ISSN 1870-4069

Hacia la predicción de pacientes con diabetes: Comparación de algoritmos de aprendizaje automático

Edgar García-Quezada, Carlos E. Galván-Tejada,
José M. Celaya-Padilla, Irma González-Curiel,
Jorge I. Galván-Tejada

Universidad Autónoma de Zacatecas,
Unidad Académica de Ingeniería Eléctrica,
México

{39207895, ericgalvan, jose.celaya,
irmacuriel, gatejo}@uaz.edu.mx

Resumen. Según la organización mundial de la salud, la diabetes mellitus es una enfermedad crónico degenerativa que aparece cuando el páncreas no secreta insulina suficiente o cuando el organismo no la utiliza de manera adecuada. Para el año 2019, en México, existían 12.8 millones de personas con diabetes. Los síntomas iniciales de la enfermedad se relacionan con la hiperglucemia e incluyen polidipsia, polifagia, poliuria y visión borrosa. El diagnostico de diabetes por parte de un médico puede ser complicado, debido a que intervienen varios factores, entre ellos el examen de sangre, que, aunque es barato, no proporciona información suficiente para realizar el diagnostico, además de ser un método invasivo. El objetivo de este artículo es analizar datos de pacientes diabéticos y personas sanas para hacer un diagnostico temprano por medio de sintomatología de la enfermedad, utilizando para esto los algoritmos: Regresión logística múltiple, Máquina de vectores de soporte y K vecinos más cercanos y evaluando cual de ellos es el mejor utilizando como métrica de evulación el área bajo la curva.

Palabras clave: Diabetes mellitus, algoritmos de aprendizaje automático, diagnóstico, área bajo la curva.

Towards the Prediction of Patients with Diabetes: Comparison of Machine Learning Algorithms

Abstract. According to World Health Organization, diabetes mellitus is a chronic degenerative disease that appears when the pancreas does not secrete enough insulin or even the body does not use it properly. By 2019, In Mexico, there were 12.8 million people with diabetes. The initial symptoms of the disease are related to hyperglycemia and include polydipsia, polyphagia, polyuria, and

blurred vision. Diagnosis of diabetes by a doctor is complicated because several factors are involved, including the blood test, which, although cheap, does not provide enough information to make the diagnosis, in addition to being an invasive method. The objective of this paper is to analyze data from diabetic patients and health people to make an early diagnosis, employing symptoms, of the disease using this the algorithms: Multiple logistic regression, Support Vector Machine and K nearest neighbors and evaluating which of them is more optimal using the area under the curve as the main indicator.

Keywords: Diabetes mellitus, machine learning algorithms, diagnosis, area under the curve.

1. Introducción

La diabetes mellitus es una alteración metabólica que se caracteriza por la presencia crónica de hiperglucemia, acompañada por alteraciones en diferentes niveles del metabolismo de hidratos de carbono, proteínas y lípidos.

Aunque esta enfermedad tiene un origen variado, todos los casos conllevan alteraciones en la secreción de insulina, en la sensibilidad a la acción de la hormona, o en ambas, en algún momento de su historia natural.

Cabe destacar que la sintomatología de la enfermedad puede llevar a dos situaciones importantes. En la primera, los síntomas son evidentes, persistentes y las cifras de glucemia suficientemente elevadas, lo que hace que el diagnóstico sea obvio en la mayoría de las ocasiones. En el segundo caso, el paciente podría ser asintomático y requerir una exploración analítica de rutina [10].

Es debido a lo mencionado anteriormente, junto con las complicaciones específicas y la presencia de otros factores asociados a la diabetes mellitus, que esta enfermedad se ha convertido en un grave problema en la actualidad [10].

En México, la diabetes es una enfermedad común en la población desde el año 2000 [22]. Según el Instituto Nacional de Salud Pública, para el 2010 la enfermedad ya había afectado a 83,000 personas [22]. Para el año 2019, la Federación Internacional de Diabetes reportó que en México había 12.8 millones de personas con diabetes [8].

Aunque el método más común para diagnosticar la enfermedad es por medio de pruebas de sangre baratas [3], este no es suficiente para una valoración completa. Por lo tanto, esta investigación se enfoca en buscar un diagnóstico no invasivo para la diabetes mellitus a través de síntomas que puedan acompañar a la enfermedad.

Para ello, se ha implementado algoritmos de aprendizaje automático en una base de datos de pacientes con diabetes tipo 2 (DT2), elaborada por el Sylhet Diabetes Hospital de Sylhet, Bangladesh [11].

El diagnóstico de la diabetes mellitus puede resultar complicado debido a las características cambiantes de la enfermedad y a los posibles errores humanos al momento de identificarla. Los análisis de laboratorio, aunque son indicadores fuertes de la patología, pueden ser interpretados de forma errónea. Por esta causa, en los últimos años se ha buscado una forma más precisa de identificar a los pacientes que padecen la enfermedad.

Fig. 1. Diferentes etapas y pasos en la implementación de los algoritmos, desde la selección de la base de datos hasta la evaluación de los resultados obtenidos.

En el estudio de Benítez, et al. [4], se habla acerca de la predicción de este padecimiento implementando Máquinas de vectores de soporte en pacientes de Baja California.

Su estudio demostró una exactitud del 99.2% en pacientes mexicanos, utilizando como indicadores el índice de masa corporal y la concentración de glucosa en la sangre. En el artículo escrito por Sisodia, D. S. [24] el objetivo es predecir diabetes utilizando varios métodos (árboles de decisión, máquina de vectores de soporte y Naive Bayes), en sus resultados se muestra que, para sus predicciones, el método Naive Bayes es el mejor, ya que tiene una exactitud del 76.3%.

En la propuesta de Liao-Li, et al. [15], el objetivo es ayudar en la prevención y diagnóstico de la enfermedad, así como poder predecir complicaciones que puedan surgir durante el control del mencionado padecimiento.

Para ello, se utilizó la Regresión Lineal Múltiple, la Regresión Logística, Bosques aleatorios y los K vecinos más cercanos. Su estudio demostró que, aunque el Bosques aleatorios obtuvo la mayor área bajo la curva, al ser un modelo no supervisado, se consideró como mejor modelo los K vecinos más cercanos.

La principal contribución de este artículo es analizar datos de pacientes con diabetes mellitus y los principales síntomas asociados a esta enfermedad, así como pacientes sanos, con el objetivo de realizar un diagnóstico temprano del padecimiento por medio de sintomatología, utilizando para esto los algoritmos de aprendizaje automático.

El presente artículo está organizado de la siguiente manera: la primera sección muestra la introducción y estudios relacionados con esta investigación; en la segunda sección se presentan los datos utilizados y la metodología para su tratamiento con el fin de predecir la diabetes; en la tercera sección se describen los resultados obtenidos, así como las comparaciones entre los métodos utilizados; y la cuarta sección tiene como objetivo presentar las conclusiones del estudio realizado.

Table 1. Descripción de atributos de la base de datos.

Atributos	Valores
Edad	20-90
Sexo	1.Masculino, 2. Femenino
Poliuria	1. Si, 2. No
Polidipsia	1. Si, 2. No
Pérdida de peso repentina	1. Si, 2. No
Debilidad	1. Si, 2. No
Polifagia	1. Si, 2. No
Candidiasis	1. Si, 2. No
Vista borrosa	1. Si, 2. No
Purito	1. Si, 2. No
Irritabilidad	1. Si, 2. No
Curación tardía	1. Si, 2. No
Paresis parcial	1. Si, 2. No
Rigidez muscular	1. Si, 2. No
Alopecia	1. Si, 2. No
Obesidad	1. Si, 2. No

2. Materiales y métodos

En la presente sección se muestra la metodología que se llevó a cabo para cada uno de los algoritmos (Fig. 1), así como la descripción de la base de datos utilizada.

Para la evaluación de los algoritmos seleccionados se utilizaron datos provenientes del repositorio de la Universidad de California, Irvine (UCI) Machine Learning. Este conjunto de datos contiene información sobre pacientes, incluyendo síntomas y signos característicos que influyen en el desarrollo de la diabetes.

Los datos se obtuvieron mediante un cuestionario aplicado a pacientes recién diagnosticados con la enfermedad, así como a personas que presentaban algunos de los síntomas aunque no la padecían. El cuestionario fue aplicado directamente a pacientes del Sylhet Diabetes Hospital de Sylhet, Bangladesh.

El conjunto de datos consta de 16 características que representan los síntomas que podrían tener pacientes con diabetes, así como la clasificación de las personas que padecen (Positivo) o no (Negativo) la enfermedad [11]. La descripción de los atributos se muestra en la Tabla 1.

2.1. Definición de los atributos

- Edad: Es la edad de los pacientes encuestados.
- Sexo: Es el sexo de los pacientes encuestados.
- Poliuria: Es la emisión de un volumen de orina superior al esperado [2].
- Polidipsia: Aumento anormal de sed que lleva a una persona a ingerir grandes cantidades de agua [2].
- Pérdida de peso repentina: Pérdida de peso, en un periodo corto de tiempo, sin explicación o causa aparente [2].

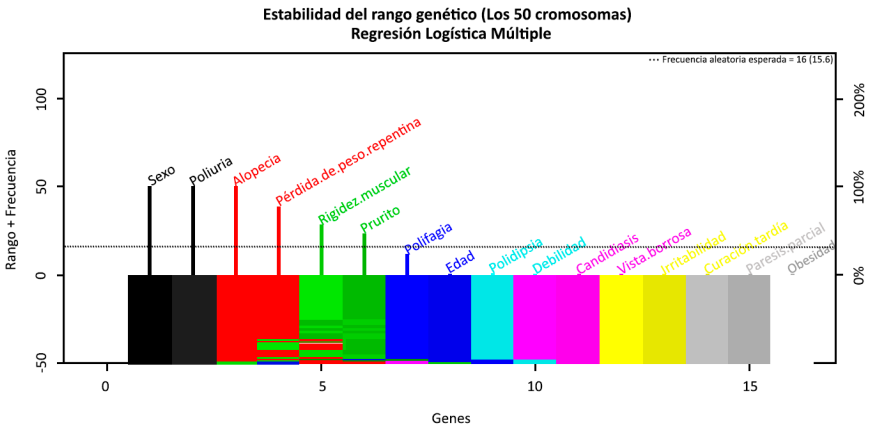


Fig. 2. Gráfica de estabilidad de genes para el modelo de regresión logística múltiple.

- Debilidad: Se refiere a la pérdida de la fuerza muscular, es decir, la persona afectada no puede mover un músculo normalmente a pesar de intentarlo con todas sus fuerzas [14].
- Polifagia: Aumento anormal de la necesidad de querer comer [25].
- Candidiasis: Es un tipo de infección fúngica causada por un hongo denominado cándida [7].
- Vista borrosa: es la pérdida de la agudeza visual, lo que hace que los objetos aparezcan fuera de foco y con opacidad [29].
- Prurito: Sensación incómoda que crea deseo de rascarse [1].
- Irritabilidad: Es un estado emocional en el que una persona tiene un temperamento explosivo y se molesta o enoja fácilmente [28].
- Curación tardía: Es cuando una herida presenta dificultad para cicatrizar o permanecer cerrada [6].
- Paresis parcial: Es la ausencia parcial de movimiento voluntario, la parálisis parcial o suave, descrito generalmente como debilidad del músculo [19].
- Rigidez muscular: Se refiere a músculos tensos o rígidos [21].
- Alopecia: Pérdida anormal de cabello [12].
- Obesidad: Es el exceso de acumulación de grasa en el cuerpo [27].

En cuanto al preprocesamiento de la base de datos, se realizó una verificación de los datos faltantes en las propiedades existentes en la primera etapa. Luego, se procedió a binarizar las 15 características y normalizar la información. Se decidió normalizar la característica Edad, que es un dato continuo, para evitar que afecte el rendimiento de los modelos predictivos. También se evaluó la distribución de los diagnósticos de los pacientes, y se encontró que había 320 casos positivos y 200 negativos, lo que indica una buena distribución entre ellos.

La base de datos se particionó en dos, utilizando el 70% de los datos, seleccionados aleatoriamente, como conjunto de entrenamiento y el 30% restante como conjunto de prueba.

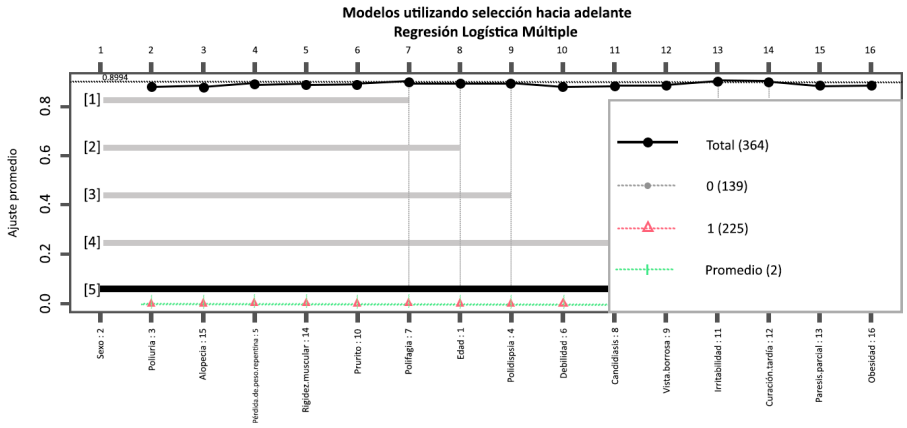


Fig. 3. Gráfica que muestra el mejor modelo usando selección hacia adelante para regresión logística múltiple.

Esto se realizó de esta manera ya que entrenar el modelo con la totalidad de los datos no nos permitiría saber si éste se esta comportando adecuadamente, ya que podría estarse prestando a un sobreajuste, es decir, que el modelo esté describiendo los datos perfectamente bien sin ser realmente adecuado.

Por eso se emplea un número de datos considerablemente grande y aleatorio para asegurarse que se describa lo mejor posible la información y lo restante se utilizará para llevar a cabo las pruebas y que se pueda asegurar que el modelo esté trabajando de forma adecuada [13].

2.2. Método de selección de variables

Para la selección de características, el procedimiento implementado es GALGO, el cual es un paquete que implementa algoritmos genéticos para la resolución de problemas de optimización, estos implican la selección de subconjuntos de variables e incluye una serie de métodos para realizar la clasificación supervisada [26].

Para cada algoritmo se construyó un modelo de selección hacia adelante, esto es un procedimiento de elección de características por pasos en el que las variables se introducen secuencialmente en el modelo.

El primer atributo considerado para la entrada en la ecuación es el que tiene mayor correlación positiva o negativa con la variable dependiente [23].

Por último, se hizo implementó una eliminación hacia atrás, esto es, un procedimiento de selección de variables en el que todas las características se introducen en la ecuación y luego se eliminan secuencialmente. El atributo con la correlación parcial más pequeña respecto a la variable dependiente se considera en primer lugar para la eliminación [20].

Table 2. Selección de características para cada modelo.

Algoritmo	Caracterśticas
Regresíon Logística Múltiple	Sexo, Poliuria, pérdida.de.peso.repentina, debilidad, Prurito, Irritabilidad, Rigidez.muscular, Alopecia.
Máquinas de vectores de soporte	Edad, Sexo, Poliuria, pérdida.de.peso.repentina, Polifagia, Rigidez.muscular, Alopecia.
K vecinos más cercanos	Poliuria, Sexo, curación.tardía, Alopecia, Polidipsia, vista.borrosa.

2.3. Algoritmos

Para desarrollar este estudio, se utilizaron tres algoritmos de aprendizaje automático: Regresión Logística Múltiple, Máquina de Soporte de Vectores y K vecinos más cercanos.

Estos algoritmos fueron seleccionados por sus diferentes enfoques de análisis. La Regresión Logística Múltiple se utiliza para predecir la probabilidad de diferentes resultados posibles de una distribución categórica, dada un conjunto de variables independientes.

La Máquina de Soporte de Vectores correlaciona los datos en un espacio de características de grandes dimensiones, lo que permite categorizar los puntos de datos incluso si no se pueden separar linealmente. Por último, K vecinos más cercanos busca las distancias entre una consulta y todos los ejemplos de los datos, seleccionando los K ejemplos más cercanos a la consulta y votando por la etiqueta más frecuente.

Regresión logística múltiple. La regresión logística múltiple (RLM) es una extensión del modelo de regresión logística simple en el que se predice una respuesta binaria en función de múltiples predictores, que pueden ser tanto continuos como categóricos [5]:

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}}, \tag{1}$$

donde $X = X_1, ..., X_p$ son los p predictores [5], como se presenta en la ecuación 1.

Máquinas de vectores de soporte. Las Máquinas de Vectores de Soporte (SVM) permiten encontrar la forma óptima de clasificar entre varias clases. La clasificación óptima se realiza maximizando el margen de separación entre las clases. Los vectores que definen el borde de esta separación son los vectores de soporte [9].

K vecinos más cercanos. El método los k vecinos más cercanos (KNN) es uno de los métodos más importantes de clasificación supervisada. En el proceso de aprendizaje no se hace ninguna suposición acerca de la distribución de las variables predictoras, es por ello que es un método de clasificación no paramétrico, que estima el valor de la función de densidad de probabilidad o directamente la probabilidad posterior de que un elemento pertenezca a la clase va partir de la información proporcionada por el conjunto de entrenamiento [17].

Table 3. Resultados de las métricas de validación para los modelos seleccionados.

Métrica	RLM	SVM	KNN
Precisión	0.8782	0.8526	0.8856
Sensibilidad	0.902	0.8750	0.9462
Especificidad	0.8333	0.8077	0.7937
Área bajo la curva	0.8645	0.8323	0.8902

El grado de cercanía entre dos tuplas $X_1 = (X_{11}...X_{1n})$ y $X_2 = (X_{21}...X_{2n})$ está basado en la distancia Euclidiana. El modelo matemático se describe en la ecuación (2):

$$dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n}(x_{1i} - x_{2i})^2}. \tag{2}$$

2.4. Validación de modelos

Con el objetivo de analizar la capacidad de clasificación de cada modelo, una vez obtenidos, se emplearon indicadores de validación. Las medidas de validación empleadas son las siguientes:

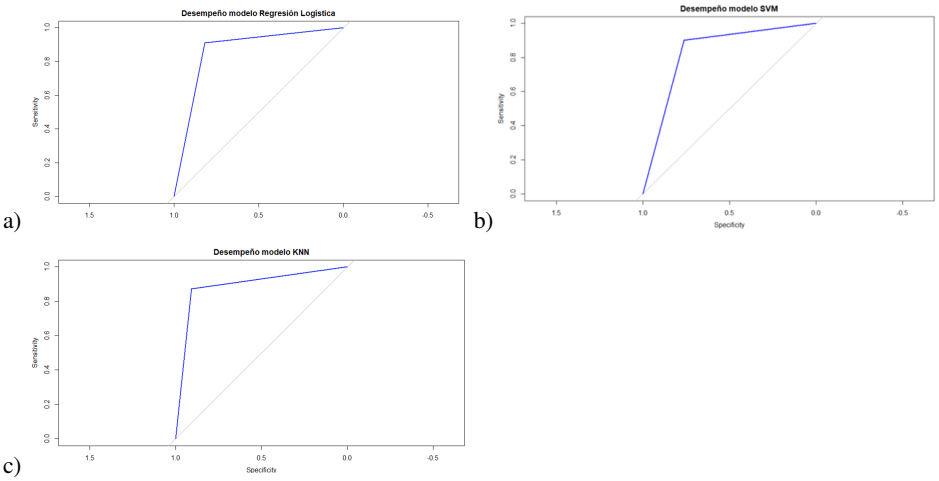
Curvas ROC
El análisis de la curva característica de operación del receptor (ROC) es una herramienta muy utilizada para el análisis del rendimiento del modelo de clasificación, ya que permite evaluar la sensibilidad y la especificidad de un modelo a diferentes umbrales de decisión.

La curva ROC representa la tasa de verdaderos positivos (sensibilidad) en función de la tasa de falsos positivos (1 - especificidad), lo que aporta información suficiente para elegir el mejor umbral y minimizar tanto los falsos positivos como los falsos negativos [16].

Área bajo la curva
El área bajo la curva (AUC), se utiliza como resumen de la rentabilidad del modelo, es decir, cuanto más esté hacia la izquierda la curva, más área habrá contenida bajo ella y, por ende, mejor será el clasificador. El clasificador aleatorio tendría una AUC de 0.5 mientras que el clasificador perfecto tendría el valor de 1 [17].

Sensibilidad y especificidad
La sensibilidad es una métrica que nos permite visualizar el desempeño del modelo, es decir, es la proporción de casos positivos que fueron correctamente identificados por el algoritmo, cuyo modelo se describe en la ecuación (3). Por otra parte, la especificidad se refiere a los casos negativos que se han clasificado correctamente, expresado en la ecuación (4). Además, expresa qué tan bien el modelo puede detectar esa clase [18]:

$$Sensibilidad = \frac{VP}{VP + FN}, \tag{3}$$



a) b) c)

Fig. 4. Resultados de las pruebas realizados para el 30% de los datos restantes utilizando validación cruzada. Para RLM (a), el área bajo la curva es de 0.8645; Para SVM (b), el área es de 0.8323 y para KNN (c) es de 0.8902.

$$Especificidad = \frac{VN}{VN + FP}, \tag{4}$$

donde VP son los verdaderos positivos; VN, los verdaderos negativos; FP, los falsos positivos y FN, los falsos negativos.

3. Resultados

De los resultados obtenidos con el entrenamiento de GALGO, utilizando el 70% de la base de datos, la gráfica 2 muestra las características ordenadas de izquierda a derecha por su grado de importancia en el modelo de predicción. Además, se puede observar la estabilidad de cada característica, lo que significa que una característica no cambiará de color si no cambia su estado (es decir, si se mantiene importante para el modelo mezclándose con otras características). Lo anterior se realiza para todos los algoritmos propuestos y para obtener de cada uno de ellos las características más descriptivas de la variable dependiente, en este caso, pacientes con o sin diabetes. Cabe mencionar que para el algoritmo KNN se utilizó K = 1 y para SVM el kernel seleccionado fue radial.

Después se construyó un modelo hacia adelante utilizando las características que se encontraron con GALGO. En las gráficas del modelo hacia adelante se puede observar cómo se seleccionan las variables más importantes al principio y se van agregando para crear el mejor modelo. Como ejemplo, para Regresión Logística Múltiple, se puede observar dicho comportamiento en la figura 3.

Para finalizar el proceso de selección de características, se implementó una eliminación hacia atrás tomando como base los modelos con mejor rendimiento en la parte de selección hacia adelante, eliminado las variables que menos aportan el resultado final; para cada modelo se muestran en la Tabla 2.

Table 4. Métricas de validación para el método de ensamble por mayoría de votos.

Métrica	Ensamble por mayoría de votos
Precisión	0.8910
Sensibilidad	0.9286
Especificidad	0.8276
ROC	0.8869

Como parte final del proceso se realizaron las pruebas de los modelos seleccionados con las características más significativas y estables arrojadas por GALGO. Las pruebas se realizaron con el 30% restante de la base de datos con una validación cruzada. Los resultados de estas pruebas se pueden observar en la Tabla 3 y las curvas ROC que se muestran en la figura 4.

Como se puede observar en la tabla 3 los modelos se comportan de manera similar, sin embargo el mejor desempeño con base a los objetivos del presente trabajo, es el modelo de los K vecinos más cercanos, pues este tiene la mejor precisión y la mayor sensibilidad, siendo esta última un apartado relevante tratándose de la salud de los pacientes, dado que es de especial importancia para la salud que a una persona que no tenga diabetes, no se le clasifique por error como enferma y que a una que tenga la enfermedad, se le clasifique erróneamente como sana.

Lo siguiente fue utilizar el método de ensamble por mayoría de votos para los resultados de los modelos y así validar la posible mejora del desempeño de estos, resultados que se muestran en la tabla 4 y en la figura 5.

En la tabla 4 se observa que la precisión mejoró aproximadamente un 1% en comparación con el mejor modelo, que fue KNN, mientras que la sensibilidad disminuyó alrededor de un 2%. Por lo tanto, el modelo de los K vecinos más cercanos sigue siendo el mejor método.

4. Conclusiones

Benítez utilizó Máquinas de Vectores de Soporte para predecir la diabetes en pacientes de Baja California, México. Su trabajo concluyó que la exactitud de este método en su base de datos demostró ser del 99.2%, utilizando el índice de masa corporal y la glucosa en sangre como indicadores. Por otro lado, Sisodia, D. S. encontró que el algoritmo Naive Bayes ofrece la mayor exactitud en la predicción de la diabetes, con un 76.3%, tras evaluar varios métodos. En su estudio, Liao-Li propuso diagnosticar la diabetes en pacientes mexicanos utilizando diferentes algoritmos. Llegó a la conclusión de que el algoritmo K vecinos más cercanos proporciona la mayor área bajo la curva y una exactitud del 87.31%.

De acuerdo con los resultados obtenidos por los modelos analizados en este estudio, el modelo K vecinos más cercanos tiene el mejor comportamiento, dada su precisión de 0.8846, una área bajo la curva de 0.8902 y la curva ROC muestra una sensibilidad de 0.9462, siendo una característica importante para este estudio, pues clasifica con exactitud a los pacientes que están enfermos. Logra este rendimiento con las siguientes características: poliuria, sexo, curación tardía, alopecia, polidipsia y vista borrosa. Estas modelan al conjunto de prueba, por lo que es posible hacer el diagnóstico inicial de forma no invasiva, cumpliendo con el objetivo general del estudio desarrollado.

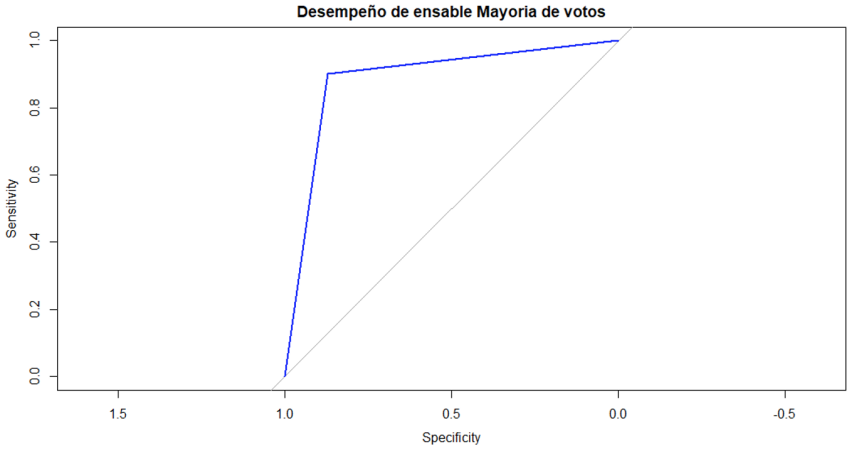


Fig. 5. Curva ROC para el modelo ensamble por mayoría de votos.

Además, de los modelos analizados, el que tiene menor desempeño fue el de máquina de vectores de soporte, pues éste cuenta con una precisión de 0.8526, un área bajo la curva de 0.8323 y la curva ROC muestra una sensibilidad de 0.8750. Sin embargo es importante mencionar que SVM requiere solamente de 7 características para lograr dicho desempeño; una característica menos que las requeridas por la Regresión Logística Múltiple.

La técnica de ensamble mostró una precisión de 0.891, una área bajo la curva de 0.8869 y una curva ROC que muestra una sensibilidad de 0.9286, aunque muestra un pequeño aumento en la precisión, disminuye el rendimiento en otras métricas, por esto no se le considera una técnica adecuada para este caso.

El resultado del modelo de K vecinos más cercanos puede ser utilizado para desarrollar una herramienta auxiliar en el diagnóstico médico, pues permitiría hacer un acercamiento al diagnóstico temprano de diabetes, siendo este corroborado después por un experto de la salud.

Como trabajo a futuro se propone implementar otros algoritmos que aborden el problema de la clasificación de los pacientes de distintas maneras, permitiendo obtener mejores resultados, además de otras técnicas de ensamble que ayuden a mejorar el rendimiento de los algoritmos por sí solos.

También se pueden explorar otras variables que no se incluyeron en este estudio, para ver si tienen un impacto significativo en la clasificación de los pacientes diabéticos. Por ejemplo, se pueden considerar variables relacionadas con el estilo de vida de los pacientes, como su nivel de actividad física, su dieta y su nivel de estrés.

Agradecimientos. Los autores agradecen al Consejo Nacional de Ciencia y Tecnología (CONACyT) y a la Universidad Autónoma de Zacatecas por el apoyo y financiamiento para la realización de este proyecto.

References

1. Alcalá-Pérez, D., Barrera-Pérez, M., Santa-Cruz, F.: Fisiopatología del prurito. Revista del Centro Dermatológico Pascua, vol. 23, no. 1, pp. 6–10 (2014)
2. Almaguer-Herrera, A., Miguel-Soca, P., Reynaldo-Será, C., Mariño-Soler, A. L., Oliveros-Guerra, R.: Actualización sobre diabetes mellitus. Correo Científico Médico, vol. 16, no. 2 (2012)
3. American Diabetes Association: Classification and diagnosis of diabetes. Diabetes Care, vol. 39, no. 1, pp. 13–22 (2016) doi: 10.2337/dc16-S005
4. Benítez, B., Castro, C., Castañeda-Martínez, R. A., Abaroa, A.: Predicción de diagnóstico de diabetes mellitus utilizando máquinas de soporte vectorial en pacientes de Baja California. In: Memorias Del XL Congreso Nacional De Ingeniería Biomédica, vol. 4, no. 1, pp. 415–418 (2017)
5. Berea-Baltierra, R., Rivas-Ruiz, R., Pérez-Rodríguez, M., Palacios-Cruz, L., Moreno, J., Talavera, J. O.: Investigación clínica XX del juicio clínico a la regresión logística múltiple. Revista Médica del Instituto Mexicano del Seguro Social, vol. 52, no. 2, pp. 192–197 (2014)
6. Berlanga-Acosta, J., Valdez-Pérez, C., Savigne-Gutierrez, W., Mendoza-Marí, Y., Franco-Perez, N., Vargas-Machiran, E., Poll-Marrón, N., Alvarez-Duarte, H., Echeverria-Requeijo, H., Perez-Aguilar, R. M.: Particularidades celulares y moleculares del mecanismo de cicatrización en la diabetes. Biotecnología Aplicada, vol. 27, no. 4, pp. 255–261 (2010)
7. Biasoli, M.: Candidiasis (2013) http://www.fbioyf.unr.edu.ar/evirtual/file.php/118/MATERIALES_2013/TEORICOS_20
8. Centro de Investigación en Alimenación y Desarrollo CIAD: La pandemia de diabetes en México (2020) https://www.ciad.mx/notas/item/2450-la-pandemia-de-diabetes-en-mexico
9. Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., Lopez, A.: A comprehensive survey on support vector machine classification: Applications, challenges and trends. Neurocomputing, vol. 408, pp. 189–215 (2020) doi: 10.1016/j.neucom.2019.10.118
10. Conget, I.: Diagnóstico, clasificación y patogenia de la diabetes mellitus. Revista Española de Cardiología, vol. 55, no. 5, pp. 528–535 (2002) doi: 10.1016/S0300-8932(02)76646-3
11. Faniqul-Islam, M. M., Ferdousi, R., Rahman, S., Bushra, H. Y.: Likelihood prediction of diabetes at early stage using data mining techniques. Computer Vision and Machine Intelligence in Medical Image Analysis, Advances in Intelligent Systems and Computing, vol. 992, pp. 113–125 (2020) doi: 10.1007/978-981-13-8798-2_12
12. Forouzan, P., Cohen, P. R.: Incipient diabetes mellitus and nascent thyroid disease presenting as beard alopecia areata: case report and treatment review of alopecia areata of the beard. Cureus, vol. 12, no. 7 (2020) doi: 10.7759/cureus.9500
13. Gholamy, A., Kreinovich, V., Kosheleva, O.: Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation. Departmental Technical Reports (CS) (2018)
14. Hasbum-Fernández, B., Ampudia-Blasco, F. J., Carmena, R.: Paciente diabético con debilidad muscular generalizada. Endocrinología y Nutrición, vol. 50, no. 5, pp. 169–170 (2003) doi: 10.1016/S1575-0922(03)74521-5
15. Liao-Li, M. K., Aparicio-Montelongo, I., Celaya-Padilla, J. M., Galván-Tejada, C. E., Cruz, M.: Implementación de algoritmos de aprendizaje automático para la predicción de pacientes con diabetes. Investigación Aplicada, un Enfoque en la Tecnología, vol. 6, no. 12, pp. 609–621 (2021)
16. Martínez-Pérez, J. A., Pérez-Martin, P. S.: La curva roc. Medicina de Familia. SEMERGEN, vol. 49, no. 1 (2023) doi: 10.1016/j.semerg.2022.101821
17. Meneses-Villegas, C., Aqueveque, D.: Diagnosis of neuropathies in diabetic patients by applying machine learning. Ingeniare: Revista Chilena de Ingenieria, vol. 29, no. 3, pp. 517–530 (2021) doi: 10.4067/S0718-33052021000300517

18. Pita-Fernández, S., Pértegas-Díaz, S.: Pruebas diagnósticas: Sensibilidad y especificidad. Cad Aten Primaria, vol. 10, no. 1, pp. 120–124 (2003)
19. Rodríguez-Maldonado, A.: Medio para la aplicación de material terapéutico para uso en adultos discapacitados que padecen algún tipo de paresia (debilidad muscular) en miembros superiores. Ph. D. thesis, Universidad de los Andes (2006)
20. Roque-López, J.: Técnicas de selección de variables en regresión lineal múltiple. Master's thesis, Universidad Internacional de Andalucía (2021)
21. Salsich, G. B., Brown, M., Mueller, M. J.: Relationships between plantar flexor muscle stiffness, strength, and range of motion in subjects with diabetes-peripheral neuropathy compared to age-matched controls. Journal of Orthopaedic & Sports Physical Therapy, vol. 30, no. 8, pp. 473–483 (2000) doi: 10.2519/jospt.2000.30.8.473
22. Secretaría de Educación, Ciencia, Tecnología e Innovación (SECTEI): México, segundo país en América Latina con prevalencia de diabetes (2021) https://sectei.cdmx.gob.mx/comunicacion/nota/mexico-segundo-pais-en-america-latina-con-prevalencia-de-diabetes
23. Singh, P., Soni, K., Nair, A. S., Singh, M.: Regression analysis of ventilation coefficient at a semi-arid IGP region using forward selection technique. MAUSAM, vol. 73, no. 3, pp. 617–626 (2022) doi: 10.54302/mausam.v73i3.5933
24. Sisodia, D., Sisodia, D. S.: Prediction of diabetes using classification algorithms. Procedia Computer Science, vol. 132, pp. 1578–1585 (2018) doi: 10.1016/j.procs.2018.05.122
25. Tiwari, D.: Polyphagia can be a side effect of diabetes. African Journal of Diabetes medicine, vol. 30, no. 5 (2022)
26. Treviño, V., Falciani, F.: GALGO: An R package for multivariate variable selection using genetic algorithms. Bioinformatics, vol. 22, no. 9, pp. 1154–1156 (2006) doi: 10.1093/bioinformatics/btl074
27. Vázquez-Morales, E., Calderón-Ramos, Z. G., Arias-Rico, J., Ruvalcaba-Ledezma, J. C., Rivera-Ramírez, L. A., Ramírez-Moreno, E.: Sedentarismo, alimentación, obesidad, consumo de alcohol y tabaco como factores de riesgo para el desarrollo de diabetes tipo 2. Journal of Negative and No Positive Results, vol. 4, no. 10, pp. 1011–1021 (2019) doi: 10.19230/jonnpr.3068
28. Vidal-Ribas, P., Brotman, M. A., Valdivieso, I., Leibenluft, E., Stringaris, A.: The status of irritability in psychiatry: A conceptual and quantitative review. Journal of the American Academy of Child & Adolescent Psychiatry, vol. 55, no. 7, pp. 556–570 (2016) doi: 10.1016/j.jaac.2016.04.014
29. Zhou, S., Carroll, E., Nicholson, S., Vize, C. J.: Blurred vision. BMJ, vol. 368 (2020) doi: 10.1136/bmj.m569