sano

Estos patrones pueden ser frecuentes si superan un umbral establecido. De igual manera, los patrones presentan una medida llamada soporte la cual se define como la razon entre el número de registros que contienen cierta condición P y todos los registros. Sin embargo, debido a que los datos médicos presentan muchos más datos de la clase normal (Sanos), resulta difícil establecer cuáles son los patrones frecuentes asociados a la clase anormal (Enfermos). Es por esto que se introduce el concepto de Soporte Local el cual se describe a continuación:

$$soporteL(P) = \frac{soporte(P \cup a)}{soporte(a)}$$

Donde P es el patrón dado y a es la clase (en este caso "enfermo"). Esta ecuación es llamada también regla y se denota $(P \rightarrow a)$ [34].

1.3. Knowledge Discovery in Databases - KDD

Se define como el proceso de descubrir conocimiento en los datos. Se refiere a la extracción de reglas, patrones, conocimiento, entre otros, los cuales están ocultos en las bases de datos. Esta información extraída puede ser útil en toma de decisiones, control de procesos y procesamiento de consultas [12]. Cabe destacar que la minería de datos es una tarea que hace parte del proceso de Extracción del Conocimiento sobre Bases de Datos (KDD), tal como se muestra en la figura 1.1.

1.3.1. Inteligencia de Negocios

Se define como Inteligencia de Negocios o Business Intelligence (BI) al proceso de recolectar información relevante para una organización, teniendo en cuenta los factores internos

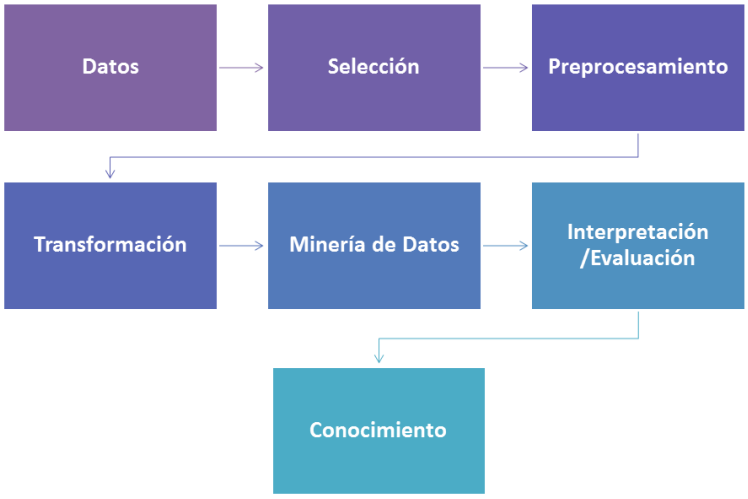

FIGURA 1.1. Pasos del Proceso KDD [17]

y externos que influyen en ella [10]. La inteligencia de negocios utiliza tecnologías para manejar adecuadamente la información así como también hacer análisis. Todo esto siempre orientado a generar ventaja competitiva y permitir a las entidades de la organización el entendimiento de sus necesidades [42]. Para una organización dedicada al cuidado de la salud, la ventaja competitiva se ve reflejada en términos de calidad de servicio y atención a sus pacientes, ya que la información organizada y enfocada a la toma de decisiones puede mejorar planes de prevención y manejo de enfermedad acorde a su población.

Básicamente la inteligencia de negocios presenta 3 fases primordiales las cuales han ido perfeccionándose con el tiempo [43], estas son:

- Fase de concentración, integración y almacenamiento de datos ETL (Extract, Transform and Load): Esta fase tuvo sus inicios con empresas que deseaban tener sus datos organizados y agrupados en un solo lugar para su fácil tratamiento. De modo que los proveedores de bases de datos mejoraron sus productos para poder almacenar y procesar estos datos de forma más eficiente. A su vez las compañías de hardware y software mejoraron las tecnologías de almacenamiento para suplir las necesidades de las vastas cantidades de información.

- Fase de análisis y reporte: Durante esta fase se puede ver la evolución en la presentación y uso de los datos integrados. Se pasa desde las simples consultas SQL al uso de técnicas de minería de datos y análisis dimensional.

- Fase de aplicación de la Inteligencia: Se comprende en esta fase todo lo relacionado con la toma de decisiones en una organización. Es posible ver que ahora las herramientas de BI proveen diferentes mecanismos de reportes y alertas que permiten a las organizaciones actuar de forma rápida ante las evidencias. De este modo se disminuye el tiempo de análisis y se logra una mayor competitividad por parte de las empresas.

Así pues, se puede ver una rápida acomodación y desarrollo en las herramientas que proveen los procesos de BI. Sin embargo, es importante reconocer que estas herramientas

deben cumplir con el cubrimiento adecuado en las áreas de BI. Dichas áreas son, a nivel general, las siguientes [16]:

- Entrega de la Información: Una herramienta de BI debe estar en capacidad de permitir la construcción, actualización y publicación de diversos reportes y alertas. Así mismo debe proveer facilidades en el uso de consultas ad-hoc.
- Integración: Toda plataforma BI debe estar en capacidad de implementar seguridad en su entorno así como permitir la creación de metadatos para el proceso de integración. También debe proveer funciones para compartir información entre usuarios.
- Análisis: Esta área comprende todo lo concerniente a los reportes y la capacidad del usuario para realizar sus propias funciones. Es importante que haya variedad en la forma de presentar la información y viabilidad para la creación de mapas estratégicos.

1.3.2. Bodegas de Datos

Las bodegas de datos son la parte principal en el proceso de inteligencia de negocios, el cual se mencionó anteriormente. De igual forma, es el componente que demanda mucho más tiempo y cuidado en su construcción puesto que debe ser diseñada con mucho cuidado y también debe estar orientada al soporte de toma de decisiones [16]. Se puede definir una bodega de datos como una *"copia de los datos transaccionales especialmente diseñados para fines de consulta y análisis"*, pero en el contexto médico seria una copia de la información de cada paciente [50]. Dentro de la arquitectura de la bodega de datos encontramos un conjunto de herramientas llamadas Back End y Front End, así como también los metadatos y las herramientas para administrar la Bodega. La figura 1.2 muestra la arquitectura básica de una bodega de datos.

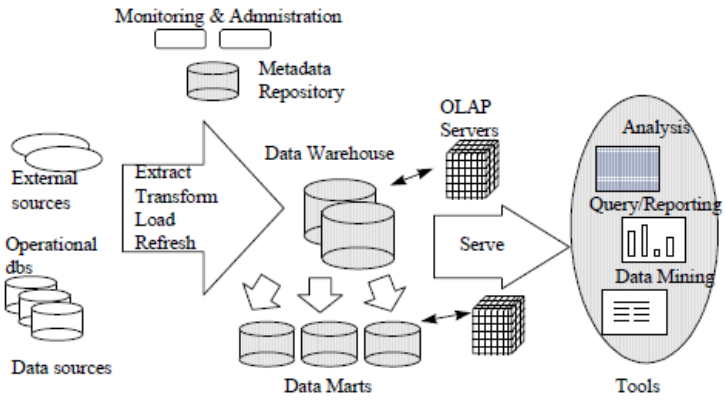

FIGURA 1.2. Arquitectura Básica de una Bodega de Datos [11]

El Back End provee herramientas que realizan trabajos de Extracción, Transformación y Carga de los datos transaccionales. Estas herramientas permiten hacer tareas de limpieza, migración, transformación, carga y actualizaciones de los datos a la bodega. Es importante tener presente que se deben realizar operaciones de creación de índices, pruebas de integridad, entre otras para garantizar datos de calidad en la Bodega. Todo esto demuestra que el proceso realizado por el Back End resulta ser el más costoso en tiempo y recursos [11].

Por su parte el Front End se encarga de presentar a los usuarios finales reportes, agregaciones, sumarizaciones, alertas, asociaciones, y todo tipo de información que pueda ser útil para el entendimiento del negocio y la toma de decisiones. Aquí se encuentran herramientas como las hojas de Excel, cubos OLAP (on-line analytical process), software de minería de datos y ambientes de consultas SQL tales como Microsoft Access [11].

Estas bodegas de datos son diseñadas utilizando principalmente el modelo multidimensional. En este esquema se almacenan los datos como hechos y dimensiones, en contraparte al modelo tradicional de bases de datos. Aquí los hechos son valores numéricos o eventos los cuales se refieren a actividades de una organización. Las dimensiones se definen como un conjunto de atributos relacionados con el evento [7]. La figura 1.3 muestra un ejemplo de modelo multidimensional.

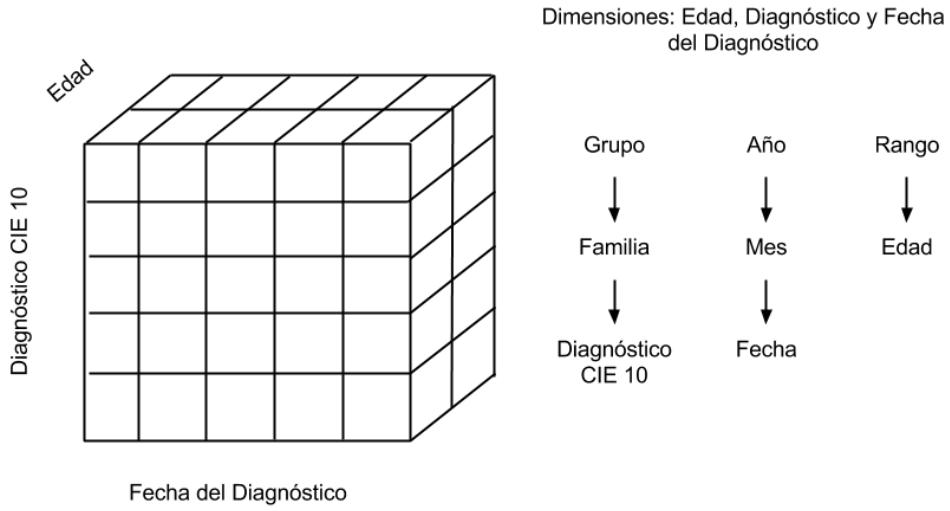

FIGURA 1.3. Modelo Multidimensional [11]

1.3.3. Procesamiento Analítico en Línea - OLAP

La tecnología de procesamiento analítico en línea OLAP permite manejar y realizar análisis complejos sobre los datos almacenados en una bodega. En este sentido, OLAP se encuentra orientado principalmente a la toma de decisiones mediante la opción de flexibilidad en la navegación sobre la bodega con el uso de vistas multidimensionales y representaciones gráficas de la información.

Las herramientas OLAP utilizan la información contenida en cubos de datos generados mediante el uso de las dimensiones y los hechos de la bodega de datos. Básicamente estas

herramientas generan consultas que permiten organizar y agregar los datos de la bodega dándole al usuario la posibilidad de navegar sobre los cubos de datos centrándose en una necesidad en particular como por ejemplo la granularidad de los hechos o los filtros que proveen las dimensiones [49].

1.3.4. Minería de Datos

Dentro de la minería de datos se enmarcan varias tareas que bien son de tipo descriptivo o de tipo predictivo. Las descriptivas presentan el comportamiento y las relaciones existentes entre los datos, mientras que las de tipo predictivo permiten suponer comportamientos a futuro basándose en los datos almacenados en las variables existentes [17]. Ejemplos de estas tareas son:

- Clasificación: Permite mediante una función el mapeo de los datos y su clasificación en una de las clases previamente establecidas sobre el conjunto de datos.
- Regresión: Funciona de igual modo que la clasificación, la diferencia radica en que su función de mapeo o clasificación devuelve un valor real.
- Reglas de Asociación y Árboles: Son mecanismos que permiten al usuario entender sus datos mediante relaciones existentes entre ellos y puntos de separación de los datos.
- Agrupación: Permite encontrar grupos de elementos dentro de los datos que son parecidos entre sí y formar grupos. Estos grupos pueden ser excluyentes o bien compartir algunos de sus atributos mutuamente.
- Cambio y detección de desviación: Esta tarea se enfoca en encontrar los cambios sobre el conjunto de datos mediante una comparación con estándares o medidas establecidas sobre ellos.

Las tareas de Minería de Datos precisan de un preprocesamiento de los datos existentes antes de la aplicación de cualquiera de sus algoritmos. Para ello existen técnicas como el muestreo simple o estratificado, métodos de reducción de dimensionalidad, manejo de valores perdidos o nulos y discretización. El muestreo simple o estratificado se realiza para reducir el conjunto de datos, sin embargo debe ser realizado con cuidado puesto que se puede incurrir en pérdida de datos valiosos para el análisis que se desee hacer. Como parte de los métodos de reducción de dimensionalidad se encuentran la Entropía y el PCA, con los cuales se pueden eliminar dimensiones que no aportan información al conjunto de datos. Por otra parte, el manejo de valores nulos o perdidos debe ser cuidadoso, ya que algunos de los algoritmos de Minería de Datos no soportan ausencia de valores y eliminarlos o darles un trato inapropiado incurriría en resultados erróneos [48].

Así mismo, las tareas de Minería de Datos deben cumplir ciertos requerimientos tales como el manejo de múltiples tipos de datos, la escalabilidad y eficiencia en los algoritmos que las soportan, así como la usabilidad y certeza de los resultados presentados [12].

1.4. Aplicaciones Clínicas de KDD

Los trabajos realizados en el campo de la salud para detección de enfermedades o factores de riesgo se pueden enmaracar en dos grupos. Primero están los trabajos estadísticos los cuales han sido usados principalmente por epidemiólogos en análisis de riesgo de enfermedades y luego se encuentran los trabajos que usan métodos computacionales para detección de enfermedades y análisis de información médica.

1.4.1. Métodos Estadísticos

Los estudios realizados sobre factores de riesgo y manejo de las enfermedades sugieren principalmente el uso de análisis exploratorio y el uso de regresiones sobre los datos médicos.

1.4.1.1. Análisis de Riesgo de Enfermedades Usando Análisis Exploratorio

El análisis exploratorio permite un primer acercamiento a los datos médicos. Haciendo uso de éste, es posible identificar datos atípicos, rangos de valores y frecuencia para cada una de ellas. Muchos de los estudios realizan este primer análisis sobre los datos para establecer cuáles variables son potencialmente relevantes para el trabajo investigativo.

Mediante el análisis exploratorio se pueden obtener medidas como la media, la desviación estándar, la moda y rangos de cada una de las variables. Como se observa en el cuadro II, los valores de media y desviación estándar permiten inferir que las diferencias entre los factores de riesgo son moderadas respecto a los dos grupos estudiados, salvo el número de cigarrillos el cual fue mayor en el grupo de hombres con autopsia.

Tabla 1.2. Media y Desviación estándar de Factores de Riesgo seleccionados para una serie de autopsia en una cohorte de hombres muertos entre 40 y 49 años [28].

Factor de Riesgo	Autopsia	Sin Autopsia	Cohorte Total
Colesterol	272.8 +/-44.7	287.5+/-56.57	280.9+/-52.0
Triglicéridos	2.73+/-1.74	2.27+/-1.07	2.48+/-1.43
Presión Sistólica	145.2+/-20.3	139.6+/-17.52	142.1+/-19.0
Presión Diastólica	92.2+/-12.5	89.5+/-13.31	90.7+/-13.0
Número de Cigarrilos	11.4+/-8.9	7.0+/-5.38	9.0+/-7.5
Puntaje de Riesgo	23.2+/-30.6	15.2+/-24.37	18.8+/-27.6

Sin duda alguna, el análisis exploratorio ha sido el método clásico utilizado sobre datos de toda naturaleza para inferir y analizar variables. Particularmente este método ha sido usado como primer paso en los estudios sobre distintas cohortes de pacientes tal como se puede ver en las investigaciones de varios países. De igual forma, se propone el análisis de correlación entre las variables como una primera aproximación al entendimiento de las relaciones existentes entre las variables (factores de riesgo) [28, 23, 24, 39].

A partir de este método, es posible utilizar otras técnicas estadisticas para encontrar relaciones que éste no puede hacer.

1.4.1.2. Regresión Logistica y Lineal

Después de la utilización de técnicas exploratorias de análisis de datos, toma importancia la relación existente entre las variables que se deseaban estudiar. Para atender esta necesidad, se han utilizado estos dos métodos estadísticos de análisis de variables en estudios sobre factores de riesgo.

Su utilidad radica en la determinación de relaciones entre variables dependientes e independientes en el conjunto de datos. La regresión Logística y la Lineal difieren en la variable de salida. Mientras que en la lineal se espera un valor continuo, en la logística se asume que este valor es binario [27]. La regresión logística se ve ampliamente utilizada en la construcción de modelos para factores de riesgo de tipo discreto como por ejemplo "Consumo de tabaco", "Realización de Ejercicio", "Consumo de Frutas" o "Consumo de Alcohol" [24]. La figura 1.3 muestra las 5 primeras combinaciones del modelo logístico para el análisis de enfermedades crónicas en jóvenes de Bogotá.

Orden	Bajo consumo de frutas y verduras	Inactividad física en tiempo libre	Consumo actual de tabaco	Consumo agudo de alcohol	n	%
1	No	Sí	No	No	305	20,8
2	No	No	No	No	199	13,6
3	No	Si	No	Sí	140	9,6
4	Si	Sí	No	No	131	8,9
5	No	No	No	Si	102	6,9

FIGURA 1.4. Combinaciones de Factores de Riesgo de comportamiento en adultos jóvenes [24]

1.4.2. Métodos Computacionales

De igual forma es posible encontrar varias aplicaciones de procesos de inteligencia de negocios y minería de datos sobre datos médicos con el fin de ayudar en la detección de factores de riesgo o enfermedades.

1.4.2.1. Inteligencia de Negocios: Bodegas y OLAP

En cuanto a bodegas de datos para salud se han desarrollado diferentes sistemas que soportan estas organizaciones. En [44] se describen los desarrollos más importantes que se han hecho en esta área. Entre los sistemas que proveen facilidades médicas (Clinical Computing, Oracle Clinical, SAS institute, MEDai, Informations Architects Inc, Shared Medical Systems, Quest informatics, Turku University Central Hospital, StanfordMedical Informatics). Cabe destacar a MEDai el cual usa inteligencia artificial y provee estudios predeterminados para diferentes enfermedades comunes. Al igual la universidad de Turky en Finlandia da un ejemplo de aplicación con una herramienta de transporte, la bodega y una herramienta propietaria de front-end y resalta las ventajas de la bodega para responder preguntas. Los otros sistemas se centran más en procesos de ayuda a la industria farmacéutica [44].

Con bodegas de datos clínicas se pueden hacer diferentes niveles de análisis: a nivel de paciente, por grupos, por ejemplo de enfermedades. También se pueden hacer investi-

gaciones médicas o mejoramiento de la calidad del servicio. Así mismo se pueden hacer análisis financieros y demográficos permitiendo analizar la rentabilidad y calidad general de los servicios prestados. Para que esto se pueda realizar la bodega debería aceptar todos los tipos de datos, desde financieros, como contratos y facturas; datos demográficos como edad y sexo; clínicos, como diagnósticos y procedimientos; numéricos, como resultado de laboratorio e imágenes como rayos x. La combinación de todos estos datos es lo que hará posible el análisis cruzado y la obtención de información [44].

En el proceso de crear la bodega se han encontrado varios problemas como que los datos clínicos originales son guardados en un tipo de formato que no es el mejor para el análisis y la revisión de los datos. En 1999 PharmaHealth technologies introduce junto con un sistema de bodega de datos el concepto ambiente controlado encaminado a apoyar el proceso de transformación por medio de la captura de metadatos. Esto permite adicionar funcionalidad extra en el proceso de transformación [50]. Entre otras cosas se creó la definición de las columnas en las tablas de análisis de la bases de datos. Así el usuario al realizar la transformación puede asegurarse que las columnas concuerdan con los estándares definidos [50].

1.4.2.2. Minería de Datos

Dentro de las técnicas computacionales utilizadas en los estudios sobre factores de riesgo, es posible distinguir las siguientes:

Asociación Las reglas de asociación se encargan de encontrar conjuntos de patrones de riesgo para cierta condición clínica. Mediante las combinaciones de estos factores es posible obtener un subconjunto de la población que se encuentra en alto riesgo de contraer dicha condición clínica. Todo esto se hace basándose en que el grupo principal de estudio es el más pequeño, es decir el que se encuentra en estado anormal o enfermo. Así mismo también es importante ajustar las medidas de soporte y confianza para que el conjunto de reglas obtenidas tenga relevancia en el estudio [20]. Este método computacional ha sido aplicado en estudios de lactancia y como pruebas para conjuntos de datos médicos [34, 26]. Se ha encontrado que esta herramienta presenta muchas ventajas puesto que ayuda en la elaboración de planes de intervención para el grupo poblacional de alto riesgo. Sin embargo, se propone el uso de herramientas estadísticas para proveer mejores resultados en su implementación.

Como un primer acercamiento a la detección de factores de riesgo interesantes, [9] discute el uso de software que soporte algoritmos de generación de reglas de asociación. Los experimentos realizados en este estudio se hacen sobre un conjunto de datos médicos previamente limpiados para su utilización en el software de Mineria de Datos. Los resultados arrojaron un conjunto de reglas entre 2.000 y 20.000. Se discute también que se seleccionó un conjunto pequeño para mostrarle al usuario. Igualmente se propone una división por ventanas de tiempo para encontrar nuevas reglas en los conjuntos de datos.

También se discute en varios trabajos que el uso de este método debe ser controlado puesto que genera demasiadas reglas y es necesario tener un conjunto que sea realmente el óptimo de patrones de riesgo [34, 33]. En su trabajo, [34] advierte que las reglas de asociación deben contener patrones que generen alto riesgo relativo en lugar de aquellas que sólo cumplan con la condición de tener una alta confianza. De igual forma propone una modificación al algoritmo de Reglas de Asociación para que en lugar de tomar un valor de

confianza tome un valor de riesgo relativo; también limita el número de reglas generadas y la forma de tratar el problema dependiendo de la población: enfermos o sanos. Al tomar como medida el riesgo relativo se asegura de mostrar a los usuarios el conjunto de reglas óptimo. Sin embargo, este conjunto de reglas puede ser muy grande y para solucionar este problema se propone el uso de la propiedad antimonótona de los conjuntos de reglas. Haciendo uso de esta, es posible reducir el conjunto generado puesto que se garantiza que para un conjunto de reglas frecuentes, sus subconjuntos también lo serán. Por último se propone una selección estratificada de las reglas de asociación encontradas para que el personal médico pueda evaluar de forma fácil y significativa las relaciones encontradas.

Como un caso de estudio, [34] aplica esta técnica para la evaluación del efecto adverso de las drogas sobre un conjunto de pacientes. La meta del experimento era encontrar el grupo de pacientes que estaba en riesgo de sufrir de angiodema después de haber consumido cierto tipo de medicamento. Para aplicar el algoritmo de generación de reglas se utilizó un soporte mínimo de 0.05, un mínimo de riesgo relativo de 2.0 y un máximo de número de atributos por regla de 4. El algoritmo se ejecutó de forma rápida y al ser evaluado se encontró que generaba reglas estadísticamente significantes para un conjunto de datos sesgado Los resultados de este experimento arrojaron un total de 417 patrones de riesgo, de los cuales 75 son significativos teniendo en cuenta su riesgo relativo. Los patrones encontrados fueron de interés para los expertos en el dominio puesto que muestran relaciones entre la edad de los pacientes y grupos de medicamentos consumidos por ellos.

En trabajos posteriores [33] continúa con la experimentación del algoritmo MORE (Mining Optimal Risk Patterns), propuesto en [34], y a la vez compara con técnicas como árboles de decisión y otros algoritmos de generación de reglas de asociación. El conjunto de datos escogido para realizar los experimentos consistía en pacientes con ingreso a urgencias y posterior hospitalización. Se requería analizar los factores que llevan a los pacientes a la hospitalización y minimizar la entrada en urgencias, debido a que el centro asistencial contaba con pocas camas. Una vez realizados los experimentos, se encontró que los árboles de decisión producían menos reglas que MORE y a su vez, estas reglas carecían de alto riesgo relativo. Particularmente se describe que la función de los árboles de decisión no es la de encontrar patrones y que por ende las reglas que genera son bastante específicas. Cabe destacar que se experimento con el árbol C4.5 y el C5.0. El número de reglas generadas por los árboles de decisión fueron un total de dos reglas, mientras que MORE encontró 131 patrones de riesgo de los cuales 75 fueron significativos. Entre los patrones de riesgo de admisión interesantes, se encuentran el vivir alejado de la ciudad, ser hombre y tener lesión en las extremidades así como también tener edad avanzada. Uno de los patrones encontrados que fueron de más interés, corresponde al de ingreso a urgencias en horas laborales los días lunes. Esto revela un problema en el modelo de atención del centro asistencial puesto que se distingue una falta de servicios disponibles en los fines de semana, lo cual puede retrasar la admisión a hospitalización de los pacientes.

Finalmente, estos trabajos [34, 33] proponen el uso de estos conjuntos de reglas y patrones para que las organizaciones médicas y los expertos puedan refinar hipótesis y hacer estudios de consumo.

Clasificación Las técnicas de clasificación más utilizadas en el campo de análisis de riesgo en salud son los árboles de decisión y las redes bayesianas. Esto gracias a la facilidad de entendimiento por parte de los clínicos de los modelos de clasificación generados [35].

Mediante la utilización de árboles de decisión es posible encontrar reglas que describen a los pacientes que padecen algún tipo de condición clínica. La generación de estas reglas son de fácil entendimiento para los clínicos y funcionan bien con datos discretos o continuos [35], lo que permite flexibilidad en las variables de tipo numérico debido a que no es necesario discretizarlas. Particularmente el árbol C4.5 se ha utilizado en estudios sobre lactancia [26] como técnica computacional debido a que presenta la ventaja de visualizar el porqué de la decisión tomada por las madres lactantes. Este método fue contrastado con la regresión logística, mencionada anteriormente, y prensentó la ventaja de no usar el conjunto completo de datos para la implementación del modelo. Además, la precisión de predicción en el C4.5 para el conjunto de datos de lactantes fue del 77.91 %. Este resultado se logró mediante una buena selección de técnicas de preprocesamiento sobre el conjunto de datos asi como las dimensiones que iban a entrar en el modelo [26].

Cabe resaltar que en algunos estudios, este método es tratado como estadístico mas no como computacional [31]. En el estudio realizado por [31] se describe el uso de árboles de decisión para encontrar subgrupos de mujeres por región geográfica que están en alto riesgo de tener un bebé con bajo peso. Como herramienta de apoyo se utilizó el software CART el cual permitió diseñar el árbol de decisión haciendo uso del índice GINI[1] para encontrar las variables de separación. Algo importante de este estudio es que se le dio importancia al resultado del clasificador, de modo que se trabajó con costos de clasificación. Los resultados del clasificador generado fueron comparados con un modelo binario logístico implementado en STATA mediante el uso de curvas ROC(Característica Operativa del Receptor) y medidas como la especificidad [2] y la sensibilidad [3]. Como conclusión del trabajo se estableció que los modelos logísticos funcionaron mejor por una diferencia mínima respecto a los modelos basados en árboles de decisión. Sin embargo, se acepta el uso de esta técnica puesto que sus resultados coinciden con investigaciones anteriores en el área. Así mismo, se establece que para mejorar el nivel de predicción de los árboles de decisión es necesario trabajar en los costos de clasificación.

En un trabajo similar realizado sobre datos de pacientes diabéticos, [8] menciona la utilidad de las técnicas de minería de datos como los árboles de decisión para encontrar asociaciones novedosas que pueden ser útiles para la comunidad médica. De igual forma, se señala la importancia de aplicar esta técnica en varias bodegas de datos de personas diabéticas para realizar comparaciones y establecer nuevas relaciones entre las variables de las poblaciones que sufren esta enfermedad.

Por último, en ambos trabajos se destaca la facilidad de entendimiento de las reglas generadas por los árboles de decisión para el personal médico y se asegura que es una técnica útil en el campo del cuidado de la salud. Además, sus resultados gráficos pueden ayudar a visualizar las relaciones entre las variables de los pacientes y también proveer información para que el personal médico brinde prevención para el grupo de personas en riesgo.

Las redes bayesianas utilizan la bien conocida teoría de probabilidades de Bayes. Esta red consta de grafos dirigidos acíclicos y captura las probabilidades condicionales entre los atributos. Cada uno de los arcos representa dependencias entre los nodos y estos a su vez representan causa y consecuencia respectivamente. Desde el punto de vista médico, estos

[1] Medida utilizada para determinar la mejor forma de dividir los registros basándose en la distribución de las clases antes y después de la división[48]

[2] Probabilidad de que un paciente sano sea correctamente clasificado

[3] Probabilidad de que una persona enferma sea diagnosticada

nodos pueden ser vistos como diagnósticos y ssintomas de los pacientes. Existen varios algoritmos que pueden crear la red utilizando los datos proporcionados. Así mismo, estos algoritmos encuentran la mejor estructura para la red bayesiana utilizando medidas como la MDL (Minimum Distance Lenght) [41]. Sin embargo, la mayoría de redes bayesianas que se han utilizado en proyectos del área del cuidado de la salud, han sido construidas a mano. Cabe destacar que la construcción de este tipo de red requiere de personas expertas en el área y con un amplio conocimiento médico [35].

Esta técnica computacional ha sido utilizada en varios trabajos de pronósticos de enfermedades como la neumonía adquirida en la comunidad (NAC) [5], en los cuales demostró tener una gran precisión para clasificar los pacientes y distinguirlos respecto a otras enfermedades. Por su parte, [35] señala que las limitaciones que se tenían con los modelos estadísticos hace algunos años ya se han superado. En trabajos como el presentado por [21] es posible ver la comparación entre un modelo de red bayesiana y uno de regresión logística. Si bien la construcción de la red resulta ser mucho más demandante de tiempo, respecto al modelo de regresión, es posible ver que los resultados de precisión en el clasificador bayesiano fueron mejores. Aunque la diferencia estadística no es muy grande, sí representa un factor importante en la aplicación a datos reales de pacientes embarazadas, puesto que de su clasificación depende el cuidado y las intervenciones que deben ser suministrados a dichas pacientes.

La metodología implementada por los trabajos mencionados consistió en la separación del grupo de datos en dos: los de entrenamiento y los de validación. Esta división fue hecha utilizando muestreo estratificado. Posteriormente se construye la red que mejor se ajusta a los datos utilizando técnicas heurísticas o algoritmos que optimizan medidas como se mencionó anteriormente. Una vez construida la red, se procede a asignar los valores de probabilidad condicional para todas las variables. Por último, se procede a validar el modelo contra los datos seleccionados en la etapa de división. Los estudios realizados utilizan la curva ROC (Caracteríistica Operativa del Receptor) al momento de comparar cuáles de las redes bayesianas fue la que mejor se comportó [5] o para ver las diferencias de eficiencia entre las regresiones logísticas y las redes bayesianas [21].

Agrupación Los algoritmos de agrupación utilizados en análisis de riesgo sugieren además de una identificación de factores de riesgo, una visión de la severidad de acuerdo a los signos o factores. Uno de los estudios presenta una agrupación de síntomas en niveles de riesgo mediante el uso de redes neuronales completamente dependientes donde cada neurona es un síntoma de la enfermedad y tiene varios niveles con sus respectivas probabilidades excitatorias, las cuales corresponden al resultado de un proceso de selección de "síntomas vecinos" y su frecuencia de ocurrencia. El algoritmo de agrupación funciona sobre las relaciones entre las neuronas y su nivel de correlación donde cada clúster se define sobre el conjunto de neuronas altamente correlacionadas [51].

En otro estudio es posible ver el uso de algoritmos de agrupación difusos mediante los cuales se puede apreciar el grado de severidad del individuo acorde a la pertenencia de múltiples síntomas. Se utiliza una versión modificada de la técnica llamada "agrupación substractiva" donde se asigna un potencial a cada uno de los datos y el máximo es escogido como el primer cluster, posteriormente se actualiza el potencial de los otros datos y se escogen los otros cluster. Una vez encontrados los centros de los clusters se asignan los datos usando la métrica de kernels inducidos y se calcula la forma de cada cluster utilizando la estimación de densidad de las máquinas de vectores de soporte. Este algoritmo permite

encontrar niveles de severidad mediante la presencia o ausencia de variables que indiquen riesgo para la enfermedad cardíaca. Sin embargo, se hace la salvedad de que para el estudio sólo se utilizaron datos numéricos [18].

En [32], los autores proponen un algoritmo de agrupación evolutivo basado en operadores genéticos adaptables por sí mismos. La ventaja de este algoritmo radica en que permite la evolución de los prototipos de agrupación mediante la mejora de las tasas de los operadores genéticos y así mismo es capaz de determinar el número apropiado de clústers. Este trabajo se basa en el agrupamiento no supervisado de niches y gracias a esto puede determinar la población que más contribuya en la conformación y permanencia de un ambiente (clúster). Este algoritmo fue probado en un conjunto de datos de diagnósticos de cáncer de mama y de diabetes para determinar su nivel de precisión; los resultados obtenidos reflejan un alto grado de detección para los pacientes.

1.5. Resumen

En este capítulo se presenta el estado del arte sobre la inteligencia de negocios utilizada para analizar riesgo de enfermedades en el sector salud. Se presenta la introducción a la enfermedad que se quiere trabajar en esta tesis desde el punto de vista médico, se describen las herramientas que presenta la metodología KDD para extraer conocimiento valioso de los datos y su aplicación en el sector salud describiendo los métodos estadísticos y computacionales más usados.

CAPÍTULO 2

Modelo KDD para el análisis de riesgo de la Diabetes Mellitus Tipo 2

Este capítulo muestra el desarrollo de un modelo de KDD para el análisis de riesgo de la Diabetes Mellitus Tipo 2. Este modelo permite aportar información importante sobre el estado de pacientes de DM2 en las EPS del país. El diseño cuenta con dos partes: un modelo de datos, compuesto por una bodega de datos clínica y un modelo de minería para la caracterización de pacientes y predicción de comorbilidades. Primero se describe de forma general el modelo de análisis de riesgo propuesto. Se muestra el diseño de la bodega de datos donde se tiene en cuenta las fuentes de datos utilizadas, el proceso de limpieza y carga de estos datos en la bodega y las variables tomadas en cuenta para la realización de la investigación, así como la selección de algoritmos de minería de datos que mejor se adapta al grupo de datos con los que se cuenta. De igual modo, se muestra la forma de describir los grupos de riesgo mediante la utilización de reglas de asociación como una primera aproximación de entendimiento sobre el comportamiento de cada nivel de riesgo de los pacientes

2.1. Gestión del Riesgo en Salud

En Colombia, la Resolución 1445 de 2006 del Ministerio de Protección Social plantea que las entidades prestadoras de salud EPS del país deben realizar planes de atención a sus usuarios basándose en la información que puedan recopilar de éstos. Ésto incluye la detección de factores de riesgo, grupos de riesgo y perfiles de pacientes los cuales permitan orientar los planes de prevención, promoción y educación de las EPS.

La información de los pacientes debe ayudar también a identificar necesidades de los pacientes y orientar el manejo del riesgo a enfermar de sus afiliados.

A continuación se describe el modelo propuesto de análisis de riesgo para la Diabetes Mellitus tipo 2 en Colombia. Este modelo realiza una aproximación mediante la inteligencia de negocios y la minería de datos a los requerimientos para generar planes de atención descritos anteriormente.

2.2. Modelo General

Para realizar el análisis de riesgo sobre la población de Diabetes Mellitus 2 se propone el modelo mostrado en la figura 2.1, el cual está basado en inteligencia de negocios y minería de datos.

El modelo se compone de una bodega de datos clínicos y de un conjunto de técnicas de minería de datos los cuales son utilizados para generar un modelo de análisis de riesgo para DM2.

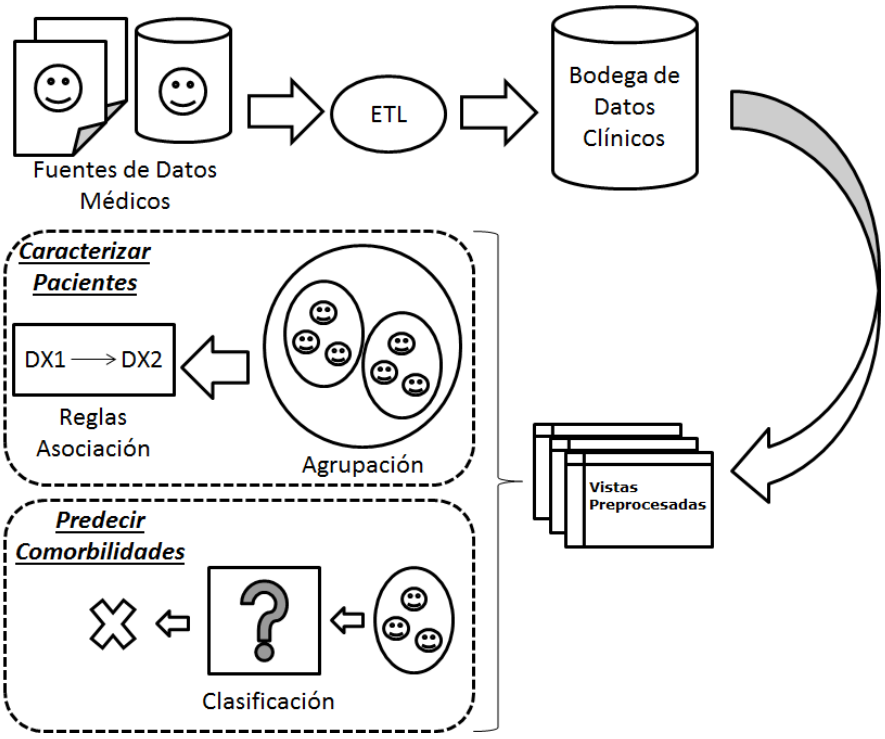

FIGURA 2.1. Modelo General de Análisis de Riesgo de DM2

La primera parte del modelo, compuesta por la bodega de datos, cumple la función de centralizar información médica y eliminar datos atípicos. Para los datos médicos es necesario tener la ayuda de un experto [1] para poder eliminar adecuadamente este tipo de datos, puesto que no todos los que parecen ser extraños resultan ser candidatos a eliminar. De este modo se garantiza que la información que se encuentra dentro de la bodega puede ser usada posteriormente sin ningún problema en análisis de datos. Otra función importante que cumple la bodega de datos, es que permite realizar análisis exploratorios rápidos a través de herramientas como OLAP, lo cual es muy útil para los expertos en riesgo en salud.

Por su parte, el modelo de minería de datos cumple la función de simular procesos de análisis de riesgo para la Diabetes Mellitus Tipo 2. Estos procesos, mencionados en la sección 2.1, incluyen la generación de perfiles de pacientes y la detección de factores de riesgo con el fin de disminuir la incidencia de la enfermedad o evitar su complicación.

[1] Liz Garavito, Directora ejecutiva de Processum LTDA e investigadora en el área de seguridad social y economía de la salud

Teniendo en cuenta esto, el modelo de minería de datos se compone de dos modelos: el de caracterización de pacientes y el de predicción de comorbilidades de DM2. El modelo de caracterización de pacientes se aborda desde la agrupación no supervisada debido a que se desconoce el número de perfiles o grupos que pueden describir a los pacientes. Por tal motivo, se opta por usar algoritmos de agrupación seleccionando cuidadosamente la mejor medida de similitud acorde a la representación del comportamiento de los pacientes. Así mismo, estos grupos generados de pacientes deben ser descritos de forma efectiva, razón por la cual se propone el uso de reglas de asociación que permitan describir los perfiles de pacientes. Por su parte, el modelo de predicción de comorbilidades permite identificar pacientes con riesgo a tener complicaciones en su enfermedad por lo que en este caso, se opta por usar algoritmos de clasificación que pudieran generar modelos de fácil visualización para los expertos clínicos. Los criterios de selección de algoritmos para cada uno de los modelos de minería y sus respectivos detalles se abordarán en la sección 2.4.3.1.

Este modelo fue diseñado de forma que se aprovechara la información que manejan las EPS en su cotidianidad. Muchos de estos datos sólo cumplen una función operativa dentro de las EPS, sirviendo luego para generar reportes para las entidades de supervisión.

La estructura y diseño del modelo general se explica en detalle a continuación, tomando en cuenta las dos grandes partes del modelo: La bodega de datos y el modelo de minería de datos.

2.3. Modelo de Datos

En esta sección se presenta el modelo de datos diseñado e implementado para almacenar la información médica, la cual posteriormente será usada por el modelo de minería de datos.

2.3.1. Bodega de Datos

La bodega de datos médicos constituye una parte fundamental del modelo propuesto en este trabajo investigativo. De su construcción e implementación depende que las variables que se van a usar para los análisis de minería de datos se encuentren limpias y libres de los varios problemas que presentan los datos médicos. Entre los problemas más comunes se encuentran los datos atípicos de variables como talla, peso, estatura y presión arterial, o el mal registro de diagnósticos como por ejemplo que un hombre tenga asociado un registro de embarazo o aborto.

Este tipo de centralización de los datos permite tener la información limpia y libre de inconsistencias y también es posible realizar análisis sencillos sobre los datos de pacientes mediante cruces rápidos para obtener información útil sobre pacientes acorde a las guías médicas establecidas para enfermedades en Colombia

2.3.2. Fuentes de Datos

Dentro de la información con la que se cuenta para la realización de este trabajo investigativo, se encuentran las recolectadas mediante fuentes de datos como el conjunto de archivos maestros que las EPS deben reportar al Sistema General de Seguridad Social en Salud (SGSSS). Estas variables contienen datos referentes a la afiliación del paciente

al SGSSS. Por otro lado, se cuenta con los reportes del Sistema de Información de Prestaciones de Salud (RIPS) que son entregados a las EPS por parte de los prestadores de servicios.

Esta información contiene datos reales de pacientes colombianos desde el año 2009 hasta el año 2012, pertenecientes a la zona urbana del centro del país y fue proporcionada por Processum LTDA [2].

A continuación se describe brevemente las generalidades sobre las fuentes de información.

2.3.2.1. Información del Paciente

Esta fuente contiene variables básicamente datos de tipo personal respecto al sistema de salud en el que se encuentra afiliado. Se destacan campos como la pertenencia étnica, tipo de cotizante, nivel de SISBEN, zona de afiliación, código del municipio de afiliación e información sobre el aportante y el cotizante.

2.3.2.2. Información de Servicios de Salud

Mediante la Resolución 3374 de 2000 se define el Registro Individual de Prestación de Servicios de Salud RIPS como " El conjunto de datos mínimos y básicos que el Sistema de Seguridad Social en Salud requiere para los procesos de dirección, regulación y control". Esta información describe cada una de las actividades realizadas por los prestadores de salud y contiene información sobre consultas, procedimientos, hospitalizaciones, recién nacidos, urgencias y otros servicios de salud.

La información recolectada por los RIPS se almacenan con códigos estandarizados. A continuación se describen los estándares usados.

- Códigos de Procedimientos en Salud: Los códigos de procedimientos clínicos son manejados usando el estándar colombiano. Este Código se llama CUPS y se encuentra disponible en la resolución 1896 de 2001 [14]. Para el momento de reporte de estos códigos, las EPS remueven los puntos que se encuentran en estos códigos.

- Códigos de Diagnósticos: Los diagnósticos se reportan usando el estándar internacional CIE10 publicado por la Organización Mundial de la Salud. Estos códigos están divididos en familias. Su forma de reporte consiste en códigos de 4 dígitos utilizando el nivel más específico de la codificación.

Por otro lado, la información disponible para un paciente contiene algunas variables que no hacen parte de los RIPS, sino que son recolectadas por historia clínica o encuestas de salud al momento de afiliación de los pacientes a una EPS. Estas variables son de tipo socio cultural y son listadas en la tabla 2.1.

[2]Empresa privada con fines de investigación aplicada para la generación de soluciones innovadoras que buscan favorecer el diseño e implementación de políticas, toma de decisión y optimización de la gestión instituciones de sectores sociales, en particular en el sector salud. Tomado de: http://processum.org/

Tabla 2.1. Variables Ecosociocultural

Variable	Tipo
Cocina con leña hace más de 10 años en recinto cerrado	Binomial
Año de abandono del Habito	Numérica
Año inicio fumador	Numérica
Correo electrónico	Categórica
Edad de la menopausia	Numérica
Número de cigarrillos al Día	Numérica
Uso de anticonceptivos en mujeres	Binomial
Numero de Gestaciones	Numérica
Edad de la menarquia	Numérica
Fuma	Categórica
Estatura	Numérica
Tensión arterial diastólica	Numérica
Peso	Numérica
Tensión arterial sistólica	Numérica
Estrato	Categórica
Código del departamento de residencia	Categórica
Código Dane de municipio Residencia	Categórica
Zona de Residencia	Categórica

2.3.3. Estructura de la Bodega de Datos

Con el fin de proporcionar una estructura que permita la fácil recuperación de datos para procesos de análisis de información, tales como la minería de datos, se ha propuesto un modelo general de centralización de esta información el cual se presenta a continuación.

2.3.3.1. Hechos y Granularidad

Los eventos que se han decidido modelar son los correspondientes al manejo de diagnósticos, manejo de procedimientos clínicos, y estado ecosociocultural del paciente. Pese a que existe el evento "entrega de medicamentos" se ha decidido no incluirlo en el modelo puesto que para el nivel de recolección de información con la que se cuenta actualmente, no aportaría significativamente en la creación de vistas de información que puedan ser analizadas. La granularidad de los hechos "diagnósticos" y "procedimientos clínicos" está dada en días, mientras que el evento "ecosociocultural" está dado en meses.

- Diagnósticos: Este hecho recopila todos los registros de CIE 10 de la población de la entidad. Mediante esta tabla de hechos se pueden distinguir los registros de incidencia epidemiológica en una enfermedad.
- Procedimientos Clínicos: En esta tabla de hechos se recopila toda la información referente a los servicios recibidos por la población de la entidad. Contiene variables como el costo y la forma de contratación del procedimiento.
- Ecosociocultural: Esta tabla de hechos maneja la información económica, social y cultural para cada mes de la población. Debido a que su forma de captura es mensual,

se debe representar en la bodega de datos como un registro del primer día del mes que se está capturando.

2.3.3.2. Dimensiones

Las dimensiones de la bodega de datos clínica contienen información especifica que permite agrupar los datos almacenados en la tablas de hechos para realizar distintos análisis sobre los datos. A continuación se describen las dimensiones definidas para la bodega de datos clínica:

- Persona: Esta dimensión contiene la información anónima de los afiliados en la EPS. Básicamente se compone de un identificador que hace referencia a un afiliado único y la fecha de nacimiento. La información es de tipo cambiante debido a que cada mes llegan nuevos pacientes a la Entidad.
- Edad: En esta dimensión se especifican las edades de los pacientes. Aunque para varios análisis de riesgo se suelen usar rangos de edad, se ha decidido que esta dimensión no tenga ese tipo de jerarquía para dar más libertad de generación de rangos de edad propios de cada una de las poblaciones presentes en las EPS. Esta dimensión se encuentra degenerada en las tablas de hecho para agilizar el cruce de información y la generación de cubos.
- Género: Esta dimensión se encuentra degenerada en cada una de las tablas de hechos modeladas en la Bodega y corresponde al sexo del paciente. Al estar degenerada en las tablas de hecho se agiliza el cruce de información y la generación de cubos.
- Fecha: Esta dimensión contiene una jerarquía de año, mes y día. Dependiendo del hecho a tratar se utilizan los diferentes niveles de jerarquía explicados anteriormente.
- CIE10: En esta dimensión se encuentran todos los códigos CIE10 discriminados en 3 niveles de jerarquía. Estos niveles fueron escogidos haciendo uso de la clasificación por categorías establecidas por la OMS. Para efectos de representación en la bodega, la primera jerarquía comprende los 2 primeros dígitos del CIE10, la segunda jerarquía contiene los 3 primeros dígitos y finalmente, la tercera jerarquía contiene el código completo de la enfermedad.
- Mercado Geográfico: Esta dimensión contiene el lugar de residencia del paciente y maneja dos niveles de jerarquía: Departamento y Municipio, siendo éste último el más bajo. Esta dimensión es estática puesto que las locaciones fueron predefinidas usando los códigos DANE (para municipios y departamentos en Colombia).
- CUPS: Esta dimensión contiene la jerarquía de los procedimientos clínicos. Se optó por usar dos niveles de jerarquía, donde el grupo describe la sección del procedimiento y el CUPS el codigo completo registrado en los RIPS.
- Comportamiento: Esta dimensión contiene todas las variables de tipo social y cultural mencionadas en la sección de fuentes de datos. También tienen jerarquía de grupo de acuerdo a su naturaleza.

El modelo estrella de la bodega de datos clínica está desglosado en las figuras 2.2., 2.3. y 2.4 en cada uno de los hechos mencionados anteriormente.

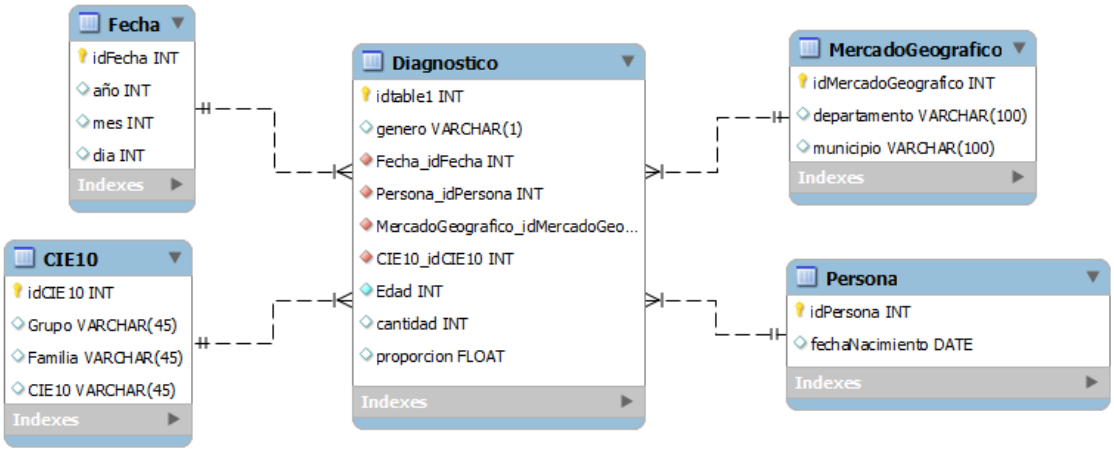

FIGURA 2.2. Hecho Diagnósticos

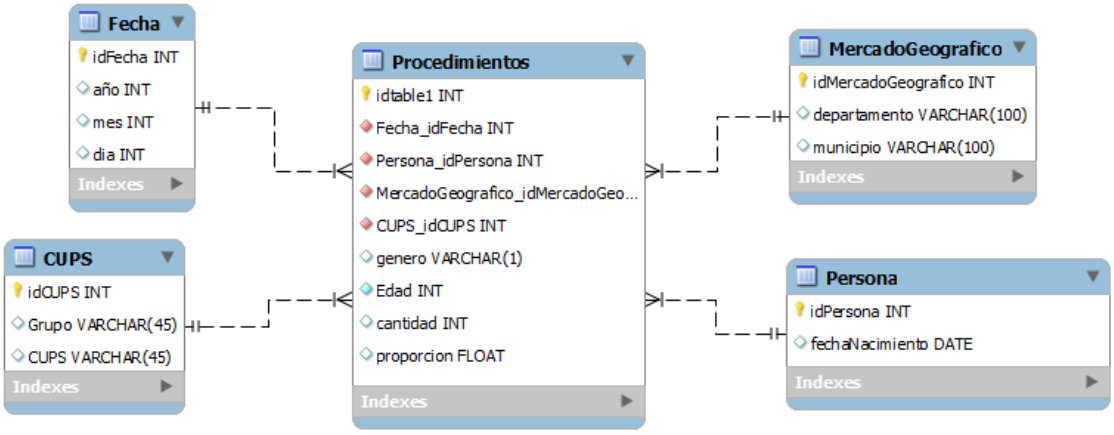

FIGURA 2.3. Hecho Procedimientos Clínicos

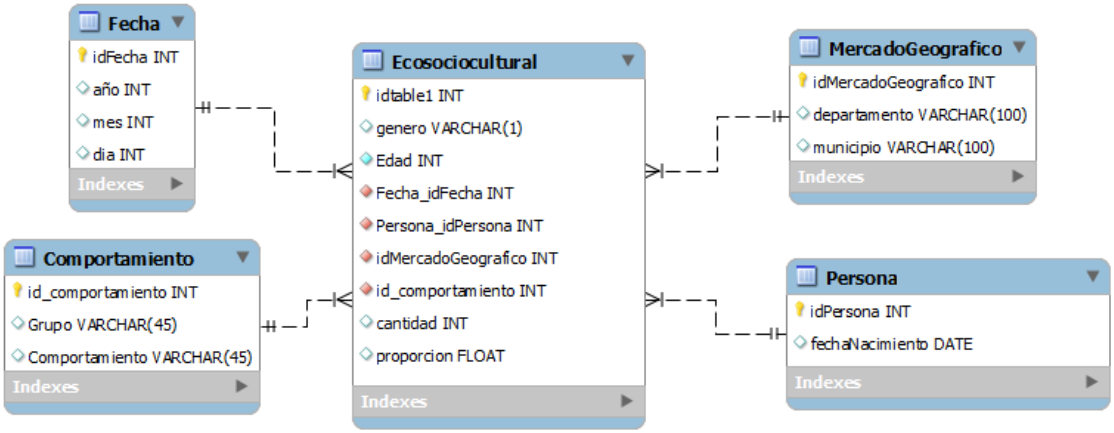

FIGURA 2.4. Hecho Ecosociocultural

2.3.3.3. Procesos ETL (Extracción-Transformación-Carga)

Los datos fuente presentan algunos problemas de registro de variables como errores en el código de los diagnósticos y los códigos de procedimientos en salud. También se encuentran

problemas de diagnósticos que no corresponden a las características de los pacientes. Esto se puede ver claramente en el reporte de diagnóstico de DM2 en población menor de 18 años. Otros problemas frecuentes corresponden a los valores atípicos de variables como talla y peso acorde a la edad de las personas y variables con valores faltantes.

Todos los procesos de extracción y transformación se realizaron con SQL directamente en la base de datos fuente.

Teniendo en cuenta lo mencionado anteriormente, a continuación se describen los principales procesamientos que fueron empleados para limpiar los datos fuente y prepararlos para su ingreso a la bodega de datos.

- Transformación de los datos de fecha mensual del ingreso de eventos socioeconómicos a granularidad día. Esto es, tomar la fecha de registro como la primera del mes.

- Realizar filtros de pacientes por edad, de modo que sólo se tomen los mayores de 18 años para el ingreso a la bodega. Este rango de edad se escogió teniendo en cuenta que la DM2 es una enfermedad que afecta principalmente a personas adultas.

- Eliminación de variables con un gran número de valores faltantes. En el caso de variables como peso y talla se optó eliminarlas en lugar de reemplazar los valores perdidos. El reemplazo puede ser realizado utilizando el promedio de los valores de cada variable, sin embargo, si esto se hacía se estaga generando información falsa para los pacientes. Esto también puede sesgar el resultado de los análisis y las estadísticas sobre los datos de los pacientes.

- Se realizaron validaciones para variables numéricas tales como peso, talla, tensiones arteriales, edad y tiempo de evolución de la enfermedad. En revisión con la epidemióloga [3] se encontró que los valores máximos y mínimos para cada variable se encontraban dentro de los límites válidos.

2.3.3.4. Análisis del modelo de datos

Gracias a la estructura generada con la bodega de datos, es posible realizar análisis exploratorios sobre los datos médicos. La generación de cubos sobre los hechos de diagnósticos y procedimientos, permite también realizar estadísticas rápidas de interés para el personal médico o los analistas de riesgo en salud.

Reportes tales como el número de pacientes por familias de enfermedad en determinados meses son de interés para ver la evolución del total de pacientes que están a su cargo. De igual forma, una EPS puede darse una idea del cumplimiento de metas preventivas y controles de sus pacientes mediante los conteos que se realizan sobre los procedimientos clínicos.

Los cubos y reportes fueron generados con la herramienta Saiku[6] y Pentaho. Algunos ejemplos de estos reportes pueden ser apreciados en el anexo A.

[3] Análisis realizado con la ayuda de Alexandra Castillo, Enfermera y Epidemióloga. Bogotá, Abril de 2013

2.4. Modelo de Minería de Datos

A partir de la centralización de datos clínicos en la bodega es posible generar vistas preprocesadas de datos [4] para realizar diferentes tipos de análisis de riesgo. Éstos pueden llevarse a cabo con varias técnicas de minería de datos, entre las cuales se destacan las reglas de asociación, agrupación y clasificación.

2.4.1. Selección de variables

Para realizar análisis sobre los datos, primero es necesario definir el grupo de estudio y las variables que se van a usar para el análisis de riesgo. En este caso, se debe generar una vista preprocesada a partir de la bodega de datos clínica, que reúna varias de las características de interés para un análisis de riesgo en salud.

Se hizo una selección de variables teniendo en cuenta que se analizará la enfermedad Diabetes Mellitus Tipo 2. Para realizar esta selección de variables se generó la tabla 2.2, la cual refiere el total de afiliados que tienen registro en la bodega de cada variable y se analizó con la epidemióloga [5]. Los últimos 7 registros de la tabla fueron seleccionados por ser variables relacionadas con la DM2 y por tener un número elevado de registros para dichos pacientes. Se descartaron algunas de las variables ecosocioculturales por no ser de interés para el análisis o bien porque tenían un gran número de valores faltantes para los pacientes de esta enfermedad.

2.4.2. Vistas Preprocesadas y preparación de datos

Con el fin de aplicar técnicas de minería de datos que permitan ayudar a generar un modelo de análisis de riesgo, se decidió generar vistas preprocesadas a partir de la bodega de datos clínica. A continuación se describe el proceso empleado para la generación de las vistas.

Estas vistas emulan el comportamiento de los pacientes, lo que quiere decir que un registro representa a un paciente con todos los diagnósticos, procedimientos y características que ha tenido a lo largo del periodo de estudio. Por esta razón, cada uno de las variables (procedimientos, diagnósticos) son transformadas de la forma "transactional" a la forma "basket". Por su parte, las variables ecosocioculturales toman el último valor registrado en el evento para cada una de ellas.

Usando el conocimiento que se tiene de datos clínicos se pueden categorizar algunas variables para que su análisis sea más interesante en el campo médico y acorde a estándares internacionales.

Gracias a la ayuda de un experto, fue posible hacer transformaciones a atributos numéricos y algunos categóricos tales como los datos de presión arterial, edad, estrato y tiempo de evolución de la enfermedad.

[4] Una vista preprocesada de datos contiene una extracción de registros y variables de la bodega de datos. Los datos contenidos en la vista son preprocesados para cumplir con los requerimientos de los algoritmos de minería de datos que van a ser utilizados

[5] Análisis realizado con la ayuda de Alexandra Castillo, Enfermera y Epidemióloga. Bogotá, Abril de 2013

Tabla 2.2. Variables Ecosociocultural

Variable	Total afiliados
Cocina con leña hace más de 10 años en recinto cerrado	27
Año de abandono del Habito	1988
Año inicio fumador	2119
Correo electrónico	4612
Edad de la menopausia	7697
Número de cigarrillos al Día	8094
Uso de anticonceptivos en mujeres	9929
Numero de Gestaciones	51471
Edad de la menarquía	52906
Fuma	53187
Estatura	53907
Tensión arterial diastólica	97635
Peso	104456
Tensión arterial sistólica	109846
Estrato	163910
Código del departamento de residencia	163967
Código Dane de municipio Residencia	163967
Zona de Residencia	163967

Se realizó la fusión de variables como TAS y TAD (tensión arterial sistólica y tensión arterial diastólica) en una sóla variable llamada TENSIÓN la cual se encuentra categorizada acorde a la guía NICE (2011). Variables como estrato, aunque son categóricas, pueden ser agrupadas según el esquema de interés y cercanía de condiciones; esto implica que se caracterizaron los estratos en 3 grupos: estratos 1 y 2, estratos 3 y 4, y estratos 5 y 6.

Aprovechando el conocimiento médico descrito anteriormente, todas las variables numéricas fueron convertidas a variables categóricas y posteriormente a variables binarias usando codificación ficticia o "Dummy Coding". Esta transformación consiste en representar cada categoría de la variable en n-1 variables dummy. Estas variables se construyen de forma artificial y sólo pueden ser representadas con valores 0 y 1, de este modo es posible representar todas las categorías de la variable como la combinación de las n-1 variables dummy. Este método permite hacer comparaciones entre las categorías y es ampliamente utilizado en análisis de datos. Por ejemplo, la variable "fumador" que puede tener las categorías "fuma", "no fuma" y "exfumador", se representa mediante las variables dummy d1 y d2 tal como se muestra en la tabla 2.3.

Tabla 2.3. Ejemplo de variables dummy creadas para una variable con 3 categorías

Fumador	D1	D2
Fuma	1	0
No fuma	0	1
Exfumador	0	0

Finalmente, se obtuvieron las siguientes vistas preprocesadas para la generación del modelo de análisis de riesgo:

- VISTA A: Pacientes enfermos de DM2 con sus respectivos registros de diagnósticos y procedimientos clínicos realizados posterior al primer diagnóstico de DM2 y variables ecosocioculturales (tensión, edad, estrato, zona de residencia, municipio, departamento, tiempo de evolución de la enfermedad, género)

- VISTA B: Pacientes enfermos de DM2 con sus respectivos registros de diagnósticos realizados después del primer diagnóstico de DM2 y antes de la incidencia de la comorbilidad. También las variables de tiempo de evolución de la enfermedad, género, edad, clase: Comorbilidad y No comorbilidad.

- VISTA C: Pacientes enfermos de DM2 con sus respectivos registros de diagnósticos realizados después del primer diagnóstico de DM2 y antes de la incidencia de la comorbilidad. Se encuentran también las variables de tiempo de evolución de la enfermedad, género, edad, clase: Micro vascular y Macro vascular.

Para la identificación de pacientes con comorbilidad y el tipo de ésta, se utilizaron los códigos CIE10 descritos en la tabla 2.4.

Tabla 2.4. Códigos CIE10 para comorbilidades

Comorbilidad	Códigos CIE10
Micro vascular	E115, E125, I250, I251, I610, I611, I612, I613, I614, I615, I616, I618, I619, I738, I739
Macro vascular	E112, E122, E114, E124, G632, G590, E113, E123, H350, H360, H280, H350, H352, H540, H541, H544

2.4.3. Modelo de categorización de pacientes DM2 y predicción de comorbilidades asociadas a DM2

El análisis de riesgo de una enfermedad de alto costo como la DM2 en Colombia requiere de una identificación de perfiles de pacientes. Así mismo, se considera de gran ayuda complementar la caracterización de esos perfiles clínicos con una herramienta que pueda ayudar a predecir futuras complicaciones o comorbilidades en los pacientes. Todo esto con el fin de brindar mejores planes de prevención y así reducir tanto la incidencia de éstas como su costo asociado.

A continuación se describen los modelos de análisis de riesgo de DM2 generados a partir de la minería de datos.

2.4.3.1. Modelo de caracterización de pacientes DM2

La generación de perfiles de pacientes para una enfermedad como la DM2 puede ayudar a orientar a las entidades prestadoras de salud sobre la eficiencia y manejo de sus planes de prevención y manejo de la enfermedad. La caracterización de pacientes permite visualizar el comportamiento natural de los pacientes y detectar posibles grupos que se están viendo afectados por falta de tratamiento o por enfermedades específicas. En estos

casos, los perfiles de pacientes generan una herramienta para los analistas de riesgo en el sentido de poder determinar si el manejo de la enfermedad se está haciendo correctamente o si, por el contrario, se deben generar nuevos planes acordes al comportamiento que están presentando los pacientes. Esto puede ser visto en términos de diagnósticos y procedimientos que están registrando los pacientes diabéticos a lo largo del desarrollo de su enfermedad.

Para la generación de perfiles de pacientes se optó por utilizar una técnica no supervisada, en este caso se usaron algoritmos de agrupación y posteriormente se describieron los grupos generados utilizando reglas de asociación. El modelo se presenta en la figura 2.2.

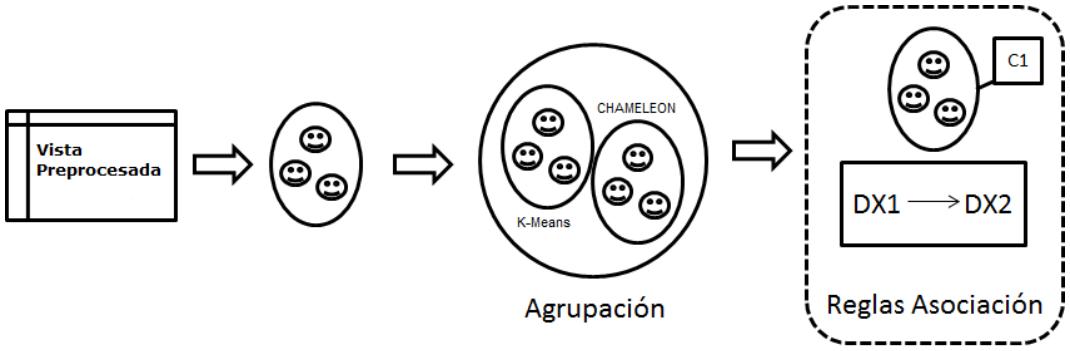

FIGURA 2.5. Modelo de caracterización de pacientes de DM2.

Los algoritmos de agrupación son útiles en determinar los diferentes perfiles que pueden generarse entre los pacientes de DM2. Sin embargo, es necesario definir qué algoritmo usar y cuál medida de similitud es apropiada para los datos clínicos.

Debido a que el preprocesamiento de los datos culminó con una gran porción de datos categóricos y binarios, se optó por escoger una medida que trabajara bien con datos binarios. En este caso, la medida escogida fue la de Jaccard puesto que gracias a su definición permite enfocarnos en los comportamientos que comparten los pacientes entre sí.

La similitud de Jaccard se define como:

$$S_{ij} = \frac{p}{p+q+r}$$

Donde p es el número de de positivos para ambos objetos, q es el número de de positivos para el i-ésimo objeto y negativos para el $j$ésimo objeto y r es el número de negativos para el i-ésimo objeto y positivos para el j-ésimo objeto.

Los algoritmos particionales y jerárquicos son los más usados en la literatura para realizar agrupación de pacientes,pero se destacan particularmente los jerárquicos aglomerativos [22, 40] donde se ha visto su buena generación de perfiles de pacientes con condiciones complejas. Por esta razon, para el modelo de caracterización de pacientes se trabajó con Hierarchical Clustering (CHAMELEON) y K-Means a modo de comparación.

El algoritmo CHAMELEON ha demostrado buenos resultados anteriormente en diferentes campos de aplicación, y su escogencia se basa en la capacidad que tiene para determinar automáticamente el número de grupos y los conceptos de cercanía basado en grafos y en medidas propias de los grupos formados, donde se toma en cuenta la densidad

de la región haciéndolo robusto a grupos con formas no gaussianas y garantizando una formación de grupos mucho más natural.

Ambos fueron ejecutados utilizando la similitud de Jaccard y los resultados sugirieron que, tanto por validación computacional como médica de los resultados, el algoritmo CHAMELEON era el apropiado para encontrar los perfiles de pacientes de DM2 (ver sección 3.1).

La ventaja adicional que presentan los algoritmos jerárquicos para la agrupación de pacientes es que, al ser aglomerativo, permite empezar a agrupar pacientes muy cercanos entre sí.

Una vez identificados los perfiles, es posible describirlos mediante relaciones interesantes entre sus datos mediante el uso de reglas de asociación. Estas reglas pueden mostrar los factores de riesgo asociados a cada uno de los perfiles. Además, la descripción de los perfiles puede servir como método de control de manejo de enfermedad de los pacientes, puesto que se pueden relacionar los procedimientos clínicos que están realizándose constantemente los pacientes.

Cabe resaltar que el hecho de que el conjunto de datos utilizado para agrupación se encontrara en forma "basket" y luego fuera binarizado, contribuyó al fácil preprocesamiento para usar los algoritmos de reglas de asociación. En este caso, se seleccionó el algoritmo FP-Growth el cual trabaja con datos de tipo "basket".

2.4.3.2. Modelo de Predicción de Comorbilidades

Los pacientes de DM2 pueden sufrir condiciones crónicas adicionales a su enfermedad primaria las cuales son llamadas comorbilidades. Éstas pueden estar directamente relacionadas con la diabetes como lo son las enfermedades micro y macrovascular o pueden desarrollar comorbilidades no relacionadas directamente con la patología, como es el caso de depresión y enfermedades musculo-esqueléticas[47]. Las enfermedades macrovasculares están relacionadas con problemas que afectan al corazón y a los vasos sanguíneos grandes, causando altos índices de mortalidad y morbilidad en los pacientes diabéticos. Por su parte, las enfermedades microvasculares tienen su origen en retinopatías diabéticas, nefropatías y neuropatías. Particularmente, los pacientes con este tipo de complicaciones terminan con diálisis, ceguera, necesidad de transplante de riñones y un alto consumo de recursos hospitalarios[52].

Las comorbilidades asociadas a la DM2 pueden llevar a sus pacientes a la muerte o altos grados de discapacidad, particularmente las comorbilidades de tipo micro y macrovascular. El tratamiento de estas complicaciones representa un alto impacto en el costo global de un paciente con DM2. En Colombia, el costo de las comorbilidades de DM2 abarcan un 86 % del costo directo de la enfermedad y un 95 % del costo indirecto [25] y enfermedades como la cardiopatia isquémica o cardiopatía hipertensiva, ambas relacionadas a comorbilidades de la DM2, se encuentran entre las enfermedades con más índice de mortalidad en el país [1].

Por lo anterior, es importante para para el panorama colombiano tratar de detectar tempranamente un paciente con riesgo de sufrir una comorbilidad bien sea de tipo macro o microvascular. En este caso, la mejor herramienta de minería de datos que se puede tener para ayudar con este problema es la clasificación.

Los algoritmos de clasificación permiten generar modelos que describen las principales características que pueden llevar a un paciente a una condición dada. Para la predicción de pacientes con comorbilidades, esto representa una gran ayuda para disminuir incidencias y costos de la enfermedad.

Las técnicas de clasificación más utilizadas en el campo médico comprenden las redes bayesianas y los árboles de decisión. Éstos últimos resultan ser los más claros para el personal médico puesto que es fácil entender el modelo y visualizar las variables que son decisivas en el resultado de la predicción.

Para la predicción de comorbilidades se realizó un clasificador, el cual consta de dos partes. La primera parte (A) utiliza árboles de decisión para determinar si un paciente puede o no sufrir de una comorbilidad. La segunda parte (B), utiliza una red bayesiana para identificar qué tipo de comorbilidad puede sufrir el paciente: micro o macro vascular. Los experimentos sobre los datos permitieron definir qué técnica era más apropiada para cada una de las partes del clasificador (ver sección 3.1) .El modelo de predicción de comorbilidades se puede apreciar en la figura 2.3.

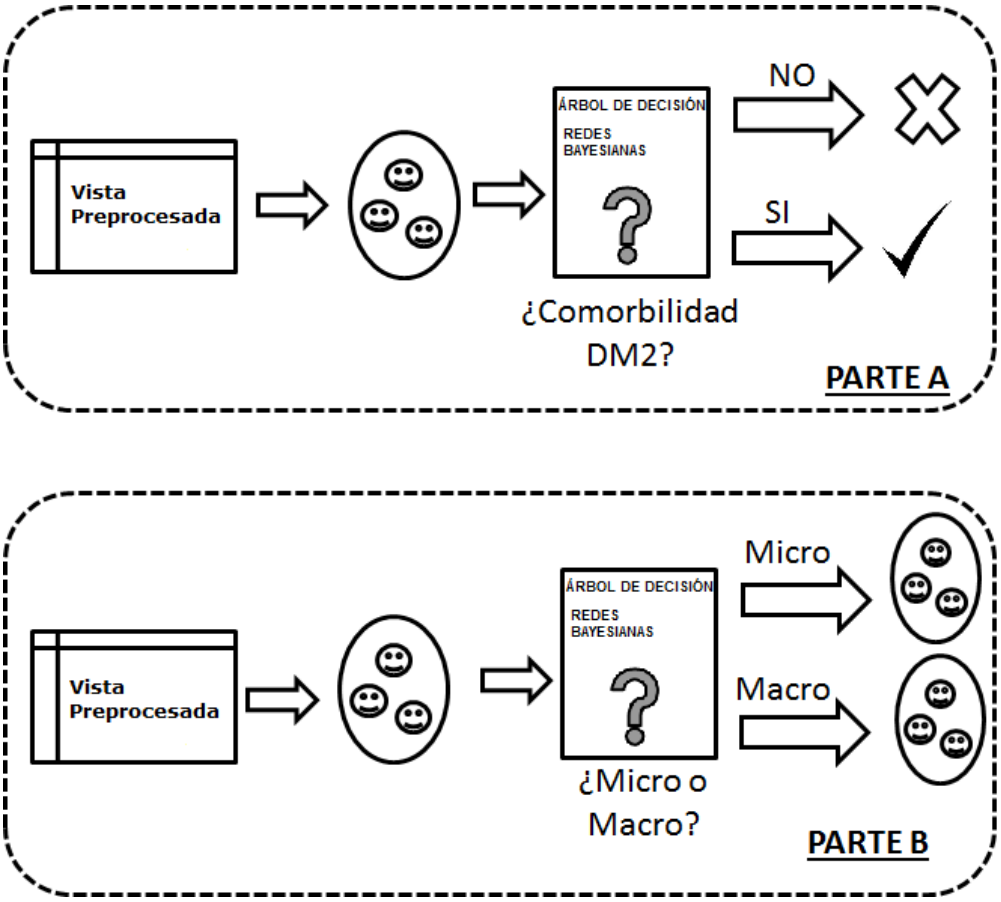

FIGURA 2.6. Modelo de predicción de comorbilidades de DM2. Parte A: Describe la posibilidad de que un paciente de DM2 tenga una comorbilidad. Parte B: Describe, de los pacientes con comorbilidad, cuál es la más probable de desarrollar entre Microvascular y Macrovascular

Primero se generan dos vistas preprocesadas donde se incluyen los diagnósticos de los pacientes, el tiempo de evolución de la enfermedad y variables como la edad y el género. Estas variables se seleccionaron para tener un universo mayor de análisis, puesto que

existe más información de servicios de salud de los pacientes que registros de variables socio económicas.

La primera vista preprocesada contiene los pacientes totales de DM2 discriminados en “COMORBILIDAD” y “NO COMORBILIDAD”. La segunda vista preprocesada contiene los pacientes de comorbilidad DM2, pero discriminados en las clases “MICRO” y “MACRO”.

Para ser consistente con el análisis de predicción de una comorbilidad, las variables seleccionadas se extrajeron de la bodega clínica en periodos anteriores al primer registro de la comorbilidad y posteriores al primer diagnóstico de DM2.

De esta forma, el modelo debe estar en capacidad de encontrar posibles pacientes que desarrollarán una comorbilidad y de qué tipo, utilizando las características del paciente y sus diagnósticos.

2.5. Resumen

En este capítulo se presentó el modelo general de análisis de riesgo el cual se encuentra compuesto por dos secciones: el modelo de datos clínicos y el modelo de minería. En el modelo de datos clínicos se describen las fuentes de datos usadas en el trabajo investigativo, el porqué de una bodega de datos clínico y su estructura.

El modelo de minería de datos describe la selección de variables realizada con concepto clínico, las vistas preprocesadas que fueron generadas para los distintos tipos de análisis así como el preprocesamiento que sufrieron y finalmente, se plantean los modelos de caracterización de pacientes de DM2 haciendo uso de los algoritmos de agrupación y el modelo de predicción de comorbilidades el cual utiliza clasificadores.

CAPÍTULO 3

Experimentos y validaciones de resultados del modelo de análisis de riesgo de DM2

En este capítulo se decriben los experimentos que se llevaron a cabo para validar el modelo propuesto en el capítulo anterior. También se explica la razón de escogencia de los algoritmos que permitieron caracterizar pacientes de DM2 y predecir comorbilidades, todo esto mediante comparación de resultados y su correspondiente validación con expertos en el área médica.

3.1. Aspectos generales de la experimentación y validación

El modelo propuesto de análisis de riesgo para la DM2 está compuesto por una bodega de datos clínica y varias técnicas de minería de datos. Las validaciones del modelo se basan en las propias de cada técnica de minería utilizado permitiendo encontrar ventajas, limitaciones y posibles mejoras sobre éstos. Respecto a los resultados obtenidos con el modelo de análisis de riesgo para DM2, éstos fueron validados tanto por una epidemióloga como por una analista de riesgo en salud [1]. Esta validación permite encaminar las buenas prácticas sobre preprocesamientos de datos, selección de algoritmos de minería de datos y entendimiento de los problemas de salud a los que se ve enfrentada la población colombiana.

Las herramientas utilizadas en este trabajo investigativo para minería de datos fueron no propietarias y son:

- RapidMiner[38]: Preprocesamientos, reglas de asociación y algoritmos de agrupación.
- Weka[36]: Algoritmos de clasificación.
- CLUTO[29]: Algoritmos de agrupación.

Acorde al concepto de los analistas clínicos, un soporte de menos del 20 % no debe ser usado para extraer reglas de asociación de un grupo de pacientes. Por ejemplo, en el caso de generación de un grupo de pacientes con 100 individuos, no es trascendente que

[1] Liz Garavito, Directora ejecutiva de Processum LTDA e investigadora en el área de seguridad social y economía de la salud

sólo 20 presenten cierta condición específica; así pues, tiene más sentido que entre el 80 % y 100 % de los individuos presente esta condición. Por esta razón, los soportes mínimos escogidos para la generación de reglas de asociación entre los perfiles de pacientes de DM2 fueron del 70 % para grupos mayores a 30 individuos y del 90 % en caso contrario. Todas las reglas fueron generadas con una confianza del 90 %.

Respecto al uso de redes bayesianas en los modelos de predicción, se optó por permitir que el algoritmo generara automáticamente la red bayesiana sin la intervención de personas expertas clínicas puesto que el modelo propuesto se encamina en descubrir factores nuevos y no en predefinirlos. Las redes bayesianas generadas a partir del conjunto de datos pueden ser visualizadas mediante grafos de causalidad.

3.2. Experimientos y validación de algoritmos de agrupación para caracterizar pacientes de DM2

Los algoritmos escogidos para realizar la caracterización de pacientes de DM2 fueron el algoritmo CHAMELEON y el K-Means. CHAMELEON es un algoritmo jerárquico y mide la similitud de dos grupos basándose en un modelo dinámico. En el proceso de fusión de los grupos obedece a la minimización de la medida de proximidad entre estos grupos [30]. Por su parte K-Means es un algoritmo particional que utiliza una técnica de refinamiento iterativo mediante la mejora del centroide del grupo [48]. La herramienta CLUTO[29] fue utilizada para ejecutar el algoritmo CHAMELEON con medida de similitud de Jaccard, mientras que para la ejecución de K-Means se utilizó la herramienta RapidMiner.

Dado que el problema de caracterizar pacientes de DM2 es no supervisado, fue necesario utilizar la validación intra-grupo. Respecto a la validación de los resultados suministrados por ambos algoritmos, se utilizaron reglas de asociación para describir los grupos de pacientes generados y de este modo el analista de riesgo en salud[2] proporcionó su opinión respecto a la consistencia de los resultados dentro del contexto de manejo de riesgo en salud.

3.2.1. Parametrización de los experimentos

Debido a que se desconoce el número de grupos o perfiles apropiado para el conjunto de pacientes, se optó por utilizar la gráfica del error cuadrático medio (ECM) por cada grupo, la cual puede ser apreciada en la figura 3.1.. El error cuadrático se define como la suma de las distancias cuadradas entre cada miembro del grupo y su centroide. Por tal motivo, el ECM puede ser considerado como una medida global de error. En términos generales, a medida que el número de grupos aumenta, el EMC disminuye debido a que los grupos empiezan a ser más pequeños en tamaño. Esto indica que es necesario seleccionar el número de grupos k a partir del cual el error cuadrático medio empieza a decrecer [30]. De este modo se determinó que el número apropiado de grupos correspondía a cuatro, puesto que a partir de este número el ECM no se encuentra altamente impactado. Adicionalmente, se observa una gran variación entre un $k = 4$ y $k = 5$ de más del 28 %. A partir de estos valores k, la variación promedio del ECM corresponde aproximadamente a un 5 % entre los distintos valores k.

[2]Liz Garavito, Directora ejecutiva de Processum LTDA e investigadora en el área de seguridad social y economía de la salud

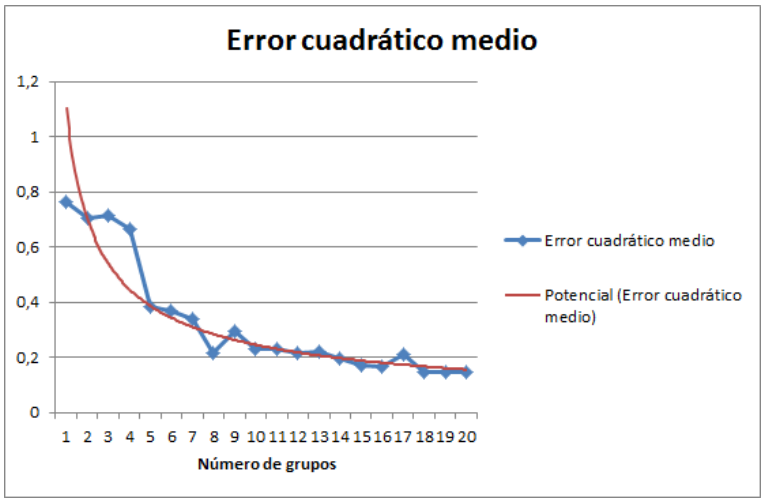

FIGURA 3.1. Cálculo del error cuadrático medio para determinar el número de grupos a usar en el algoritmo K-Means

Para la caracterización mediante el algoritmo K-Means se utilizó la herramienta RapidMiner. El máximo número de 10 de ejecuciones del algoritmo con una inicialización aleatoria y un máximo de 100 iteraciones por ejecución del algoritmo. Respecto a la configuración de parámetros del algoritmo CHAMELEON en la herramienta CLUTO[29], se utilizó un valor inicial de k igual a 4 y debido a que requiere la especificación del número de particiones iniciales del conjunto de datos, a partir de las cuales va a empezar a agrupar. Para este caso, se escogió como iniciación el valor 30. Para ambos algoritmos, tanto K-Means como CHAMELEON, se utilizó la medida de similitud de Jaccard.

3.2.2. Caracterización usando K-Means

Los resultados que se obtuvieron para la generación de 4 grupos con K-Means se observan en la tabla 3.1.

TABLA 3.1. Número de grupos generados por K-Means y distancia intra-grupo

grupo	Tamaño	Distancia Intra-grupo
1	60	0,010
2	1406	0,006
3	210	0,008
4	198	0,006

Debido a que la distancia intra-grupo es pequeña, se puede decir que los grupos de pacientes generados son compactos. Sin embargo, es necesario entrar a revisar cada uno de los grupos con el fin de descubrir si tienen sentido dentro del contexto clínico.

Las reglas de asociación aplicadas para cada grupo fueron indispensables para su descripción. Las descripciones obtenidas mediante las reglas generadas son:

- Grupo 1: Pacientes en su mayoría de Cundinamarca, con consultas a especialista asociada a diagnósticos como la hipertensión arterial y controles propios de la diabe-

tes como exámenes A1C y Glicemia Basal. El tiempo de evolución de la enfermedad para estos pacientes es de menos de 5 años.

- Grupo 2: Pacientes en su mayoría de Bogotá, con consultas a especialista asociadas a hipertensión arterial y trastornos de refracción y acomodación del ojo como miopía, presbiscia y astigmatismo. Pacientes que presentan controles A1C y Glicemias basales.

- Grupo 3: Pacientes en su mayoría de Bogotá, con presencia de caries y daños en la estructura dental. Consumos de especialista y controles de odontología. Pacientes con control de glicemia y A1C.

- Grupo 4: Pacientes en su mayoría de Cundinamarca, con controles de glicemia y A1C. Presentan consultas a especialista asociada a hipertensión arterial y tiempo de evolución menor a 5 años.

3.2.3. Caracterización utilizando CHAMELEON

Los resultados de los grupos se muestran en la tabla 3.2.

TABLA 3.2. Número de grupos generados por CHAMELEON y similitud intra-grupo

Grupo	Tamaño	Similitud Intra-grupo
1	39	1
2	38	1
3	38	1
4	28	1
5	7	1
6	45	0,83
7	46	0,79
8	15	0,588
9	5	0,678
10	72	0,314
11	76	0,283
12	90	0,208
13	101	0,162
14	110	0,137
15	256	0,074
16	310	0,034
17	552	No Aplica

Como se puede apreciar en la tabla anterior, los grupos de pacientes generados por CHAMELEON son, en general, muy compactos con excepción de los grupos más grandes en los cuales la similitud intra-grupo no es tan alta. Nótese que el último grupo no tiene un valor para la similitud intra-grupo, esto se debe a que este conjunto de pacientes no fue utilizado por el algoritmo para ser agrupado. Esta situación se presenta debido a que estos pacientes pueden no compartir ningun atributo (vértice) con el resto de pacientes. Sin embargo, esto hace interesante analizar el grupo 17 en búsqueda de patrones que describan su comportamiento atípico respecto al conjunto en general de pacientes de DM2.

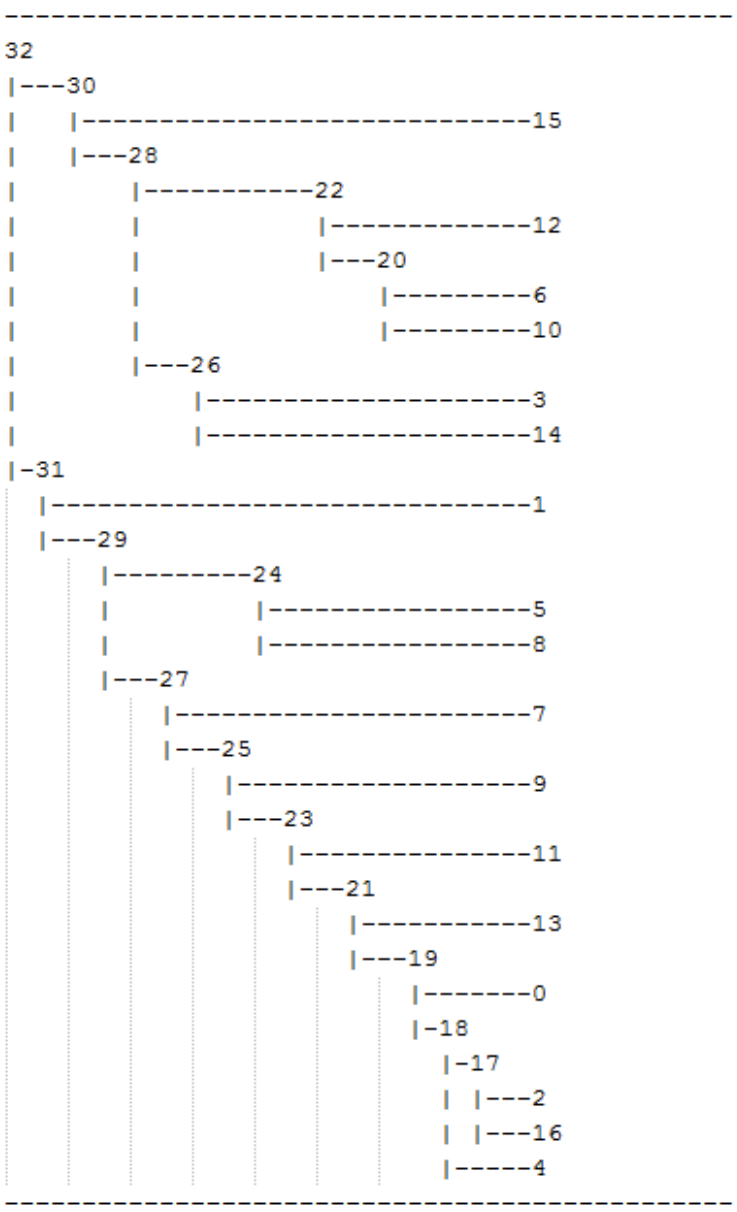

FIGURA 3.2. Dendograma de los grupos generados con CHAMELEON para pacientes de DM2

Con el fin de describir los grupos y encontrar las relaciones de características entre los pacientes de cada uno de ellos, se generaron reglas de asociación las cuales tuvieron en cuenta el criterio médico de interés sobre el número de pacientes en cada grupo para especificar el soporte y la confianza de cada uno de estos grupos. Tal como se menciona en la sección 3.1, se propone que entre más pequeño sea el grupo, mayor debe ser el soporte para garantizar que se capturarán las relaciones más significativas que describen al grupo. La tabla 3.3 muestra los valores de soporte y confianza determinados para cada grupo.

Las reglas de asociación permitieron describir cada uno de los grupos encontrados por CHAMELEON. La caracterización de pacientes acorde a esta agrupación es:

TABLA 3.3. Valores de soporte y confianza para las reglas de asociación

Grupo	Soporte	Confianza
1	0.7	0.9
2	0.7	0.9
3	0.7	0.9
4	0.8	0.9
5	0.9	0.9
6	0.7	0,9
7	0.7	0,9
8	0.7	0,9
9	0.9	0,9
10	0.9	0,9
11	0.7	0,9
12	0.7	0,9
13	0.7	0,9
14	0.7	0,9
15	0.7	0,9
16	0.7	0,9
17	0.7	0,9

- Grupo 1: Pacientes mujeres de estratos 1 y 2 entre los 44 y 60 años con rango de tensión normal. Presentan consulta de medicina especializada y controles A1C y Glicemias. Tiempo de evolución menor a 5 años.

- Grupo 2: Pacientes hombre de estratos 3 y 4 entre los 18 y 44 años con rango de tensión normal. Presentan controles A1C y Glicemias, el tiempo de evolución es de menos de 5 años.

- Grupo 3: Pacientes hombres de estratos 1 y 2 con más de 60 años y rango de tensión normal. Presentan consumos de medicina especializada, controles A1C y de Glicemias.

- Grupo 4: Pacientes mujeres de estratos 3 y 4 entre los 18 y 44 años con rango de tensión normal. Presentan controles A1C y Glicemias.

- Grupo 5: Pacientes mujeres de estratos 3 y 4 mayores de 60 años con rango de tensión HTA Grave. Consumen servicios de medicina especializada y presentan controles A1C y Glicemias.

- Grupo 6: Pacientes mujeres de estratos 1 y 2 mayores de 60 años con rango de tensión normal. Consumen servicios de medicina especializada y presentan controles A1C y Glicemias.

- Grupo 7: Pacientes hombres de estratos 3 y 4 mayores de 60 años con tension HTA Sistólica Aislada. Presentan consulta de medicina especializada y controles A1C y Glicemias.

- Grupo 8: Pacientes hombres de estratos 1 y 2 entre los 44 y 60 años con tensión normal. Presentan consulta de medicina especializada y controles A1C y Glicemias.

- Grupo 9: Pacientes mujeres de estratos 5 y 6 mayores de 60 años con tensión normal. Consumen servicios de medicina especializada y presentan controles A1C y Glicemias.

- Grupo 10: Pacientes hombres de zona rural con consumos de medicina especializada y controles A1C y glicemias. Presencia de diagnósticos asociados a trastornos de refracción y acomodación del ojo como miopía, presbiscia y astigmatismo.

- Grupo 11: Pacientes mujeres de estratos 3 y 4 mayores de 60 años con tension HTA Sistólica Aislada. Presentan consulta de medicina especializada y controles A1C y Glicemias.

- Grupo 12 y Grupo 13: Estos dos grupos se pueden fusionar debido a que lo único que los diferencia es el género de los pacientes. El grupo 12 está compuesto por pacientes femeninas mientras que el 13 por pacientes masculinos. Sin embargo, las demás características son compartidas entre ambos grupos. Éstas son: ser pacientes de estratos 3 y 4 mayores de 60 años con tensión en rango normal. Presentan consulta de medicina especializada y controles A1C y Glicemias.

- Grupo 14 y Grupo 15: Pacientes hombres de estratos 3 y 4 entre 44 y 60 años con tensión normal. Consumen medicina especializada y controles A1C y glicemias.

- Grupo 15: Pacientes mujeres de estratos 3 y 4 entre 44 y 60 años con tensión normal. Consumen medicina especializada y controles A1C y glicemias.

- Grupo 16: Pacientes de Cundinamarca que presentan consulta de medicina especializada y controles A1C y Glicemias.

- Grupo 17: Pacientes bogotanos con consulta de medicina especializada y controles A1C y Glicemias. Presentan exámenes de microalbuminuria.

Los resultados de validación con los expertos clínicos sugieren que es de más importancia la agrupación jerárquica realizada con CHAMELEON, ya que permite ver características muy particulares en ciertos grupos los cuales pueden describir varias tendencias de riesgo en la población de pacientes de DM2.

Los resultados de CHAMELEON resultaron ser consistentes respecto a lo esperado en cuanto a comportamientos para los pacientes de DM2. Por su parte, los resultados encontrados por K-Means no distinguieron aspectos claves del comportamiento de los pacientes, como es el control de la tensión y la discriminación de poblaciones puntuales a pesar de mostrar resultados superiores en cuanto a medidas de similitud interna.

Los grupos que más interés generaron por sus características son los primeros 9, los cuales también presentan una buena medida de similitud intra-grupo, a pesar de ser grupos con números de pacientes reducidos. Los demás grupos, del 10 al 17, no presentan características aisladas por lo que muchos de ellos pueden ser fusionados entre sí. Adicionalmente, estos últimos grupos representan la mayoría de los pacientes, y de acuerdo a las reglas generadas sobre estos se puede inferir que sus características corresponden al comportamiento promedio de los pacientes de DM2.

En gran parte de los grupos de pacientes generados por el algoritmo jerárquico es posible encontrar que la enfermedad se encuentra controlada, esto se puede ver porque los niveles de tensión se encuentran normales, no se presentan consumos de medicina

especializada significativos y se están cumpliendo los controles de A1C y glicemias. Sin embargo, es posible ver grupos que pese a que se cumplen los controles de A1C y glicemias, la tensión se encuentra en un estadio grave y además se están utilizando frecuentemente consultas de medicina especializada (grupos 5, 7 y 11). Este último escenario sugiere que los pacientes pertenecientes a esos grupos pueden presentar problemas de adherencia debido a que no toman el tratamiento. Debido a que son pacientes mayores de 60 años, se puede decir también que estas características se deben al deterioro físico por la edad o bien a falta de buenas prácticas de autocuidado del paciente de DM2.

3.3. Experimentos y validación de algoritmos para predicción de comorbilidades

La predicción de comorbilidades se basó en dos grandes preguntas: ¿Va a enfermarse un paciente de DM2 de una comorbilidad directamente asociada con esta enfermedad? y ¿Qué tipo de comorbilidad es más susceptible a desarrollar?

Con base en estas preguntas, se plantearon dos clasificadores los cuales fueron entrenados con dos conjuntos de datos acordes a su finalidad. Ambos se describen en las secciones 3.3.2. y 3.3.3. respectivamente

3.3.1. Parametrización de los experimentos

Los dos tipos de clasificadores seleccionados fueron los árboles de decisiones y las redes bayesianas. Particularmente se trató con el árbol J48 y el algoritmo de redes bayesianas respectivamente. La herramienta utilizada para generar los modelos de los clasificadores fue WEKA.

El árbol de decisión fue configurado usando la opción de poda y con un valor de confianza de 0.25 con el fin de favorecer la poda del árbol. Por su parte, la red bayesiana fue configurada para utilizar el algoritmo K2 de búsqueda para generar los valores de probabilidad directamente de los datos.

La evaluación del desempeño de los modelos de clasificación generados se realizó mediante las matrices de confusión, las cuales proveen información de las instancias correcta e incorrectamente clasificadas. También se usaron medidas como la *Precisión* la cual se define como la relación entre las instancias correctamente clasificadas y el total de predicciones[30].

3.3.2. Predictor de comorbilidad en pacientes DM2

El conjunto de datos utilizado para este clasificador presenta desbalance de clases. El 83.29 % pertenecen a la clase “No Comorbilidad”, mientras que el 16.71 % es etiquetado como “Comorbilidad”. En este caso, por ser una de las clases de más interés en determinar se realizaron experimentos utilizando el total de los datos, un balanceo por muestreo aleatorio simple y una matriz de costos. Los valores de la matriz de costos se observan en la tabla 3.4. y aplican la máxima penalización para los falsos negativos, mientras que la penalización por predecir falsos positivos es menor.

Tabla 3.4. Matriz de Costos

	Pred. No Comorbilidad	Pred. Comorbilidad
Real No Comorbilidad	0	1
Real Comorbilidad	5	0

Los resultados obtenidos para este predictor se muestran en las tablas de la 3.5 a la 3.11.

Tabla 3.5. Resultados generales de los clasificadores

Clasificador	Exactitud	Área ROC
Red Bayesiana (Total)	81.42 %	0.831
Red Bayesiana (Muestreo)	75.42 %	0.828
Red Bayesiana (Costo)	78.91 %	0.832
Árbol de Decisión (Total)	90.42 %	0.844
Árbol de Decisión (Muestreo)	87.27 %	0.859
Árbol de Decisión (Costo)	88.09 %	0.842

Tabla 3.6. Matriz de confusión para la Red Bayesiana (Total)

	Pred. No Comorbilidad	Pred. Comorbilidad	Precisión de la clase
Real No Comorbilidad	10342	1492	90 %
Real Comorbilidad	1147	1226	45.1 %
Recuerdo de la clase	87.4 %	51.7 %	

Tabla 3.7. Matriz de confusión para la Red Bayesiana (Muestreo)

	Pred. No Comorbilidad	Pred. Comorbilidad	Precisión de la clase
Real No Comorbilidad	3291	709	79.3 %
Real Comorbilidad	857	1516	68.1 %
Recuerdo de la clase	87.4 %	51.7 %	

Tabla 3.8. Matriz de confusión para la Red Bayesiana (Costo)

	Pred. No Comorbilidad	Pred. Comorbilidad	Precisión de la clase
Real No Comorbilidad	9565	2269	92.9 %
Real Comorbilidad	726	1647	42.1 %
Recuerdo de la clase	80.8 %	69.4 %	

Tabla 3.9. Matriz de confusión para el Árbol de Decisión (Total)

	Pred. No Comorbilidad	Pred. Comorbilidad	Precisión de la clase
Real No Comorbilidad	11115	719	94.5 %
Real Comorbilidad	641	1732	70.7 %
Recuerdo de la clase	93.9 %	73 %	

Los resultados nos sugieren que el mejor clasificador para este problema es el árbol de decisión, particularmente el que trabaja con una matriz de costos. Debido a que existe un

TABLA 3.10. Matriz de confusión para el Árbol de Decisión (Muestreo)

	Pred. No Comorbilidad	Pred. Comorbilidad	Precisión de la clase
Real No Comorbilidad	3637	363	89 %
Real Comorbilidad	448	1925	84.1 %
Recuerdo de la clase	90.9 %	81.1 %	

TABLA 3.11. Matriz de confusión para el Árbol de Decisión (Costo)

	Pred. No Comorbilidad	Pred. Comorbilidad	Precisión de la clase
Real No Comorbilidad	10577	1257	96.1 %
Real Comorbilidad	434	1939	60.7 %
Recuerdo de la clase	89.4 %	81.2 %	

problema de desbalanceo de clases, las opciones de tratar con el muestreo de clases y con la matriz de costos son las más apropiadas para el problema. Adicionalmente, debido al contexto médico es más importante trabajar con una matriz de costos que le de prioridad a la correcta detección de la clase comorbilidad reduciendo los falsos negativos. Para efectos de selección de poblaciones a intervenir para planes de prevención, no tiene mayor importancia que se incremente el número de falsos positivos puesto que al fin y al cabo serán intervenidos y descartados.

En revisión con la analista [3] determinó que el modelo visual generado por el árbol de decisión es de fácil entendimiento pese a su tamaño. En este caso, los diagnósticos y las demás características atribuidas a los pacientes resultan ser las ramas del árbol de decisión.

Haciendo uso de la visualización del árbol de decisión es posible ver algunos aspectos importantes respecto a las variables que influencian un diagnóstico de comorbilidad para un paciente de DM2. Algunos de estos, interesantes para la analista [4], fueron:

- La edad de los pacientes y su género no son decisivos en la predicción de diagnóstico de una comorbilidad. En las ramas de los árboles no aparecen en ramas superiores el género de los pacientes ni la edad, por lo cual se deduce que no entregan información representativa para el modelo de clasificación.

- El tiempo de evolución de la DM2 es un factor importante para ayudar a predecir comorbilidades. En el modelo se pueden distinguir ramas superiores donde aparece el tiempo de evolución superior a 4 años.

- Se encuentran que los pacientes de DM2 que son clasificados con comorbilidad por el modelo, presentan diagnósticos de disfunción sexual así como desórdenes de ansiedad, enfermedades respiratorias, transtornos de personalidad y enfermedades digestivas.

- Se pueden observar relaciones entre enfermedades crónicas como el VIH y la DM2. Dos ramas principales del árbol cuentan con la presencia de un diagnóstico de VIH previo a la decisión de comorbilidad. En estas ramas se encuentra que la predicción de

[3] Liz Garavito, Directora ejecutiva de Processum LTDA e investigadora en el área de seguridad social y economía de la salud

[4] Liz Garavito, Directora ejecutiva de Processum LTDA e investigadora en el área de seguridad social y economía de la salud

comorbilidad para pacientes con VIH está relacionada con diagnósticos de obesidad y desórdenes de los glóbulos blancos. De estos dos diagnósticos, se destaca el de obesidad por ser un factor de riesgo conocido de la DM2. Así pues, el modelo muestra una relación entre dos enfermedades crónicas y puede propiciar un análisis más profundo de pacientes multisindrómicos.

- La mayoría de diagnósticos que se encuentran en las ramas que conducen al diagnóstico de comorbilidad, son consistentes con lo citado en la literatura médica. Entre estos diagnósticos se destacan los trastornos de la retina, la enfermedad cardíaca isquémica, insuficiencia renal, neuropatías e insuficiencia cardíaca [52].

3.3.3. Predictor de tipo de comorbilidad en pacientes DM2

El conjunto de datos utilizado para determinar el tipo de comorbilidad en pacientes de DM2 se encontraba con un desbalanceo leve, por lo que se optó por realizar un muestreo aleatorio simple para balancear las clases. El total de registros inicalmente correspondía a 2373 pacientes de los cuales el 64 % pertenecía a la clase "Microvascular" mientras que el 35 % a la clase "Macrovascular". Una vez realizado el muestreo, el total de registros finales fue de 1600 donde el 50 % pertenecían a la clase "Microvascular" y el otro 50 % a la clase "Macrovascular". Para este caso en particular no se trabaja una matriz de costos, puesto que no hay una clase positiva como tal. Esto implica que no es necesario penalizar los errores de clasificación de falsos negativos o positvos.

Los resultados obtenidos para este predictor se muestran en las tablas de la 3.12 a la 3.14.

Tabla 3.12. Resultados generales de los clasificadores

Clasificador	Exactitud	Área ROC
Red Bayesiana	69.81 %	0.77
Árbol de Decisión	67.75 %	0.72

Tabla 3.13. Matriz de confusión para Red Bayesiana

	Pred. Macro vascular	Pred. Micro vascular	Precisión de la clase
Real Macro vascular	533	267	71.2 %
Real Micro vascular	216	584	68.6 %
Recuerdo de la clase	66.6 %	73 %	

Tabla 3.14. Matriz de confusión para el árbol de decisión

	Pred. Macro vascular	Pred. Micro vascular	Precisión de la clase
Real Macro vascular	552	248	67.3 %
Real Micro vascular	268	532	68.2 %
Recuerdo de la clase	69 %	66.5 %	

Los resultados para este predictor sugieren que el modelo de red bayesiana presenta alrededor de 2 % más exactitud que el del árbol de decisión. Debido a que este algoritmo trabaja básicamente con medias y medidas estadísticas, la solución no muestra un modelo visual de fácil entendimiento. Adicionalmente, el grafo de causalidad utilizado para

describir la red bayesiana no reveló relaciones importantes que ayudaran a entender la predicción de enfermedades micro o macrovasculares.

Por otro lado, el modelo visual generado por el árbol de decisión fue revisado por la analista [5] , dando como resultado las siguientes conclusiones:

- Los pacientes con dolores en garganta y pecho son más susceptibles a desarrollar enfermedades macro vasculares.
- Los pacientes con edades superiores a 65 años se relacionan con enfermedades macro vasculares.
- El desarrollo de enfermedad micro vascular tiene relación con el diagnóstico de alcalosis metabólica.
- Es posible encontrar diagnósticos relacionados con las enfermedades micro vasculares, los cuales han sido citados en la literatura, tales como la falla renal crónica y el deterioro de la retina.

Acorde a la validación computacional tanto la red bayesiana como el árbol de decisión se acercan bastante a la meta de predecir comorbilidades para la DM2. Sin embargo, la validación con expertos encuentra que por su visualización es mucho más útil un árbol de decisión.

Los resultados obtenidos en las matrices de confusión sugieren que existe un margen de error de cerca del 30 % en todos los clasificadores para las clases. Estos resultados son buena aproximación en términos generales para un país como Colombia, en donde actualmente no se cuenta con herramientas computacionales de predicción para comorbilidades.

Los modelos generados por los árboles de decisión pueden ayudar a identificar posibles factores de riesgo que están directamente relacionados con el desarrollo de una comorbilidad en un paciente diabético. Ésto se puede lograr identificando los elementos pertenecientes a las ramas que conducen a las hojas con resultado de comorbilidad. Este mismo criterio de detección de factores de riesgo se puede aplicar en el predictor del tipo de comorbilidad.

3.4. Resumen

En este capítulo se describieron los experimentos hechos sobre el modelo de análisis de riesgo para la DM2 en Colombia. Se presentan las validaciones tanto computacionales como médicas a las que fueron sometidas las dos principales partes del modelo: caracterización de pacientes y predicción de comorbilidades.

Para el modelo de caracterización de pacientes se experimentó con los algoritmos K-Means y CHAMELEON usando en ambos la medida de similitud de Jaccard. El resultado de perfiles de pacientes generado por CHAMELEON fue el más relevante para la analista de riesgo [6] y en general mostró buenos resultados de similitud intra-grupo.

[5] Liz Garavito, Directora ejecutiva de Processum LTDA e investigadora en el área de seguridad social y economía de la salud

[6] Liz Garavito, Directora ejecutiva de Processum LTDA e investigadora en el área de seguridad social y economía de la salud

Respecto al modelo de predicción de comorbilidades, los algoritmos de redes bayesianas y árboles de decisión fueron analizados y probados. Ambos obtuvieron buenos niveles de predicción, sin embargo para los analistas médicos [7] [8] fue de más fácil entendimiento los modelos de clasificación generados por los árboles de decisión.

[7] Liz Garavito, Directora ejecutiva de Processum LTDA e investigadora en el área de seguridad social y economía de la salud

[8] Alexandra Castillo, Enfermera y Epidemióloga.

APÉNDICE A

Reportes OLAP

Los reportes presentados en este anexo fueron generados a partir de la bodega de datos médicos. Para realizar estos reportes se utilizaron las herramientas Schema Workbench de PENTAHO para la generación de cubos y posteriormente, se utilizó la herramienta Saiku[6] para su análisis.

Saiku es conocido como la herramienta de análisis de Pentaho y permite crear reportes de manera fácil y amigable. La interfaz de Saiku se puede apreciar en la figura A.1.

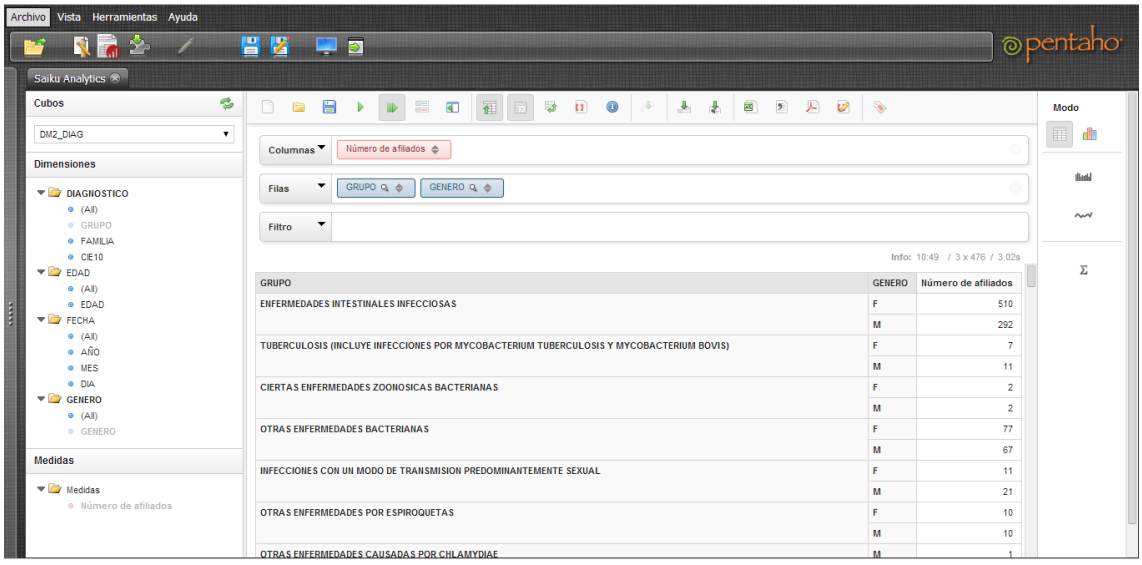

FIGURA A.1. Interfaz gráfica de Saiku para generar reportes

Los reportes generados fueron revisados por la analista de riesgo [1] y la epidemióloga [2]. Se encontró que los reportes proporcionados por la bodega de datos médicos fueron de interés y permitieron sacar estadísticas y conclusiones sobre el comportamiento de los pacientes de DM2.

[1] Liz Garavito, Directora ejecutiva de Processum LTDA e investigadora en el área de seguridad social y economía de la salud

[2] Alexandra Castillo, Enfermera y Epidemióloga.

Fue posible realizar estadísticas de enfermedades o laboratorios que requieren control en los pacientes de DM2. Un ejemplo de este tipo de reporte pueden ser apreciados en las figuras A.2 y A.3.. La forma de realizar filtros de este estilo en Saiku se puede observar en la fugura A.4..

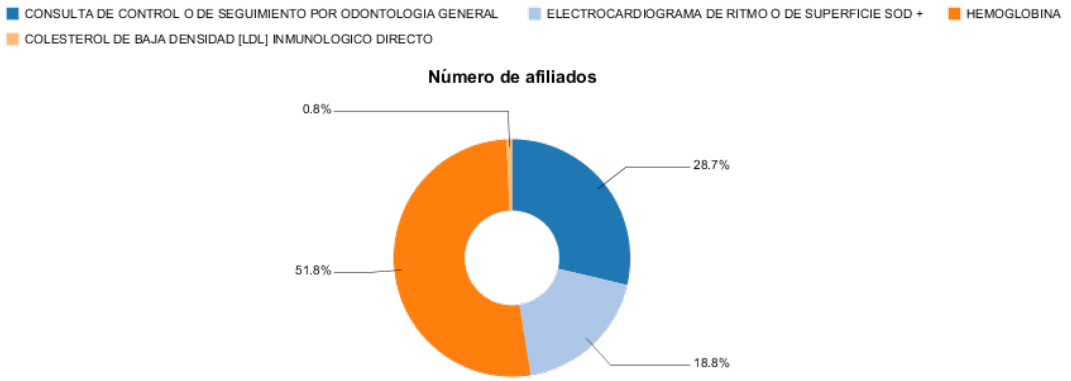

FIGURA A.2. Reporte del total de pacientes con CUPS de control de DM2

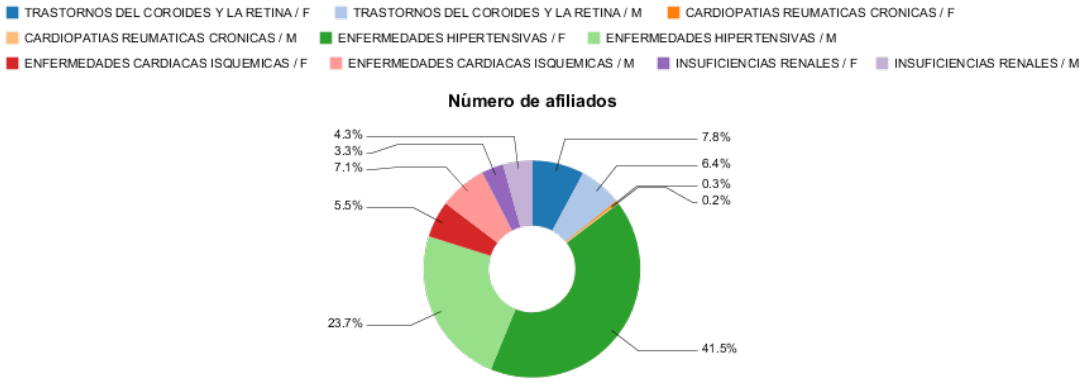

FIGURA A.3. Reporte de conteo de pacientes en grupos de enfermedades discriminado por género

En la figura A.3. se puede apreciar el número de pacientes que ha realizado sus controles a lo largo del periodo de análisis (2009 a 2012). Este tipo de reportes son útiles para realizar actividades de control de tratamiento y seguimiento a los pacientes de una enfermedad crónica. Es posible ver que la cobertura de controles de hemoglobina registrados para los pacientes de DM2 corresponde al 51 %, lo que indica que se desconoce el progreso de la enfermedad en un gran porcentaje de la población. En la figura A.5. se puede observar este reporte exportado en formato PDF y adicionalmente discriminado por género.

El reporte presentado en la figura A.4. muestra el número total de pacientes discriminado por género y grupo de enfermedad. Es posible ver que un más del 60 % de la población de DM2 padece de hipertensión y que la proporción de enfermedad renal se encuentra entre las más bajas. Este tipo de reportes pueden ayudar a entender la prevalencia de ciertos diagnósticos en la población de una manera rápida y visual.

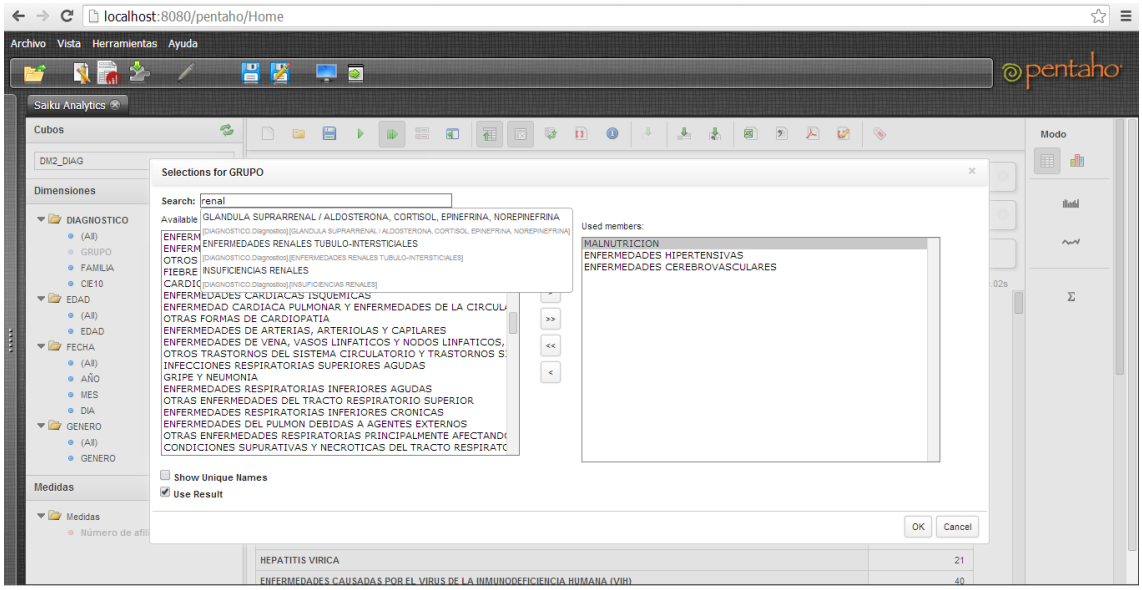

FIGURA A.4. Aplicación de filtros en Saiku para generar reportes específicos

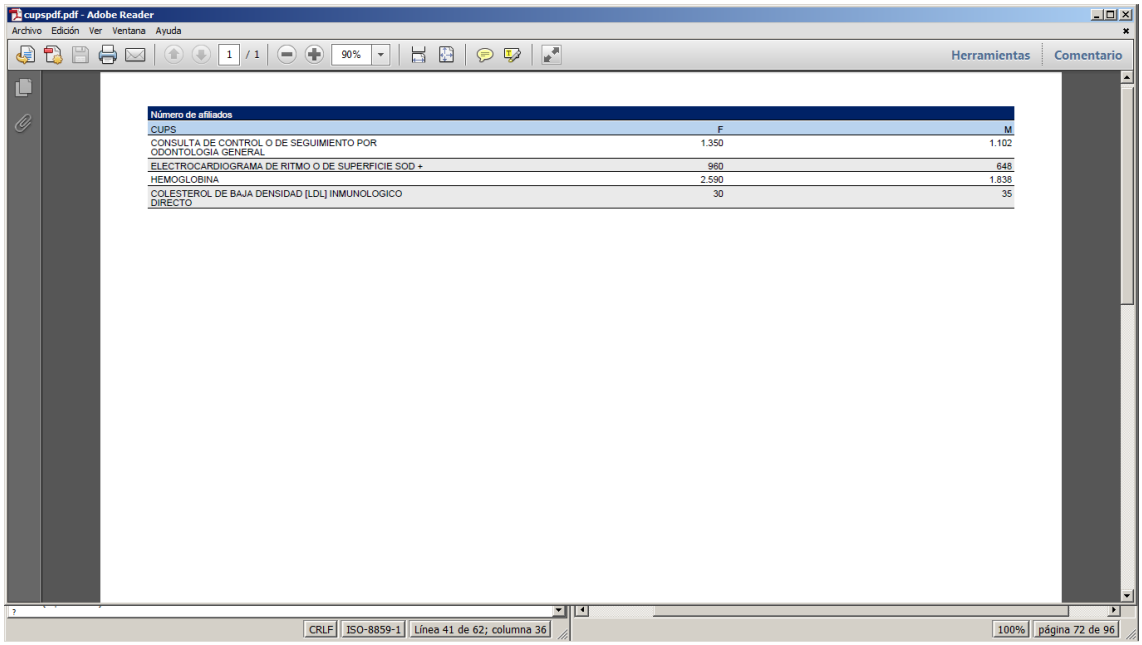

FIGURA A.5. Exportación en formato PDF del reporte del total de pacientes con CUPS de control de DM2 discriminado por género

APÉNDICE B

Artículo: "An approach to the risk analysis of diabetes mellitus type 2 in a health care provider entity of Colombia using business intelligence"

El articulo titulado "An approach to the risk analysis of diabetes mellitus type 2 in a health care provider entity of Colombia using business intelligence" fue presentado en la sexta versión de la Euro American Conference on Telematics and Information Systems (EATIS), en la ciudad de Valencia, España. Mayo de 2012.

En el artículo se abordó el uso de la inteligencia de negocios para centralizar datos médicos y poder aplicar fácilmente técnicas de minería de datos. Particularmente se utilizaron reglas de asociación para encontrar factores de riesgo de pacientes de DM2.

An Approach to the Risk Analysis of Diabetes Mellitus Type 2 in a Health Care Provider Entity of Colombia Using Business Intelligence

Una Aproximación al Análisis de Riesgo de Diabetes Mellitus Tipo 2 en una Entidad Prestadora de Salud de Colombia Usando Inteligencia de Negocios

Angela María Franco Pérez
Universidad Nacional de Colombia
Bogotá, Colombia
amfrancop@unal.edu.co

Elizabeth León Guzmán
Universidad Nacional de Colombia
Bogotá, Colombia
eleonguz@unal.edu.co

ABSTRACT

Business intelligence provides organizations with the ability to maintain a competitive advantage in the market. This property can be used in wide fields such as health care where lower costs and early prevention of patients are the main goals. This article develops an approach to the use of business intelligence to achieve a centralized clinical data and apply data mining, particularly the use of association rules. All this in order to find risk factors associated with diabetes mellitus type 2 (DM2) and the way healthcare providers perform the management of this disease. The experiment was conducted using a database of patients with DM2 treated by a health care provider entity in Colombia.

Keywords

Business Intelligence, risk analysis, diabetes mellitus type 2, data mining, association rules

ABSTRACT

La inteligencia de negocios brinda a las organizaciones la posibilidad de mantener una ventaja competitiva en el mercado. Esta propiedad puede ser utilizada en amplia medida en campos como el cuidado de la salud, siendo la disminución de costos y prevención oportuna de pacientes las principales metas. En este artículo se desarrolla una aproximación al uso de la inteligencia de negocios para lograr una centralización de datos clínicos, acompañada del uso de técnicas de mineria de datos, particularmente las reglas de asociación, con el objetivo de descubrir factores de riesgo relacionados con la Diabetes Mellitus Tipo 2 (DM2) y la forma en la que se realiza el manejo de esta enfermedad. La experimentación se realizó utilizando una base de datos de pacientes de DM2 tratados por una entidad prestadora de salud en Colombia.

EATIS 2012, Valencia, España

Keywords

Inteligencia de Negocios, análisis de riesgo, diabetes mellitus tipo 2, minería de datos, reglas de asociación

1. INTRODUCCIÓN

Hoy en día, las organizaciones, incluyendo las del cuidado de la salud, están adoptando el uso de la inteligencia de negocios para proporcionarles una ventaja competitiva en el mercado y así tener un mejor control sobre sus recursos y decisiones[15]. Esto para sacarle un mejor provecho a sus datos operacionales e implementar mejoras en sus negocios.

Para 1998 la implementación de un sistema de inteligencia de negocios movía 15 mil millones de dolares, creciendo de 10 % al 15 % anual, con un estimado de que el 95 % de las 1000 empresas americanas más grandes y diferentes gobiernos la usan para dar sentido a los grandes volúmenes de datos[6]. Para las organizaciones dedicadas al cuidado de la salud es primordial el tema de prevención e intervención oportuna de enfermedades, es por esto que toma importancia la identificación de factores de riesgo para poder hacer seguimiento sobre poblaciones expuestas. Gracias a que últimamente esta información se encuentra disponible en bases de datos médicas, es posible aplicar técnicas de análisis sobre ellas de una manera mucho más fácil haciendo uso de tecnologías de bases de datos, estadística y minería de datos.

Este artículo desarrolla una aproximación al uso de la inteligencia de negocios sobre datos clínicos. A partir de la centralización de dicha información se pueden realizar análisis de minería de datos para encontrar factores de riesgo para una enfermedad de alto costo en Colombia como la Diabetes Mellitus tipo 2. Así mismo, mediante el análisis de reglas de asociación también es posible hacer seguimiento a poblaciones y a la forma como las entidades prestadoras de salud realizan el manejo de esta enfermedad.

En el presente artículo, la sección 2 describe porqué es importante el uso de la inteligencia de negocios y la minería de datos en el análisis de riesgo de enfermedades como la Diabetes Mellitus 2 (DM2). La sección 3 recopila los trabajos más representativos de estas dos áreas enfocados a temas de salud. La sección 4 describe la naturaleza de los datos clínicos de una Entidad Prestadora de Salud en Colombia sobre los cuales se aplicará el proceso de inteligencia de negocios. La sección 5 presenta las generalidades del modelo

de bodega de datos clínicos propuesto que centralice toda la información de la que disponen dichas entidades. Finalmente, en la sección 6 se muestra un primer avance de los resultados en el análisis de los datos clínicos haciendo uso minería de datos, específicamente reglas de asociación.

2. MOTIVACIÓN

La diabetes mellitus tipo 2 (DM2) es una enfermedad crónica caracterizada por una hiperglucemia y trastornos en el metabolismo de las grasas, hidratos de carbono y proteínas de forma tal que genera defectos en la producción y acción de la insulina en el cuerpo. Esta enfermedad presenta complicaciones crónicas que deterioran la calidad de vida de los pacientes y aumentan significativamente el riesgo de muerte[5]. Los pacientes de diabetes utilizan con mucha más frecuencia servicios especializados como odontología, optometría y nutricionistas. Adicionalmente el costo derivado del tratamiento de la diabetes supone casi el doble de un paciente no diabético y además consumen entre 2 y 6 veces más recursos que los pacientes de edad y sexo similares con otras enfermedades crónicas[1].

En Colombia no existen muchos estudios que permitan conocer el comportamiento real de la enfermedad así como tampoco el del manejo de los factores de riesgo asociados a ella. Sin embargo, [18] provee datos de prevalencias de la enfermedad y señala el aumento de los factores de riesgo asociados a ella. No obstante, se puede evidenciar una preocupación nacional la cual se ve reflejada en el Plan Nacional de Saludo 2007-2010 en el cual se propone la disminución de enfermedades crónicas como la DM2 mediante el diagnóstico temprano, la prevención y el control, así como la identificación de la población en riesgo de desarrollar enfermedad renal crónica[4].

La correcta identificación de factores de riesgo asociados a una enfermedad crónica en Colombia como lo es la Diabetes Mellitus 2, puede ayudar a su oportuna prevención. Mediante la creación de niveles de riesgo usando estos factores, las Entidades de Salud pueden saber qué población requiere planes específicos de prevención. También al tener un control sobre la población sana, es posible disminuir la incidencia de la enfermedad y por ende se reducirían también los costos derivados del tratamiento de la enfermedad. Actualmente, en Colombia se trabaja con la aplicación de técnicas estadísticas sobre datos de pacientes clínicos para encontrar los factores de riesgo enfermedades de interés[14]. Cabe resaltar que en estudios relacionados con el tema se tratan áreas como la estadística descriptiva y métodos de regresión lineal y logística, de modo que se puede descubrir conocimiento sobre una población. Sin embargo, al no encontrarse muchas de las relaciones entre los datos, se hace necesario el uso de técnicas inteligentes para descubrir patrones ocultos entre ellos. La minería de datos se ha utilizado sobre datos clínicos de otros países para encontrar factores de riesgo asociados a enfermedades de interés [17]. Particularmente se ha trabajado con árboles de decisión y redes bayesianas como clasificadores de pacientes con enfermedades asociadas a distintos factores de riesgo [11][9], y reglas de asociación para encontrar las posibles relaciones entre dichos factores[3]. Sin embargo, es necesario centralizar los datos médicos y realizar las transformaciones necesarias sobre éstos para que no solo cumplan con una labor administrativa sino que la información que poseen pueda ser utilizada en análisis de riesgo en salud. De esta forma, será posible aplicar fácil y efectivamente los algoritmos de minería de datos que permitan identificar factores de riesgo en poblaciones y niveles de severidad de las enfermedades en los pacientes.

3. TRABAJOS PREVIOS

Son varios los trabajos relacionados con almacenamiento y procesamiento de la gran cantidad de información con la que cuentan actualmente las organizaciones. Esta sección describirá los trabajos más relevantes relacionados la forma en que las organizaciones dedicadas al área de la salud pueden obtener ventaja competitiva mediante el uso apropiado de mecanismos innovadores de almacenamiento y procesamiento adecuado de sus datos.

3.1 Bodegas de datos clínicas

En cuanto a bodegas de datos para salud se han desarrollado diferentes sistemas que soportan estas organizaciones. [16] describe los desarrollos más importantes que se han hecho en esta área. Entre los sistemas que proveen facilidades médicas (Clinical Computing, Oracle Clinical, SAS institute, MEDai, Informations Architects Inc, Shared Medical Systems, Quest informatics, Turku University Central Hospital, StanfordMedical Informatics). Cabe destacar a MEDai el cual usa inteligencia artificial y provee estudios predeterminados para diferentes enfermedades comunes. Al igual la universidad de Turky en Finlandia da un ejemplo de aplicación con una herramienta de transporte, la bodega y una herramienta propietaria de front-end y resalta las ventajas de la bodega para responder preguntas. Los otros sistemas se centran más en procesos de ayuda a la industria farmacéutica[16]. Con bodegas de datos clínicas se pueden hacer diferentes niveles de análisis: a nivel de paciente, por grupos, por ejemplo de enfermedades. También se pueden hacer investigaciones médicas o mejoramiento de la calidad del servicio. Así mismo se pueden hacer análisis financieros y demográficos permitiendo analizar la rentabilidad y calidad general de los servicios prestados. Para que esto se pueda realizar la bodega debería aceptar todos los tipos de datos, desde financieros, como contratos y facturas; datos demográficos como edad y sexo; clínicos, como diagnósticos y procedimientos; numéricos, como resultado de laboratorio e imágenes como rayos x. La combinación de todos estos datos es lo que hará posible el análisis cruzado y la obtención de información [16]. En el proceso de crear la bodega se han encontrado varios problemas como que los datos clínicos originales son guardados en un tipo de formato que no es el mejor para el análisis y la revisión de los datos. En 1999 PharmaHealth technologies introduce junto con un sistema de bodega de datos el concepto ambiente controlado encaminado a apoyar el proceso de transformación por medio de la captura de metadatos. Esto permite adicionar funcionalidad extra en el proceso de transformación [20]. Entre otras cosas se creó la definición de las columnas en las tablas de análisis de la bases de datos. Así el usuario al realizar la transformación puede asegurarse que las columnas concuerdan con los estándares definidos [20].

3.2 Minería de datos: Asociación

Dentro de la minería de datos se enmarcan varias tareas que bien son de tipo descriptivo o de tipo predictivo. Las descriptivas presentan el comportamiento y las relaciones existentes entre los datos, mientras que las de tipo predic-

tivo permiten suponer comportamientos a futuro basándose en los datos almacenados en las variables existentes[7]. Es así como la asociación se enmarca dentro de las tareas descriptivas ya que permiten al usuario entender sus datos mediante relaciones existentes entre ellos y puntos de separación de los datos.

3.2.1 Reglas de Asociación

Las relaciones ocultas entre los datos pueden ser representadas mediante las llamadas reglas de asociación. Éstas son implicaciones del tipo $X \rightarrow Y$, donde X y Y son conjuntos de ítems disyuntos. Dentro de las reglas de asociación se manejan dos medidas las cuales determinan qué tan fuerte es, estas son el soporte y la confianza.

El soporte de una regla se define como qué tan frecuente es dicha regla dado un conjunto de ítems, mientras que la confianza mide la frecuencia con la que los ítems del conjunto Y aparecen en las transacciones que contienen X. Estas dos medidas se utilizan con el fin de minimizar el hecho de encontrar patrones por casualidad. Para que la regla sea considerada fuerte, debe sobrepasar los parámetros establecidos de soporte y confianza determinados por el experto para el conjunto de datos.

Adicionalmente, el análisis de interpretación de las reglas de asociación debe hacerse cuidadosamente ya que algunas reglas que puedan carecer de sentido a simple vista, realmente puede que para el experto del conjunto de datos representen valor como un patrón desconocido e interesante[19].

3.2.2 Algoritmos

El algoritmo Apriori presentado en [2] es el más común para descubrir reglas de asociación y funciona bajo la base del principio "Apriori" el cual dice que si un conjunto de items es frecuente, entonces sus subconjuntos también lo son[19]. Este algoritmo calcula las frecuencias de cada item y toma los que superan el valor de soporte definido por el usuario. Luego genera nuevos conjuntos de ítems por cada iteración haciendo uso de los ya seleccionados, calcula sus frecuencias dentro del conjunto de items y selecciona nuevamente las que superan el umbral de soporte. Una vez generado el conjunto de items frecuentes, evalúa las reglas que pueden surgir de este conjunto y toma las que superan el umbral de confianza definido.

Cabe destacar que existe otra aproximación la cual es el algoritmo FP-Growth [19] el cual optimiza la obtención de conjuntos de items frecuentes mediante la codificación de las transacciones en una estructura llamada Árbol-FP (FP-Tree) donde cada una de ellas está representada en una rama del árbol y cada nodo contiene la frecuencia del item. El acceso a esta estructura hace que encontrar los items frecuentes sea mucho más rápido y así mismo la generación de las reglas de asociación.

3.2.3 Aplicaciones en el sector salud

Las reglas de asociación se encargan de encontrar conjuntos de patrones de riesgo para cierta condición clínica. Mediante las combinaciones de estos factores es posible obtener un subconjunto de la población que se encuentra en alto riesgo de contraer dicha condición clínica. Todo esto se hace basándose en que el grupo principal de estudio es el más pequeño, es decir el que se encuentra en estado anormal o enfermo. Así mismo también es importante ajustar las medidas de soporte y confianza para que el conjunto de reglas obtenidas tenga relevancia en el estudio[8]. Este método computacional ha sido aplicado en estudios de lactancia y como pruebas para conjuntos de datos médicos[13, 10]. Se ha encontrado que esta herramienta presenta muchas ventajas puesto que ayuda en la elaboración de planes de intervención para el grupo poblacional de alto riesgo. Sin embargo, se propone el uso de herramientas estadísticas para proveer mejores resultados en su implementación. Como un primer acercamiento a la detección de factores de riesgo interesantes, [3] discute el uso de software que soporte algoritmos de generación de reglas de asociación. Los experimentos realizados en este estudio se hacen sobre un conjunto de datos médicos previamente limpiados para su utilización en el software de minería de datos. Los resultados arrojaron un conjunto de reglas entre 2.000 y 20.000. Se discute también que se seleccionó un conjunto pequeño para mostrarle al usuario. Igualmente se propone una división por ventanas de tiempo para encontrar nuevas reglas en los conjuntos de datos. También se menciona en varios trabajos que el uso de este método debe ser controlado puesto que genera demasiadas reglas y es necesario tener un conjunto que sea realmente el óptimo de patrones de riesgo[13, 12]. En su trabajo, [13] advierte que las reglas de asociación deben contener patrones que generen alto riesgo relativo en lugar de aquellas que sólo cumplan con la condición de tener una alta confianza. De igual forma propone una modificación al algoritmo de Reglas de Asociación para que en lugar de tomar un valor de confianza tome un valor de riesgo relativo; también limita el número de reglas generadas y la forma de tratar el problema dependiendo de la población: enfermos o sanos. Al tomar como medida el riesgo relativo se asegura de mostrar a los usuarios el conjunto de reglas óptimo. Por último se propone una selección estratificada de las reglas de asociación encontradas para que el personal médico pueda evaluar de forma fácil y significativa las relaciones encontradas. Como un caso de estudio, [13] aplica esta técnica para la evaluación del efecto adverso de las drogas sobre un conjunto de pacientes. La meta del experimento era encontrar el grupo de pacientes que estaba en riesgo de sufrir de angiodema después de haber consumido cierto tipo de medicamento. Los resultados de este experimento fueron bastante buenos puesto que al hacerles un estudio con expertos se llegó a la conclusión de que las reglas generadas eran de su interés. Así mismo, el algoritmo se ejecutó de forma rápida y al ser evaluado se encontró que generaba reglas estadísticamente significantes para un conjunto de datos sesgado. Este tipo de trabajos proponen el uso de estos conjuntos de reglas y patrones para que las organizaciones médicas y los expertos puedan refinar hipótesis y hacer estudios de consumo.

4. NATURALEZA DE LOS DATOS CLÍNICOS

En las bases de datos de las Entidades Prestadoras de Salud (EPS) de Colombia es posible encontrar los siguientes tipos de información:

4.1 Información del Individuo en el Sistema General de Seguridad Social en Salud

Mediante la resolución 812 de 2007, el Ministerio de Protección Social estableció la información que deben remitir las instituciones prestadoras de salud en el país al administrador Fiduciario del Fosyga para la actualización de la

Base de Datos Única de Afiliados, BDUA. Todas las variables exigidas por el Ministerio se encuentran debidamente documentadas con sus tipos y tamaños estándar. La información exigida por este decreto contempla básicamente datos de tipo personal respecto al sistema de salud en el que se encuentra afiliado. Se destacan campos como la pertenencia étnica, tipo de cotizante, nivel de SISBEN, zona de afiliación, código del municipio de afiliación e información sobre el aportante y el cotizante.

4.2 Información de Servicios de Salud

Respecto a la parte de servicios de atención, se establecen las variables según la Resolucion 3374 de 2000 mediante la cual se establecen ciertos tipos de archivos y sus correspondientes variables. Estos archivos contienen, a groso modo, información sobre transacciones, usuarios de los servicios de salud, descripción agrupada de dichos servicios, consulta, procedimientos clínicos, hospitalizaciones, urgencias, recién nacidos, medicamentos y otros servicios.

Cabe aclarar que las bases de datos de las EPS con las que se tiene convenio para el uso de datos, manejan un archivo maestro en el cual se condensan los archivos de consulta, procedimientos, hospitalización, urgencias y medicamentos principalmente.

Respecto a los códigos diagnósticos, de medicamentos y procedimientos clínicos, las bases de datos de las EPS hacen uso de los siguientes estándares:

- Códigos de Diagnósticos: Se reportan usando el estándar internacional CIE10 (Clasificación Internacional de Enfermedades, décima versión) publicado por la Organización Mundial de la Salud.
- Códigos de Procedimientos Clínicos: son manejados usando el estándar colombiano. Este código se llama CUPS (Clasificación Única de Procedimientos en Salud) y se encuentra disponible en la resolución 1896 de 2001.
- Códigos de Medicamentos: Se manejan los códigos del Sistema de clasificacion química, terapéutica y anatómica, ATC (siglas en inglés).

4.3 Variables Ecosocioculturales

Estas variables son de tipo económico, social y cultural de la población de cada EPS. Estas variables están actualmente en proceso de recolección y depuración por lo que sólo algunas de ellas se encuentran lo suficientemente pobladas para realizar análisis con ellas. La figura 1 muestra un ejemplo del número de variables con registro en una de las EPS.

Las variables marcadas en Azul son las más pobladas para los pacientes de DM2. Sin embargo, existen variables como "Gestante" que, a pesar de tener muchos registros, deben ser depuradas y por eso no se marcan para ser utilizadas en los análisis de datos.

5. CENTRALIZACIÓN DE LA INFORMACIÓN

Con el fin de cumplir con el requerimiento de análisis de datos clínicos para modelos de riesgo en enfermedades de alto costo como la DM2, es necesario centralizar dichos datos. Es por esto, que la construcción de una bodega de datos clínicos que albergue no solo información administrativa de una EPS, sino también información relativa a riesgo en enfermedades, es imprescindible para lograr mejores resultados en prevención y atención por parte de la Entidad. A continuación se describe el modelo general propuesto para la bodega.

Figura 1: Variables Ecosocioculturales

Variable Ecosociocultural	Tipo Variable	Número de Pacientes con registro
Año inicio fumador	Numérico	67
Condición de Salud	Nominal	15
Código Dane de municipio Residencia	Nominal	20383
Código del departamento de residencia	Nominal	20383
Edad de la menarquía	Numérico	837
Edad de la menopausia	Numérico	10
Enfermedad con Antecedentes Familiares	Nominal	3
Escolaridad	Nominal	20235
Estado civil	Nominal	19170
Estatura	Numérico	11585
Fuma	Nominal	71
Gestante	Binomial	11789
IBC	Numérico	14267
Numero de Gestaciones	Numérico	2067
Número de cigarrillos al Día	Numérico	67
Ocupación Laboral	Nominal	16935
Peso	Numérico	12607
Tensión arterial diastólica	Numérico	15395
Tensión arterial sistólica	Numérico	15396
Uso de anticonceptivos en mujeres	Nominal	4
Zona de Residencia	Nominal	20383

5.1 Hechos y Granularidad

Los eventos que se han decidido modelar son los correspondientes al manejo de diagnósticos, manejo de procedimientos clínicos, entrega de medicamentos y estado ecosociocultural del paciente. La granularidad de los hechos "diagnósticos" y "procedimientos clínicos" está dada en días, mientras que el hecho "ecosociocultural" está dado en meses.

- Diagnósticos: Este hecho recopila todos los registros de CIE 10 de la población de la entidad. Mediante esta tabla de hechos se pueden distinguir los registros de incidencia epidemiológica en una enfermedad.
- Procedimientos Clínicos: En esta tabla de hechos se recopila toda la información referente a los servicios recibidos por la población de la entidad. Contiene variables como el costo y la forma de contratación del procedimiento.
- Ecosociocultural: Esta tabla de hechos maneja la información económica, social y cultural para cada mes de la población.

5.2 Dimensiones

- Afiliado: Esta dimensión contiene los datos personales de cada uno de los afiliados en la EPS. La información es de tipo cambiante debido a que cada mes llegan nuevos pacientes a la Entidad.
- Edad: En esta dimensión se especifican las edades de los pacientes. Aunque para varios análisis de riesgo se suelen usar rangos de edad, se ha decidido que esta dimensión no tenga ese tipo de jerarquía para dar más libertad de generación de rangos de edad propios de cada una de las poblaciones presentes en las EPS. Esta dimensión esta degenerada en las tablas de hecho.

- Sexo: Esta dimensión se encuentra degenerada en cada una de las tablas de hechos modeladas en la Bodega y corresponde al sexo del paciente.
- Fecha: Esta dimensión contiene una jerarquía de año, mes y día. Dependiendo del hecho a tratar se utilizan los diferentes niveles de jerarquía explicados anteriormente.
- Diagnóstico: En esta dimensión se encuentran todos los códigos CIE10 discriminados en 3 niveles de jerarquía. Estos niveles fueron escogidos haciendo uso de la clasificación por categorías establecidas por la OMS. Para efectos de representación en la bodega, la primera jerarquía comprende los 2 primeros dígitos del CIE10, la segunda jerarquía contiene los 3 primeros dígitos y finalmente, la tercera jerarquía contiene el código completo de la enfermedad.
- Mercado Geográfico: Esta dimensión contiene el lugar de residencia del paciente y maneja dos niveles de jerarquía: Departamento y Municipio, siendo éste último el más bajo. Esta dimensión es estática puesto que las locaciones fueron predefinidas usando los códigos DANE (para municipios y departamentos en Colombia).

5.3 Análisis de Datos

Haciendo uso de la bodega de datos clínica es posible tener de manera fácil y rápida reportes de indicadores importartes para el análisis de riesgo de enfermedades de alto costo, que no son necesariamente la DM2. El cruce de información permite tener listados rápidos y agregados de pacientes y sus condiciones clínicas, bien sea diagnósticos, procedimientos o sus condiciones ecosocioculturales.

Dentro de los reportes destacados y de interés para los analistas de riesgo en salud se encuentran los siguientes ejemplos:

- Listado de mujeres gestantes que presentan diagnóstico de VIH.
- Pacientes con Enfermedad Pulmonar Obstructiva Crónica (EPOC) que son fumadores posterior a su diagnóstico de EPOC.
- Número de intervenciones preventivas realizadas para enfermedades de alto costo en el mes, trimestre o semestre. Bien sean pruebas ELISA para VIH, espirometrías para EPOC, glicemias para DM2 o mamografías en el caso de Cáncer de Mama.
- Sumarización de los costos de cada paciente por servicios consumidos (Hospitalizaciones o procedimientos clínicos).
- Reporte histórico de los datos de presión arterial sistólica/diastólica, peso y estatura posteriores al diagnóstico de DM2 para los pacientes de esta enfermedad.
- Total de los costos de intervenciones preventivas o diagnósticas para el mes, trimestre o semestre.
- Listado de mujeres gestantes que tienen reporte de fumador.
- Conteo de personas asociadas a un diagnóstico por periodos.

Todos estos reportes obtenidos mediante la creación de cubos dimensionales para la visualización y consulta de la bodega de datos clínica, pueden ser utilizados en el análisis de manejo de enfermedades crónicas dentro de las EPS. También se pueden realizar seguimientos al cumplimiento de metas de cobertura de intervenciones realizadas por las EPS, así como ayuda en la generación de reportes para auditorías.

6. REGLAS DE ASOCIACIÓN A PARTIR DE LA BODEGA CLÍNICA

Una vez centralizados y depurados los datos clínicos, se procedió a realizar experimentos de reglas de asociación sobre el conjunto de pacientes de DM2 registrados hasta el momento en una EPS.

El criterio de selección de estos pacientes fue que contaran con un registro de CIE 10 comprendido entre los siguientes códigos E110 y E129, y también el E149, los cuales fueron acordados con la Entidad como los códigos de identificación de sus pacientes de DM2. El total de pacientes seleccionado fue de 21936.

Para realizar las pruebas de los algoritmos de asociación se escogieron 3 diferentes conjuntos de datos sobre una de las EPS, los cuales serán llamados A, B y C.

El primer conjunto de datos (A) reune distintas variables asociadas a los pacientes de DM2 como lo son: sexo, peso, diagnósticos registrados después de la incidencia en DM2, fecha del diagnóstico, edad al momento del diagnóstico, procedimientos clínicos realizados posterior a la incidencia en DM2, resultado del procedimiento, costo del procedimiento, fecha de realización del procedimiento, forma de contratación del procedimiento y departamento en el que vive el paciente.

El segundo conjunto de datos (B) contempla todos los diagnósticos de los pacientes de DM2 registrados posterior a la incidencia de la enfermedad, y finalmente, el tercer conjunto de datos (C) reune todos los procedimientos realizados a los pacientes de DM2 con fecha posterior a la incidencia de la enfermedad.

Para la generación de las reglas de asociación sobre los conjuntos de datos fueron utilizados los algoritmos Apriori (conjunto A) y FP-Growth (Conjuntos B y C). Debido a la cantidad de datos, se consideró un 0.2 como valor apropiado para el soporte y un 0.85 como el valor adecuado para la confianza. Estos valores fueron determinados debido a previas experimentaciones y además mostraron los resultados más interesantes.

6.1 Preprocesamiento de los datos

El preprocesamiento de los conjuntos de datos permitirá el uso adecuado de los algoritmos de agrupación. A continuación se describirá la forma en la que cada conjunto de datos fue tratado de acuerdo a sus propias variables.

6.1.1 Conjunto de Datos A

Los valores de campo numérico fueron discretizados por la técnica de binning simple donde el rango de valores numéricos es dividido en segmentos del mismo tamaño donde cada uno de estos segmentos representa un "bin". Los valores numéricos son asignados al bin representante del segmento que contiene el valor. Esto se realizó particularmente para las pruebas con el algoritmo Apriori donde se probaron diferentes números de bins para lograr obtener mejores reglas de asociación.

Dentro de estas variables discretizadas se encuentran:

- Edad: Para esta variable se realizaron varias pruebas de binning que incluían 4 y 5 bins. Es importante destacar que las edades pueden ser discretizadas utilizando el criterio epidemiológico de rangos de edades el cual es aplicado por las EPS en los análisis de riesgo. Estos rangos son: de 0 a 1 año, de 1 a 4 años, de 4 a 14 años, de 14 a 44 años, de 44 a 60 años y de 60 años en adelante.
- Peso: Esta variable también tuvo la prueba de binning de 4 y 5 bins para obtener resultados en las pruebas del algoritmo de asociación. Sin embargo, esta variable no presenta ningún tipo predefinido de discretización por lo cual es mucho más libre de manejar que otras variables.
- Costo Asociado del Procedimiento Clínico: La discretización de esta variable se realizó usando la técnica de binning para generar 4 rangos de costos.

Respecto a los datos de tipo nominal en su definición, pero recolectados de tipo numérico (zona de residencia, ocupación laboral, código DANE, código del departamento, escolaridad) fue necesario convertirlos a tipo nominal para poder aplicar algoritmos de asociación.

Cabe destacar que para estos afiliados existen también variables de atenciones como lo son los diagnósticos que han presentado después de su incidencia en la enfermedad, así como los procedimientos, sus resultados y su costo asociado. Tanto los códigos diagnósticos como los de procedimientos son de tipo nominal, mientras que los costos son de tipo numérico. El resultado del procedimiento puede ser de tipo numérico o categórico según sea el procedimiento clínico, por esto se decidió dejarlo como tipo nominal para las primeras pruebas del algoritmo de asociación.

6.1.2 *Conjuntos de Datos B y C*

Para estos dos conjuntos de datos se realizó un proceso de binarización tanto de los diagnósticos (conjunto B) como de los procedimientos clínicos (conjunto C) y posteriormente a que se tuviera la matriz binaria se aplicó el algoritmo FP-Growth.

Cabe resaltar que se realizaron dos preprocesamientos para el conjunto B de diagnósticos debido a que se decidió probar a dos diferentes niveles de detalle esta variable. Esto se pudo hacer ya que la variable diagnóstico presenta una jerarquía establecida desde su definición como código CIE10.

6.2 Resultados y Validación de Reglas de Asociación

Una vez generadas las reglas de asociación se procedió a analizar los resultados obtenidos y validarlos con la ayuda de un analista de riesgo en salud. Estos resultados se presentan a continuación para cada uno de los conjuntos de datos.

6.2.1 *Conjunto A*

Los resultados obtenidos en las reglas de asociación para este conjunto de datos permiten ver tendencias de las entidades aseguradoras en la atención de los pacientes diabéticos. Es posible inferir mediante estas reglas los periodos del año en los cuáles los pacientes toman los exámenes de control como lo es el mes de Julio y también es posible ver que las mujeres son las que más hacen uso de los servicios de atención de la EPS. Las reglas más interesantes se pueden ver en la tabla 1.

Tabla 1: Reglas de asociación Conjunto A

Antecedente	Consecuente	Confianza
Procedimiento = Hemoglobina Glicosilada	Fecha CUP = 07/2010	1
Sexo=F y Procedimiento =Hemoglobina Glicosilada	Departamento = BOGOTÁ	0.91
Fecha Dx=06/2010 y Procedimiento =Hemoglobina Glicosilada y Departamento = BOGOTÁ	Forma Contratación= EVENTO	0.98
Edad Dx= [52 - 65] y Sexo=F y Departamento = BOGOTÁ	Fecha CUP = 07/2010	0.98

Se puede ver en los resultados que las mujeres consultantes son de edad relativamente avanzada. Para los hombres no se encontró una relación entre la edad y el control de la enfermedad, lo que hace suponer que las mujeres entre los 52 y 65 años son quienes más se preocupan por realizar los controles. Las relaciones encontradas en este conjunto de datos son básicamente descriptivas del comportamiento de los pacientes y su demografía.

6.2.2 *Conjunto B*

Debido a que este conjunto contenía los diagnósticos posterior a la fecha de incidencia de la DM2, es posible encontrar entre los resultados muchas de las enfermedades asociadas a la DM2 las cuales se encuentran descritas en las guías clínicas existentes. Entre ellas están las enfermedades cardíacas y renales. Sin embargo, lo observado en los experimentos hechos a los dos niveles mostró una relación entre la DM2 y las enfermedades del sistema digestivo, particularmente con la gastritis. Esta relación resulta interesante puesto que en las guías clínicas no se evidencia una estrecha relación entre estas dos enfermedades. Adicionalmente, se puede inferir de esta relación que los pacientes de DM2 de la EPS pueden estar presentando reacciones adversas a los medicamentos suministrados para el tratamiento de otras enfermedades asociadas a DM2.

Mediante la experimentación a un nivel jerárquico alto fue posible encontrar más reglas que cuando se utilizó un nivel bajo de jerarquía. Las reglas más interesantes generadas para el nivel jerárquico alto se muestran en la tabla 2.

Tabla 2: Reglas de asociación Conjunto B (Jerarquía a 2 dígitos de CIE10)

Antecedente	Consecuente	Confianza
(K20-K31) Enfermedades del esófago, estómago y del duodeno	(E10-E14) Diabetes mellitus. y (I10-I15) Enfermedades hipertensivas	0.8
(J40-J47) Enfermedades respiratorias inferiores crónicas	(E10-E14) Diabetes mellitus.	1
(A00-A09) Enfermedades intestinales infecciosas	(E10-E14) Diabetes mellitus.	1
(E00-E07) Desórdenes de la glándula tiroides.	(E10-E14) Diabetes mellitus.	1

Sin embargo, los resultados obtenidos a nivel jerárquico bajo soportan en su mayoría lo obtenido en el nivel mayor haciendo posible detallar exactamente qué enfermedad es la que está generando la relación específica con la DM2. Algunas de las reglas más interesantes encontradas en esta jerarquía se presentan en la tabla 3.

Tabla 3: Reglas de asociación Conjunto B (Jerarquía a 3 dígitos de CIE10)

Antecedente	Consecuente	Confianza
(K29) Gastritis y duodenitis	(I10) Hipertensión arterial esencial (primaria)	0.8
(K29) Gastritis y duodenitis y (E03) Otros hipotiroidismos	(E14) Otras diabetes mellitus sin especificar y (I10) Hipertensión arterial esencial (primaria)	0.82
(M54) Dorsalgia	(E14) Otras diabetes mellitus sin especificar	0.91
(J44) Otras enfermedades pulmonares obstructivas crónicas	(E14) Otras diabetes mellitus sin especificar	0.92

Se puede observar que la relación entre la Enfermedad Pulmonar Obstructiva Crónica (EPOC) y la DM2 es nueva y fuerte. Las guías de manejo de enfermedad de DM2 para Colombia no se encuentra esta relación, así como tampoco la relación con las enfermedades de la glándula tiroides.

6.2.3 Conjunto C

Este conjunto permitió ver las relaciones entre los procedimientos clínicos que se les suministran a los pacientes de DM2. Principalmente se pueden encontrar relaciones que ayudan a validar si la entidad aseguradora está ofreciendo los tratamientos adecuados a dichos pacientes. Adicionalmente se pudo distinguir una relación no convencional entre los procedimientos clínicos "Hemoglobina Glicosilada" la cual es de control para los pacientes diabéticos y "Radiografía de Tórax" la cual no se encuentra entre los procedimientos de rutina que deben ser aplicados a los pacientes. Esta relación podría ser analizada más a fondo por la EPS para encontrar la razón de por qué sus pacientes están recibiendo este procedimiento. La tabla 4 presenta las reglas más representativas del experimento.

Tabla 4: Reglas de asociación Conjunto C (CUPS)

Antecedente	Consecuente	Confianza
Radiografía de tórax y Potasio	Sodio	0.97
Electrocardiograma de ritmo o de superficie sod y Consulta externa de complejidad mediana sod	Hemoglobina glicosilada por anticuerpos monoclonales	0.85
Creatinina en suero, orina u otros y Sodio	Potasio y Nitrógeno uréico [bun]	0.83
Microalbuminuria por nefelometría	Hemoglobina glicosilada por anticuerpos monoclonales	0.81

En los resultados también se puede ver que los pacientes efectivamente están recibiendo los controles y tratamientos más relevantes para los pacientes de DM2 que presentan comorbilidades. Estos son los Ecocardiogramas, los controles de alteración renal (Microalbuminuria, creatinina y nitrógeno uréico), así como los tratamientos con sodio y potasio.

7. CONCLUSIONES

La centralización y depuración de los datos clínicos mediante una bodega de datos facilita ampliamente su utilización en diversos análisis de riesgo para enfermedades de alto costo. Particularmente las técnicas de minería de datos permiten ayudar en la labor de identificación de factores de riesgo sobre su población para una determinada enfermedad, como lo es la DM2. La mayoría de los estudios hechos sugieren que los factores de riesgo que se presentan son de fácil control por parte de las organizaciones prestadoras de salud. De modo que las actividades de prevención serían bastante efectivas y ayudarían a disminuir los costos derivados de tratamientos por padecimiento de la enfermedad[14].

La construcción de la bodega de datos clínica permitió extraer de forma más sencilla conjuntos de datos para ser usados en análisis de minería de datos. Debido a que ya se encontraban limpios y organizados, los preprocesamientos correspondientes a la generación de reglas de asociación fue fácil y rápido. También se puede observar que la centralización de datos favorece la generación de indicadores de riesgo básicos como lo son totales de poblaciones, proporciones, promedios y listados de los mismos.

Los experimentos con reglas de asociación llevadas a cabo en el desarrollo de este artículo, sugieren que es posible identificar factores propios para la EPS como la edad, el sexo o condiciones de salud que afectan el desarrollo de la DM2 y hace a ciertas poblaciones más vulnerables a tener bajos índices de controles y tratamiento adecuado de la enfermedad. También es posible, a partir de estas técnicas, proveer a la EPS una herramienta de control de sus procesos de prevención y atención de la enfermedad mediante las relaciones establecidas por los consumos realizados por sus pacientes de DM2. El descubrimiento de nuevos factores de riesgo como EPOC, enfermedades respiratorias y de la glándula tiroides pueden ayudar en la identificación de nueva población de riesgo para realizar intervenciones preventivas de la enfermedad en la EPS.

Finalmente, se propone la continuidad de exploración de las diversas técnicas de minería de datos sobre una bodega de datos clínica, particularmente la clasificación y agrupación con el fin de ayudar a identificar oportunamente pacientes con complicaciones derivadas de la DM2.

8. AGRADECIMIENTOS

Este trabajo ha sido apoyado por Processum LTDA. y la Universidad Nacional de Colombia.

9. BIBLIOGRAFÍA

[1] ADA. Economic costs of diabetes in the u.s. in 2002. *Diabetes Care*, 26(3):917–932, 2003.

[2] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *SIGMOD '93: Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, pages 207–216, New York, NY, USA, 1993. ACM.

[3] Stephen E Brossette, Alan P Sprague, J Michael Hardin, Ken B Waites, Warren T Jones, and Stephen A Moser. Association rules and data mining in hospital infection control and public health surveillance. *Journal of the American Medical Informatics Association*, 5(4):373–381, 1998.

[4] Ministerio de la Protección Social. Plan nacional de salud pública 2007 - 2010. decreto 3039 de 2007, 2007.

[5] Ministerio de Sanidad y consumo. Guía de práctica clínica sobre diabetes tipo 2. Servicio central de publicaciones del gobierno vasco. Vitoria, 2008. Guía de Práctica Clínica en el SNS.

[6] Environmental Systems Research Institute (ESRI). Spatial data warehousing for hospital organizations, March 1998.

[7] Usama Fayyad, Gregory Piatetsky-shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17:37–54, 1996.

[8] Liqiang Geng and Howard J. Hamilton. Interestingness measures for data mining: A survey. *ACM Comput. Surv.*, 38(3):9, 2006.

[9] O. Gevaert, F. De Smet, E. Kirk, B. Van Calster, T. Bourne, S. Van Huffel, Y. Moreau, D. Timmerman, B. De Moor, and G. Condous. Predicting the outcome of pregnancies of unknown location: Bayesian networks with expert prior information compared to logistic regression. *Human Reproduction*, 21(7):1824–1831, 2006.

[10] Hongxing He, Huidong Jin, Jie Chen, Damien McAullay, Jiuyong Li, and Tony Fallon. Analysis of breast feeding data using data mining methods. In *AusDM '06: Proceedings of the fifth Australasian conference on Data mining and analystics*, pages 47–52, Darlinghurst, Australia, Australia, 2006. Australian Computer Society, Inc.

[11] Panagiota Kitsantas, Myles Hollander, and Lei Li. Using classification trees to assess low birth weight outcomes. *Artif. Intell. Med.*, 38(3):275–289, 2006.

[12] Jiuyong Li, Ada Wai-chee Fu, and Paul Fahey. Efficient discovery of risk patterns in medical data. *Artif. Intell. Med.*, 45(1):77–89, 2009.

[13] Jiuyong Li, Ada Wai-chee Fu, Hongxing He, Jie Chen, Huidong Jin, Damien McAullay, Graham Williams, Ross Sparks, and Chris Kelman. Mining risk patterns in medical data. In *KDD '05: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 770–775, New York, NY, USA, 2005. ACM.

[14] Diana Paola Córdoba R. Martha Díaz Perilla. Identificación de factores de riesgo de enfermedad cardiovascular presentes en los pacientes que ingresan al hospital san ignacio. *LECTURAS SOBRE NUTRICIÓN*, 10:51–58, 2003.

[15] Zeljko Panian. Return on investment for business intelligence. In *MCBE'07: Proceedings of the 8th Conference on 8th WSEAS Int. Conference on Mathematics and Computers in Business and Economics*, pages 205–210, Stevens Point, Wisconsin, USA, 2007. World Scientific and Engineering Academy and Society (WSEAS).

[16] Torben Bach Pedersen and Christian S. Jensen. Clinical data warehousing- a survey, 1998.

[17] Jonathan C. Prather, David F. Lobach, Ph. D, Linda K. Goodwin, Ph. D, Joseph W. Hales, Ph. D, Marvin L. Hage, W. Edward Hammond, and Ph. D. Medical data mining: Knowledge discovery in a clinical data warehouse. In *In 1997 Annual Conference of the American Medical Informatics Association*, pages 101–105, 1997.

[18] Peñaloza E Eslava J Gómez LC Sánchez H Amaya JL Arenas R Botiva Y Rodríguez J, Ruiz F. Encuesta nacional de salud 2007. resultados nacionales.

[19] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2005.

[20] Peter Villiers. Clinical data warehouse functionality. In *The Clinical Data Warehouse*, 1998.

APÉNDICE C

Artículo: "Meta-classifier for Type 2 Diabetes Mellitus comorbidities in Colombia"

El artículo titulado "Meta-classifier for Type 2 Diabetes Mellitus comorbidities in Colombia" fue presentado en la IEEE 15th International Conference on e-Health Networking, Applications and Services (Healthcom 2013), en la ciudad de Lisboa, Portugal, en Octubre de 2013.

En el artículo se propuso un meta-clasificador para comorbilidades de la DM2 basado en inteligencia de negocios y técnicas de minería de datos. Este modelo de clasificación de comorbilidades consta de dos partes. La primera sección del meta clasificador está enfocada en predecir qué pacientes pueden llegar a sufrir una comorbilidad asociada a la DM2, mientras que la segunda sección pretende determinar qué tipo de comorbilidad puede ser adquirida (micro vascular o macro vascular).

2013 IEEE 15th International Conference on e-Health Networking, Applications and Services (Healthcom 2013)

Meta-classifier for Type 2 Diabetes Mellitus comorbidities in Colombia

Angela Franco
Systems and Industrial
Engineering Department
Universidad Nacional de Colombia
MIDAS research group
Email: amfrancop@unal.edu.co

Elizabeth León
Systems and Industrial
Engineering Department
Universidad Nacional de Colombia
MIDAS research group
Email: eleonguz@unal.edu.co

***Abstract*—This article presents a general meta classifier model for Type 2 Diabetes Mellitus (T2DM) comorbidities which is based on business intelligence and data mining techniques. The proposed meta classifier has two phases: i) the model predicts whether a patient can develop a comorbidity and ii)the model predicts which kind of comorbidity could be: micro or macro vascular. Experiments were carried out with a cohort of 14162 T2DM patients from 2009 to 2012. 3459 of them were comorbidity patients. Obtained results show an accuracy of 87% in the first phase of the meta-classifier and an accuracy of 68% in the second phase.**

I. Introduction

Healthcare organizations are using new technologies of business intelligence to improve the way they manage the information of a pool of patients as the correct use of these technologies can lead to a competitive advantage. Some examples of this are the organization of patient data into clinical data warehouses and the development of several healthcare models based on data mining techniques.

These models are called classifiers and are defined as a task of data mining. The goal of classifiers is to assign a category, also called class, to an object. Classifiers can perform as descriptive or predictive depending on a class label set to given or not, respectively [1]. Their construction is based on a learning algorithm which decides the best model. Classification techniques can be useful in healthcare risk systems to predict future diagnosis of chronic diseases such as Type 2 Diabetes Mellitus (T2DM).

Comorbidities are defined as chronic conditions associated with a primary disease, and this occurs commonly in patients diagnosed with diabetes. These conditions can have an impact on the patient's cost since they tend to generate a higher use of clinical resources, i.e. specialized supervision and hospital care.

Patients with diabetes can suffer from diabetes related conditions (micro and macro vascular diseases) or non-diabetes related comorbidities which have nothing to do with the natural development of the pathology (depression and musculoskeletal diseases) [2]

In Colombia, T2DM is among the top 10 diseases responsible for death and disability [3]. Hypertensive cardiomyopathy and ischemic cardiopathy, both related as macro vascular comorbidities of T2DM, have even higher mortality rates.

The annual cost of T2DM in Colombia is estimated in U.S. $2.708 million from the societal perspective, however from the government health entities the direct annual cost of health care was U.S. $911 million. The treatment of diabetes and complications related to comorbidities represents around 86% of the direct cost and 95% of the indirect cost[4].

That is why having an appropriate way to detect those diabetes patients that are more likely to develop a comorbidity, will help target possible new incidences and release prevention plans over that population. An early detection of comorbidities will reduce the cost of diabetes patients for a given healthcare organization.

In this paper, a new model to predict T2DM comorbidities in Colombia is proposed. The model is composed by two phases. First one will allow to predict whether a patient may or may not suffer T2DM related comorbidity. On the other hand, part two of meta-classifier will predict which type of comorbidity it can be.

This paper is organized as follows. Section 2 shows the previous work about classifiers applied in healthcare issues. Section 3 describes the general model of the proposed classifier. Data sources, variables and pre-processing are discussed in section 4. Finally, section 5 presents the results obtained with the classifiers and the validation made by an epidemiologist.

II. Previous Work

Type 2 Diabetes mellitus (T2DM) is a chronic disease characterized by hyperglycemia and disorders in the metabolism of fat, carbohydrates and proteins in a manner that produces defects in the production and action of insulin in the body. This disease presents chronic complications that deteriorate patients life quality and significantly increase the risk of death. In Colombia, it is clear that ensuring early detection and intervention for type 2 diabetes mellitus should be a priority.

Healthcare studies have been using data mining techniques to find risk factors about a disease or to discover rules that can lead to a condition or diagnose. Principal concepts and results of some techniques used in healthcare prediction are described below.

978-1-4673-5800-2/13/$26.00 ©2013 IEEE

A. Decision Tree Induction

A decision tree is a hierarchical structure of nodes and directed edges, where a node can be seen as a crafted question that when answered will lead to another node. The terminal node, also known as leaf node, is assigned a class label[1].

By using decision trees, it is possible to find rules which describe the patients of a given pathology. These rules are usually easy to understand for clinical analysts and work well with discrete or continuous data[5].

A C4.5 tree was used in [6] to predict what method of feeding will women choose at 3 month postpartum. The classification accuracy reached over 75% by using properly selected feature subsets, demonstrating that the success of this technique depends on the careful choice of data inputs and their correct pre-processing. The knowledge discovered from the results with the decision tree may be used by doctors and nurses to estimate the risk of new mothers about type of feeding for their children. In the same study, the results of decision tree were compared against logistic regression and showed higher accuracy results over the statistical method.

In [7], decision trees and logistical regression techniques are used to predict when a woman will give birth to an underweight child. The results of classifiers show that regression techniques were superior over the decision trees. However, a cost sensitive analysis performed over the decision tree helped to improve the results of the classifier.

A data mining study in [8] over a diabetic data warehouse was performed to find out what patient of diabetes will have a bad HgbA1c values. The rules generated by the model are of interest for the medical community and their graphical representation makes it easier to understand the given relationships.

B. Bayesian Modeling

This classifier uses the well-known Bayes theorem and builds a probabilistic directed acyclic graphical model. Each edge represents a dependency between nodes as they do for causes and effects. For example, they can be seen as symptoms and diseases. These networks have been built with the help of doctors and experts in epidemiology[5].

This technique has been used in [9] to predict pneumonia and the result was accurate in classifying the patients with this condition. Even if the Bayesian models require more time to be built, their outcomes are better than statistical models such as regression [10].

The metodology used in the mentioned studies with bayesian modeling is described as follows. First, the data is split into validation and training data sets by using stratified sampling. The network is built by using heuristic or algoritms whose optimize measures. Once the network is built, the conditional probabilities are assigned to the variables. The model generated by the bayesian network on the training data set is tested on the validation data. Finally, the results are measured via ROC area to compare the performance of bayesian modeling over logistic regression.

III. General Model of T2DM Comorbidities Detection

A general model of a classifier for T2DM Comorbidities is proposed as shown in figure 1.

In a risk management system it is important to give useful and timely information to healthcare analysts in order to control costs increases in a group of patients. The comorbidities meta-classifier is embed into the general model to help with that task, as is shown in figure 1.

In the proposed model, the data sources comes from Healthcare Organizations and then are centralized in a clinical data warehouse following the design proposed in[11].

Once the information is organized and cleaned, then the T2DM patient data is sent to the classifier.

Figure 2 details the general model of T2DM comorbidities detection.

The first phase of the classifier, part A, will be able to determine potential T2DM comorbidity patients. After they are detected, the second phase, part B, of the classifier will help predicting which class of it can appear.

Finally, the output wil be the list of patients with risk of having T2DM comorbidities.

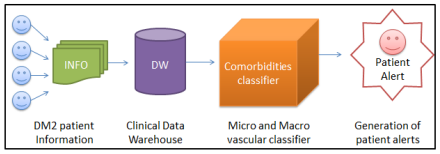

Fig. 1. General Model of Risk Analysis

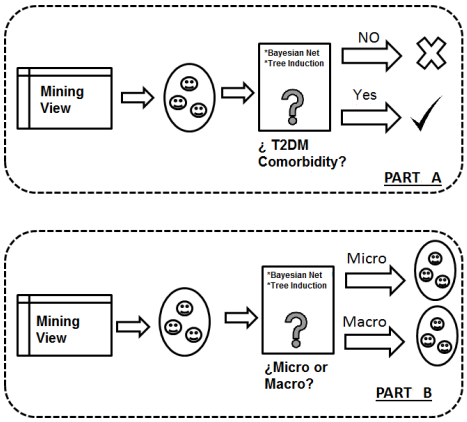

Fig. 2. Detailed Model of Comorbidities Meta-Classifier

IV. Model of Comorbidities detection in Colombia

A. Data sources

The group of patients consists of 14162 patients from which 3459 have comorbidity. The information of the patients was provided by Processum LTDA, and the cohort of patients goes from 2009 to 2012 The variables used in this work were chosen according to the normal way doctors identify comorbidity in a T2DM patient. These are the diagnoses codes (ICD-10) and disease evolution time. Social variables of patients such as age and gender were also included in the analysis. ICD-10 Codes used to identify patients with comorbidities are shown in Table I.

TABLE I. ICD-10 Codes for Comorbidities

Comorbidity	ICD-10 Codes
Micro vascular	E115, E125, I250, I251, I610, I611, I612, I613, I614, I615, I616, I618, I619, I738, I739
Macro vascular	E112, E122, E114, E124, G632, G590, E113, E123, H350, H360, H280, H350, H352, H540, H541, H544

The diagnoses and social variables were taken from years 2009 to 2012, and particularly, the date of diagnoses were taken after the registered incidence of T2DM and before the incidence of either macro or micro vascular comorbidity. That way, it is possible to have the clinical behavior of a patient just right before the incidence of a comorbidity.

B. Mining Views

The T2DM patients data was obtained from the clinical data warehouse and were divided as follows:

- Mining view A: Total patients of T2DM with all variables mentioned before. In this data set, the class corresponds to "Comorbidity" and "No Comorbidity". No comorbidity class represents 83.29 % meanwhile only the 16.71 % of data is labeled as Comorbidity Class.
- Mining view B: Total patients of T2DM with comorbidities with all variables mentioned before. The class to this data set corresponds to "Micro vascular" and "Macro vascular"

By dividing the data into these groups, it will be possible to test classifiers to first predict if the patient will have the comorbidity, and after that telling what kind of comorbidity they are most likely to develop.

C. Pre-processing

The ICD-10 codes were taken to a higher hierarchy by selecting only the first 3 digits of the codes. This was made in order to reduce dimensionality.

Additionally, ICD-10 codes Z00 and Z01 were removed because they refer only to people encountering health services for examination and investigation and not to a disease itself. Codes from E11 to E12 and E14 were deleted because they are used to diagnose T2DM.

Outlier detection and missing values management were not necessary for the mining views. Both views were obtained from the clinical data warehouse where the patients information was previously cleaned in the ETL process.

For experiments in mining view A, the classes were balanced using random sampling due to "No Comorbidity" class representing more than 80% of entries.

V. Experiments and discussion

This section will show the results of classifiers applied to each mining view. Experiments were performed over all mining views using bayes nets and tree induction. Since the classification problem of T2DM was intended to be unsupervised, the Bayesian modeling was performed without using a previous expert network.

The classifiers were measured using accuracy, ROC area and confusion matrix. A 10-fold stratified cross-validation was used on each classifier.

A. Mining view A: Can the patient develop a comorbidity?

This mining view was used to train classifiers in order to predict whether a patient can develop a comorbidity.

Since the classes were unbalanced, a sampling and a cost sensitive analysis were performed over the data set. The cost matrix used in the experiments is shown in table II.

TABLE II. Cost Matrix

	Pred. No Comorbidity	Pred. Comorbidity
Real No Comorbidity	0	1
Real Comorbidity	5	0

The following tables show the results obtained with classifiers.

TABLE III. Overall results for Mining view A Classifiers

Classifier	Accuracy	ROC Area
Bayesian Net (Unbalanced dataset)	81.42%	0.831
Bayesian Net (Class Balance by Sampling)	75.42%	0.828
Bayesian Net (Class Balance by Cost Sensitive analysis)	78.91%	0.832
Decision Tree (Unbalanced dataset)	90.42%	0.844
Decision Tree (Class Balance by Sampling)	87.27%	0.859
Decision Tree (Class Balance by Cost Sensitive analysis)	88.09%	0.842

TABLE IV. Confusion matrix for Bayesian Network (Unbalanced dataset)

	Pred. No Comorbidity	Pred. Comorbidity	Class Precision
Real No Comorbidity	10342	1492	90%
Real Comorbidity	1147	1226	45.1%
Class Recall	87.4%	51.7%	

TABLE V. Confusion Matrixfor Bayesian Network (Class Balance by Sampling)

	Pred. No Comorbidity	Pred. Comorbidity	Class Precision
Real No Comorbidity	3291	709	79.3%
Real Comorbidity	857	1516	68.1%
Class Recall	82.3%	63.9%	

TABLE VI. CONFUSION MATRIXFOR BAYESIAN NETWORK (CLASS BALANCE BY COST SENSITIVE ANALYSIS)

	Pred. No Comorbidity	Pred. Comorbidity	Class Precision
Real No Comorbidity	9565	2269	92.9%
Real Comorbidity	726	1647	42.1%
Class Recall	80.8%	69.4%	

TABLE VII. CONFUSION MATRIX FOR DECISION TREE (UNBALANCED DATASET)

	Pred. No Comorbidity	Pred. Comorbidity	Class Precision
Real No Comorbidity	11115	719	94.5%
Real Comorbidity	641	1732	70.7%
Class Recall	93.9%	73%	

TABLE VIII. CONFUSION MATRIX FOR DECISION TREE (CLASS BALANCE BY SAMPLING)

	Pred. No Comorbidity	Pred. Comorbidity	Class Precision
Real No Comorbidity	3637	363	89%
Real Comorbidity	448	1925	84.1%
Class Recall	90.9%	81.1%	

TABLE IX. CONFUSION MATRIX FOR DECISION TREE (CLASS BALANCE BY COST SENSITIVE ANALYSIS)

	Pred. No Comorbidity	Pred. Comorbidity	Class Precision
Real No Comorbidity	10577	1257	96.1%
Real Comorbidity	434	1939	60.7%
Class Recall	89.4%	81.2%	

According to results showed in tables III, IV, V, VI, VII, VIII and IX, the best classifier was decision tree using cost sensitive classification. The advantage of cost sensitive decision tree is that true positive predictions are given higher importance in the classification task by penalising the prediction of false positve patients.

In the decision tree generated, the diagnoses and other characteristics of patients become the branches of the decision tree. Using the visualization of decision tree, risk analysts or doctors can target the main risk factors directly related with the prediction of comorbidity for the T2DM.

The results obtained with the decision tree for this data set were highly interesting for the epidemiologist. The main reason for this interest is the fact that there are no such tools to help doctors or healthcare risk analysts to detect possible cases of comorbidities for T2DM.

The branches of the tree were analyzed by the epidemiologist and the most relevant conclusions are listed below:

- Gender and age of the patient is not decisive to predict comorbidities.
- Disease evolution time of T2DM and the diagnoses are the main factors that help to predict comorbidities.
- Patients developing comorbidities have other conditions such as anxiety disorders, diseases of the digestive and respiratory system and sexual dysfunction. These conditions are not common in patients with T2DM comorbidities.
- It was possible to detect interesting relationships between HIV and T2DM. Patients with HIV, T2DM and obesity diagnose are more likely to develop a comorbidity.
- Most of the related factors of comorbidities quoted in the literature can be found in the decision tree structure.

B. Mining view B: What comorbidity can the patient develop?

This mining view was used to trained classifiers in order to predict whether a patient targeted as possible comorbidity patient can develop macro vascular or micro vascular conditions.

The following tables show the results obtained with classifiers.

TABLE X. OVERALL RESULTS FOR MINING VIEW B CLASSIFIERS

Classifier	Accuracy	ROC Area
Bayes Network	69.81%	0.77
Decision Tree	67.75%	0.72

TABLE XI. CONFUSION MATRIX FOR BAYES NETWORK

	Pred. Macro vascular	Pred. Micro vascular	Class Precision
True Macro vascular	533	267	71.2%
True Micro vascular	216	584	68.6%
Class Recall	66.6%	73%	

TABLE XII. CONFUSION MATRIX FOR DECISION TREE INDUCTION

	Pred. Macro vascular	Pred. Micro vascular	Class Precision
True Macro vascular	552	248	67.3%
True Micro vascular	268	532	68.2%
Class Recall	69%	66.5%	

The best classifier for Mining view B is Bayes network. The precision of classes in Bayes Network is higher than decision tree.

Since this classifier works with means and probabilities of variables, it will not provide a graphic view of any rules over the attributes that contribute to predict comorbidity. In order to counteract the lack of a visual model from the bayesian model, a recovering of causal structure from Bayesian network was performed. The generated graph did not reveal important relationships among diagnoses and prediction of macro or micro vascular disease.

However, the model provided by decision tree allowed the epidemiologist to analyze the influence of some previous diagnoses on comorbidity prediction. The most relevant conclusions are listed below:

- Patients with throat and chest pain are more likely to develop macro vascular diseases.

- Ages above 65 are highly related to macro vascular disease.
- Patients with micro vascular disease have presence of alkalosis diagnosis.
- It is possible to find risk factors of micro vascular disease quoted in the literature. Some examples are chronic renal failure and retinal detachments and breaks, being the renal failure the most related to micro vascular in decision tree modeling.

VI. Conclusions and future work

The results on experiments conducted with the T2DM patients data illustrate the promise of new tools to prevent high costs diseases. In Colombia, it is a high priority goal to reduce incidence and costs related with T2DM comorbidities.

It was possible to compare two well-known classification techniques for healthcare issues on a T2DM comorbidity prediction problem. According to the results, both of those techniques were helpful to generate a comorbidity meta-classifier for T2DM patients.

Unbalanced classes are a well known issue in healthcare classification problems. However, sampling and cost sensitive analysis approaches can help to get better results to the classifiers. In the first phase of T2DM comorbidities prediction classifier, a cost sensitive approach was used in order to improve the results of correctly classify the positive class.

Decision tree worked better on first phase of meta-classifier, obtaining high precision on determining whether a patient can develop comorbidity. On the other hand, bayesian modeling obtained better results in the second phase of meta-classifiers by showing better precision than decision tree to predict the kind of comorbidity a patient can suffer.

The first part of the meta-classifier described by decision tree model was easy and intuitive to understand for epidemiologist. For the expert, it was possible to code up a series of questions and observations about the tree structure.

The analysis made by the epidemiologist suggested new and interesting diagnoses related to prediction of comorbidity. Some examples of them are an anxiety disorders, disorders of white blood cells and some diseases of the respiratory system.

To futher validate the work done with meta-classifiers for T2DM comorbidities, we have identified the next directions and tasks to keep working on:

- Improving the accuracy and precision of both classifiers (first and second phase) by adding information about clinical procedures related to T2DM disease management.
- Implement a software tool to visualize the list of patients with probability of develop comorbidities.

Acknowledgment

This work has been supported by Universidad Nacional de Colombia and Processum LTDA.

References

[1] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining, (First Edition)*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2005.

[2] S. F. W. G. V. d. B. G. Struijs J, Baan C, "Comorbidity in patients with diabetes mellitus: Impact on medical health care utilisation," *BMC Health Services Research*, 2006.

[3] R. G. J. Acosta Ramrez N, Pealoza RE, "Carga de enfermedad colombia 2005: Resultados alcanzados," Cendex-Pontificia Universidad Javeriana, Tech. Rep., 2008.

[4] J. C. Gonzlez, J. H. Walker, and T. R. Einarson, "Cost-of-illness study of type 2 diabetes mellitus in colombia," *Revista Panamericana de Salud Pblica*, vol. 26, pp. 55 – 63, 07 2009.

[5] P. Lucas, "Bayesian analysis, pattern analysis, and data mining in health care," *Current Opinion in Critical Care*, vol. 10, pp. 399–403, 2004.

[6] H. He, H. Jin, J. Chen, D. McAullay, J. Li, and T. Fallon, "Analysis of breast feeding data using data mining methods," in *AusDM '06: Proceedings of the fifth Australasian conference on Data mining and analystics*. Darlinghurst, Australia, Australia: Australian Computer Society, Inc., 2006, pp. 47–52.

[7] P. Kitsantas, M. Hollander, and L. Li, "Using classification trees to assess low birth weight outcomes," *Artif. Intell. Med.*, vol. 38, no. 3, pp. 275–289, 2006.

[8] J. L. Breault, C. R. Goodall, and P. J. Fos, "Data mining a diabetic data warehouse," *Artificial Intelligence in Medicine*, vol. 26, no. 1-2, pp. 37–54, 2002.

[9] D. Aronsky and P. J. Haug, "Diagnosing community-acquired pneumonia with a bayesian network," 1998.

[10] O. Gevaert, F. De Smet, E. Kirk, B. Van Calster, T. Bourne, S. Van Huffel, Y. Moreau, D. Timmerman, B. De Moor, and G. Condous, "Predicting the outcome of pregnancies of unknown location: Bayesian networks with expert prior information compared to logistic regression," *Human Reproduction*, vol. 21, no. 7, pp. 1824–1831, 2006.

[11] A. Perez and E. Guzman, "An approach to the risk analysis of diabetes mellitus type 2 in a health care provider entity of colombia using business intelligence," in *Telematics and Information Systems (EATIS), 2012 6th Euro American Conference on*, 2012, pp. 1–8.

Conclusiones y Trabajo Futuro

Las principales conclusiones obtenidas a partir de la realización de este trabajo son:

- En este trabajo se propuso el desarrollo de un modelo de análisis de riesgo de la Diabetes Mellitus Tipo 2 (DM2) en Colombia, basado en inteligencia de negocios y minería de datos. Su aplicación y validación fue realizada sobre un conjunto de datos reales y se contó con la ayuda de expertos clínicos para dar su concepto en relación a los resultados obtenidos con el modelo. El modelo general consta de dos partes: un modelo de datos clínicos y un modelo de minería de datos. En el modelo de datos clínicos se estudian las fuentes de datos y se propone una estructura de bodega de datos que pueda almacenar y limpiar datos clínicos. El modelo de minería de datos propone un análisis de riesgo para los pacientes de DM2 mediante la caracterización de los pacientes, descripción y predicción de comorbilidades.

- La generación de la bodega de datos permitió centralizar información clínica y limpiarla. Con la ayuda de personas expertas fue posible encontrar procesos de extracción, transformación y carga (ETL) que garantizaran un conjunto de datos limpio y consistente. La bodega de datos se modeló utilizando información diagnóstica, de procedimientos clínicos y de registros de variables socioculturales de los pacientes, ya que estos son los datos utilizados para la generación de guías de prevención para diferentes enfermedades.

- Los reportes generados por la bodega de datos permitieron realizar análisis exploratorios de manera fácil sobre los datos de los pacientes. Las estadísticas generadas por los cubos son de utilidad para expertos clínicos puesto que les facilita la extracción de indicadores sobre la población, como conteos y discriminaciones por sexo o ámbitos socioculturales.

- El proceso de generación de vistas minables para la experimentación requerida por el modelo de minería de datos fue facilitada por la bodega de datos. El comportamiento clínico de cada paciente pudo ser emulado gracias a estas vistas minables y el preprocesamiento que sufrieron. De este modo, cada paciente representa un registro donde los diagnósticos, procedimientos clínicos y variables socioculturales son condiciones que puede o no presentar.

- El modelo de caracterización de pacientes de DM2 se compuso de un proceso de agrupación y de la generación de reglas de asociación para describir los grupos de pacientes generados. En el caso de las reglas de asociación, se determinó con ayuda de las analistas de riesgo en salud que un soporte menor al 20 % no debe ser tenido

en cuenta a la hora de describir cualquier grupo identificado, puesto que los comportamientos tan específicos relacionados a un porcentaje bajo de pacientes no generan preocupación en un ente asegurador o de prevención puesto que son casos que un médico podría manejar sin problemas. Los algoritmos de agrupación usados fueron el algoritmo jerárquico aglomerativo (CHAMELEON) y el algoritmo particional K-Means.

- Los resultados obtenidos para la caracterización de pacientes con el algoritmo CHAMELEON fueron superiores a los resultados proporcionados por K-Means en cuanto a relevancia de resultado. En general, ambos presentan buenos resultados a nivel de medida de similitud intra-grupo.Sin embargo, la validación realizada por las analistas de riesgo en salud sobre la descripción de cada grupo generada con las reglas de asociación, concluye que los resultados obtenidos por CHAMELEON son muy consistentes con el desarrollo de la enfermedad y con ciertas condiciones de la población colombiana.

- El modelo de predicción de comorbilidades se componía de dos partes: la predicción de diagnóstico de comorbilidad y la predicción del tipo de comorbilidad. Se utilizaron dos clasificadores ampliamente usados en el campo médico: redes bayesianas y árboles de decisión. Ambos algoritmos tuvieron en general un buen desempeño como clasificadores para ambas tareas en cuanto a precisión de clases. Sin embargo, los modelos generados por el árbol de decisión para las dos partes del modelo de predicción resultaron ser de mayor entendimiento para las analistas de riesgo en salud. Utilizando el modelo visual de los árboles de decisión, se pudieron establecer comportamientos que desencadenan el diagnóstico de una comorbilidad o el desarrollo de un tipo de ésta en particular.

- La primera parte del clasificador de comorbilidades presentaba el problema de desbalanceo de clases. Por esta razón, técnicas como muestreo y matrices de costos fueron utilizadas para mejorar la precisión del clasificador. Esto permite que el clasificador tenga penalización en la detección de falsos negativos, los cuales deben ser mínimos debido al contexto médico.

- El modelo de caracterización de pacientes está enfocado a generar grupos de pacientes con características muy simalares que puedan ser útiles a entidades aseguradoras o de prevención en salud en desarrollos de planes de prevención y manejo de la enfermedad. Los perfiles de pacientes que se generan con este modelo pueden ayudar a entender niveles de adherencia a diversos tratamientos, establecer niveles de control de la enfermedad en grupos poblacionales de riesgo o vigilar el cumplimiento de las guías de manejo de enfermedad establecidas por los entes de salud.

- El modelo de predicción de comorbilidades ilustra la promesa del desarrollo de nuevas herramientas que permitan detectar grupos de pacientes con alta posibilidad de sufrir una comorbilidad asociada a la DM2 en Colombia. De acuerdo a la opinión de las analistas de riesgo en salud, no se encuentra en el momento una herramienta que permita identificar estos enfermos potenciales.

Para trabajos futuros, se proponen las siguientes actividades que pueden complementar el trabajo ya realizado:

- Complementar el modelo de datos con la integración de información referente a medicamentos para poder ampliar las posibilidades de análisis sobre los pacientes.

- Generar un prototipo de software que permita visualizar los pacientes pertenecientes a cada perfil encontrado en el modelo de caracterización, así como automatizar el paso del predictor de comorbilidades al predictor del tipo de comorbilidad obteniendo como resultado un listado de posibles pacientes que pueden sufrir una comorbilidad y de qué tipo.
- Realizar seguimiento de pacientes en el tiempo haciendo uso de técnicas de minería de datos que permitan analizar la evolución de variables a través del tiempo.
- Generar más modelos usando minería de datos que permitan analizar los riesgos de pacientes para diferentes enfermedades o circunstancias.

Bibliografía

[1] Rodríguez García J Acosta Ramírez N, Peñaloza RE, *Carga de enfermedad colombia 2005: Resultados alcanzados*, Tech. report, Cendex-Pontificia Universidad Javeriana, 2008.

[2] ADA, *Economic costs of diabetes in the u.s. in 2002*, Diabetes Care **26** (2003), no. 3, 917–932.

[3] Ruth Navarro-Artieda Antoni Sicras-Mainar, Alexandra Prados-Torres, *Opinión de los médicos de atención primaria sobre el uso de un sistema de ajuste de riesgos: Los adjusted clinical groups*, Revista Peruana de Medicina Experimental y Salud Pública **30** (2013).

[4] Ruth Navarro-Artieda Antonio Sicras-Mainar, *Los adjusted clinicals groups: Un sistema de clasificación de pacientes por ajuste de riesgos*, Revista Peruana de Medicina Experimental y Salud Publica **30** (2013).

[5] D. Aronsky and P. J. Haug, *Diagnosing community-acquired pneumonia with a bayesian network*, 1998.

[6] Tom Barber and Paul Stoellberger, *Saiku*, http://meteorite.bi/saiku, 2010.

[7] Ladjel Bellatreche, Kamalakar Karlapalem, and Mukesh Mohania, *Some issues in design of data warehousing systems*, 2002.

[8] Joseph L. Breault, Colin R. Goodall, and Peter J. Fos, *Data mining a diabetic data warehouse*, Artificial Intelligence in Medicine **26** (2002), no. 1-2, 37–54.

[9] Stephen E Brossette, Alan P Sprague, J Michael Hardin, Ken B Waites, Warren T Jones, and Stephen A Moser, *Association rules and data mining in hospital infection control and public health surveillance*, Journal of the American Medical Informatics Association **5** (1998), no. 4, 373–381.

[10] Claudia Georgeta Carstea, Ioan-Gheorghe Ratiu, Lucian Patrascu, and Nicoleta David, *New approach for business intelligence and optimization in practice*, CEA'10: Proceedings of the 4th WSEAS international conference on Computer engineering and applications (Stevens Point, Wisconsin, USA), World Scientific and Engineering Academy and Society (WSEAS), 2009, pp. 110–114.

[11] Surajit Chaudhuri and Umeshwar Dayal, *An overview of data warehousing and olap technology*, SIGMOD Record **26** (1997), no. 1, 65–74.

[12] Ming-Syan Chen, Jiawei Han, and Philip S. Yu, *Data mining: An overview from a database perspective*, IEEE Trans. on Knowl. and Data Eng. **8** (1996), no. 6, 866–883.

[13] Asociación Colombiana de Endocrinología, *Fascículos de endocrinología - fascículo diabetes*, ch. 4, 2011.

[14] Ministerio de Salud, *Resolucion numero 1896 de 2001*, 2001.

[15] Ministerio de Sanidad y consumo, *Guía de práctica clínica sobre diabetes tipo 2*, Servicio central de publicaciones del gobierno vasco. Vitoria, 2008, Guía de Práctica Clínica en el SNS.

[16] Carlo Dell'Aquila, Francesco Di Tria, Ezio Lefons, and Filippo Tangorra, *Business intelligence systems: a comparative analysis*, WSEAS Trans. Info. Sci. and App. **5** (2008), no. 5, 612–621.

[17] Usama Fayyad, Gregory Piatetsky-shapiro, and Padhraic Smyth, *From data mining to knowledge discovery in databases*, AI Magazine **17** (1996), 37–54.

[18] Ariel L. Garcia Gamboa, Miguel Gonzalez Mendoza, Rodolfo E. Ibarra Orozco, Jaime Mora Vargas, and Neil Hernandez Gress, *Hybrid fuzzy-sv clustering for heart disease identification*, Proceedings of the International Conference on Computational Inteligence for Modelling Control and Automation and International Conference on Intelligent Agents Web Technologies and International Commerce (Washington, DC, USA), IEEE Computer Society, 2006, pp. 121–.

[19] Miguel García, *Factores de riesgo: una nada inocente ambigüedad en el corazón de la medicina actual*, Aten Primaria **22** (1998), 585–595.

[20] Liqiang Geng and Howard J. Hamilton, *Interestingness measures for data mining: A survey*, ACM Comput. Surv. **38** (2006), no. 3, 9.

[21] O. Gevaert, F. De Smet, E. Kirk, B. Van Calster, T. Bourne, S. Van Huffel, Y. Moreau, D. Timmerman, B. De Moor, and G. Condous, *Predicting the outcome of pregnancies of unknown location: Bayesian networks with expert prior information compared to logistic regression*, Human Reproduction **21** (2006), no. 7, 1824–1831.

[22] Ronald J Gironda and Michael E Clark.

[23] Pinto Masis D-Gil Laverde JFA Rondón Sepúlveda M Díaz Granados N. Gómez-Restrepo C, Bohórquez A, *Prevalencia de depresión y factores asociados con ella en la población colombiana.*, Rev Panam Salud Publica. **16** (2004), 378–386.

[24] Diego Iván; GIRON VARGAS Sandra Lorena y ESPINOSA GARCIA Gladys. GOMEZ GUTIERREZ, Luis Fernando; LUCUMI CUESTA, *Conglomeración de factores de riesgo de comportamiento asociados a enfermedades crónicas en adultos jóvenes de dos localidades de bogotá, colombia: importancia de las diferencias de género.*, Rev. Esp. Salud Publica [online]. **78** (2004), 493–504.

[25] Juan Camilo González, John H. Walker, and Thomas R. Einarson, *Cost-of-illness study of type 2 diabetes mellitus in colombia*, Revista Panamericana de Salud Pública **26** (2009), 55 – 63 (en).

[26] Hongxing He, Huidong Jin, Jie Chen, Damien McAullay, Jiuyong Li, and Tony Fallon, *Analysis of breast feeding data using data mining methods*, AusDM '06: Proceedings of the fifth Australasian conference on Data mining and analystics (Darlinghurst, Australia, Australia), Australian Computer Society, Inc., 2006, pp. 47–52.

[27] Lemeshow S. Hosmer D, *Applied logistic regression*, John Wiley & Sons, 2000.

[28] A Helgeland I Hjermann P Leren PG Lund-Larsen LA Solberg I Holme, SC Enger and JP Strong, *Risk factors and raised atherosclerotic lesions in coronary and cerebral arteries. statistical analysis from the oslo study*, Arterioscler Thromb Vasc Biol **1** (1981), 250–256.

[29] George Karypis, *Cluto a clustering toolkit*, Tech. Report 02-017, University of Minnesota, Department of Computer Science, Minneapolis, MN 55455, November 2003.

[30] George Karypis, Eui-Hong (Sam) Han, and Vipin Kumar, *Chameleon: Hierarchical clustering using dynamic modeling*, Computer **32** (1999), no. 8, 68–75.

[31] Panagiota Kitsantas, Myles Hollander, and Lei Li, *Using classification trees to assess low birth weight outcomes*, Artif. Intell. Med. **38** (2006), no. 3, 275–289.

[32] E. León, O. Nasraoui, and J. Gómez, *Scalable evolutionary clustering algorithm with self adaptive genetic operators*, Evolutionary Computation (CEC), 2010 IEEE Congress on, july 2010, pp. 1 –8.

[33] Jiuyong Li, Ada Wai-chee Fu, and Paul Fahey, *Efficient discovery of risk patterns in medical data*, Artif. Intell. Med. **45** (2009), no. 1, 77–89.

[34] Jiuyong Li, Ada Wai-chee Fu, Hongxing He, Jie Chen, Huidong Jin, Damien McAullay, Graham Williams, Ross Sparks, and Chris Kelman, *Mining risk patterns in medical data*, KDD '05: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining (New York, NY, USA), ACM, 2005, pp. 770–775.

[35] Peter Lucas, *Bayesian analysis, pattern analysis, and data mining in health care*, Current Opinion in Critical Care **10** (2004), 399–403.

[36] Geoffrey Holmes Bernhard Pfahringer Peter Reutemann Ian H. Witten Mark Hall, Eibe Frank, *The weka data mining software: An update*, SIGKDD Explorations **11** (2009) (Versión 3.7).

[37] Diana Paola Córdoba R. Martha Díaz Perilla, *Identificación de factores de riesgo de enfermedad cardiovascular presentes en los pacientes que ingresan al hospital san ignacio*, LECTURAS SOBRE NUTRICIÓN **10** (2003), 51–58.

[38] Ingo Mierswa, Michael Wurst, Ralf Klinkenberg, Martin Scholz, and Timm Euler, *Yale: Rapid prototyping for complex data mining tasks*, KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (New York, NY, USA) (Lyle Ungar, Mark Craven, Dimitrios Gunopulos, and Tina Eliassi-Rad, eds.), ACM, August 2006, pp. 935–940.

[39] Beatriz Gracia Martha Liliana Cruz Andrés Felipe Sánchez Cecilia Aguilar de Plata Mildrey Mosquera, Alberto Pradilla, *Relacion entre los factores de riesgo de enfermedades crónicas no trasmisibles y la sensibilidad a la insulina en adultos jóvenes de*

18 a 39 años de la ciudad de cali-colombia, Sociedad Latinoamericana de Nutrición **57** (2007), 1.

[40] S.R. Newcomer, J.F. Steiner, and E.A. Bayliss, *Identifying subgroups of complex patients with cluster analysis.*, Am J Manag Care **17** (2011), no. 8, e324–32.

[41] Po S. Ngan, Man L. Wong, Wai Lam, Kwong S. Leung, and Jack C. Cheng, *Medical data mining using evolutionary computation*, Artificial Intelligence in Medicine **16** (1999), no. 1, 73–96.

[42] Zeljko Panian, *Return on investment for business intelligence*, MCBE'07: Proceedings of the 8th Conference on 8th WSEAS Int. Conference on Mathematics and Computers in Business and Economics (Stevens Point, Wisconsin, USA), World Scientific and Engineering Academy and Society (WSEAS), 2007, pp. 205–210.

[43] ———, *Expected progress in the field of business intelligence*, AIKED'09: Proceedings of the 8th WSEAS international conference on Artificial intelligence, knowledge engineering and data bases (Stevens Point, Wisconsin, USA), World Scientific and Engineering Academy and Society (WSEAS), 2009, pp. 170–175.

[44] Torben Bach Pedersen and Christian S. Jensen, *Clinical data warehousing- a survey*, VIII Mediterranean Conference on Medical and Biological Engineering and Computing (Medicon98), 1998.

[45] A.M.F. Perez and E.L. Guzman, *An approach to the risk analysis of diabetes mellitus type 2 in a health care provider entity of colombia using business intelligence*, Telematics and Information Systems (EATIS), 2012 6th Euro American Conference on, 2012, pp. 1–8.

[46] Jonathan C. Prather, David F. Lobach, Ph. D, Linda K. Goodwin, Ph. D, Joseph W. Hales, Ph. D, Marvin L. Hage, W. Edward Hammond, and Ph. D, *Medical data mining: Knowledge discovery in a clinical data warehouse*, In 1997 Annual Conference of the American Medical Informatics Association, 1997, pp. 101–105.

[47] Schellevis F Westert G Van den Bos G Struijs J, Baan C, *Comorbidity in patients with diabetes mellitus: Impact on medical health care utilisation*, BMC Health Services Research (2006).

[48] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, *Introduction to data mining, (first edition)*, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2005.

[49] Monica Chiarini Tremblay, Robert Fuller, Donald Berndt, and James Studnicki, *Doing more with more information: Changing healthcare planning with {OLAP} tools*, Decision Support Systems **43** (2007), no. 4, 1305 – 1320, ¡ce:title¿Special Issue Clusters¡/ce:title¿.

[50] Peter Villiers, *Clinical data warehouse functionality*, The Clinical Data Warehouse, 1998.

[51] Jie Wang, Yanwei Xing, Janxin Chen, and Yonghong Gao, *Discovering syndromes in coronary heart disease by cluster algorithm based on random neural network*, Bioinformatics and Biomedical Engineering , 2009. ICBBE 2009. 3rd International Conference on, june 2009, pp. 1 –4.

[52] R. Williams, L. Van Gaal, and C. Lucioni, *Assessing the impact of complications on the costs of type ii diabetes*, Diabetologia **45** (2002), no. 7, S13–S17 (English).