Algoritmo para la detección temprana de diabetes mellitus tipo 2 en mujeres mayores de 21 años

Lemuz Fuentes Omar Alejandro

16 de junio de 2020

Resumen

El presente artículo propone un sistema de clasificación supervisado basado en el algoritmo de minería de datos bayesiano ingenuo para la detección temprana de la diabetes mellitus tipo 2 en mujeres mayores de 21 años con base en el análisis de ocho variables tales como número de embarazos, concentración de glucosa en plasma, presión sanguínea diastólica, espesor cutáneo del tríceps, insulina sérica, índice de masa corporal, función de pedigrí de diabetes y edad. La metodología propuesta entrena y prueba el sistema propuesto con base en 768 muestras tomadas de pacientes con y sin diabetes. El sistema se validó a través de la división de las muestras en dos grupos: 80% de las muestras para entrenarlo y el 20% restante para probarlo. Esto indicó que, en el 92.2% de los casos, el clasificador bayesiano ingenuo logra detectar de manera correcta esta enfermedad tomando en cuenta las variables mencionadas anteriormente.

Palabras clave: algoritmo, diabetes mellitus, mujeres, detección temprana, minería de datos, bayes ingenuo, clasificación supervisada.

Abstract

This article proposes a supervised classification system based on the naive bayes data mining algorithm for the early detection of type 2 diabetes mellitus in women over 21 years based on the analysis of eight variables such as number of pregnancies, concentration plasma glucose, diastolic blood pressure, triceps skin thickness, serum insulin, body mass index, diabetes pedigree function and age. The proposed methodology trains and tests the proposed system based on 768 samples taken from patients with and without diabetes. The system was validated dividing the samples into two groups: 80% of the samples to train it and the remaining 20 % to test it. The results indicate that, considering all the variables mentioned above, this disease was correctly detected by the the naive bayes classifier in 92.2% of the cases.

Keywords: algorithm, diabetes mellitus, women, early detection, data mining, naive bayes, supervised classification.

1 Introducción

La diabetes se define como una condición crónica que se presenta cuando los niveles de glucosa en la sangre son elevados debido a que el cuerpo no es capaz de producir insulina suficiente. La insulina es una hormona producida por el páncreas, que se encarga de llevar glucosa hasta las células del cuerpo donde esta es convertida en energía. La falta de insulina provoca lo que se conoce como hiperglucemia, que se traduce como altos niveles de glucosa en la sangre. Este padecimiento puede causar daños en diferentes órganos del cuerpo, lo cual conlleva al desarrollo de problemas de salud e incluso la muerte. No obstante, el tratamiento adecuado de esta enfermedad puede retrasar o prevenir dichos problemas. [1]

En la actualidad, según la Federación Internacional de Diabetes, 425 millones de personas alrededor del mundo viven con diabetes. Tan solo en México, existen aproximadamente 11.5 millones de personas con esta enfermedad [2], sin embargo, la Federación Mexicana de Diabetes [3], estima que esta cifra podría aumentar considerablemente al agregar a aquellas personas que tienen diabetes pero no han sido diagnosticadas aún.

Esta enfermedad produce una gran cantidad de complicaciones en quienes las padecen, y van desde problemas renales y cardiovasculares hasta la muerte. [4] Actualmente existen diversos métodos que permiten detectarla, sin embargo, todos ellos requieren del uso de análisis clínicos y pruebas de laboratorio que ayuden a determinar los niveles de distintas sustancias presentes en nuestra sangre, tales como la glucosa [5]. Aunque este tipo de métodos se han ido refinando con el paso del tiempo con el fin de que cada vez sean más certeros en cuanto a los resultados, tienen algunas desventajas: requieren de cierto tiempo para producirse, pueden llegar a ser lo suficientemente costosos como para no ser accesibles para ciertas personas, requieren de una inversión monetaria importante por parte del gobierno en instituciones públicas [6] y pueden no estar disponibles en zonas específicas en donde los servicios de salud son deficientes o no existen. Un ejemplo claro de esto se ve en México [7], la ineficiencia de las instituciones de salud pública en cuanto a su diagnóstico y tratamiento, indican que necesario implementar nuevas técnicas que sean capaces de proporcionar un diagnóstico temprano sobre los pacientes enfermos con el fin de que sean tratados de forma eficiente y rápida.

De esta forma, con el fin de poder crear un método distinto y más accesible comparado con los estudios de laboratorio, el presente trabajo propone un algoritmo basado en minería de datos e inteligencia artificial (aprendizaje automatizado) para realizar diagnósticos tempranos de diabetes mellitus, llamado clasificador

Bayesiano Ingenuo. Los algoritmos de minería de datos analizan los datos en busca de patrones o tendencias con el propósito de crear modelos predictivos [8], que para este caso en específico, ayudarían a producir diagnósticos de pacientes con y sin diabetes para que a futuro se logre llevar a cabo este mismo análisis de forma predictiva, y así, automatizar hasta cierto punto, los métodos con los que se cuenta hasta ahora para detectar esta enfermedad en mujeres mayores de 21 años; esto debido al tipo de variables que se tomarán en cuenta, las cuales juegan un papel muy importante dentro de esta investigación, ya que, son los factores que permiten detectar la presencia (o ausencia) de la enfermedad en las pacientes dentro del clasificador supervisado bayesiano ingenuo mediante el uso de las variables que se consideran dentro del conjunto de datos utilizado.

Dicho lo anterior, los factores o variables que pueden determinar la detección temprana de la diabetes mellitus en el tipo de pacientes mencionado son el número de embarazos, la concentración de glucosa en la sangre, la presión diastólica sanguínea, el grosor de la piel en el área del tríceps, la insulina sérica, el índice de masa corporal, la función de pedigrí de la diabetes y la edad [9]. Estos factores, mediante su análisis y clasificación, a través del uso del algoritmo bayesiano ingenuo, son capaces de dar un diagnóstico positivo o negativo de esta enfermedad.

La clasificación supervisada permite, a partir de ciertas clases dadas, encontrar una regla para clasificar una nueva observación dentro de las clases existentes, es decir, obtener el valor más probable de una variable hipótesis dados los valores de otras variables predictoras [10]. Por esta razón, debido a que se cuenta con una variable clase que determina el estado positivo o negativo de la enfermedad dentro del conjunto del datos, es necesario aplicar un método de clasificación supervisada para crear las reglas necesarias para clasificar nuevas observaciones y poder brindar una predicción en cuanto al diagnóstico de la diabetes mellitus tipo 2; y dado que, la clasificación supervisada, vista desde un enfoque bayesiano, consiste en asignar a un objeto (diagnóstico) descrito por un conjunto de variables o atributos a una clase (positivo o negativo) [11], el algoritmo se ajusta perfectamente al enfoque que se busca dar dentro del tratamiento de las variables.

2 Antecedentes

Con el paso del tiempo y el surgimiento de diferentes tecnologías se han intentado implementar diferentes métodos que permitan la detección temprana o predicción de la diabetes mellitus en diferentes tipos de pacientes, sin embargo, los parámetros que suelen ser utilizados e incluso las características de los pacientes en los que se busca realizar la predicción pueden presentar grandes variaciones.

Por ejemplo, según el trabajo "Sistema Bayesiano para la Predicción de la Diabetes", publicado en [12]; para poder predecir la diabetes mellitus tipo 2, los autores crearon un sistema bayesiano de predicción en el que se incluyeron las mismas variables que serán contempladas en el presente trabajo y que se mencionaron anteriormente en la sección de "fuente de datos". En él, se incluyen tanto la metodología seguida como los resultados obtenidos.
Mediante el análisis previo de las variables, estas fueron divididas en tres distintos grupos como se muestra a continuación:

> Primer grupo: número de embarazos, presión sanguínea diastólica, espesor del pliegue cutáneo del tripces e índice de masa corporal.
>
> Segundo grupo: número de embarazos, glucosa en plasma, espesor del pliegue cutáneo del triceps, índice de masa corporal y edad.
>
> Tercer grupo: presión sanguínea diastólica, espesor del pliegue cutáneo del triceps, insulina en suero, índice de masa corporal, función de pedigrí de diabetes y edad.

Las pruebas finales fueron realizadas con base en los grupos mencionados anteriormente y muestran el porcentaje de aciertos obtenidos considerando tanto a las pacientes con diabetes como a las que no tienen la enfermedad. De esta forma, el sistema bayesiano propuesto en el artículo obtuvo los siguientes resultados:

> Para el grupo uno se obtuvo en promedio un 87.69% de aciertos.
>
> Para el grupo dos se obtuvo en promedio un 92.31% de aciertos.
>
> Para el grupo tres se obtuvo en promedio un 96.92% de aciertos.

Siendo el grupo tres en el que se obtuvo un mayor porcentaje de aciertos.

Otro de los trabajos realizados dentro de este campo es el titulado "2-Aminoadipic acid is a biomarker for diabetes risk", publicado en [13]. En este se describe otro método de predicción de la diabetes mellitus tipo 2 basado en una única variable o biomarcador conocido como ácido 2-aminoadípico (2-AAA). Según el estudio, se analizaron 376 personas; 188 que desarrollaron diabetes y 188 sin la enfermedad. De esta forma, las personas que tenían concentraciones de 2-AAA en sus organismos presentaban un mayor riesgo de desarrollar diabetes mellitus tipo 2 durante el periodo de seguimiento del estudio (10 años) comparados con los pacientes cuyos niveles de este marcador biológico era más bajo. Cabe mencionar que la forma de medición principal para este caso en específico son análisis clínicos.

Por su parte, el estudio publicado en [14], que lleva como título "Predicción de Diagnóstico de Diabetes Mellitus utilizando Máquinas de Soporte Vectorial en Pacientes de Baja California", menciona que los factores o variables de mayor importancia para producir un diagnóstico de esta enfermedad son edad, índice

de masa corporal y la concentración de glucosa en la sangre. Además, busca dar un diagnóstico temprano haciendo uso de una máquina de soporte vectorial (conjunto de algoritmos de aprendizaje supervisado) cuyo diagnóstico podría ser: sin diabetes, con predisposición a diabetes o con diabetes. Los resultados obtenidos por este método tuvieron una exactitud de 99.2% con pacientes mexicanos y una exactitud de 65.6% con pacientes de un origen étnico distinto.

Además de los trabajos mencionados anteriormente, existen otros que si bien, no están enfocados directamente a la diabetes mellitus, hacen uso de diferentes métodos estadísticos y algoritmos de minería de datos para poder brindar un diagnóstico de diversas enfermedades mediante el análisis previo de las bases de datos correspondientes.

En el artículo publicado en [15], muestra el desempeño de diferentes algoritmos basados en árboles de decisión para evaluar el desempeño de clasificación de datos médicos que puede brindar cada uno de estos con el fin de determinar si esta método de clasificación es realmente una herramienta de soporte y ayuda eficaz en el tratamiento y diagnóstico médico. Dentro de este trabajo se tomaron en cuanta los algoritmos ID3, J48 y Bayes Ingenuo; a partir de estos, se procesaron los datos correspondientes a dos experimentos. En el primero, se dividi ó en dos conjuntos una base de datos de forma aleatoria: 462 elementos para la fase de entrenamiento y 230 para la fase de prueba, obteniendo así los siguientes resultados:

- ID3: 93.04% de casos clasificados de manera correcta.
- J48: 91.73% de casos clasificados de manera correcta.
- Árbol de Bayes Ingenuo: 94.35% de casos clasificados de manera correcta.

Siendo así, el algoritmo de Bayes Ingenuo aplicado a árboles el que obtuvo un mayor porcentaje en cuanto a la clasificación correcta de los casos, aunque los otros dos métodos usados presentaron resultados por encima del 90%, por lo que se pueden considerar bastante buenos también.
En el segundo experimento se usó el mismo conjunto de 462 datos de entrenamiento, sin embargo, se cambió el conjunto de datos de prueba por otro que contaba con 322 datos en total provenientes de una segunda base de datos distinta a la usada en el primer experimento. Este segundo experimento arrojó los siguientes resultados:

- ID3: 82.60% de casos clasificados de manera correcta.
- J48: 81.98% de casos clasificados de manera correcta.
- Árbol de Bayes Ingenuo: 85.71% de casos clasificados de manera correcta.

El algoritmo de Bayes Ingenuo fue de nuevo el que obtuvo mejores resultados, siendo en este caso el mejor de los tres, ya que se encuentra más cerca al 90%

de éxito.

Estos no han sido los únicos trabajos que se han realizado con el fin de poder brindar un diagnóstico temprano o predicción acerca de la diabetes mellitus tipo 2 u otras enfermedades. Como estos, existen muchos otros que utilizan diferentes métodos para lograr su cometido, sin embargo, dicho cometido se trata de un objetivo común entre todos los autores de este tipo de estudios, que es brindar un diagnóstico temprano de la enfermedad a partir de la implementación de diferentes métodos y técnicas relacionadas con tecnologías emergentes como lo es la minería de datos y el aprendizaje automatizado.

Este tipo de estudios surgen a partir de la necesidad de crear alternativas más accesibles y sencillas con respecto a las que se tienen actualmente que requieren de estudios clínicos más exhaustivos; el hecho de contar con una herramienta que de forma automatizada pueda capturar los datos del paciente y arrojar un diagnóstico tentativo en cuestión de segundos puede significar el siguiente paso en cuanto a la creación de diagnósticos tempranos de diversas enfermedades, ya que, debido al conocimiento con el que cuenta el algoritmo gracias a su previo entrenamiento, será posible dar un veredicto con los datos obtenidos en una etapa temprana de la enfermedad en cuestión con el fin de poder tomar decisiones con respecto al tratamiento a seguir de los pacientes para evitar evoluciones negativas y rápidas de la enfermedad o descartar cualquier sospecha de su existencia.

Por esta razón, se busca enfocar la solución propuesta dentro de este artículo a diversas instituciones médicas con el fin de dar un diagnóstico temprano de la diabetes mellitus tipo 2 en mujeres mayores de 21 años para evitar que la enfermedad pueda ser controlada a tiempo y no llegue a una fase terminal por no ser diagnosticada en una fase temprana y así intentar reducir gastos tanto de la parte de las instituciones médicas como de las pacientes con respecto al tratamiento de la enfermedad en sí y de las complicaciones futuras que se pueden presentar debido a esta.

3 Método

La metodología utilizada se compone por los siguientes pasos: 1) Definición de la base de datos a utilizar; 2) Análisis de correlación de las variables contenidas dentro de la base de datos; 3) Elección del algoritmo de clasificación; y 4) Validación.

Paso 1. Definición de la base de datos a utilizar. A partir de la fuente de datos obtenida de [16], es como se definió el objetivo principal del presente artículo de investigación. Dentro de los datos que la componen, se encuentran 768 instancias; cada una representa a una mujer mayor de 21 años, de las cuales 500 pertenecen a la clase 0 (diagnóstico de diabetes negativo) y 268 a la clase 1 (diagnóstico de diabetes positivo). Además, cada una de las instancias posee

ocho atributos que corresponden a los siguientes parámetros o variables:

1. Número de embarazos.
2. Concentración de glucosa en plasma a dos horas en una prueba oral de tolerancia a la glucosa
3. Presión sanguínea diastólica (mm Hg)
4. Grosor del pliegue de la piel del tríceps (mm)
5. Insulina sérica de dos horas (μIU/mL)
6. Índice de masa corporal (peso[kg]/(altura[m])^2)
7. Función de pedigrí de diabetes
8. Edad en años

Estas variables se utilizaron dentro del algoritmo como variables dependientes, es decir, fueron esenciales para llevar a cabo la predicción de la variable independiente *class* que es la que define el estado positivo o negativo del diagnóstico de diabetes mellitus.

Paso 2. Análisis de correlación de las variables contenidas dentro de la base de datos. Para poder determinar si era necesario discriminar alguna de las variables contenidas en la base de datos para reducir la dimensionalidad de esta, fue necesario llevar a cabo un análisis de correlación entre todas las variables dependientes mediante el uso de la herramienta R y su biblioteca *Rattle*. A través de dicho análisis se obtuvieron los siguientes resultados:

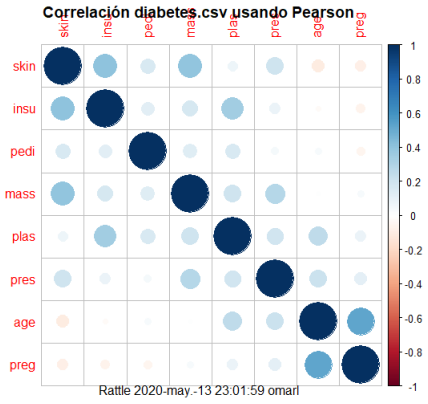

Figura 1: Análisis de correlación de variables gráfico

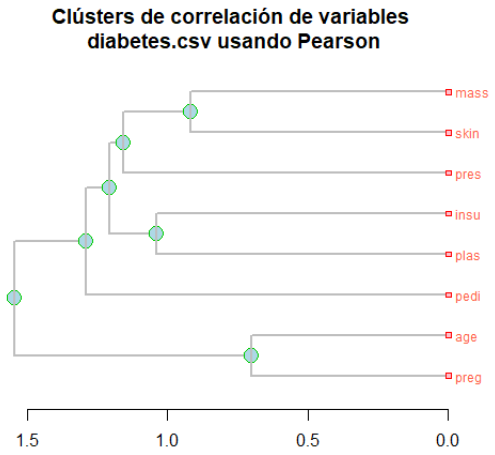

Figura 2: Análisis de correlación de variables jerárquico

Como se puede observar, los resultados obtenidos muestran que no existe una relación lo suficientemente estrecha entre las variables que fueron analizadas para discriminar alguna, por lo cual, se decidió tomar en cuenta todas dentro del algoritmo seleccionado.

Paso 3. Elección del algoritmo de clasificación. Tras determinar que todas las variables serían utilizadas, se determinó el algoritmo que sería utilizado para procesarlas y clasificarlas de la mejor manera. Después de haber realizado diferentes pruebas usando diferentes algoritmos de regresión logística, árboles de decisión y redes neuronales, el algoritmo de clasificación bayesiano ingenuo fue el elegido debido a que presentó el mejor comportamiento en cuanto al tratamiento de las variables tal y como se encuentran estructuradas dentro de la base de datos.
La clasificación supervisada, vista desde un enfoque bayesiano, consiste en asignar a un objeto descrito por un conjunto de variables o atributos, $X_1, X_2, ..., X_n$, a una clase de m clases posibles $(c_1, c_2, c_3, ..., c_m)$ tal que la probabilidad de la clase, según los atributos dados, se maximiza: [17]

$$Argc[MaxP(C|X_1, X_2, ..., X_n)] \qquad \text{(Eq. 1)}$$

Este clasificador toma como base la regla de Bayes para calcular la probabilidad posterior de la clase según los atributos dados:

$$P(C|X_1, X_2, ..., X_n) = \frac{P(C)P(X_1, X_2, ..., X_n|C)}{P(X_1, X_2, ..., X_n)} \qquad \text{(Eq. 2)}$$

Y, debido a que es posible decir que $X = X_1, X_2, ..., X_n$, la Eq. 2 puede reescribirse de la siguiente forma:

$$P(C|X) = \frac{P(C)P(X|C)}{P(X)} \quad \text{(Eq. 3)}$$

Por lo tanto, el problema de clasificación que se basa en la Eq. 1, puede escribirse como:

$$Argc\left[Max\left[P(C|X) = \frac{P(C)P(X|C)}{P(X)}\right]\right] \quad \text{(Eq. 4)}$$

De esta forma, para poder resolver el problema de clasificación a través del uso de un enfoque bayesiano, es necesario contar con la probabilidad a priori de cada clase, $P(C)$, y la probabilidad de los atributos según la clase dada, $P(X|C)$ (*verosimilitud*); para obtener la probabilidad posterior $P(C|A)$. Así, para que este clasificador logre aprender a partir de un conjunto de datos, es necesario estimar las probabilidades mencionadas anteriormente (a priori y verosimilitud), a partir de los parámetros o variables del clasificador. Una alternativa para resolver este problema es considerar relaciones de independencia mediante el clasificador bayesiano ingenuo. [18]

El clasificador bayesiano ingenuo se encuentra basado en la suposición de que todas las variables son independientes dada la clase; por esta razón es que tuvo éxito en el caso de estudio del presente artículo, además de que se reafirma con el hecho de que se encontró una baja correlación entre las variables utilizadas. Matemáticamente hablando, se dice que cada variable o atributo X_i es condicionalmente independiente de todos los demás atributos según la clase dada:

$$P(X_1|X_i, C) = P(X_i|C), \forall j \neq i \quad \text{(Eq. 5)}$$

Al usar este método se obtiene una reducción considerable de la complejidad en cuanto a espacio y tiempo de ejecución comparado con el clasificador bayesiano convencional [17]. Así, dadas las ecuaciones que proporciona el teorema explicado anteriormente en conjunto con este último argumento, fue posible obtener las probabilidades del objeto diagnóstico dadas todas las variables predictoras del conjunto de datos utilizado para determinar a que clase pertenece el diagnóstico; positivo o negativo. Esto con el fin de poder crear un modelo predictor a partir del algoritmo bayes implementado con ayuda de las herramientas y funciones contenidas dentro del lenguaje de programación conocido como R.

Paso 4. Validación. El grupo de muestras contenido dentro de la base de datos definida en el paso 1, fue evaluado a través del modelo generado a partir

de las ecuaciones proporcionadas por el algoritmo definido en el paso 3. A partir del modelo generado, fueron clasificados todas y cada una de las muestras de datos disponibles dentro de alguna de las dos clases posibles: diagnóstico positivo o diagnóstico negativo. Posteriormente, los resultados generados por el clasificador fueron comparados con el respectivo diagnóstico (positivo o negativo) que le fue asignado a cada paciente dentro de la base de datos, es decir, el diagnóstico real. Esto con el fin de conocer el número de aciertos y errores del clasificador y poder establecer el porcentaje de aciertos del modelo.

El sistema de clasificación elegido (bayesiano ingenuo), se encarga de generar una función de probabilidad por cada una de las clases de interés [19]; en este caso, la clase se trata del diagnóstico positivo o negativo. Estas funciones son generadas a partir del análisis y procesamiento del conjunto de datos introducido para crear el modelo que contiene las funciones de probabilidad mencionadas con anterioridad. De esta forma, en caso de que se presenten nuevas entradas referentes a pacientes que no pertenecían originalmente al conjunto de los datos, pueden ser evaluadas a través de las funciones de probabilidad (a priori) contenidas dentro del modelo de predicción del algoritmo utilizado; así, se supone que esta nueva paciente obtendrá un diagnóstico médico (positivo o negativo), condicionado al valor de probabilidad mayor que brindan las funciones del modelo, lo cual se busca que sea implementado de forma temprana con el propósito de poder tomar decisiones importantes con respecto al tratamiento futuro de la paciente en cuestión. [20]

4 Resultados

Del conjunto de 768 muestras con el que se contaba, se llevó a cabo una división de estas en dos grupos:

1. ***Grupo de entrenamiento***: Compuesto por el 80% de los datos (614 muestras).
2. ***Grupo de prueba***: Compuesto por el 20% de los datos (154 muestras).

Ambos grupos se formaron de forma aleatoria e incluyen muestras con diagnósticos tanto positivos como negativos.

```
> data <- diabetes
> sample <- sample.int(n = nrow(data), size = floor(.80*nrow(data)), replace = F)
> train <- data[sample, ]
> test  <- data[-sample, ]
```

Figura 3: Lectura y división del conjunto de datos mediante R

Tras haber realizado la división de los datos, fue necesario construir el modelo de predicción bayes ingenuo utilizando el grupo de entrenamiento.

```
> bayesModel <- naiveBayes(class ~ ., data = train)
> bayesModel

Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
        0         1
0.6596091 0.3403909
```

Figura 4: Creación del modelo bayes ingenuo mediante R

Como se puede observar en la figura 4, la función utilizada muestra las probabilidades a priori del modelo correspondientes a las dos clases disponibles (0 como diagnóstico negativo y 1 como diagnóstico positivo), de las cuales, la probabilidad de que el diagnóstico sea negativo es mucho mayor a su contra parte.

A partir del modelo generado, se llevó a cabo la fase de prueba con el 20% de datos restantes y se evaluaron los resultados mediante el uso de una matriz de confusión.

```
> prediction <- predict(bayesModel, test)
> table(prediction, test$class)

prediction  0  1
         0 87  4
         1  8 55
```

Figura 5: Evaluación del modelo de predicción mediante una matriz de confusión

Con la matriz obtenida fue posible evaluar el desempeño del modelo, cuyos resultados son presentados en la siguiente tabla:

Diagnóstico	Aciertos	Errores
Positivo	55	8
Negativo	87	4

Tabla 1: Resultados de la matriz de confusión

Esto indica que el modelo logró predecir de mejor forma los diagnósticos negativos que positivos, y que debido a que la cantidad de falsos positivos y falsos negativos es pequeña, se puede deducir que es un buen modelo de predicción. Con respecto a la evaluación exhaustiva de cada una de las características que produce la matriz de confusión, se puede decir que el modelo cuenta con una exactitud del 92.2%; precisión del 93.2%; tasa de error del 7.8%; sensibilidad del 87%; y especificidad del 95.6%.

Estos resultados muestran que, al tomar en cuenta todas las variables contenidas dentro de cada una de las muestras del conjunto de datos, de obtiene un buen

número de aciertos, y por lo tanto, una buena exactitud por parte del modelo generado, ya que se encuentra por encima del 90% (92.2%). Por esta razón, el modelo propuesto muestra un hallazgo importante mediante el cual es posible realizar una predicción acertada del diagnóstico de la diabetes mellitus tipo 2 en mujeres mayores de 21 años, con la cual se busca detectar la enfermedad de forma temprana para evitar que evolucione y traiga consigo diversas complicaciones. Esto con el fin de poder disminuir el costo que genera esta enfermedad tanto en las pacientes como en las instituciones de salud que las atienden y, a su vez, de generar una alternativa que busca ser más accesible y rápida comparada con los métodos con los que se cuenta actualmente, de los cuales, muchos de ellos requieren de pruebas de laboratorio.

5 Conclusiones y trabajo futuro

A partir de los resultados obtenidos, el clasificador bayesiano ingenuo para identificar mujeres mayores de 21 años con diabetes mellitus tipo 2, se propone como una alternativa para la detección temprana de esta enfermedad con el fin de evitar análisis clínicos y de laboratorio de las características que se utilizan para detectar la enfermedad. De forma precisa, el nivel de acierto del modelo propuesto es del 92.2% tomando en cuenta las ocho características que se incluyen dentro de la base de datos analizada. Lo cual indica que este modelo podría llegar a ser una alternativa confiable y eficiente en cuanto a la detección temprana de esta enfermedad.

Según el programa de acción para la prevención y control de la diabetes mellitus publicado por el Centro Nacional de Programas Preventivos y Control de Enfermedades para los años 2013 al 2018 [21], de los 6.4 millones de la población adulta que padece diabetes en México, solo el 9.2% tenía un diagnóstico previo. Hablando únicamente de mujeres, se estima que de los 3.65 millones de mujeres mexicanas que padecen diabetes, solo el 9.67% tiene un diagnóstico previo. Por esta razón, es necesario incluir nuevas alternativas en cuanto a la detección temprana de distintos padecimientos médicos con el fin de evitar pérdidas humanas, evoluciones de la enfermedad y gastos médicos generales. Conforme a esta misma fuente [21], la atención y tratamiento a esta enfermedad consume el 15% del total de los recursos del sistema mexicano de salud, lo cual representa otro punto importante para buscar alternativas tecnológicas que permitan disminuir este porcentaje mediante la implementación de herramientas tales como el modelo presentado a lo largo del presente artículo.

A futuro, este modelo podría ser expandido para que realice diagnósticos a pacientes del sexo masculino, así como también, su constante alimentación y entrenamiento provocarían una mayor certeza en cuanto a los resultados que pueda brindar. Y, aunque esta versión se trata únicamente de un prototipo, podría ser refinado al grado de establecer una plataforma web que incluya el algoritmo para que este sea más gráfico y sencillo de utilizar para aquel personal

que no esté familiarizado con las herramientas utilizadas en su construcción original.

Referencias

[1] I. D. Federation, *Atlas de la DIABETES de la FID*. Chaussée de La Hulpe 166 B-1170 Bruselas, Bélgica: International Diabetes Federation, eighth ed., 2017.

[2] I. D. Federation, *Atlas de la DIABETES de la FID*. Chaussée de La Hulpe 166 B-1170 Bruselas, Bélgica: International Diabetes Federation, seventh ed., 2015.

[3] F. M. de Diabetes, "Diabetes en méxico," oct 2014.

[4] C. E. de Vigilancia Epidemiológica y Control de Enfermedades, *Reporte Diabetes Mellitus*, 2013.

[5] J. Medina, E. D. de León, G. Troncoso, *et al.*, *Diagnóstico y Tratamiento de Diabetes Mellitus en el Adulto Mayor Vulnerable*. Instituto Mexicano del Seguro Social, Durango 289- 1A Colonia Roma Delegación Cuauhtémoc, 06700 México, Ciudad de México, 2013.

[6] M. Barranza, M. Guajardo, J. Picó, *et al.*, *Carga Económica de la Diabetes Mellitus en México, 2013*. Funsalud, primera ed., jun 2015.

[7] M. Jiménez, "Sistema de salud pública, un problema más para los mexicanos," oct 2016.

[8] B. Beltrán, "Minería de datos." Definición consultada en la página 18 del documento.

[9] A. Lifshitz, "Diabetes mellitus."

[10] E. Sucar, "Sesión 6: Clasificadores bayesianos." Definición consultada en la página 18 del documento.

[11] P. Oscar and R. Casillas, "Aprendizaje bayesiano.".

[12] O. Castrillón, W. Sarache, and E. Castaño, "Sistema bayesiano para la predicción de la diabetes," *Información Tecnológica*, vol. 28, pp. 161–168, diciembre 2017.

[13] T. Wang, D. Ngo, N. Psychogios, *et al.*, "2-aminoadipic acid is a biomarker for diabetes risk," *The Journal of Clinical Investigation (JCI)*, vol. 123, oct 2013.

[14] B. Benítez, C. Castro, R. Castañeda-Martínez, and A. Abaroa, "Predicción de diagnóstico de diabetes mellitus utilizando máquinas de soporte vectorial en pacientes de baja california," *Memorias del Congreso Nacional de Ingeniería Biomédica*, vol. 4, no. 1, pp. 415–418, 2017.

[15] R. Barrientos, N. Cruz, H. Acosta, *et al.*, "Árboles de decisión como herramienta en el diagnóstico médico," *Revista Médica de la Universidad Veracruzana*, pp. 20–24, jul 2009.

[16] D. Dheeru and G. Casey, "UCI machine learning repository," 2017.

[17] L. Sucar, "Clasificadores bayesianos: de datos a conceptos."

[18] I. Hernández, *Clasificador Bayesiano Ingenuo en RapidMiner*. PhD thesis, Benemérita Universidad Autónoma de Puebla, México, Puebla, sep 2016.

[19] E. Morales and H. Escalante, "Aprendizaje bayesiano."

[20] J. García, "Clasificadores bayesianos."

[21] CENAPRECE, *Prevención y Control de la Diabetes Mellitus 2013-2018*, 2013.

[22] Microsoft, "Algoritmos de minería de datos (analysis services: Minería de datos)," mar 2017.