

APLICACIÓN DE LA INTELIGENCIA ARTIFICIAL PARA LA DETECCIÓN DE PERSONAS EN RIESGO DE DIABETES

APPLYING ARTIFICIAL INTELLIGENCE TO DETECT INDIVIDUALS AT RISK OF DIABETES

BRENDA MARGARITA SALAS RAMOS¹

IRMA DELIA GARCÍA CALVILLO²

JESÚS ALEJANDRO NAVARRO ACOSTA²

GLORIA ISELA MENDOZA FRÍAS³

RESUMEN

La diabetes mellitus es una enfermedad crónica caracterizada por la hiperglucemia, derivada de una producción insuficiente de insulina o de la incapacidad del organismo para utilizarla de forma eficaz. Esta condición genera complicaciones graves como enfermedades cardiovasculares, insuficiencia renal y neuropatías. Según la Organización Mundial de la Salud, el número de personas que viven con diabetes pasó de 200 millones en 1990 a más de 830 millones en 2022, lo que subraya su impacto en la salud pública. La detección temprana es esencial para reducir la progresión y complicaciones de la enfermedad. El presente estudio tiene como objetivo predecir el riesgo de diabetes en personas a partir de datos clínicos y de encuestas recolectados en centros de salud del municipio de Saltillo, México, utilizando algoritmos de *Machine Learning* (ML). La base de datos fue proporcionada por el Centro de Investigación en Matemáticas Aplicadas de la Universidad Autónoma de Coahuila. Se implementaron cinco algoritmos: *Support Vector Machine*, *Logistic Regression*, *Random Forest*, *Multilayer Perceptron* y *K-Nearest Neighbors* evaluando su capacidad predictiva. Los modelos mostraron un desempeño adecuado en la clasificación de individuos sin riesgo de diabetes y una menor sensibilidad en la identificación de personas con riesgo. Esta limitación se asocia principalmente al desbalanceo de clases y a la naturaleza de los datos basados en encuestas. Actualmente, se desarrolla una nueva base de datos balanceada con registros clínicos que permitirá optimizar el rendimiento de los modelos. Los resultados confirman el potencial de las técnicas de ML como herramientas de apoyo en la detección temprana de diabetes. Esta investigación busca generar un modelo predictivo adaptado al contexto local de Saltillo para ofrecer una alternativa innovadora que podría ser escalable a otros entornos de atención médica.

Palabras clave: diabetes; inteligencia artificial; *Machine Learning*.

1. Centro de Estudios e Investigaciones Interdisciplinarias, Unidad Sureste, UAdeC.
2. Centro de Investigación en Matemáticas Aplicadas, Unidad Sureste, UAdeC.
3. Hospital Universitario "Dr. Gonzalo Valdés Valdés", Unidad Sureste, UAdeC.

Correspondencia
b_salas@uadec.edu.mx
<https://orcid.org/0009-0009-5218-7712>
Fecha de recepción
5 de junio de 2024.
Fecha de aceptación
22 de octubre de 2025.

ABSTRACT

Diabetes mellitus is a chronic disease characterized by hyperglycemia, resulting from insufficient insulin production or the body's inability to use it effectively. This condition leads to serious complications such as cardiovascular diseases, kidney failure, and neuropathies. According to the World Health Organization, the number of people living with diabetes increased from 200 million in 1990 to more than 830 million in 2022, highlighting its impact on public health. Early detection is essential to reduce disease progression and complications. The present study aims to predict the risk of diabetes in individuals based on clinical and survey data collected from health centers in the municipality of Saltillo, Mexico, using Machine Learning (ML) algorithms. The database was provided by the Center for Research in Applied Mathematics of the Autonomous University of Coahuila. Five algorithms were implemented: Support Vector Machine, Logistic Regression, Random Forest, Multilayer Perceptron, and K-Nearest Neighbors, evaluating their predictive capacity. The models showed adequate performance in classifying individuals without risk of diabetes, while exhibiting lower sensitivity in identifying those at risk. This limitation is mainly associated with class imbalance and the survey-based nature of the data. A new balanced database with clinical records is currently being developed to optimize model performance. The results confirm the potential of ML techniques as support tools for the early detection of diabetes. This research seeks to generate a predictive model adapted to the local context of Saltillo, offering an innovative alternative that could be scalable to other healthcare settings.

Keywords: diabetes, artificial intelligence, Machine Learning.

INTRODUCCIÓN

La diabetes mellitus es una de las enfermedades crónicas más prevalentes a nivel mundial, con un impacto significativo en la salud pública. Existen dos tipos de diabetes principalmente, la diabetes tipo 1 y la diabetes tipo 2. La diabetes tipo 1 es una enfermedad autoinmune que suele presentarse en la infancia o adolescencia y se caracteriza por la destrucción de las células pancreáticas, lo que provoca una deficiencia absoluta de insulina y la necesidad de tratamiento sustitutivo de por vida. La diabetes tipo 2 se asocia

principalmente con resistencia a la insulina y producción pancreática insuficiente, vinculada a factores de riesgo como obesidad, sedentarismo y antecedentes familiares, y generalmente se presenta en los adultos (ADA, 2022, WHO, 2023). Según la Organización Mundial de la Salud (WHO, 2023), en 2022 más de 830 millones de personas vivían con esta enfermedad, cifra que se ha cuadruplicado desde 1990. En México, la situación es alarmante: de acuerdo con la Encuesta Nacional de Salud y Nutrición (ENSANUT), la diabetes es una de las principales causas de morbilidad y mortalidad en el país (Rojas-Martínez y col., 2024). Uno de los mayores retos radica en la alta proporción de casos no diagnosticados o que son detectados en etapas avanzadas, lo que incrementa la aparición de complicaciones graves (enfermedades cardiovasculares, neuropatías y daño renal) y genera una carga significativa para el sistema de salud (Arredondo y col., 2020; Barquera y col., 2021).

Ante este panorama, resulta fundamental desarrollar herramientas que permitan mejorar la detección temprana de la diabetes. En los últimos años, los algoritmos de *Machine Learning* (ML) han mostrado resultados prometedores en el análisis de grandes volúmenes de información clínica y demográfica, con aplicaciones exitosas en la identificación temprana de diversas enfermedades (Oladimeji y col., 2021; Muhammad y col., 2020; Gulshan y col., 2016). Estas tecnologías permiten identificar patrones complejos en los datos que no siempre son evidentes mediante métodos estadísticos tradicionales, lo que las convierte en un recurso valioso para el diagnóstico temprano y la predicción de riesgos en salud.

La presente investigación tiene como objetivo aportar en el desarrollo de una herramienta predictiva, aplicando algoritmos de ML a una base de datos previamente recolectados en una clínica de salud del municipio de Saltillo, México. Dicha base fue generada mediante encuesta validada por médicos del Instituto Mexicano del Seguro Social (IMSS) y aplicadas por médicos residentes (De la Rosa-De León y col., 2024), coordinada por el Centro de Investigación en Matemáticas Aplicadas (CIMA) de la Universidad Autónoma de Coahuila. Se presentan los resultados y el desempeño de estos modelos en la clasificación de pacientes con y sin riesgo de diabetes, con el fin de apoyar la capacidad diagnóstica del sector salud local.

MATERIALES Y MÉTODOS

En el presente trabajo, se utilizó un enfoque sistemático basado en los pasos para el desarrollo de un modelo predictivo descrito por Guleria (2019), como se muestra en la figura 1. Este enfoque proporciona una guía estructurada para la preparación de los datos, la selección de algoritmos, el entrenamiento de los modelos y la evaluación de su desempeño.

El análisis se realizó sobre la base de datos generada a partir de encuestas a pacientes de una clínica de salud del municipio de Saltillo, la cual incluyó preguntas sobre características demográficas, antecedentes médicos y hábitos de salud relevantes para la diabetes (De la Rosa-De León y col., 2024).

La base de datos contiene información de 1903 pacientes: 976 de sexo femenino y 915 de sexo masculino, con edades comprendidas entre 15 y 84 años. Se registraron variables como género, edad, ciudad de origen, escolaridad, ocupación, peso, estatura, índice de masa corporal, presión arterial, antecedentes de diabetes, niveles de glucosa y hemoglobina glucosilada, consumo de refrescos, triglicéridos, colesterol, síntomas relacionados con diabetes y actividad física. En esta base de datos existe un desbalanceo en la cantidad de personas reportadas con diabetes ($\approx 37\%$) y las que no tienen diabetes ($\approx 63\%$).

Los datos fueron sometidos a un preprocesamiento que incluyó la imputación estadística de valores faltantes, normalización y estandarización de variables numéricas, y codificación de variables categóricas para garantizar la compatibilidad con los algoritmos de *Machine Learning*. Posteriormente, la base se dividió en conjuntos de entrenamiento, validación y prueba, asegurando la representatividad de las clases.

Se seleccionaron cinco algoritmos de *Machine Learning* ampliamente utilizados: Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), Multilayer Perceptron (MLP) y K-Nearest Neighbors (KNN). Cada modelo fue entrenado y optimizado mediante validación cruzada y ajuste de hiperparámetros.

Los registros de la base de datos fueron divididos aleatoriamente en conjunto de entrenamiento (80 %) y conjunto de prueba (20 %). El conjunto de entrenamiento se empleó para ajustar los parámetros del modelo y hacer validación cruzada. El conjunto de prueba se mantuvo aislado durante el

entrenamiento y es el que proporciona una estimación de la capacidad predictora del modelo, mediante métricas como precisión, exhaustividad, exactitud y F1-Score.

La programación se realizó en el lenguaje Python (versión 3) utilizando librerías especializadas, la ejecución de modelos se llevó a cabo en un equipo con sistema operativo Windows 10 Pro de 64 bits, procesador Intel Core i7 DE 2.8 GHZ, 16 GB de memoria RAM.

El desempeño de los algoritmos se evaluó mediante algunas métricas de clasificación, permitiendo comparar la capacidad de cada técnica para identificar correctamente a pacientes con y sin riesgo de diabetes.

RESULTADOS

El desempeño de los modelos de ML se evaluó mediante las métricas precisión, exhaustividad, F1-Score y exactitud, considerando por separado la clase 0 (sin riesgo de diabetes) y la clase 1 (con riesgo de diabetes).

Para la clase 0, los modelos mostraron un buen desempeño. La precisión se mantuvo por debajo del 80 % en todas las técnicas, mientras que la exhaustividad superó el 90 % en todos los algoritmos, excepto en K-Nearest Neighbors (KNN). El F1-Score se ubicó por encima del 80 % y la exactitud estuvo por encima del 70 %, esto se muestra en la Figura 1.

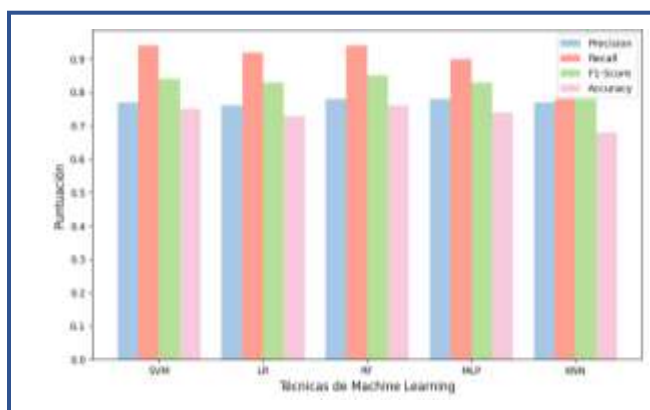


Figura 1. Gráfica de comparación de porcentajes obtenidos en las métricas (precisión, exhaustividad, F1-Score y exactitud) para los cinco modelos de ML para la clase 0 (sin riesgo de diabetes).

En contraste, para la clase 1, los resultados fueron considerablemente más bajos. La precisión osciló entre 40 y 60 %, mientras que la exhaustividad se situó entre 25 y 35 %, indicando que un gran número de pacientes con riesgo no fue identificado. El F1-Score se encontró entre 30 y 40 %, y la exactitud global se mantuvo por encima del 70 %, esto se muestra gráficamente en la Figura 2.

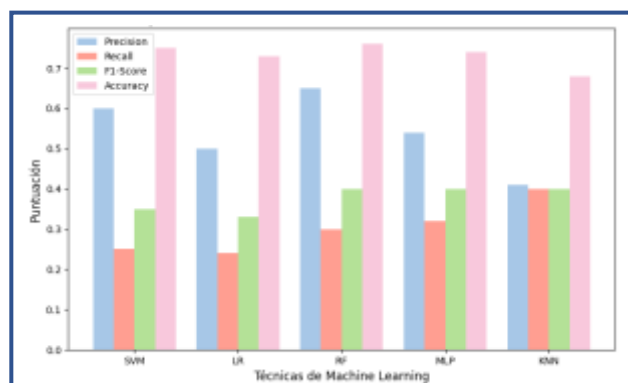


Figura 2. Gráfica de comparación de porcentajes obtenidos en las métricas (precisión, exhaustividad, F1-Score y exactitud) para los cinco modelos de ML para la clase 1 (con riesgo de diabetes).

DISCUSIÓN

Los resultados obtenidos en este estudio muestran que las técnicas de ML evaluadas fueron eficaces para identificar correctamente a los pacientes sin riesgo de diabetes en una base de datos desbalanceada y obtenida mediante encuestas. Sin embargo, todas las técnicas presentaron un desempeño limitado para clasificar correctamente a los pacientes con riesgo de diabetes, con valores de exhaustividad y F1-Score bajos, lo que refleja la dificultad de los modelos al trabajar con clases minoritarias. Este hallazgo coincide con estudios previos que destacan que los modelos de ML tienden a favorecer la clase mayoritaria cuando se entrenan con datos desbalanceados (Chawla y col., 2002; He & Garcia, 2009).

El bajo desempeño en la identificación de pacientes con riesgo de diabetes subraya la importancia de contar con bases de datos equilibradas y de calidad, con información obtenida directamente de archivos clínicos, que proporcionen medidas objetivas y precisas. Estudios recientes en predicción de diabetes con datos clínicos han mostrado que técnicas como *Random Forest*

y *XGBoost* pueden obtener mejores resultados cuando la base está balanceada y se aplican estrategias de manejo de desbalanceo, como sobremuestreo o ponderación de clases (Rana y col., 2020; Zhang y col., 2021).

Estos hallazgos motivan la continuación de la investigación utilizando técnicas de ML que tengan un buen desempeño en bases desbalanceadas, así como la recolección de datos de archivos clínicos en otra institución de salud en Saltillo, buscando una base balanceada de pacientes diabéticos y no diabéticos.

CONCLUSIONES

Las técnicas de ML evaluadas fueron eficaces para identificar pacientes sin riesgo de diabetes, pero mostraron un bajo desempeño al clasificar pacientes con riesgo, especialmente en bases desbalanceadas.

Los resultados destacan la necesidad de bases de datos balanceadas y de calidad clínica, así como de la aplicación de técnicas de ML adaptadas a datos desbalanceados, para garantizar una clasificación precisa.

El estudio sienta las bases para futuras investigaciones orientadas a crear una herramienta predictiva eficiente, que permita al sector salud de la región detectar a pacientes con riesgo de diabetes de manera temprana y asistir a los médicos en la toma de decisiones para prevenir complicaciones derivadas de la enfermedad.

REFERENCIAS

- American Diabetes Association (ADA). (2022). Standards of Medical Care in Diabetes -2022. *Diabetes Care*, 45: S1-S2. <https://doi.org/10.2337/dc22-SINT>
- Arredondo, A., Orozco, E., & Alcalde-Rabanal, J. (2020). Health care costs and financial consequences of epidemiological changes in chronic diseases in Latin America: Evidence from Mexico. *BMC Public Health*, 20: 1–9. <https://doi.org/10.1186/s12889-020-09315-2>

- Barquera, S., Hernández-Barrera, L., Trejo-Valdivia, B., Shamah-Levy, T., Campos-Nonato, I., & Rivera-Dommarco, J. (2021). Diabetes in Mexico: Cost and management of care. *Annals of Global Health*, 87:1–9. <https://doi.org/10.5334/aogh.3174>
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321–357. <https://doi.org/10.1613/jair.953>
- De la Rosa-De León, H., Navarro-Acosta, J. & García Calvillo, I. (2024). Aprendizaje automático aplicado a la detección temprana de Diabetes mellitus tipo 2: Caso Saltillo, México. *Revista Internacional de Investigación e Innovación Tecnológica*, 12: 1-24.
- Guleria, K., Prasad, D. & Kadyan, V. (2019). Detection of diabetic patterns using supervised learning. *International Journal of Engineering and Advanced Technology*, 9: 1169–1173.
- Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P. C., Mega, J. L., & Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316: 2402–2410. <https://doi.org/10.1001/jama.2016.17216>
- He, H., & García, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21: 1263-1284. <https://doi.org/10.1109/TKDE.2008.239>
- Muhammad, L., Algehyne, E. & Usman, S. (2020) Predictive Supervised Machine Learning Models for Diabetes Mellitus. *SN Computes Science*, 1:240. <https://doi.org/10.1007/s42979-020-00250-8>
- Oladimeji, O. Oladimeji, A & Oladimeji, O. (2021) Classification models for likelihood prediction of diabetes at early stage using feature selection. *Applied Computing and Informatic*, 21: 279-286.
- Rana, M., Lee, Y., & Lee, S. (2020). Prediction of type 2 diabetes using machine learning algorithms in a large population-based cohort. *Scientific Reports*, 10: 21038. <https://doi.org/10.1038/s41598-020-77659-1>
- Rojas-Mtz, R., Escamilla-Nuñez, C., Castro-Porras, L., Gómez-Velasco, D., Romero-Martínez, M., Hernández-Serrato, M., López-Ridaura, R., Díaz, M. & Aguilar-Salinas, C. (2024) Detección oportuna de prediabetes y diabetes. *Salud pública de México*. 66: 520-529. <https://doi.org/10.21149/15837>
- World Health Organization (WHO). (2023). Diabetes. WHO. <https://www.who.int/news-room/fact-sheets/detail/diabetes>
- Zhang, Z., Beck, M. W., Winkler, D. A., Huang, B., Sibanda, W., & Goyal, H. (2021). Opening the black box of neural networks: methods for interpreting model prediction. *BMJ*, 372: 159. <https://doi.org/10.1136/bmj.n159>