

Aplicación del proceso de descubrimiento del conocimiento para la detección de diabetes

Application of the knowledge discovery process for diabetes screening

DOI: 10.46932/sfjdv3n2-060

Received in: February 15th, 2022

Accepted in: March 1st, 2022

Irma Yazmín Hernández Báez

Institución: Universidad Politécnica del Estado de Morelos

Dirección: Boulevard Cuauhnáhuán 566 Jiutepec, Morelos

Correo electrónico: ihernandez@upemor.edu.mx

Sandra Elizabeth León Sosa

Institución: Universidad Politécnica del Estado de Morelos

Dirección: Boulevard Cuauhnáhuán 566 Jiutepec, Morelos

Correo electrónico: lsandra@upemor.edu.mx

Miguel Ángel Ruiz Jaimes

Institución: Universidad Politécnica del Estado de Morelos

Dirección: Boulevard Cuauhnáhuán 566 Jiutepec, Morelos

Correo electrónico: mruiz@upemor.edu.mx

Alma Delia Nieto Yáñez

Institución: Universidad Politécnica del Estado de Morelos

Dirección: Boulevard Cuauhnáhuán 566 Jiutepec, Morelos

Correo electrónico: anieto@upemor.edu.mx

Roberto Enrique López Díaz

Institución: Universidad Politécnica del Estado de Morelos

Dirección: Boulevard Cuauhnáhuán 566 Jiutepec, Morelos

Correo electrónico: rlopezd@upemor.edu.mx

RESUMEN

En este trabajo de investigación se presenta el uso de la ciencia de datos en el proceso de clasificación de pacientes diabéticos o no diabéticos, analizando diversos factores como: El índice de masa corporal, los niveles de glucosa, el grosor de la piel, los niveles de insulina, la edad y la presión sanguínea. El alcance principal del proyecto es el desarrollo de un sistema web que permite realizar las diferentes fases del proceso KDD para realizar clasificación, se implementaron diferentes funciones en cada una de las fases del proceso KDD, para permitir al usuario sintonizar los diferentes parámetros y así poder realizar experimentos para obtener varios resultados y, posteriormente, evaluar cual de esos resultados es el mejor. Se presentan los resultados de la experimentación realizada comparando los resultados de la aplicación desarrollada vs. los resultados obtenidos con la aplicación comercial RapidMiner. Se comparan resultados utilizando los algoritmos: KNN, clasificador bayesiano, árbol de decisión y el perceptrón multicapa de ambos sistemas.

Palabras clave: descubrimiento de conocimiento, diabetes y redes neuronales.

ABSTRACT

This research work presents the use of data science in the process of classifying diabetic or non-diabetic patients, analyzing various factors such as: Body mass index, glucose levels, skin thickness, levels insulin levels, age and blood pressure. The main scope of the project is the development of a web system that allows to perform the different phases of the KDD process to perform classification, different functions were implemented in each of the phases of the KDD process, to allow the user to tune the different parameters and thus be able to perform experiments to obtain various results and subsequently evaluate which of those results is the best. The results of the experimentation carried out are presented comparing the results of the developed application vs. the results obtained with the commercial RapidMiner application. Results are compared using the algorithms: KNN, Bayesian classifier, decision tree and the multilayer perceptron of both systems.

Keywords: kdd, diabetes, neural networks.

1 INTRODUCCIÓN

Uno de los campos científicos de la informática es la minería de datos, la cual es el proceso de descubrir conocimiento interesante de grandes cantidades de datos almacenadas en bases de datos, u otro repositorio de información [1].

Una de las áreas que cuenta con grandes cantidades de información y donde esta técnica ha demostrado su utilidad en la extracción de conocimiento, es el área de la salud, de la cual derivan muchas enfermedades que, si se detectan a tiempo, pueden prolongar la vida de la persona.

La enfermedad de la diabetes, aqueja a gran parte de la población de México, esta no se ve obstaculizada por el nivel socio-económico, características físicas, edad ni sexo. Un gran número de científicos han dedicado investigaciones para analizar esta enfermedad, descubrir sus causas y revisar posibles tratamientos.

Esta enfermedad es un padecimiento en el cual el azúcar en la sangre se encuentra en un nivel elevado, esto se debe a que el cuerpo no produce o no utiliza adecuadamente la insulina. Sin la suficiente insulina se altera todo el mecanismo regulador, como consecuencia, las células, no producen energía suficiente y alteran sus funciones.

El 14 por ciento de los adultos en México tienen diabetes, además es la primera causa de ceguera prevenible en el adulto en edad productiva, insuficiencia renal terminal, amputaciones no traumáticas y de infartos al miocardio.

Se estiman 80,000 muertes por año y algo que no se aprecia correctamente, es que muchas de estas muertes, además de ser prematuras, las precede un periodo largo de discapacidad severa y costosa. La Diabetes Mellitus es un problema de la salud pública prioritaria en México debido a su tendencia creciente y a su relación con la obesidad [2].

Al menos el 50% de las personas diabéticas ignoran que lo son y el 80% de la Diabetes tipo 2 es prevenible mediante la adopción de una dieta saludable y el incremento de la actividad física. Por lo tanto, si la diabetes tipo 2 puede ser prevenible, lo ideal sería detectar la enfermedad a tiempo, mediante el estudio de los registros médicos con los que cuentan los hospitales.

Hoy en día los hospitales tienen grandes cantidades de registros médicos de pacientes con diabetes, pero el ser humano no es capaz de procesar y analizar tales cantidades de información, por lo tanto, se necesitan emplear tecnologías como el proceso KDD que nos ayude a extraer conocimiento de los datos históricos con los que cuentan los hospitales

2 OBJETIVOS

El objetivo general del presente proyecto es aplicar el proceso KDD para detectar si una persona es o no diabética mediante un sistema web experimentador y RapidMiner, utilizando datos de prueba y datos sintéticos.

Los objetivos específicos que llevarán al cumplimiento del objetivo general son:

- Implementar el experimentador web del proceso KDD.
- Implementar un generador de datos sintéticos de prueba.
- Identificar los requisitos funcionales del sistema.

En México los cálculos del costo de atención por paciente diabético van desde 700 hasta 3200 dolares anuales, esto es aproximadamente el 14% del gasto en salud destinado a la atención de esta enfermedad, el poder detectar a tiempo la diabetes tipo 2, disminuiría tal gasto y además permitiría realizar un plan de contingencia para modificar el estilo de vida y las características socio-ambientales que, unidas a factores genéticos, constituyen las principales causas desencadenantes de la diabetes.

El sistema experimentador, podrá permitir aplicar el proceso KDD no solo para realizar la clasificación de la enfermedad de la diabetes, sino de muchas otras más enfermedades, o inclusive, de problemas de otras áreas que no tengan que ver con el área de la salud.

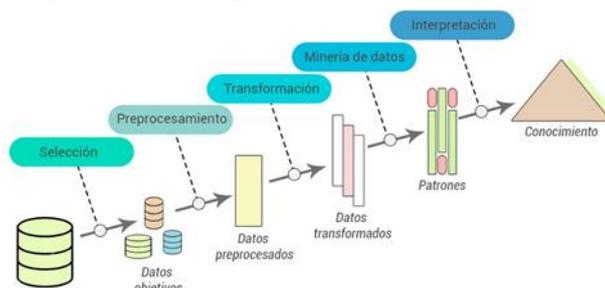
3 METODOLOGÍA

Para este proyecto, en la parte de minería de datos, se utilizó la metodología KDD, la cual es un proceso metodológico para encontrar un “modelo” válido, útil y entendible que describa patrones de acuerdo a la información, y como modelo entendemos que es la representación que intenta explicar ese patrón en los datos [3]. Ya que esta metodología permite retrocesos entre varias de sus fases, se puede aprovechar para volver a analizar fases anteriores dependiendo de los resultados obtenidos. Además, el proyecto se puede llegar a tornar cíclico, pues este no se termina una vez que se encuentre una solución y

como se planea experimentar con los diferentes resultados posibles, se adecua para volver a fases anteriores del ciclo de esta metodología.

El proceso KDD que se muestra en la Figura 1 es interactivo e iterativo, involucra numerosos pasos con la intervención del usuario en la toma de muchas decisiones. Se resume en las siguientes etapas:

Figura 1. Metodología KDD. Fuente: Adaptado de [3]



Selección de datos: En esta etapa se determinan las fuentes de datos y el tipo de información a utilizar. Es la etapa donde los datos relevantes para el análisis son extraídos desde la o las fuentes de datos.

Preprocesamiento: Esta etapa consiste en la preparación y limpieza de los datos extraídos desde las distintas fuentes de datos en una forma manejable, necesaria para las fases posteriores. En esta etapa se utilizan diversas estrategias para manejar datos faltantes o en blanco, datos inconsistentes o que están fuera de rango, obteniéndose al final una estructura de datos adecuada para su posterior transformación.

Transformación: Consiste en el tratamiento preliminar de los datos, transformación y generación de nuevas variables a partir de las ya existentes con una estructura de datos apropiada. Aquí se realizan operaciones de agregación o normalización, consolidando los datos de una forma necesaria para la fase siguiente.

Minería de datos: Es la fase de modelamiento propiamente, en donde métodos inteligentes son aplicados con el objetivo de extraer patrones previamente desconocidos, válidos, nuevos, potencialmente útiles y comprensibles y que están contenidos u “ocultos” en los datos.

Interpretación y Evaluación: Se identifican los patrones obtenidos y que son realmente interesantes, basándose en algunas medidas y se realiza una evaluación de los resultados obtenidos [4].

3.1 DISEÑO DE LA SOLUCIÓN

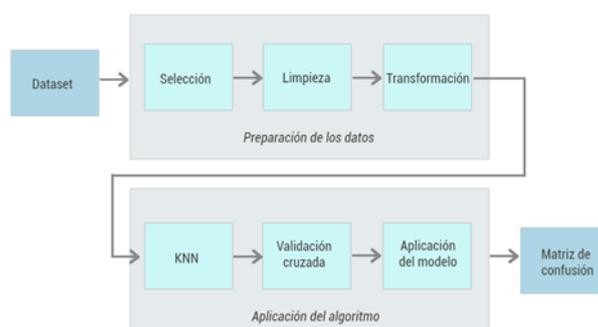
Para el proyecto se siguió la arquitectura flujo de datos, ya que se utiliza cuando hay transformaciones en secuencia sobre ciertos datos, donde cada módulo funciona como un filtro y las conexiones entre cada módulo se basan en el concepto de tuberías y éstas transportan de manera encapsulada, los datos generados por cada módulo.

La arquitectura lógica para este proyecto se basa principalmente en seis procesos muy apegados a las fases del proceso KDD, la **Figura 2** muestra la arquitectura empleada.

Selección

Este módulo de selección es el encargado de leer el archivo del dataset, el archivo debe estar en formato .xlsx y cada columna debe contener los datos de un atributo, el módulo debe ser capaz de obtener las cabeceras y mostrarlas al usuario para que seleccione las columnas con las que desea trabajar.

Figura 2. Arquitectura lógica del proyecto.

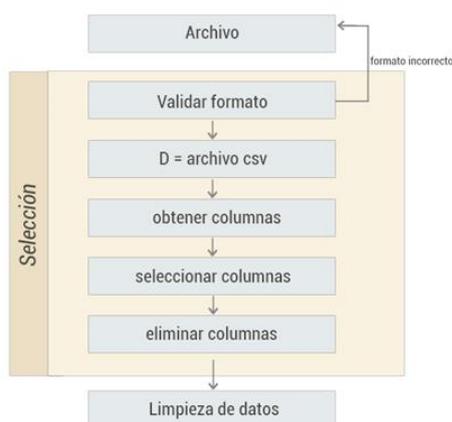


Fuente: Elaboración propia

Como se puede apreciar en la Figura 3 el módulo realiza el siguiente proceso:

1. Recibe por parámetro la ruta donde se encuentra el archivo.
2. Valida que el formato sea xlsx, en caso de que no sea, notificará que el archivo puede generar problemas durante el proceso KDD.
3. Almacena en un objeto el contenido del archivo con el fin de manipularlo con mayor facilidad.
4. Obtiene las cabeceras y las envía al usuario.
5. Para cada columna que se seleccione se saca el valor máximo, mínimo y la media y se le muestra al usuario.
6. Recibe un objeto con las columnas que el usuario seleccionó.
7. Elimina las columnas.

Figura 3. Módulo de selección



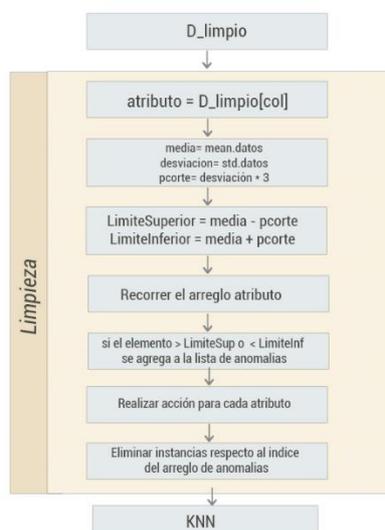
Limpieza

El módulo de limpieza es el que reajusta los datos para reparar o eliminar imperfecciones que puedan generar ruido al momento de aplicar el algoritmo de minería. Para este ejemplo, se muestra el proceso de detección de outliers mediante la desviación estándar.

Como se puede apreciar en la Figura 4 el módulo realiza el siguiente proceso:

1. Recibe el objeto de la dataset sin las columnas que el usuario eliminó en la fase de selección.
2. Crea un arreglo para cada una de las columnas.
3. Obtiene la media, la desviación y el punto de corte, por lo general, se utiliza el número 3, ya que a partir del 4 se aleja del punto máximo de la desviación estándar y se considera como anomalía.
4. Se recorre cada atributo y si el elemento es mayor al límite superior o menor al límite inferior, se añade a la lista de anomalías.
5. Se eliminan las instancias respecto al índice de anomalías.

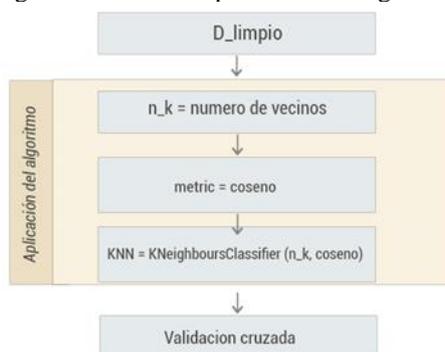
Figura 4. Módulo: limpieza.



Aplicación del algoritmo

Como se puede apreciar en la Figura 5 el módulo es el encargado de aplicar el algoritmo, se utiliza como ejemplo el algoritmo KNN, el cual recibe el objeto con los datos ya tratados, y recibe el número de vecinos y a través del cálculo de distancias de coseno, y almacenará los índices de las menores distancias.

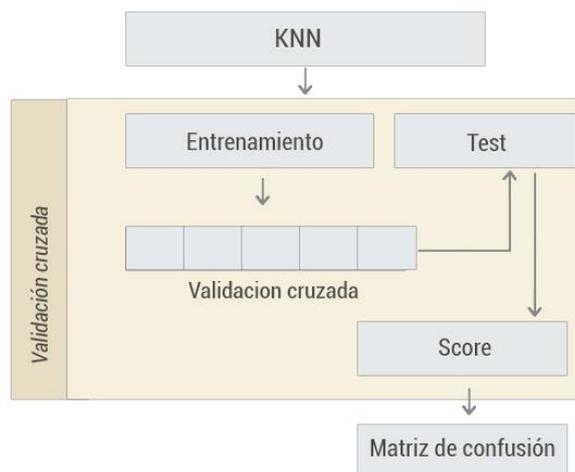
Figura 5. Módulo: Aplicación del algoritmo.



Validación cruzada

Como se puede apreciar en la Figura 6 el módulo para la validación cruzada se dividen los datos en k subconjuntos, y se entrena en k-1 uno de esos subconjuntos. Lo que hace es mantener el último subconjunto para la prueba y se puede hacer para cada uno de los subconjuntos.

Figura 6. Módulo: Validación cruzada

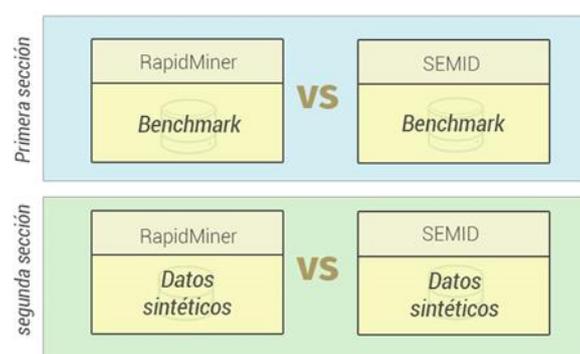


4 RESULTADOS

En esta sección se describe el proceso de evaluación del rendimiento del sistema, comparándolo con el sistema comercial RapidMiner, se describen los datos utilizados para la evaluación, los algoritmos, y la comparativa de los resultados obtenidos por ambos sistemas.

Las pruebas se realizaron en dos secciones, la primera sección se encarga de experimentar con los datos del benchmark, probando cuatro algoritmos con RapidMiner y cuatro algoritmos con el sistema web desarrollado (SEMID). Por otra parte, en la segunda sección se realizan pruebas con los datos generados de manera sintética, de igual manera probando cuatro algoritmos con RapidMiner y cuatro algoritmos con el sistema web SEMID tal como se puede observar en la Figura 7.

Figura 7. Esquema del plan de experimentación
Plan de experimentación



Para la experimentación se utilizaron dos conjuntos de datos los cuales se describen a continuación.

El primer conjunto de datos original elegido para realizar las pruebas de la primera sección fue “Pima Indians Diabetes Database” disponible de manera libre desde la base de datos de UCI Machine Learning en <https://www.kaggle.com/uciml/pima-indians-diabetes-database>, el cual cuenta con 767 instancias con dos clases, y ocho atributos [5], véase Tabla 1.

Tabla 1. Definición de relaciones y atributos dataset “1” (Benchmark)

Atributo	Descripción
Embarazo	Este atributo se refiere al número de veces que el paciente se ha embarazado.
Glucosa	El atributo corresponde a el nivel de concentración plasmática de glucosa a 2 horas en una prueba oral de tolerancia a la glucosa.
Presión Sanguínea	Este atributo corresponde a la presión arterial diastólica (mm Hg)
Grueso de la piel	Tríceps espesor de pliegues cutáneos (mm)
Insulina	2 horas de insulina en suero (mu U / ml)
IMC	El atributo IMC corresponde al Índice de masa corporal (peso en kg / (altura en m) ^ 2)
Diabetes Pedigree	El atributo Diabetes pedigree, corresponde a una función hereditaria de diabetes, es decir, que probabilidad de que en sus genes este la enfermedad de diabetes.
Años	Este atributo corresponde a la edad en años del paciente diabético.
Salida	la variable de clase (0 o 1) 268 de 768 son 1, las otras son 0

El segundo conjunto de datos fue generado en el sistema web SEMID, tratando de igualar los datos del primer conjunto de datos, con la finalidad de demostrar que el generador de datos puede ser útil para generar datos de prueba cuando no se tiene datos originales o estadísticos, en la Tabla 2 se describen los

atributos y las distribuciones de probabilidad que se utilizaron para cada uno de dichos datos, donde en la columna distribución se indica la distribución de probabilidad que sigue ese atributo y el valor 1 y valor 2 son los parámetros necesarios para generar datos con tal distribución.

Los resultados de la primera sección de la experimentación realizada con el benchmark, se muestran en la Tabla 3. Como se puede observar, para los experimentos en RapidMiner, para el algoritmo KNN es mejor trabajarlo con $k = 7$, para los árboles de decisión es mejor trabajar con los datos discretizados, igualmente para el clasificador bayesiano, y para el perceptrón multicapa es mejor no aplicar transformación de datos.

Tabla 1. Definición de relaciones y atributos *dataset* “2” (Sintético)

Atributo	Dataset	Distribución	Tipo	Valor 1	Valor 2
Clase	Diabetes	Binomial	Clase	1.0	0.36
Edad	Diabetes	Poisson	Atributo	40.0	1.0
Embarazos	Diabetes	Poisson	Atributo	7.0	1.0
Glucosa	Diabetes	Normal	Atributo	140.0	20.0
GrosorPiel	Diabetes	Laplace	Atributo	61.0	5.0
MC	Diabetes	Normal	Atributo	44.0	8.0
Insulina	Diabetes	Laplace	Atributo	450.0	50.0
Pedigree	Diabetes	Normal	Atributo	1.8	0.8
Presión	Diabetes	Normal	Atributo	68.0	13.0

Tabla 2. Resultados para el benchmark (RapidMiner)

Algoritmo	Transformación	Parámetros	Resultado
KNN	Ninguna	K=7	73.39%
KNN	Ninguna	K=12	73.27%
tree decisión	Ninguna	máx. depth= 10	68.71%
tree decisión	Discretización	máx. depth= 10	73.12%
Random tree	Ninguna	15 folds	67.32%
Random tree	Discretización	15 folds	71.12%
Naive Bayes	Discretización	8 folds 7 binas	77.88%
Naive Bayes	Ninguna	8 folds	76.79%
Naive Bayes	Normalización	8 folds	76.79%
MLP	Ninguna	10 generaciones k=10	77.07%
MLP	Discretización	10 generaciones	76.10%

En la Tabla 4 se presentan los resultados de los experimentos realizados en SEMID, se puede observar que para el KNN es mejor trabajar con la medida de los pesos por distancias que de manera uniforme, para el árbol de decisión es mejor no aplicar ningún filtro ni transformación, y para el clasificador bayesiano y perceptrón multicapa brindan mejores resultados cuando los datos se encuentran discretizados.

Tabla 3. Resultados para el benchmark (SEMID)

Algoritmo	Transformación	Parámetros	Resultado
KNN	Ninguna	K=12 uniforme	72.58%
KNN	Normalización	K=12 uniforme	73.57%
KNN	Ninguna	K=12 distancias	74.16%
KNN	Normalización	K=12 distancias desviaciones=2	75.82%
tree decision	Ninguna	Ninguna	71.99%
tree decision	Ninguna	desviaciones=3	69.60%
tree decision	Normalización	desviaciones=3	69.60%
tree decision	Discretización	3 binas	64.75%
Naive Bayes	Ninguna	Ninguna	73.96%
Naive Bayes	Ninguna	desviaciones=3	72.90%
Naive Bayes	Discretización	binas= 3 desviaciones=3	75.34%
Naive Bayes	Discretización	binas= 3 desviaciones=2	78.02%
MLP	Ninguna	Ninguna	65.08%
MLP	Discretización	desviaciones=2 binas= 3	73.62%

A continuación, se muestra la experimentación realizada con los datos que fueron generados de forma sintética. En la Tabla 5 se muestran los resultados de la experimentación realizada sobre RapidMiner.

Tabla 4. Resultados para datos sintéticos (RapidMiner)

Algoritmo	Transformación	Parámetros	Resultado
KNN	Ninguna	K=7	61.13%
KNN	Ninguna	K=12	60.04%
KNN	Discretización	K=12 binas=5	61.82%
tree decision	Ninguna	máx. depth= 10	65.25%
tree decision	Discretización	máx. depth= 10	65.11%
Naive Bayes	Discretización	8 folds 7 binas	64.12%
Naive Bayes	Ninguna	8 folds	65.25%
Naive Bayes	Normalización	8 folds	65.25%
MLP	Ninguna	10 generaciones k=10	65.53%
MLP	Discretización	10 generaciones	64.42%

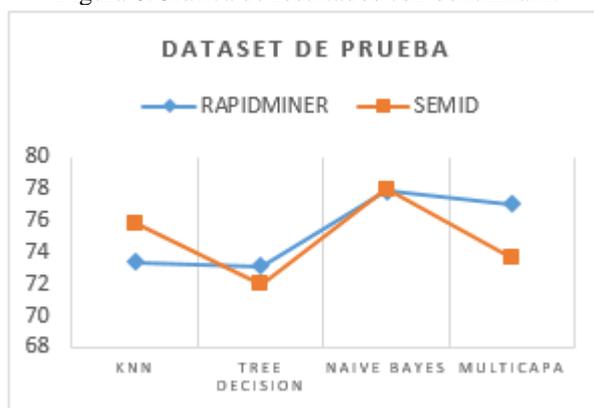
En la Tabla 6 se muestran los resultados de la experimentación realizada con el sistema web SEMID.

Tabla 5. Resultados para datos sintéticos (SEMID).

Algoritmo	Transformación	Parámetros	Resultado
KNN	Ninguna	K=12 uniforme	62.32%
KNN	Normalización	K=12 uniforme	63.90%
KNN	Ninguna	K=12 distancias	61.56%
KNN	Normalización	K=12 distancias desviaciones=2	63.87%
tree decision	Ninguna	Ninguna	52.66%
tree decision	Ninguna	desviaciones=3	52.16%
tree decision	Normalización	desviaciones=3	53.81%
tree decision	Discretización	3 binas	53.40%
Naive Bayes	Ninguna	Ninguna	62.52%
Naive Bayes	Ninguna	desviaciones=3	62.88%
Naive Bayes	Discretización	binas=3 desviaciones=3	65.36%
Naive Bayes	Discretización	binas=3 desviaciones=2	63.58%
MLP	Ninguna	Ninguna	65.08%
MLP	Discretización	desviaciones=2 binas= 3	63.58%

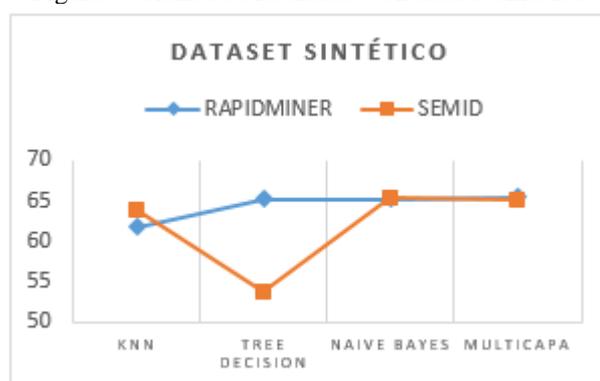
En la Figura 8 se puede observar la comparativa de las pruebas realizadas con el benchmark en ambos sistemas de minería, en donde los resultados no hubo mucha diferencia, que inclusive en el algoritmo Naive Bayes se obtuvieron casi, los mismos resultados.

Figura 8. Gráfica de resultados con benchmark.



Por otro lado, en la Figura 9 se puede observar la comparativa de las pruebas realizadas con el dataset generado de forma sintética en ambos sistemas de minería, en donde los resultados se muestran aún más similitud que en la primera sección, en esta segunda sección tres algoritmos fueron bastante parecidos, y solo hubo una diferencia notable en el árbol de decisión.

Figura 9. Gráfica de resultados con dataset sintético.



5 CONCLUSIONES

El principal aporte del desarrollo de este sistema web es el poder otorgar una herramienta de uso libre para minar datos, a la cual se podrá acceder desde internet sin tener que estar instando algún programa en el ordenador, si bien existen herramientas que dominan en el mercado como RapidMiner, ésta se tiene que instalar y además pagar una licencia.

En cuanto a los resultados obtenidos, si bien se utilizó una librería de uso libre para Python, sus resultados no están muy alejados del gigante de RapidMiner, por lo que SEMID puede ser una herramienta

alternativa capaz de minar datos, además de contar con un módulo que le permite generar datos de prueba, esto también puede ser bastante útil, ya que si no se tienen datos estadísticos o reales, se pueden generar datos sintéticos con diferentes distribuciones de probabilidad y ajustarlas a un problema de la vida real.

En cuanto a la experimentación para detectar pacientes diabéticos, definitivamente el algoritmo Naive Bayes es el indicado para clasificar pacientes, que si bien los resultados estuvieron muy aproximados al 80% no significa que la experimentación haya fallado, si no que los datos del benchmark no eran del todo utilizables, ya que atributos como el embarazo no generaban mucho aporte a la experimentación.

Al realizar nuevos datos de prueba la experimentación de nuevos escenarios sin ninguna necesidad de recolectar grandes cantidades de datos, simplemente obteniendo algunas estadísticas, se podrían generar más datos para realizar otras pruebas no solo para pacientes con diabetes, si no para otras enfermedades también, y esto sería de gran ayuda para el sector salud.

Para la validación de los datos y evitar dar un diagnóstico incorrecto se aplicó la validación de entrenamiento y prueba además de generar una matriz de confusión para poder evaluar cuántos falsos negativos, existen ya que estos son los que pueden generar más riesgo.

REFERENCIAS

- [1] J. Han, M. Kamber y J. Pei, *Data Mining: Concepts and Techniques*, Waltham, MA: Morgan Kaufmann, 2012.
- [2] Instituto de Seguridad y Servicios Sociales de los Trabajadores del Estado, «Diabetes, uno de los principales problemas de salud en México,» 2019. [En línea]. Available: <https://www.gob.mx/issste/es/articulos/diabetes-uno-de-los-principales-problemas-de-salud-en-mexico?idiom=es>.
- [3] U. Fayyad y E. Simoudis, *Data Mining and Knowledge Discovery in Databases*, 1997.
- [4] M. Hofmann y R. Klinkenberg, *Rapid Miner Data Mining: Use Cases and Business Analytics Applications*, CRC Press, 2013.
- [5] J. Smith, J. Everhart, W. Dickson, W. Knowler y R. Johannes, «Using the ADAP learning algorithm to forecast the onset of diabetes mellitus,» *Proceedings of the Symposium on Computer Applications and Medical Care*, pp. 261-265, 1988.