Comparativa de modelos multivariados basados en aprendizaje automático para la predicción del riesgo de diabetes en etapas tempranas.

Zayra Ramírez Gaytán[1*], Alen Francisco Luévano Lara[1], Vanessa Alcalá-Rmz[1].
[1] Unidad Académica de Ingeniería Eléctrica, Universidad Autónoma de Zacatecas, Ramón López Velarde 801, Centro, Zacatecas 98000, Zac., México.

** zay.rg98@gmail.com*

Abstract— **Diabetes is one of the fastest-growing, life-threatening, chronic degenerative diseases. According to the International Diabetes Federation (IDF), it has affected 463 million adults worldwide in 2019, 1 in 2 (232 million) people with diabetes were undiagnosed and 4.2 million people died. In this work, a data set of 520 instances has been used. The data set has been analyzed with the next three algorithms: logistic regression, decision tree and random forest. The results show that the decision tree algorithm had better performance with an AUC of 98.47% before feature selection and 97.97% after feature reduction. Also, it was found the most common symptoms that a person with a risk of diabetes presents are polyuria, polydipsia and sudden weight loss.**

Palabras clave— **Diagnóstico Asistido, Diabetes, Bosques Aleatorios, Árboles de decisión, Regresión Logística, Riesgo de diabetes, Síntomas, Signos.**

I. Introducción

Según la Federación Internacional de Diabetes (IDF) la diabetes es una de las enfermedades crónicas potencialmente mortales, que ha tenido un rápido aumento en la población, debido a que ha afectado a 463 millones de personas en todo el mundo, de las cuales la mitad de ellas no habían sido diagnosticadas [1], además se define como una enfermedad crónica de larga duración en la cual el cuerpo no es capaz de regular la cantidad de glucosa en la sangre [2]**.** La glucosa en la sangre es la principal fuente de energía y proviene de los alimentos que consumimos, ésta es catalizada por una hormona llamada insulina, esta hormona es producida por el páncreas para controlar la glucosa en la sangre [3]**.** La diabetes puede ser causada por la poca producción de insulina, resistencia a la insulina o ambas [4]**.** Además, la prevalencia ha tenido un aumento notable, principalmente en los países que tienen ingresos bajos y medios. La detección oportuna de la diabetes es relevante, debido a que en etapas avanzadas se pueden desarrollar diferentes complicaciones las cuales van desde ceguera, insuficiencia renal, ataques cardiacos, derrames cerebrales hasta amputación de miembros inferiores [2]**.** La diabetes es actualmente la primera causa de mortalidad en México y su tendencia muestra un incremento progresivo en los últimos años [5]**.** Debido a las complicaciones generales que se producen a consecuencia de la diabetes, se han realizado estudios indicativos de prevención o moderación de la enfermedad, sin embargo, la enfermedad no se suele diagnosticar en etapas tempranas debido a la falta de conocimiento de los síntomas relacionados [6]. El objetivo del presente trabajo es resaltar los síntomas y signos relacionados a la prevalencia de la diabetes, cabe destacar que una detección temprana conlleva un conjunto de ventajas que abarcan desde tratamientos más económicos y sencillos hasta tratamientos menos traumáticos y/o invasivos [7]. En este trabajo se presenta un conjunto de modelos, que pueden implementarse en herramientas de diagnóstico asistido, los cuales consisten en procedimientos que ayudan a los profesionales en la interpretación de distintos datos. Los sistemas de diagnóstico asistido utilizan algoritmos para reconocer patrones en datos de pacientes, de esta manera se proporciona un soporte a los especialistas al momento de realizar un diagnóstico [8].

II. Metodología

Las técnicas de clasificación de aprendizaje automático han tenido una gran aceptación por los investigadores cuando se trata de modelar el riesgo de padecer una enfermedad [9]. Para el desarrollo del trabajo se llevó a cabo la metodología mostrada en la Fig. 1.

dx.doi.org/10.24254/CNIB.21.62

Fig. 1 Metodología del trabajo.

Adquisición de la base de datos: La base de datos contiene la información requerida para trabajar con la predicción de riesgo de diabetes en etapas tempranas, los pacientes pueden estar etiquetados como pacientes con riesgo de padecer diabetes o paciente sin riesgo. Cabe destacar que según la IDF 1 de cada 5 personas mayores de 65 años tiene diabetes [1] y la mayoría de pacientes en la base de datos se encuentran entre los 45 y 55 años de edad, indicando un posible riesgo y una posible detección temprana de la enfermedad. Las variables utilizadas para el desarrollo del modelo corresponden a síntomas y signos relacionados con la diabetes. La información presente en la base de datos ha sido recopilada mediante cuestionarios directos de los pacientes del Sylhet Diabetes Hospital en Sylhet, Bangladesh [10].

La distribución de los datos corresponde a un 38.5 % de personas con riesgo de padecer diabetes y un 61.5 % de personas sin riesgo. La base de datos se compone de 16 características binarias, las cuales son: edad y género del paciente, irritabilidad, paresia parcial, alopecia, obesidad, polidipsia, poliuria, debilidad, polifagia, candidiasis genital, visión borrosa, prurito, rigidez muscular, pérdida de peso repentina y curación retardada [10].

Pre-procesamiento: El conjunto de datos es dividido en dos subconjuntos, el primero pertenece al entrenamiento con un 80% de datos del conjunto total y el segundo corresponde al de prueba y contiene el 20% restante. Posteriormente, se normalizan los subconjuntos de datos por medio del método del puntaje Z, el cual consiste en transformar los datos a una distribución con una media 0 y una desviación estándar de 1, este método tiene el propósito de definir una misma escala numérica para los datos.

Implementación de algoritmos: En esta investigación se implementaron tres algoritmos de aprendizaje automático los cuales se describen a continuación:

Árboles de decisión: Es un algoritmo comúnmente utilizado en problemas de clasificación, el algoritmo analiza los datos y toma decisiones basándose en una serie de preguntas, en la etapa del aprendizaje el modelo administra la ganancia de información en el nodo dado [11].

Regresión logística: Se centra en encontrar las relaciones entre la variable dependiente y la independiente utilizando la función logística para las probabilidades [12].

Bosques aleatorios: Se trata de un algoritmo utilizado constantemente en el área médica, consiste en crear múltiples árboles de decisión para así generar el llamado bosque [13].

Además, se realizó una selección de características para encontrar aquellas que aportan mayor información al modelo, para este paso, se optó por implementar 'Boruta', un algoritmo que proporciona criterios para la selección de atributos importantes, el cual funciona agregando más aleatoriedad en el sistema. Se basa en realizar una copia aleatoria del sistema, combinar la copia con el original y crear un clasificador para este sistema extendido, para evaluar la importancia de la variable en el sistema original se compara con la de las variables aleatorias. Solamente las variables cuya importancia es superior a la de las aleatorias se consideran importantes [14], esta herramienta consiste en visualizar las características más significativas, a través de la asignación de un nivel dentro de un rango, según la relación con las características analizadas [15].

Validación: Para validar y conocer el desempeño de los algoritmos se utilizaron diferentes métricas, las cuales son presentadas a continuación:

La curva ROC (receiver operating characteristic curve) se utiliza para visualizar el desempeño de los clasificadores, ésta se complementa con el área bajo la curva, la cual representa la probabilidad de que una muestra aleatoria positiva sea clasificada correctamente. Para calcular el área de la curva ROC se necesitan dos valores que definen el trayecto de la misma: la sensibilidad y la especificidad [16].

La sensibilidad o tasa de verdaderos positivos se refiere a la proporción de sujetos con una condición positiva que fueron correctamente clasificados y se calcula con la Ecuación (1), donde VP son los verdaderos positivos y FN son los falsos negativos [17].

$$Sensibilidad = \frac{VP}{VP+FN} \quad (1)$$

La especificidad corresponde a la proporción de verdaderos negativos, es decir, los sujetos con una condición negativa que fueron correctamente clasificados y se calcula con la Ecuación (2), donde VN son los verdaderos negativos y FP son los falsos positivos [17].

$$Especificidad = \frac{VN}{VN+FN} \quad (2)$$

La curva ROC se grafica tomando en cuenta la unión de distintos puntos de corte, donde el eje Y representa a la sensibilidad y el eje X a (1-especificidad) de cada uno de ellos [16].

De igual manera se calculó la precisión del modelo, como se muestra en la ecuación (3). Ésta indica la confiabilidad del modelo al clasificar los datos dentro de una clase midiendo el número de términos correctamente reconocidos respecto al total de términos predichos, sean verdaderos o falsos [18]. Por otro lado, la exactitud calcula el rendimiento promedio de los algoritmos, como se muestra en la ecuación (4), el propósito de esta métrica es calcular el porcentaje de muestras que son clasificadas correctamente [17].

$$Precisión = \frac{VP}{VP+FP} \quad (3)$$

$$Exactitud = \frac{VP+VN}{VP+VN+FP+FN} \quad (4)$$

III. RESULTADOS

Como primer paso se optó por utilizar los árboles de decisión para realizar una comparativa respecto a la selección de características realizada con Boruta.

En la Tabla 1 se muestra la comparativa de los resultados en los árboles de decisión. La segunda columna muestra los resultados obtenidos para las 16 variables originales de la base de datos. La tercera columna corresponde a los resultados del modelo con 12 características relevantes según el análisis realizado con Boruta, las variables eliminadas fueron: obesidad, debilidad, candidiasis genital y rigidez muscular. Finalmente se desarrolló otro modelo dejando solamente las 10 características más importantes, donde las dos variables eliminadas respecto al conjunto mencionado anteriormente fueron: retraso de curación y visión borrosa.

TABLA 1.
COMPARATIVA DE LOS RESULTADOS EN LOS 3 MODELOS DE ÁRBOL DE DECISIÓN.

Métricas	16 variables	12 variables	10 variables
Exactitud	98.07%	97.11%	98.07%
Precisión	100.00%	96.92%	98.44%
Área bajo la curva	98.46%	96.72%	97.97%
Sensibilidad	96.92%	98.44%	98.44%

Los resultados muestran que el conjunto de datos que cuenta con 16 variables obtuvo 98.46% de área bajo la curva, seguido del modelo con 10 variables el cual obtuvo un 97.97%, finalmente el modelo con 12 características obtuvo un menor desempeño con 96.92%. Es importante resaltar que se tomará en cuenta el conjunto de 10 características, debido a que la diferencia en las métricas es mínima comparada con el modelo de 16 características, lo que significa que con una cantidad menor de variables se logran obtener resultados similares al del conjunto de datos completo, permitiendo así, disminuir los costos computacionales y de tiempo al momento de entrenar los algoritmos. Después se implementó regresión logística y bosques aleatorios con el conjunto de datos que contiene 10 variables, obteniendo lo siguiente:

TABLA 2.
COMPARATIVA DE LOS RESULTADOS EN LOS 3 MODELOS DESARROLLADOS.

Métricas	Regresión Logística	Árbol de Decisión	Bosque Aleatorio
Exactitud	90.38%	98.07%	97.11%
Precisión	92.00%	98.44%	95.58%
Área bajo la curva	96.25%	97.97%	96.15%
Sensibilidad	92.00%	98.44%	100.00%

En la Tabla 2 se observa la comparativa entre los tres modelos desarrollados, los cuales lograron resultados estadísticamente significativos con porcentajes superiores al 90% en todas las métricas. Además, se observa que el modelo de árboles de decisión logró un mejor desempeño en todas las métricas obtenidas excepto en la sensibilidad que fue superado por el modelo de bosque aleatorio con un 100%. Las métricas de la Tabla 2 fueron calculadas a partir de los valores obtenidos en la matriz de confusión mostrados en la Tabla 3.

TABLA 3.
MATRIZ DE CONFUSIÓN DE LOS 3 MODELOS DESARROLLADOS

Resultado de Prueba	Regresión Logística	Árbol de Decisión	Bosque Aleatorio
Verdaderos positivos	60	63	65
Falsos positivos	5	1	3
Verdaderos Negativos	34	39	36
Falsos Negativos	5	1	0

Por último, en la Fig. 2 se muestran las curvas ROC de los 3 modelos y el valor del área bajo la curva.

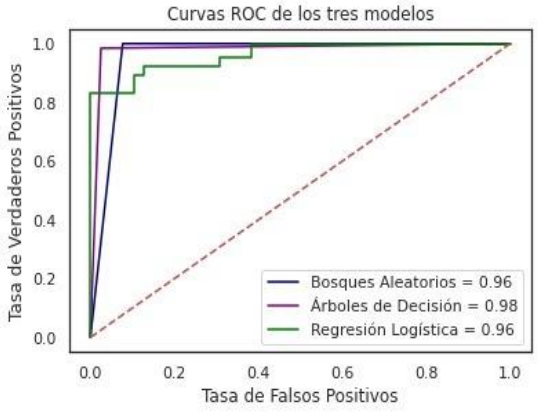

Fig. 2 Curva ROC y el valor del área bajo la curva.

IV. Discusión

La base de datos utilizada para llevar a cabo el presente trabajo fue dividida en dos subconjuntos, uno para entrenamiento que corresponde al 80% de los datos y el 20% restante para el subconjunto de prueba. En la sección III se observa la importancia de la implementación de los métodos de selección de características, en este caso Boruta, ya que se logró reducir en aproximadamente un 38% el conjunto de datos original, sin afectar de manera significativa el desempeño del modelo obtenido, lo cual indica que al momento de entrenar o implementar algún modelo con este conjunto de datos, los costos computacionales serán menores en contraste a utilizar el conjunto de datos original. Cabe destacar que el algoritmo utilizado para la selección de características tiene la capacidad de ordenar en un rango ascendente-descendente las variables con más aporte al análisis. Como se mostró en la Tabla 1, se presenta un descenso cuando se descartan algunas de las variables de la base de datos original, debido a esto se hizo una comparativa con la literatura de la Organización Mundial de la Salud, ésta menciona que entre los síntomas principales de diabetes se incluyen la excreción excesiva de orina (poliuria), sed (polidipsia), hambre constante o polifagia, pérdida repentina de peso, trastornos visuales y debilidad, características que no se descartaron en el modelo con 10 variables [19]. Los síntomas pueden aparecer de forma súbita, retomando así por qué el desempeño disminuye al momento en que algunas de estas características son eliminadas. Después de decidir el conjunto de datos óptimo se procede a implementar los

algoritmos de regresión logística, bosque aleatorio y árbol de decisión, con la finalidad de conocer el desempeño de cada uno de ellos al determinar si un paciente tiene riesgo de padecer diabetes. Con base en los resultados obtenidos, se determina que los tres modelos lograron resultados estadísticamente significativos, pero el modelo de árboles de decisión mostró un mejor desempeño, alcanzando un 97.97% en la métrica correspondiente al área bajo la curva, lo que significa que aproximadamente el 98% de las veces el modelo fue capaz de determinar si un paciente tiene riesgo de padecer o no diabetes. Es importante mencionar que cuando se logra diagnosticar los problemas de salud en el momento inicial los tratamientos siempre van a ser más eficaces, sencillos y económicos, una vez que se establecen los tratamientos, el proceso de la enfermedad se va a ver ralentizado. Si no se actúa ante la aparición de los primeros síntomas el deterioro en la calidad de vida del paciente se vuelve inminente. Otro beneficio de la prevención de dicha enfermedad es el ahorro en gastos médicos mayores, sobre todo en los hospitales de México que el costo total anual de los pacientes con diabetes mellitus para el IMSS fue de US$452 064 988 durante el periodo 2002-2004, correspondiente a 3.1 % del gasto de operación [20].

V. CONCLUSIONES

En relación a la correlación indicada en este trabajo y la literatura obtenida de la Organización Mundial de la Salud se concluye que existen síntomas y signos significativos para predecir la diabetes en etapas tempranas y este trabajo sirve como un primer acercamiento para lograr desarrollar herramientas predictivas de diabetes que apoyen a los profesionales de la salud en la detección de dicha enfermedad [19]. La importancia del diagnóstico asistido en la vida cotidiana es relevante para que los profesionales de la salud puedan tomar una decisión óptima [1].

En el presente trabajo se implementan tres modelos los cuales obtuvieron resultados estadísticamente significativos. Además, se logra comprobar la eficacia del método de selección de características Boruta. Por otro lado, con esta aportación se logra dar un paso importante para ayudar en la detección temprana de la diabetes, debido a que en la base de datos se encuentran pacientes con riesgo de padecer o no esta enfermedad, de esta manera se aporta a la posibilidad de un diagnóstico temprano, evitando así complicaciones que se pueden desarrollar posteriormente en el paciente. Es importante mencionar que con los modelos que se presentan se pueden desarrollar herramientas que tengan un margen de error mínimo en la predicción del riesgo de padecer diabetes.

REFERENCIAS

1. International Diabetes Federation,“About Diabetes”, 2021, Available: https://idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html
2. Brent Wisse, “American Diabetes Association. 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes-2021,” Diabetes Care. 2021, pp. 36-37 Accessed on: Jun, 2nd, 2021, DOI: 10.2337/dc21-S002, [Online].
3. Centers for Disease Control and Prevention, “National Diabetes Statistics Report, 2020,” CDC, Atlanta, GA, U.S. Department of Health and Human Services, 2020.
4. Chausmer, Arthur B. "Zinc, insulin and diabetes," Journal of the American College of Nutrition, no. 17.2, pp. 109-115, 1998, DOI: 10.1080/07315724.1998.10718735, Accessed on: June, 3st, 2021, [Online].
5. Quinde, Cristobal Franco, et al. "Prevalencia y factores de riesgo de diabetes tipo II," RECIMUNDO: Revista Científica de la Investigación y el Conocimiento, no. 2.1, pp. 530-549, 2018, DOI: 10.26820/recimundo/2.1.2018.530-549, Accessed on: May, 27th, 2021, [Online].
6. J. Escobedo-de la Peña, L. V. Buitrón-Granados, J. C. Ramírez-Martínez, R. Chavira-Mejía, H. Schargrodsky y B. M. Champagne, “Diabetes en México. Estudio CARMELA,” Cirugía y cirujanos, vol. 79, n °5, pp. 424-431, 2011, Accessed on: June, 1st, 2021, [Online]. https://www.medigraphic.com/pdfs/circir/cc-2011/cc115f.pdf
7. Carlos Vassallo Sella, “La vuelta a la pandemia en doce semanas,” 1a ed. Ciudad Autónoma de Buenos Aires, Argentina: CVS, 2020, ch. 1, sec. 1, pp. 133–149. [Online]. Available: http://www.epidemiologia.anm.edu.ar/wp-content/uploads/2021/01/La-vuelta-a-la-pandemia-en-12-semanas.pdf
8. Rafael Llobet Azpitarte, “Aportaciones al Diagnóstico de Cáncer Asistido por Ordenador,” RLA, DSIC, UPV, Comunidad Valenciana, Valencia, España, Julio 2006.
9. Kolachalama, V.B., Garg, P.S. “Machine learning and medical education.” npj Digital Med vol. 1, no. 54 September, 27th, 2018, Accessed on: September, 3, 2021 DOI: 10.1038/s41746-018-0061-1 [Online].
10. M. F. Islam, R. Ferdousi, S. Rahman y H. Y. Bushra, “Likelihood prediction of diabetes at early stage using data mining techniques”, in Computer Vision and Machine Intelligence in Medical Image Analysis, Singapore, Springer, 2020, pp. 113-125, [Online]. Available: https://www.kaggle.com/ishandutta/early-stage-diabetes-risk-prediction-dataset
11. Alam, F., Mehmood R., Katib, I. Comparison of decisión tres and Deep learning for object classification in autonomius driving. In Smart Infrastructure and Applications; Springer: Cham, Switzerland, 2020; pp. 135–158.
12. Hilbe, J.M. Logistic Regression Models; CRC Press: Boca Raton, FL, USA, 2009.
13. Speiser, J.L., Miller, Michael E., Tooze, J., Ip, E. A comparison of random forest variable selection methods for classification prediction modeling. Expert Syst. Appl. 2019, 134, pp.93-101
14. Miron B. Kursa∗, Aleksander Jankowski, Witold R. Rudnicki, “Boruta – A System for Feature Selection,” Fundamenta Informaticae 101 (2010) 271–285,DOI. 10.DOI 10.3233/FI-2010-288.
15. Miron Bartosz Kursa, Witold Remigiusz Rudnicki, “Package ‘Boruta’,” PB. CRAN., USA, Rep. Mayo, 21, 2020 UTC.
16. J. Cerda y L. Cifuentes, “Uso de curvas ROC en investigación clínica: Aspectos teórico-prácticos”, Revista chilena de infectología, vol. 29, n° 2, págs. 138-141, 2012, DOI: 10.4067/S0716-10182012000200003, Accessed on: June, 1st, 2021, [Online].
17. Alcalá-Rmz, Vanessa, et al. "Identification of People with Diabetes Treatment through Lipids Profile Using Machine Learning Algorithms." Healthcare. Vol. 9. No. 4. Multidisciplinary Digital Publishing Institute, Mexico City, Mexico April, 2021, Accessed on: July, 10, 2021, DOI: 10.3390/healthcare9040422, [Online].
18. Corso, Cynthia Lorena, “Aplicación de algoritmos de clasificación supervisada usando Weka. Córdoba” Universidad Tecnológica Nacional, Facultad Regional Córdoba, 2009, Accessed on: September, 07, 2021. Available: https://www.investigacion.frc.utn.edu.ar/labsis/Publicaciones/congresos_labsis/cynthia/CNIT_2009_Aplicacion_Algoritmos_Weka.pdf [Online].
19. Organización Mundial de la Salud,“Diabetes”,2021, Available:https://www.who.int/es/news-room/fact-sheets/detail/diabetes
20. R. d. l. Á. Rodríguez Bolaños, L. M. Reynales Shigematsu, J. A. Jiménez Ruíz, S. A. Juárez Márquezy y M. Hernández Ávila, “Costos directos de atención médica en pacientes con diabetes mellitus tipo 2 en México: análisis de microcosteo,” Revista panamericana de salud pública, vol. 28, pp. 412-420, 2010, Accessed on: March, 25, 2021, DOI: 10.1590/s1020-49892010001200002, [Online].