

Desarrollo de una aplicación Web para predecir la aparición de complicaciones en pacientes con diabetes tipo II

Sahir Omar Burciaga Sosa, Beatriz A. Olivares Zepahua, Giner Alor Hernández,
Ana María Chávez Trejo, Hilarión Muñoz Contreras
División de Estudios de Posgrado e Investigación,
Instituto Tecnológico de Orizaba,
Orizaba, Veracruz, México
sburciaga@acm.org, bolivares@ito-depi.edu.mx, galor@ito-depi.edu.mx,
achavezt@prodigy.net.mx, hilarionmunozcontreras@gmail.com

Resumen. Las personas con diabetes son propensas a desarrollar complicaciones médicas, especialmente si no tienen un control adecuado de su enfermedad. Dentro de las complicaciones crónicas más comunes en los diabéticos se encuentran la retinopatía diabética, la nefropatía diabética y el pie diabético. Estas complicaciones afectan la calidad de vida del paciente y, sin el tratamiento adecuado, pueden conducir a su muerte. En el presente proyecto se trabaja en el desarrollo de una aplicación Web que, utilizando Minería de Datos, realiza la predicción de la aparición de complicaciones en pacientes con Diabetes Mellitus Tipo II.

Palabras Clave: Minería de Datos, Diabetes, Clasificación, Predicción, WEKA.

1 Introducción

En años recientes la incidencia de la diabetes en la población mexicana aumentó de forma alarmante [1] [2] [3]. El padecimiento de esta enfermedad incrementa el riesgo de desarrollar complicaciones crónicas como la Nefropatía Diabética (ND), la Retinopatía Diabética (RD) y el Pie Diabético (PD). La ND es una de las complicaciones más temidas y es la responsable de la mayoría de los casos de insuficiencia renal crónica terminal; la RD representa una de las principales causas de ceguera y el PD es una de las principales causas de amputación. Estas complicaciones pueden presentarse de forma asintomática, es decir, el paciente no siente dolor, malestar o algún síntoma que advierta de la presencia de la enfermedad incluso en etapas avanzadas cuando el riesgo de tales complicaciones es mayor por lo que resulta indispensable obtener un diagnóstico oportuno.

Por otro lado, diversas técnicas de Minería de Datos (MD) se emplean exitosamente para el diagnóstico, identificación de riesgos y generación de tratamientos de diversas enfermedades [4] [5] [6] [7].

En este trabajo se presenta el desarrollo de una aplicación Web que, utilizando MD, realice la predicción de la aparición de complicaciones en pacientes con Diabetes Mellitus Tipo II (DMTII). El trabajo se enfoca en la aplicación de la tecnología de MD en el sector de la salud para ofrecer una solución computacional que ayude a los médicos a la identificación de la aparición de complicaciones en pacientes diabéticos.

2 Trabajos relacionados

Existen múltiples investigaciones donde se hace uso de técnicas de MD en el área de la salud. Por ejemplo en [5] se comparan diversas técnicas de MD para el diagnóstico de la enfermedad cáncer de mama. En [8] se emplea Minería de Textos para identificar factores de riesgo relacionados con la aparición de RD. En [4] [9] [10], [11], [12] se aplican diversas técnicas de MD para el diagnóstico de la diabetes pero en ellos no se aborda el tema de las complicaciones asociadas como la RD, PD o ND. En [13] y [14] se reporta la aplicación de MD a la predicción de condiciones específicas de hemodiálisis (necesidad de hospitalización, diálisis) sin relacionarla directamente con variables de DMTII. En [15] se predice la probabilidad de que los pacientes con DMTII padezcan una enfermedad cardíaca. En [16] se usa MD para predecir oportunamente la aparición de nefropatía diabética en pacientes con DMTII. Por otro lado, los estudios formales reportados en México se centran en el análisis estadístico de la presencia de la enfermedad [1] [2] [3] y las complicaciones que de ella derivan, como son la gestión de tratamientos de largo plazo, la necesidad de unidades médicas especializadas, los costos de hospitalización y el aumento de pensiones por incapacidad entre otros [17].

3 Minería de Datos

Mehmed Kantardzic define a la MD como “el proceso para encontrar información nueva, valiosa y no trivial en grandes volúmenes de datos” [18]. En otras palabras, la MD permite extraer conocimiento útil y desconocido que se encuentra oculto en los datos. La MD se aplica en una gran variedad de tipos de repositorios como son base de datos relacionales, almacenes de datos y archivos planos entre otros [19]. Las tareas principales de la MD son la clasificación/predicción y la descripción [19].

3.1 Clasificación y predicción

La clasificación y la predicción permiten analizar los datos y extraer modelos que ayudarán a predecir el comportamiento de nuevos datos. La principal diferencia entre estos conceptos radica en que la clasificación se utiliza cuando se desea obtener valores discretos y la predicción cuando se desea conocer valores continuos [19]; sin embargo

algunos autores, e incluso herramientas como las de Microsoft, usan solamente el término *predicción* sin importar el tipo específico de valor.

3.2 Modelo de Minería de Datos

Un modelo de MD se refiere no solamente a un algoritmo, si no a un conjunto de datos, estadísticas y patrones que se utilizan para realizar las tareas de MD como predicciones o clasificaciones [20]. Una estructura de MD contiene la información que representa el origen de los datos. El modelo de MD obtiene los datos a partir de esta estructura y posteriormente la analiza y procesa usando un algoritmo de MD. Una vez que se concluye este proceso, se almacenan los resultados, patrones, metadatos y enlaces de la estructura de MD en el modelo. De esta manera es posible construir diversos modelos con distintos algoritmos y estructuras de datos obteniendo como resultado un conjunto diferente de metadatos y patrones que permitirán realizar predicciones con diversos resultados [20].

3.3 Técnicas de Minería de Datos

Existen diversas técnicas de MD que se desarrollaron y probaron a través de los años en diferentes proyectos. En [21] se clasifica a los algoritmos de MD en dos grupos: supervisados y no supervisados. Los supervisados son aquellos que predicen el valor de un atributo *etiqueta* con base en un conjunto de datos conocido, es decir, funcionan para la clasificación o para la predicción. A partir de los valores donde la *etiqueta* se conoce, se realiza una relación entre la *etiqueta* y los otros valores y es por medio de estas relaciones que se realiza la predicción sobre una etiqueta desconocida. Por otro lado, los métodos no supervisados descubren patrones y tendencias en los datos. Los algoritmos no supervisados buscan similitud entre los datos con el fin de establecer grupos. El presente proyecto se enfoca en técnicas de aprendizaje supervisado. Algunos ejemplos de algoritmos de aprendizaje supervisado son a) *Métodos Bayesianos*: como Naive Bayes, b) *Árboles de decisión*: como CART, ID3, C 4.5, c) *Regresión*: tanto Lineal como No Lineal, d) *Aprendizaje perezoso*: como K-Vecino más cercano y e) *Reglas de asociación*: CBA, CMAR, CPAR entre otros.

4 Aplicación Web para explotación de modelos de MD

Para hacer uso de los modelos de MD que predicen la aparición de complicaciones en pacientes con DMTII, se desarrolló una aplicación Web bajo la tecnología de JSF (*JavaServer Faces*). JSF es la tecnología más reciente y el estándar oficial que ofrece Java para la construcción de aplicaciones Web. JSF implementa el modelo arquitectónico

MVC (Modelo-Vista-Controlador), el cual permite separar la lógica del negocio (modelo), de la presentación (vista), obteniendo así aplicaciones escalables y de fácil mantenimiento.

En la figura 1 se muestra la arquitectura de la aplicación, al tratarse de una aplicación Web, ésta se encuentra alojada en un servidor y un cliente accede a ella a través de un *browser*. Los componentes que integran la arquitectura del sistema se describen a continuación:

Controlador: contiene el FacesServlet, es decir el *servlet* de JSF cuya función principal es coordinar a la vista y al modelo controlando el flujo de navegación, gestionando las diferentes peticiones de recursos y generando las respuestas correspondientes a la vista.

Vista: está compuesta por archivos XHTML que representan la interfaz del usuario, su principal función es el despliegue y la captura de información. Para realizar dicha tarea se ayuda del uso de las etiquetas propias de JSF. También se incluye la biblioteca de PrimeFaces que es un API para JSF que permite crear aplicaciones Web enriquecidas.

Modelo: En el modelo se encuentran los *beans* administrados (*Managed Beans*) que dan acceso a los componentes de la interfaz del usuario y permiten el paso de información a los *beans* de modelo (*Model Beans*). Los *Model Beans* representan a las clases importantes en el dominio del problema y hacen uso de la API de Weka que es una biblioteca para Java que permite realizar diversas tareas de MD y se usa para acceder a los modelos de MD representados por archivos `.model`, y de la API de JDBC para acceder al gestor de base de datos PostgreSQL, donde se encuentra almacenada la base de datos que contiene la información perteneciente a los médicos y un registro histórico de las diversas predicciones que se han realizado, con el fin de que en un futuro sirvan para reentrenar al modelo.

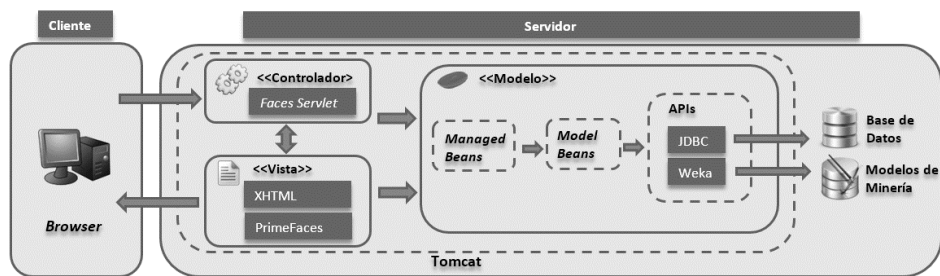


Fig. 1. Arquitectura de la aplicación.

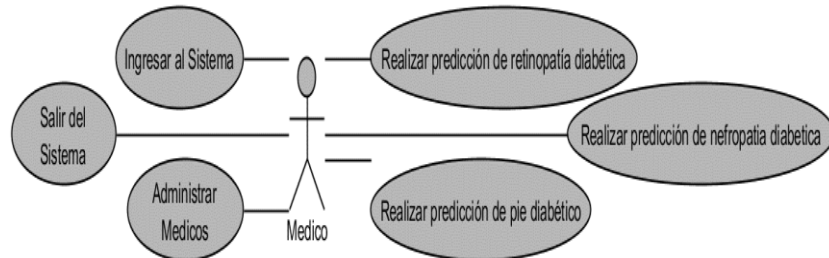


Fig. 2. Funcionalidad de la aplicación expresada a través un diagrama de casos de uso.

En la figura 2 se observa la funcionalidad de la aplicación. Como se aprecia en la imagen, sólo existe un actor: Médico, que representa al profesional de la salud que se encarga de manejar el sistema. En total se identifican 6 casos de uso: uno de ellos referente a la gestión de los médicos que contiene las típicas operaciones de registro, bajas y modificaciones. Dos de ellos a la autenticación del sistema (Ingresar al Sistema, Salir del Sistema) y tres con respecto a la realización de las predicciones (predicción de RD, predicción de ND, predicción de PD) que engloban todas las tareas necesarias para predecir la aparición de complicaciones de la diabetes.

En la figura 3 se modelan las relaciones entre clases que integran al sistema y la distribución de éstas en sus respectivos paquetes. En total existen 5 paquetes que contienen todas las clases del sistema, a continuación se describen brevemente cada uno de ellos:

WEB: representa prácticamente a la vista del sistema, aquí se encuentran todas las páginas XHTML de la aplicación.

Filtros: contiene a la clase *FiltroAcceso* que se comunica con el *servlet* de la aplicación y restringe a los usuarios el acceso a ciertos recursos.

ManagedBeans: son las clases que implementan la lógica del negocio, a) *LoginJB*, invoca al modelo para que un usuario inicie o cierre sesión, b) *PrediccionesJB*, invoca al modelo para que se realice la predicción, c) *AdministrarMedicosJB*, invoca al modelo para que se registren, eliminen, consulten o modifiquen los médicos y d) *EdicionCampos*: modifica las propiedades en los campos de un formulario.

ModelBeans: compuesto por las clases que son importantes en el dominio del problema, a) *Medico*: representa a una entidad médico con todos sus atributos y contiene los métodos para iniciar sesión, cerrar sesión, registrar, eliminar y modificar médicos; b) *Paciente*: representa a una entidad paciente con todos sus atributos y contiene los métodos para predecir la aparición de nefropatía, retinopatía y pie diabético.

Datos: incluye dos clases, a) AccesoDatos, contiene los métodos para acceder, manipular y consultar la base de datos y b) AccesoMinería, contiene los métodos para acceder y consultar los modelos de minería de datos.

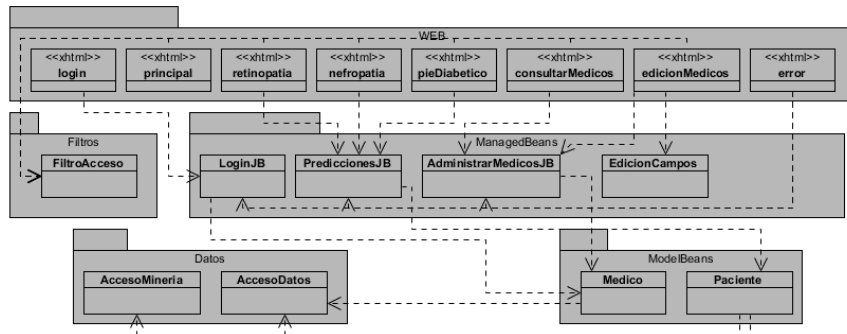


Fig. 3. Diagrama de clases de la aplicación.

En la vista se implementaron elementos de JSF y de PrimeFaces para lograr un buen grado de usabilidad. En la figura 4 se observa un formulario que permite la captura de las 53 variables involucradas en la predicción de la aparición de ND.

Fig. 4. Pantalla de captura de datos para predecir la aparición de nefropatía diabética.

La aplicación funciona con un modelo de MD provisional debido a que, de momento, se trabaja en la captura de información real de los modelos de MD que predicen la aparición de complicaciones de DMTII; una vez finalizada la captura y la validación de los modelos, éstos se integrarán a la aplicación Web anteriormente descrita.

5 Metodología para MD

Para el desarrollo de los modelos de MD se sigue el proceso de Descubrimiento de Conocimiento en Datos, conocido por sus siglas en inglés como KDD (*Knowledge Discovery from Data*). En la figura 6 se observan las 5 fases que conforman KDD.

En este proyecto se elaboraron tres modelos de minería de datos. Uno para la predicción de RD, otro para ND y otro para PD. Para el análisis de los datos se emplea WEKA, que contiene una colección de algoritmos de aprendizaje automático para tareas de MD. Dentro de las tareas para las que se usa WEKA están el preprocesamiento de datos, clasificación, visualización, y selección de características.

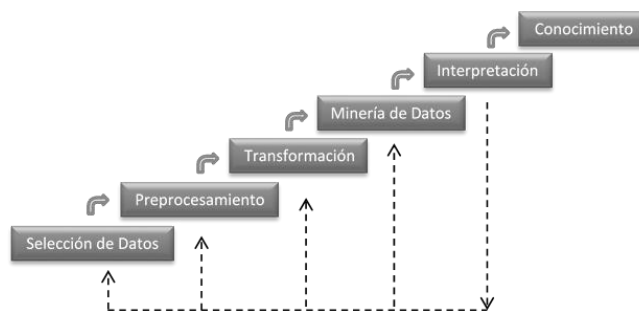


Fig. 6. Fases en el proceso de KDD.

Para la fase de KDD “Selección de Datos”, en la figura 7 se observan las tareas que se realizaron. Como primer paso se definieron las características de la población y el tamaño de la muestra. Posteriormente, con la ayuda de un médico, se seleccionaron las variables que se involucran en la aparición de complicaciones de la DMTII. Después se procedió a capturar digitalmente en Excel la información de los expedientes clínicos de los pacientes. Y, por último, se procedió a importar el documento de Excel a una base de datos para su manipulación en fases posteriores.

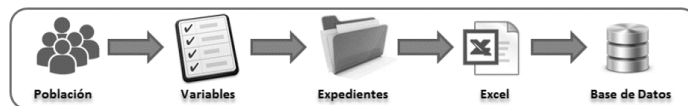


Fig. 7. Actividades realizadas en la fase de selección de datos.

La población se obtuvo del Hospital Regional de Río Blanco (HRRB) ubicado en el municipio de Río Blanco, Veracruz, México. El universo de trabajo está conformado por pacientes con diagnóstico de DMTII, en cualquier etapa de la enfermedad y con presencia de RD o ND o PD. Y también pacientes con diagnóstico de DMTII, en cualquier etapa de la enfermedad pero sin presencia de complicaciones, es decir, sin diagnóstico de RD, ND y/o PD. El tamaño de la muestra es de un total de 200 pacientes con DMTII que acudieron a consulta externa, cirugía o urgencias en el HRRB en el periodo comprendido del año 2012 a 2014.

Para la selección de las variables se consultaron diferentes guías de la práctica clínica elaboradas por instituciones de salud como el Instituto Mexicano del Seguro Social (IMSS) y Asociación Latinoamericana de Diabetes (ALAD) entre otras y también se consultó a un médico experto que se encargó de validar y corregir las variables seleccionadas, además de agregar algunas otras que consideraba de importancia para la realización del estudio. Al final se seleccionaron 53 variables entre las cuales se encuentran: edad, sexo, duración de la diabetes, índice de masa corporal (IMC), tipo hipoglucemiante, hipertensión arterial (HTA), tabaquismo, alcoholismo, Hemoglobina Glucosilada (HbA1c), glucosa promedio, nitrógeno ureico promedio, entre otras. Una vez definidas las variables se procedió a capturarlas en un documento de Excel a partir de los expedientes clínicos de los pacientes.

Cabe mencionar que una de las mayores dificultades por la que se ha atravesado en el desarrollo de este proyecto es la obtención de información en el sector de la salud. Esto debido principalmente a que los expedientes de los pacientes no están sistematizados y presentan información incompleta y bajo distintos formatos, lo cual ralentiza considerablemente el proceso de recolección de datos. Actualmente se sigue trabajando en la captura de la información pero una vez concluida esta actividad se podrá continuar con las siguientes fases de KDD. En la etapa de "*Minería de Datos*" los principales algoritmos de clasificación que se emplearán son Bayes Ingenuo, Árboles de Decisión (ID3, C4.5), Vecino Más Cercano, Perceptrón Multicapa y Maquinas de Vectores de Soporte, todos ellos permiten trabajar tanto con valores nominales como con valores numéricos a excepción del clasificador ID3 que sólo acepta valores nominales, por lo que las variables como la edad, duración de la diabetes, IMC, glucosa, entre otras será necesario transformarlas a sus valores nominales, dicha tarea se llevará a cabo en la fase de "*Transformación*". Por otro lado, todos estos algoritmos también permiten trabajar con valores perdidos sólo que cada uno de ellos emplea una estrategia particular para tratar con ellos. En la etapa de "*Preprocesamiento*" se detectarán y tratarán este tipo de valores, en el caso de las variables que cuentan con muy pocas mediciones como sucede con la Hemoglobina Glucosilada (HbA1c) que se encuentra ausente en más del 90% de las tuplas se procederá a eliminarla del conjunto de entrenamiento, cabe mencionar que la HbA1c se había contemplado inicialmente dentro de las variables a medir debido a que permite conocer los niveles de glucosa en los últimos 3 meses en un paciente diabético, pero en base los resultados obtenidos en la recolección de datos se considera poco relevante incluirla en el estudio. Quizás una de las razones por la que existen pocos registros de esta variable es debido a que el precio de realizar esta prueba es relativamente alto en

comparación a la medición de glucosa tradicional. Siguiendo con el preprocesamiento también se ignorarán aquellas tuplas que presenten mucha información faltante como es el caso de aquellos pacientes que no tienen registro de ninguna prueba de laboratorio. La intención es que la muestra de pacientes contenga registros completos y así mantener la información lo más real posible, aquellas tuplas que presenten pocos valores faltantes se les asignará un valor a juicio médico. Finalmente en la etapa de *Interpretación* se evaluarán los modelos con base en su nivel de certeza, sensibilidad y especificidad, en caso de que no sean los óptimos se regresará a etapas anteriores hasta que se logre el mejor desempeño. De igual forma se utilizará la técnica de selección de características para medir la influencia de las variables en la predicción con el objetivo de eliminar aquellas que sean irrelevantes.

6 Problemática encontrada

La relevancia de incorporar aspectos tecnológicos para la gestión, diagnóstico y tratamiento de enfermedades constituye una propuesta importante para el apoyo de problemas de salud. Una vez terminada e instalada la aplicación, se logrará analizar información correspondiente a pacientes de la región que de otra forma sería difícil de evaluar por un profesional de la salud y que generará modelos de comportamiento, permitiendo así predecir la aparición de complicaciones en pacientes diabéticos.

Una de las complicaciones por las que atraviesa el desarrollo de este trabajo, es la dificultad que implica la recolección de la información. Esto debido principalmente a que la mayoría de los hospitales en México, no cuenta con sistemas que implementen el expediente clínico electrónico (ECE), por lo que la recolección de datos se realiza de forma manual a partir de la revisión de expedientes físicos, lo que ralentiza el avance del proyecto; la revisión de un expediente puede tardar hasta una hora y aun así presentar información incompleta. Otras dificultades que ocurren es el cambio de personal médico que ocasiona que existan diferentes formatos, estilos de redacción de notas médicas y pruebas de laboratorio entre otras cosas. Y por último, existen muchos pacientes que acuden de forma irregular al médico por lo que no se cuenta con su expediente clínico completo. Todos estos factores mencionados anteriormente conllevan a una irregular e incompleta información clínica por lo que se hace difícil obtener información valiosa de los datos.

De momento aún queda mucho trabajo de MD por realizar en el proyecto y el aporte que éste realice a los profesionales de la salud mucho dependerá de los resultados que se obtengan al entrenar los modelos, con base en su nivel de certeza, sensibilidad y especificidad, aunque cabe mencionar que los modelos nunca serán lo suficientemente precisos o completos para reemplazar las decisiones tomadas por los profesionales de la salud. En este aspecto, el principal objetivo de este trabajo es sugerir una alternativa computacional que proporcione información útil a los médicos para planear estrategias apropiadas en el control de las complicaciones de la diabetes.

7 Conclusiones y trabajo a futuro

Es importante valorar el impacto de las Tecnologías de Información y Comunicación en el sector de la salud. La MD ofrece alternativas computacionales para el diagnóstico e identificación de diversas enfermedades como son el cáncer de mama, problemas cardiacos, diabetes, problemas renales, entre otros. La MD permite analizar diversas variables que se relacionan con la aparición de complicaciones en pacientes diabéticos que de otra forma sería difícil de evaluar por un profesional de la salud. Además permite generar modelos de comportamiento que sirven para predecir la aparición de complicaciones en pacientes diabéticos.

Con el desarrollo de esta aplicación se espera proporcionar a los profesionales de la salud una alternativa computacional en el proceso de toma de decisiones y a la identificación de complicaciones en pacientes diabéticos.

Se espera que en un futuro el número de hospitales que implementen el sistema de ECE aumente, lo que conducirá a una más eficiente recolección de datos y hará más fácil la aplicación de técnicas de MD en el área de la salud pública.

Es importante señalar que, dado el gran impacto que tiene la diabetes en México, es necesario hacer todos los esfuerzos posibles en cada uno de los diferentes sectores con el objetivo de disminuir el problema de representa esta enfermedad.

Como trabajo a futuro está el terminar la construcción de los modelos de MD. Para esto se debe de continuar con la captura de la información de los expedientes clínicos y posteriormente realizar las actividades que propone KDD en sus distintas fases, tal y como se describió en el apartado 5 y una vez construidos integrar los modelos a la aplicación Web descrita en el apartado 4.

Agradecimientos

Al Consejo Nacional de Ciencia y Tecnología, CONACyT, por el apoyo económico otorgado para la realización del trabajo; al Hospital Regional de Río Blanco, HRRB, por facilitar la obtención de los datos y por la asesoría médica prestada.

Referencias

1. Secretaría de Salud, «Programa de Acción Específico 2007 – 2012,» Unidades de Especialidades Médicas UNEMES, (2007)
2. G. Olaiz Fernández, R. Rojas, C. A. Aguilar Salinas, J. Rauda y S. Villalpando, «Diabetes mellitus en adultos mexicanos. Resultados de la Encuesta Nacional de Salud 2000,» Salud Publica Mex, vol. 49, n° s3, (2007)
3. J. Escobedo de la Peña y B. Rico Verdín, «Incidencia y letalidad de las complicaciones agudas y crónicas de la diabetes mellitus en México; Incidence and fatlity rates of short and long-term diabetes mellitus complications in Mexico,» Salud pública Méx, vol. 8, n° 4, pp. 236-42, (1996)
4. K. Harleen y K. W. Siri, «Empirical study on applications of data mining techniques in healthcare,» Journal of Computer Science, vol. 2, n° 2, pp. 194-200, (2006)

5. M. F. Bin Othman y T. M. Shan You, «Comparison of different classification techniques using WEKA for breast cancer,» In 3rd Kuala Lumpur International Conference on Biomedical Engineering, pp. 520-523, (2007)
6. N. Chalortham, M. Burranarach y T. Supnithi, «Information and Creativity Support Systems,» de Ontology Development for Type II Diabetes Mellitus Clinical Support System, (2009)
7. A. A. Aljumah, M. Gulam Ahamad y M. Khubeb Siddiqui, «Application of Data Mining: Diabetes Health Care in Young and Old Patients,» de Journal of King Saud University-Computer and Information Sciences, (2012)
8. S. Imambi y T. Sudha, «Building Classification System to Predict Risk factors of Diabetic Retinopathy Using Text mining,» International Journal, vol. 2, n° 7, pp. 2309-2312, (2010)
9. K. Rajesh y V. Sangeetha, «Application of Data Mining Methods and Techniques for Diabetes Diagnosis,» International Journal of Engineering and Innovative Technology (IJEIT), vol. 2, n° 3, (2012)
10. M. Cedeno, A. y D. Andina, «Data mining for the diagnosis of type 2 diabetes,» de In World Automation Congress (WAC), (2012)
11. H. C. Koh y G. Tan, «Data mining applications in healthcare,» Journal of Healthcare Information Management, vol. 19, n° 2, p. 65, (2011)
12. X.-H. Meng, Y.-X. Huang, D.-P. Rao, Q. Zhang y Q. Liu, «Comparison of three data mining models for predicting diabetes or prediabetes by risk factors,» Kaohsiung Journal of Medical Sciences, vol. 29, n° 2, pp. 93-9, (2013)
13. J.-Y. Yeh, T.-H. Wu y C.-W. Tsao, «Using data mining techniques to predict hospitalization of hemodialysis patients,» Decis. Support Syst, vol. 50, n° 2, pp. 439-448, January (2011)
14. T.-C. Lu y C.-Y. Tseng, «Hemodialysis key features mining and patients clustering technologies,» Adv. Artif. Neu. Sys, p. 6, (2012)
15. G. Parthiban, A. Rajesh y S. Srivatsa, «Diagnosis of Heart Disease for Diabetic Patients using Naive Bayes Method,» International Journal of Computer Applications, vol. 24, n° 3, pp. 7-11, 2011.
16. B. Hwan Cho, H. Yu, K.-W. Kim, T. Hyun Kim, I. Young Kim y S. I. Kim, «Application of irregular and unbalanced data to predict diabetic nephropathy using visualization and feature selection methods,» Artif. Intell. Med, vol. 42, n° 1, pp. 37-53, (2008)
17. A. Arredondo y E. de Icaza, «Costos de la diabetes en América Latina: evidencias del caso Mexicano,» Value Health, vol. 14, n° 5, pp. s85-8, (2011)
18. M. Kantarzdziec, Data Mining: Concepts, Models, Methods, and Algorithms (2nd ed.), Wiley-IEEE Press, (2011)
19. J. Han, M. Kamber y J. Pei, Data Mining: Concepts and Techniques (3rd ed.), San Francisco, CA: Morgan Kaufmann Publishers Inc., (2011)
20. Microsoft Corporation, «MSDN Library», [En línea]. Available: <http://msdn.microsoft.com/es-es/library/cc645779.aspx#feedback>. Agosto (2013)
21. S. M. Weiss y N. Indurkha, Predictive Data Mining: A Practical Guide, San Francisco, CA: Morgan Kaufmann Publishers Inc, (1998)