Aplicación de minería de datos a información de pacientes prediabéticos

Application of data mining information prediabetes

Henry Jesús Hernández Gómez
Universidad Juárez Autónoma de Tabasco
henryhernandez20@outlook.com

Resumen

El objetivo del presente trabajo fue obtener patrones de comportamiento de los expedientes clínicos de pacientes prediabéticos, utilizando técnicas de minería de datos, como apoyo a la toma de decisiones para el control de la diabetes. Para el logro de las metas trazadas se aplicó software de minería de datos Weka, que se caracteriza por tener las funciones necesarias que permitieron realizar transformaciones sobre los datos, así como tareas selección de atributos, clasificación, clustering para la extracción del conocimiento en las bases de datos. Las bases de datos sobre las que se trabajó son Adultos 20 años o más de la Encuesta Nacional de Salud y Nutrición (ENSANUT) 2012, Glucosa y Lípidos de ENSANUT 2006. La metodología de minería de datos que se utilizó fue CRISP-DM (Cross Industry Standard Processfor Data Mining).Según el problema propuesto y el objetivo planteado la investigación se encuadró bajo un enfoque cualitativo, así como en el método hermenéutico digital y etnográfico. El impacto de esta investigación se ve reflejado en la extracción de conocimiento en grandes almacenes de datos como son los expedientes clínicos para proveerdatos útiles que funcionen como apoyo a la toma de decisiones siendo estos la parte mas importante.

Palabras claves: Minería de datos, Extracción de conocimiento, Weka, Patrones, Comportamiento, Prediabéticos, Diabetes.

Abstract

The aim of this work was to obtain behavior patterns of the clinical records of prediabetes, using data mining techniques, to support decision making for the control of diabetes. To achieve the goals mining software Weka data, which is characterized by having the necessary functions that allowed perform transformations on the data as well as tasks feature selection, classification, clustering for knowledge extraction was applied on the basis of data. The databases on which we worked are adults 20 years or more from the National Survey of Health and Nutrition (ENSANUT) 2012 Glucose and Lipids ENSANUT 2006. The methodology of data mining was CRISP-DM was used (Cross Data Mining Industry Standard Processfor) .According to the proposed and objective investigation raised the issue framed a qualitative approach, as well as the digital and ethnographic hermeneutical method. The impact of this research is reflected in the knowledge extraction in large data stores such as clinical records for useful proveerdatos function as support for decision making these being the most important part.

Key words: Data mining, knowledge extraction, Weka, patterns, behavior, pre-diabetes, Diabetes.

Fecha recepción: Enero 2016 **Fecha aceptación:** Abril 2016

Introducción

En la actualidad las enfermedades no trasmisibles como la diabetes tiene gran prevalencia en México. Según datos de la Encuesta Nacional de Salud y Nutrición 2012 (ENSANUT) la razón de adultos con diagnostico médico previo de diabetes es de 9.2% en comparación con ENSANUT 2006 que era de 7%, lo que muestra que este tipo de padecimientos va en aumento (ENSANUT, 2012).

La Encuesta Nacional de Salud y Nutrición 2012 específica que la prevalencia mas alta de edades que presentan diabetes esta en un rango de 60 a 79 años, siendo mayor en mujeres con un 26.3% y 24.15% para los hombres. (ENSANUT 2012).

En busca de generar soluciones que sirvan a el Sistema Nacional de Salud como apoyo para toma de decisiones para mitigar y prevenir enfermedades, surge lo que es el Programa de Acción Especifica que tiene como objetivo promover y fomentar la generación de información y conocimiento relevantes respecto a tecnologías para la salud.

La Evaluación de Tecnologías en Salud es una actividad compleja, que comprende conocimientos y prácticas provenientes de diversas áreas a saber: investigación básica y aplicada, epidemiología, ingeniería etc. con el propósito fundamental de apoyar la toma de decisiones (Secretaria de Salud, 2012).

El Plan Nacional de Desarrollo 2013-2018 establece como línea de acción el Instrumentar mecanismos que permitan homologar la calidad técnica e interpersonal de los servicios de salud. Así como instrumentar acciones para la prevención y control del sobrepeso, obesidad y diabetes (Plan Nacional de Desarrollo, 2013).

Hoy en día, la Secretaría Salud (SS) tiene la necesidad de realizar la recopilación de información a través de sistemas de información y encuestas aplicadas a los pacientes, para evaluar las condiciones de salud en áreas básicas como las inmunizaciones, la atención a padecimientos crónicos, así como sobre los retos en salud como el control de la hipertensión arterial, el sobrepeso, la obesidad y la diabetes.

Buscando aportar soluciones tecnológicas que permitan evaluar y conocer los patrones de comportamiento que presentan los grandes conjuntos de datos de la secretaria de salud surgió la ideade aplicar técnicas de minería de datos a expedientes clínicos de pacientes prediabéticosque permitan extraer el conocimiento oculto en los datos, que sirva como apoyo a la toma de decisiones.

Contenidos

Materiales, métodos y metodologías

Las metodologías utilizadas para el desarrollo de esta investigación fueron las siguientes:Según el problema propuesto y el objetivo planteado, la investigación se encuadro bajo un enfoque cualitativo con un método hermenéutico digital y etnográfico.

Hernández, R., Fernández, C. y Baptista, P. (2003) menciona que un enfoque cualitativo se basa en la recolección de datos sin medición numérica para descubrir o afinar preguntas de investigación y puede o no probar hipótesis en su proceso de interpretación. De acuerdo a la definición la investigación es cualitativa, porque permitió conocer contextos y comportamientos predominantes a través de la descripción exacta de las acciones, objetos y procesos.

Para Arráez M. et al (2006) la hermenéutica es una disciplina que se dedica a interpretar y develar el sentido de los mensajes haciendo que su comprensión sea posible. La investigación se enmarco dentro del tipo hermenéutico porque determinadamente estuvo enfocado a la interpretación y comprensión de datos digitales, para descubrir contextos nuevos en los grandes volúmenes de información.
De la misma forma se inserto en el método etnográfico que según Galeano M. (2003) se concibe como la descripción, registro sistemático y análisis de un campo de la realidad social especifico, de una escena cultural, de patrones de interacción social. La investigación se enfatizó en este método por que se realizó la recolección de todo tipo de datos accesibles para poder mostrar conocimiento sobre el comportamiento de los datos de pacientes prediabéticos.

Para darle validez se realizó el proceso de **triangulación metodológica** utilizando hermenéutica digital y etnográfica. Para aportar confiabilidad al estudio se realizó la **triangulación de los de datos** obtenidos de manuales, médicos internistas y resultados de la base de datos.

Se determinaron los sujetos más adecuados para el estudio, usando el muestreo por segmentación.

El muestreo en investigaciones cualitativas se debe elegir la menor cantidad de sujetos que proporcionen la mayor cantidad de datos ricos para el estudio (Patton, 2002). De acuerdo a lo mencionado se tomo una muestra cualitativa de cuatro médicos internistas quienes fueron los colaboradores de la investigación.
El universo o población de estudio de esta investigación lo constituyen los pacientes prediabéticos. Las bases de datos contempladas para esta investigación fueron ENSANUT2006 y 2012.

La **metodología de minería de datos** utilizada fue CRISP-DM (*Cross Industry Standard Processfor Data Mining*, por sus siglas en ingles), por tener las características de flexibilidad y personalización, necesarias para obtener un proceso de minería de datos. Esta metodología de minería de datos está constituida por seis fases:

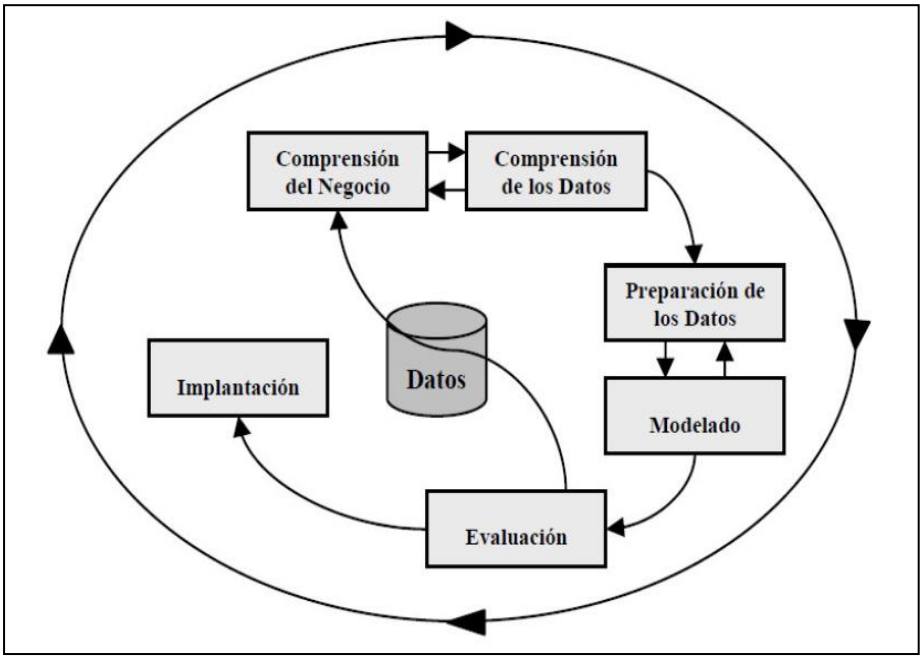

Ilustración 1. Metodología CRISP-DM
(Arancibia, 2010)

Las fases que conforman este modelo se describen a continuación:

- *Comprensión del negocio o análisis del problema:* en esta fase se realizó la identificación de las expectativas y requerimientos. Permitieron tener una visión general del problema y realizar un plan preliminar en el cual se pueda dar solución y alcanzar el objetivo planteado.
- *Comprensión de los datos:* Se realizó la recolección de los datos, lo cual permitió establecer un contacto inicial con el problema. Esta fase tiene como finalidad evitar problemas inesperados durante la fase de preparación de los datos.
- *Preparación de los datos:* En esta fase se llevó acabo todas las tareas (Selección, limpieza y transformación de los datos) con el objetivo de construir el conjunto final de datos que serán utilizados en la herramienta de minería de datos.
- *Modelado:* Se realiza la selección de las técnicas de minería de datos que mas se apropiaron y que en un final permitieron alcanzar el objetivo del proceso de minería de datos. También se realizaron generación de pruebas, evaluación del modelo y construcción del mismo.
- *Evaluación:* En la fase de evaluación, se evaluó el modelo teniendo en cuentas los criterios de éxito del problema. Aquí se realizó la documentación y presentación de los resultados de manera comprensible para lograr proporcionar el conocimiento obtenido.
- *Implementación:* Se trata de explotar la potencialidad de los modelos, intégralos en los proceso de toma de decisiones en la organización y difundir informes del conocimiento extraído.

Resultados

A continuación se describen cada hallazgo de acuerdo a el algoritmo utilizado en el proceso de minería de datos aplicado a las Bases de datos (Adultos, Glucosa y Lípidos) de la Encuesta Nacional de Salud y Nutrición 2006 y 2012 .

Algoritmo de clasificación (J48)

El comportamiento de los datos con el algoritmo de clasificación fue el siguiente: Se descubrió que las personas a las que algún médico les ha dicho que Si tienen diabetes o azúcar alta en la sangre, presentan

presión alta en un tiempo después del diagnostico. Aun que no es causa para el desarrollo de la patología, debido a que existen personas con diabetes y no presentan presión alta. Pero si puede considerarse como factor de riesgo para quienes presentan diabetes. Las reglas de clasificación halladas son las siguientes:

- Existen personas con diabetes desde hace 10 años pero desconocen si tienen presión alta.
- Hay personas que tienen presión alta desde hace 6 años, y llevan 10 años con diabetes.
- Quienes llevan 2 años con presión alta, tienen diabetes desde hace 10 años.
- Se descubrió que quienes asumen que tienen 10 años con presión alta, tiene 5 años con diabetes.
- Las personas que no tienen presión alta, si presentan diabetes desde hace 10 años.

De igual forma el algoritmo de clasificación fue aplicado a tres variables que almacenan algunos de los exámenes de rutina de un paciente con diabetes. Se encontró el promedio de veces que realizan dicho examen, a continuación se muestran.

- Se analizo la variable que representa el examen general de orina, las personas con diabetes se realizan esta prueba 13.5 veces en 12 meses.
- El automonitoreo es realizado 27.31 veces en 12 meses, lo que se puede decir que las personas con diabetes se realizan dos vez por mes automonitoreo.
- Por lo general el examen de Microalbuminuria es realizado en un promedio de 7.64 veces en 12 meses, lo que se puede decir que las personas con diabetes se realizan una prueba de control cada 2 meses.

Clustering (Agrupación- Simple Kmeans)

Base de datos adultos 20 años o más

De la aplicación del algoritmo de agrupación se obtuvieron grupos (Clúster). A continuación se describen algunas características que presentan los grupos:

- **Clúster 0.-** Es prevalecido por Mujeres en edad de 38 años, ningún médico le ha dicho que tiene presión alta, no tienen diabetes y por consecuente no presentado ningún padecimiento relacionado con la patología.
- **Clúster 1.-** Conformado en su mayoría por Mujeres en edad de 36 años, no presentan diabetes, tampoco presentan presión alta o colesterol.
- **Clúster 2.-** Al igual que el clúster 0 predominan las mujeres en edad de 38 años, estas han presentado colesterol alto, así como triglicéridos en limite establecido para considerarlo como normal.
- **Clúster 3.-** Se constituye por hombres en edad de 20 años, no presentan diabetes, tampoco padecen colesterol o triglicéridos.

Sobre las agrupaciones realizadas se puede expresar que en algunas mujeres en edad 38 años en adelante empiezan a presentar elementos que pueden llevar a detonar la diabetes. Mientras que los hombres en la misma edad no presentan ningún criterio que lo lleve a considerarlos en riesgo de desarrollo de la patología. De estas agrupaciones se eligieron aleatoriamente a 4 personas, 2 con diabetes y 2 sin la patología, a continuación se describen las características que presentan los pares de personas:

Clúster 1. Las mujeres en edad de 45 a 60 años son diabéticas, empiezan a manifestar otros tipos padecimientos como presión arterial alta, triglicéridos altos. Por otra parte quienes no son diabéticas el rango de edad esta entre 21 a 35 años muestran ciertas características como el colesterol alto. Estas particulares se pueden empezar a considerar como factores de riesgo que llevan a la detonación de la diabetes.

Clúster 2. Las Mujeres en edad de 29 años presentan diabetes, así como triglicéridos altos. Quienes no son diabéticos presentan dos elementos (Colesterol y triglicéridos altos) que pueda ser considerado como factor de riesgo en la presencia de la diabetes.

Clúster 3. Las características de los participantes elegidos que si presentan diabetes se encuentran en un rango de edad de 55 a 65 años, su presión alta es alta en uno de los casos, triglicéridos y colesterol están en el rango normal. Los participantes que representan a quienes no tiene diabetes tienen criterios como edad entre 44 y 69 años, presentan triglicéridos en estado normal, así como un nivel de colesterol normal.

Clúster 4. Los elegidos del clúster se caracterizaron por ser hombres en su totalidad, en primer instancia se describen a los que presentan diabetes estos tienen presión arterial, triglicérido y colesterol en un nivel alto . La edad de quienes poseen diabetes esta entre 50 a 66 años. Describiendo a quienes no presentan diabetes su edad esta entre 45 a 49 años, su colesterol esta en estado normal.

Clúster (Agrupación-Simple Kmeans)

Base de datos Glucosa y Lípidos

De la aplicación del algoritmo de agrupación se obtuvieron grupos (Clúster). A continuación se describen algunas características que presentan los grupos :

Clúster 0. Esta agrupación es prevalecida por hombres quienes no están confirmados como diabéticos, su colesterol esta en 150 mg/dL, la hemoglobina es de 13.3 mg/dL, el colesterol determinado por inmunizador es de 189.2706 mg/dL, la proteína C reactiva determinada mediante nefelómetro es de 1.8833, la glucosa es 105.5089, el Colesterol de alta densidad es de 37.2074 mg/dL. Aquí se encuentran el 52 % de los registros

Clúster 1. Conformado por mujeres no confirmadas como diabéticas, su colesterol esta en 150 mg/dL, la hemoglobina es de 13.3 mg/dL, el colesterol determinado por inmunizador es de 201.3676 mg/dL, la proteína C reactiva determinada mediante nefelómetro es de 4.5726, la glucosa es 106.7048 mg/dL, el colesterol de alta densidad es de 40.8583 mg/dL. En esta agrupación están el 48% de los datos.

Se realizó un proceso de visualización de las variables Confirmado Diabético y Colesterol, se obteniendo lo siguiente:

Las características encontradas dentro de los grupos son que quien no especifico que saber si tiene o no la patología sus niveles de triglicéridos oscilan en 128.21 mg/dl, el colesterol es de 179.63 mg/dl, la glucosa 115.69 mg/dl y su insulina es de 14.0 microU/ml. Quienes especificaron tener diabetes sus niveles de triglicéridos fluctúan en 83.59 mg/dl, el colesterol es de 257.64 mg/dl, la glucosa 378.89 mg/dl y su insulina

es de 6.0 microU/ml. Por otra parte las mujeres que en su embarazo fueron confirmadas con diabéticas cuentan con las siguientes características: triglicéridos oscilan en 104.43 mg/dl, el colesterol es de 220.44 mg/dl, la glucosa 263.10 mg/dl y su insulina es de 6.5 microU/ml.

Discusión

Es indudable que el uso de las TI actualmente se han incorporado en todos las áreas de la vida cotidiana, en especial en la medicina impactando significativamente en su uso. De forma que el uso de herramientas tecnológicas como la minería de datos permite conocer cuales personas están en riesgo de presentar algún padecimiento, así como disminuir los tiempos en el proceso de análisis de la información y apoyar la toma de decisiones es un aspecto fundamental para las instituciones. Para sector salud es necesario añadir las tecnologías que permitan extracción del conocimiento en áreas como (Epidemiología, Enfermedades crónicas como IRC, Accidentes, Ginecología, Prevención médica).

Conclusión

Cabe mencionar dos aspectos que intermedian en la conclusión de este trabajo: el primero es dar respuesta a la pregunta de investigación y el segundo es el cumplimiento del objetivo planteado.

Para concluir con el primer aspecto es dar la respuesta a la pregunta de investigación:

¿Qué técnicas de minería de datos, permitirán obtener los patrones de comportamiento en los grandes volúmenes de información de pacientes Pre-diabéticos?

Las técnicas que permitieron obtener patrones de comportamiento fueron los árboles de decisión específicamente la técnica J48 que están en la categoría de los algoritmos supervisados o predictivos, así como la técnica de clustering (agrupación) en la que se utilizó el algoritmo SimpleKmeans que esta en la categoría de los no supervisados o descubrimiento de conocimiento que permitió determinar las características por grupo o individuo.

El objetivo de esta investigación se cumple para el caso de estudio de detección de patrones de comportamiento de pacientes Pre-diabéticos y Diabéticos, esto porque permitió descubrir y conocer sus características como edad, sexo, niveles de colesterol, triglicéridos, glucosa e insulina representadas a través de los arboles de decisión y clusters mostrados por la herramienta de minería de datos.

Los patrones de comportamiento encontrados a través de la herramienta de minería de datos (Weka) permite llegar a la conclusión que el utilizar tecnologías dedicadas a la extracción del conocimiento es una solución de mucha utilidad para descubrir contextos ocultos en los grande volúmenes de información.

Se concluye que el modelo CRISP-DM fue de mucha ayuda debido a que permitió orientar los objetivos del plan del proyecto, ya que suministra un delineación de un ciclo de vida para minería de datos. Integrado por actividades y tareas propias de la evolución del proyecto.

Sin duda alguna la solución tecnológica presentada en este documento, puede funcionar de manera útil en la medicina, como apoyo en verificación de riesgos que presentan cierto grupo de personas con enfermedades no trasmisibles como la diabetes. También puede ser utilizado como punto de partida para tomar decisiones y formular estrategias de prevención y control rutinario de diabetes.

Bibliografía

Ángeles, M. I., & Santillán, A. M. (2004). Minería de datos: Concepto, características, estructura y aplicaciones. Recuperado de http://www.ejournal.unam.mx/rca/190/RCA19007.pdf

Arancibia, J. A. G. (2010). Metodología para el Desarrollo de Proyectos en Minería de Datos CRISP-DM. Recuperado de http://oldemarrodriguez.com/yahoo_site_admin/assets/docs/Documento_CRISP-DM.2385037.pdf

Arráez, M., Calles, J., Moreno, L. (2006). La Hermenéutica: Una actividad interpretativa. Recuperado de http://www.redalyc.org/pdf/410/41070212.pdf

BI Analytics. Inteligencia de negocios [BI Services]- Construcción de modelos de minería de datos [BIA DataMining]. Recuperado de http://bianalytics.biz/index.php/bi-analytics-inteligencia-de-negocios-bi-services/75-construccion-mineria-datos

Cervantes, A. M., López, V. G., &Gayosso, G. Y. (2010). Minería de Datos. Recuperado de http://www.ingenieria.buap.mx/DOCUMENTOS/REVISTA/REV_11/art_4.pdf

Césari, M. (2002). Minería de Datos, 1-42. Recuperado de http://ai.frm.utn.edu.ar/micesari/files/01_Matilde_Mineria_datos.pdf

Dueñas, M. X. (2009). Minería de datos espaciales en búsqueda de la verdadera información *. *Energy*, *13*(1), 137-156. Recuperado de http://redalyc.uaemex.mx/pdf/477/47711998007.pdf

Elmasri, R., &Navathe, S. B. (2005). *Sistemas de bases de datos (Conceptos fundamentales)*. (2° ed.) (A. WESLEY, Ed.)

ENSANUT (2012). Encuesta Nacional de Salud y Nutrición- Resultados Nacionales. Recuperado de http://ensanut.insp.mx/doctos/ENSANUT2012_Sint_Ejec-24oct.pdf

Galeano, M. E.(2004). Diseño de proyectos en la investigación cualitativa. Recuperado de http://books.google.com.mx/books/about/Dise%C3%B1o_de_proyectos _en_la_investigaci%C3%B3.html?hl=es&id=ufsZQkjMUFEC&redir_esc=y

Gobierno de la Republica (2013). Plan Nacional de Desarrollo 2013-2018. Recuperado de http://pnd.gob.mx

Hernández, R., Collado, C.F, & Baptista, L. (2003). Metodología de la Investigación. (3° ed.) McGraw-Hill.

Limite, C. (2012). Minería de Datos. RetrievedMay 22, 2012, Recuperado de http://ciclolimite.com/mineria-de-datos/