Revista Internacional de Investigación e Innovación Tecnológica

Página principal: www.riiit.com.mx

Aprendizaje automático aplicado a la detección temprana de Diabetes mellitus tipo 2: Caso Saltillo, México

Automated learning applied to early detection of type 2 Diabetes mellitus: The Case of Saltillo, México

De la Rosa-De León, H., Navarro-Acosta, J.A., García-Calvillo, I.D.*

Centro de Investigación en Matemáticas Aplicadas, Universidad Autónoma de Coahuila, CP 25250, Saltillo, Coahuila, México.

hectorleon@uadec.edu.mx; alejandro.navarro@uadec.edu.mx; irma.garcia@uadec.edu.mx*

Innovación Tecnológica: Aplicación de técnicas de aprendizaje automático como apoyo para la detección oportuna de Diabetes mellitus.

Área de aplicación industrial: Sector salud, apoyo al diagnóstico de padecimientos médicos.

Recibido: 09 octubre 2023
Aceptado: 28 junio 2024

Abstract

Over the last few decades, health systems around the world have registered a significant growth in the number of people diagnosed with Diabetes Mellitus, which can be described as a condition associated with unhealthy living conditions such as obesity and/or overweight, a sedentary life and a diet rich in sugars and fats, as well as a predisposition due to genetic factors. Given the lack of early detection strategies, those affected are diagnosed once they have developed the disease and, therefore, present obvious symptoms. This article presents the results of a study, where clinical information was collected on people over 15 years of age with risk factors for Type 2 Diabetes Mellitus, treated at the Mexican Institute of Social Security (IMSS), in the City of Saltillo, Coahuila, México. This database of 1820 records, made it possible to explore different Automated Learning tools that allow early detection of diseases, such as Supervised Learning models like Naive Bayes and Random Forest, which, when trained, achieved more than 73% sensitivity in the prediction of this disease.

Keywords: Type 2 Diabetes Mellitus, Automated Learning, Predictive models.

Resumen

A lo largo de las últimas décadas, los sistemas de salud en el mundo han registrado un importante crecimiento en el número de personas con diagnóstico de Diabetes Mellitus, el cual se puede describir como un padecimiento asociado a condiciones de vida poco saludables tales como la obesidad y/o sobrepeso, vida sedentaria y alimentación rica en azúcares y grasas, así como a la predisposición por factores genéticos. Ante la falta de estrategias de detección temprana, los afectados son diagnosticados una vez que han desarrollado la enfermedad y, por tanto, presentan síntomas evidentes. En este artículo se presentan los resultados de un estudio, donde se llevó a cabo el levantamiento de información clínica sobre personas mayores de 15 años con factores de riesgo a la Diabetes Mellitus Tipo 2, atendidos en el Instituto Mexicano del Seguro Social en la Ciudad de Saltillo, Coahuila, México. Dicha base de datos, de 1820 registros, permitió explorar diferentes herramientas de *Machine Learning* que permiten la detección temprana de enfermedades, tal es el caso de los modelos de Aprendizaje supervisado como Naive Bayes y Random Forest, los cuales, al ser entrenados alcanzaron más de 73% de sensibilidad en la predicción de esta enfermedad.

Palabras clave: Diabetes mellitus tipo 2, Machine Learning, Modelos predictivos.

I. Introducción

De acuerdo con la Encuesta Nacional de Salud y Nutrición (ENSANUT, 2022), en la población de adultos en México en el año 2022, la prevalencia de diabetes diagnosticada y no diagnosticada fue de 12.6% y 5.8%, respectivamente, para una prevalencia total de 18.3%, que representa 14.6 millones de personas (Basto-Abreu et al. 2023). El número de personas con diabetes no diagnosticadas de alrededor de 5 millones de mexicanos, representa un reto importante para el control y la prevención, se requieren acciones para un diagnóstico oportuno.

La Organización Mundial de la Salud (OMS, 2016) afirma que México presenta la prevalencia de Diabetes más alta entre los países que integran la Organización para la Cooperación y el Desarrollo Económico y registra el consumo per cápita más alto de refrescos a nivel mundial. Prueba de lo anterior es la prevalencia de sobrepeso y obesidad superior al 33% en niños y cerca del 70% en adultos.

En Basto-Abreu *et al.* (2020), con base en los resultados de la encuesta ENSANUT 2016, se comenta que la Diabetes Mellitus tipo 2 (DM2) en México es la segunda causa de muerte y la primera causa de años de vida saludable perdidos, ocasionando que el mismo año esta enfermedad fuera declarada emergencia epidemiológica nacional.

En Escamilla (2021), se afirma que la DM2 clínicamente puede ser silenciosa, hecho que complica el diagnóstico de casos nuevos y, además, casi no existen datos de calidad en los sistemas de salud que reflejen la verdadera incidencia, por ello es más significativo hablar de *prevalencias*, debido a que el diagnóstico se realiza cuando el paciente acude a consulta médica o a realizarse estudios, es decir, cuando ya hay presencia de síntomas o se manifiesta alguna complicación clínica.

De acuerdo con cifras oficiales del Gobierno del Estado de Coahuila (2020), la prevalencia de la DM2 en el estado es de 12.3% en población de más de 20 años con diagnóstico médico previo. En el mismo sentido, son

detectados anualmente cerca de 13,000 coahuilenses con dicho padecimiento, lo que equivale a 33 casos diarios, afectando por igual a hombres y mujeres. Por otro lado, son cerca de 100 los casos identificados en menores de 17 años debido a la mala alimentación y falta de actividad física. A finales del año 2020, la DM2 se posicionó como la tercera causa de muerte en Coahuila, de acuerdo con la misma fuente. Ante la falta de estadísticas municipales sobre la población que vive actualmente con tal enfermedad, se estima una alta prevalencia en la Región Sureste de Coahuila, considerando que concentra el 34.68% de la población total del Estado. Resulta importante destacar las ventajas de la detección temprana de la DM2 para mantener un estilo de vida saludable.

Para Herman *et al*. (2015), las pruebas tempranas para detectar la DM2, seguidas de un tratamiento oportuno, dan lugar a una reducción del riesgo de enfermedades del corazón (cardiovasculares) y de muerte dentro de un período de seguimiento de cinco años, en comparación con los pacientes que no se hacen las pruebas.

Un reto fundamental en el control de la enfermedad es la detección temprana, para el caso de la ENSANUT 2006, la mitad de los mexicanos con Diabetes no habían sido diagnosticados al momento de las entrevistas. Existe evidencia de la evolución de la Diabetes, la cual es lenta y puede mantenerse sin ser detectada hasta que se hacen presentes las primeras complicaciones. De 2006 a 2022 la prevalencia de la DM2 ha aumentado de 14.4% (Villalpando et al. 2010) a 18.3% (Basto-Abreu et al. 2023) en México.

Debido a lo anterior, se propone la utilización de técnicas de aprendizaje automático (*Machine Learning*) para coadyuvar en el diagnóstico temprano de la DM2, resaltando la alta prevalencia de personas con sintomatología y/o complicaciones de dicha enfermedad en el estado de Coahuila de Zaragoza. La detección anticipada evitaría la aparición de este padecimiento y sus complicaciones a una corta edad, asimismo, evitaría la aparición del factor de riesgo relacionado con infecciones virales y bacterianas, así como la aparición de infartos y otros trastornos.

El objetivo de este trabajo es desarrollar y validar mediante técnicas de Machine Learning un modelo predictivo que facilite la detección temprana de la Diabetes Tipo 2 en el Municipio de Saltillo, Coahuila, México. Los objetivos específicos son identificar factores de riesgo para ayudar a la detección temprana de la DM2, recopilar información clínica de personas residentes del Municipio de Saltillo, llevar a cabo el procesamiento necesario de los datos obtenidos, aplicar y validar técnicas de Machine Learning para el desarrollo del modelo predictivo.

Machine Learning y la predicción de enfermedades

Para Sidey-Gibbons y Sidey-Gibbons (2019), el *Machine Learning* es una reciente disciplina que combina un conjunto de técnicas matemáticas, estadísticas y computacionales, mismas que permiten a las computadoras aprender y llevar a cabo actividades de procesamiento complejas. Dichas características brindan a las máquinas la habilidad de hacer predicciones utilizando distintos orígenes de datos. Kaur y Kumari (2020) llevaron a la práctica una interesante investigación, con distintas técnicas de *Machine Learning*, el objetivo fue identificar tendencias, así como detectar patrones de riesgo asociados a la Diabetes. En Mejía et al. (2023) se presenta un estudio en Colombia donde aplicaron modelos basados en técnicas de aprendizaje automático como apoyo al diagnóstico temprano de Diabetes utilizando información socieconómica y ambiental sin la dependencia de toma de muestras clínicas.

Con los recientes avances del *Big Data* en áreas como la Biomedicina y el cuidado de la salud, se generan grandes beneficios gracias a la detección temprana de enfermedades y mejora en la atención de los pacientes. No obstante, se reduce la precisión de los análisis cuando los registros de información médica se encuentran incompletos (Chen, 2017).

En la tabla del Anexo A *Recopilación documental de trabajos científicos* se presenta una recopilación de artículos publicados en los últimos años, que muestran la aplicación de modelos predictivos relacionados principalmente con la detección temprana de la DM2, enfermedad bajo estudio desde la perspectiva del *Machine Learning*. Destacan países como India y China con un mayor número de publicaciones. Dichos trabajos varían en la selección de variables o atributos, no obstante, algunos de ellos coinciden en características como *género, edad, nivel de presión arterial, índice de masa corporal, factor familiar, hambre excesiva, sed excesiva, problemas de la vista* y, desde luego, *nivel de glucosa en sangre*.

Por otro lado, se aprecian tamaños variados en las bases de datos utilizadas, se identificó un rango de muestras que va desde 200 hasta 290,000 observaciones. Destacan, además, la utilización de modelos de Aprendizaje supervisado tales como Naive Bayes, Support Vector Machines, Linear Discriminant Analysis, Random Forest, *k* Nearest Neighbors, Logistic Regression y Artificial Neural Networks, así como otros enfoques donde se utilizan ensambles de clasificadores. Finalmente, algunos de los modelos de aprendizaje supervisado que obtuvieron mejor desempeño en su nivel de *Accuracy* fueron el Naive Bayes con 93.00% (Mujumdar y Vaidehi, 2019), Random Forest con 98.48% (Reddy *et al*., 2020) y *k* Nearest Neighbors con 90.38% (Das *et al*., 2020).

Con estos estudios como antecedentes, se presentan en este trabajo los resultados al aplicar técnicas de Machine Learning a una base de datos de pacientes de una clínica del Instituto Mexicano del Seguro Social (IMSS) en la Ciudad de Saltillo, Coahuila, México, para el diagnóstico temprano de la DM2. El resto del artículo está organizado de la siguiente forma, en la sección II se presenta la metodología utilizada, así como la descripción de la base de datos con la que se trabajó. La sección III los algoritmos utilizados, la sección IV presenta una discusión de los resultados obtenidos, finalmente, la sección V presenta las conclusiones seguida de las referencias bibliográficas.

II. Materiales y equipo

La metodología empleada en este trabajo se compone de manera general en tres fases: 1) adquisición y procesamiento de datos, 2) implementación de modelos de aprendizaje automático para la detección de DM2 y 3) validación del desempeño de dichos modelos, como se aprecia en figura 1.

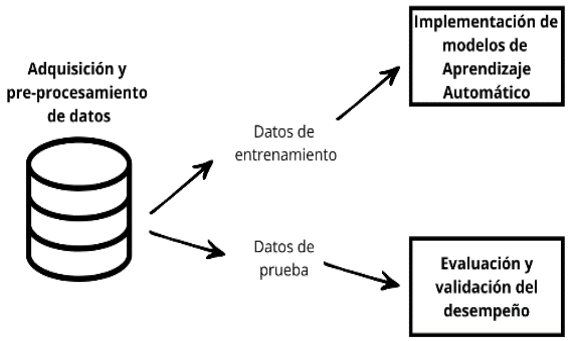

Figura 1. *Esquema de la metodología propuesta.*
Fuente: elaboración propia.

Para esto, se hizo uso de las siguientes tecnologías informáticas:

- *Lenguaje de programación PHP (Hypertext Pre-Processor)* para Windows versión 7.4.29.
- *MySQL (Structured Query Language)*:

Software para el manejo de bases de datos en Windows versión mysql-8.0.28- winx64.

- *Anaconda*: Distribución de Anaconda para computación científica y Aprendizaje automático versión conda 4.11.0.

La Figura 2 muestra la secuencia de uso de dichas tecnologías. La recopilación y el preprocesamiento de los datos se llevó a cabo en una computadora DELL Inspiron 5584 con procesador Intel(R) Core(TM) i7-8550U CPU @ 1.80GHz 1.99 GHz, memoria RAM instalada de 8.00 GB y Sistema operativo de 64 bits, procesador basado en x64.

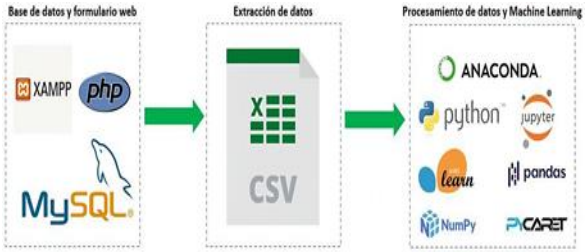

Figura 2. *Ambientes y plataformas utilizados.*
Fuente: elaboración propia.

Obtención de datos

Tras la firma de un convenio de colaboración entre el Instituto Mexicano del Seguro Social y la Universidad Autónoma de Coahuila, personal directivo de dicha institución médica analizó y validó la propuesta inicial del cuestionario (Anexo B: instrumento de investigación) y, posteriormente, realizó ajustes a algunos de los ítems clínicos identificados para el diagnóstico de la DM2. Más adelante, dicho instrumento fue codificado y adaptado en un formulario en línea para el levantamiento de la información en campo. Los trabajos de colaboración, permitieron la conformación de un equipo de médicos residentes de la Unidad de Medicina Familiar No. 82, quienes recopilaron la base de datos durante la segunda quincena del mes de diciembre de 2021. Dicho grupo de residentes estuvo conformado por 5 personas del sexo femenino y 3 del sexo masculino, quienes realizaron una serie de entrevistas a derechohabientes del mismo instituto durante sus visitas para recibir atención médica o gestionar algún servicio en la referida clínica. Durante el proceso de espera antes de ingresar a la consulta médica, se informó a la población entrevistada acerca del desarrollo del proyecto de investigación y se les invitó a participar de manera voluntaria, dejando de manifiesto la no utilización de datos personales, así como la confidencialidad y objetividad de los datos clínicos recabados.

Durante las conversaciones con las autoridades médicas del hospital, se propuso llevar a cabo un levantamiento de muestra robusto que pudiera asegurar un entrenamiento de calidad, las autoridades se comprometieron a recabar la información únicamente durante la segunda quincena del mes de diciembre de 2021, de acuerdo a sus posibilidades operativas, por lo que los registros capturados y proporcionados por el equipo médico estuvieron limitados en cuanto al tamaño y calidad de la muestra, sobre todo por el corto periodo de tiempo dedicado al estudio. Para el levantamiento de información no se consideró la estratificación del muestreo, únicamente se respetó el criterio de edad donde los médicos residentes aplicaron el cuestionario a pacientes mayores de 15 años.

El formulario en línea se desarrolló mediante el lenguaje de programación PHP, y se efectuaron las configuraciones necesarias para conectar dicho formulario con un servidor de la Universidad Autónoma de Coahuila, sitio de almacenamiento institucional donde fue alojada la información general y clínica de la población encuestada haciendo uso de la tecnología MySQL. La interfaz de captura en su pantalla principal proporcionó una serie de recomendaciones al personal de salud con el propósito de apoyar y orientar a los pacientes en cada una de las preguntas planteadas en la encuesta. La segunda pantalla del formulario en línea incluyó preguntas de información general

y clínicas planteadas a los derechohabientes del IMSS. Algunas de ellas fueron de opción múltiple, donde el personal médico planteó una serie de opciones de respuesta, mientras que otras fueron capturadas en formato numérico o de texto. Para una mejor referencia consulte el Anexo B *Instrumento de investigación*.

El número total de pacientes contenidos en el dataset original fue de 1,896 personas en el rango de edad de 15 a 85 años, como se muestra en la siguiente figura (3). De ellos, 976 corresponden al sexo femenino y 914 al sexo masculino, se tuvo evidencia de 6 personas a las cuales no les fue registrado su sexo.

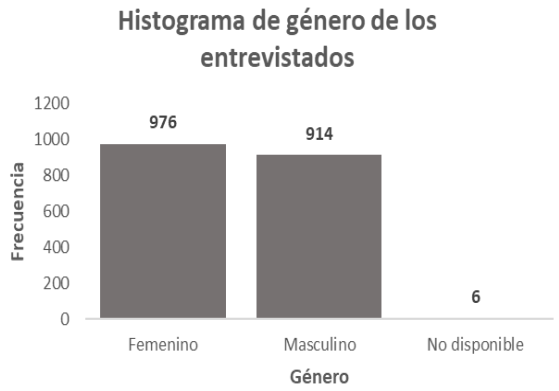

Figura 3. *Género de los entrevistados.*
Fuente: elaboración propia.

Dicha población encuestada mostró una distribución de 166 pacientes en edad joven de 15 a 24 años, 1,289 personas se concentraron en la etapa adulta de 25 a 59 años, mientras que 435 se ubicaron en la etapa adulto mayor de 60 años en adelante, respectivamente. En el atributo edad, se identificó a 6 personas a las cuales no les fue recabada dicha información.

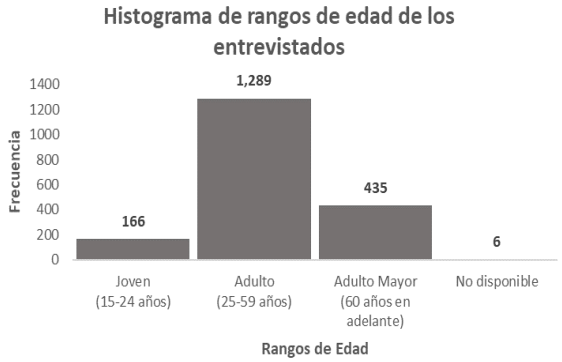

Figura 4. *Rango de edad de los entrevistados.*
Fuente: elaboración propia.

Por otro lado, la muestra registró una media aritmética de 47 años de edad y una desviación estándar de 15 años, lo cual refleja una ligera dispersión de los datos donde, la mayoría de los pacientes entrevistados se ubicó entre los 32 y 62 años de edad.

III. Métodos experimentales

A continuación, se presenta la descripción de las fases de procesamiento de datos y la experimentación implementando los modelos de aprendizaje automático para la detección de la DM2.

Limpieza de Datos

Para el proceso de limpieza de datos se ejecutó un script en lenguaje Python. Con el objetivo de focalizar los esfuerzos en una base de datos más adecuada para su posterior análisis mediante algoritmos de aprendizaje automático, se omitieron del conjunto de datos aquellas variables con información general y socioeconómica como *Escolaridad, Ocupación, Fuente de ingresos*, *Ingreso Mensual*, *Encuestador*, *Evento* y *Fecha del Evento*. Dicha información es valiosa para otros análisis fuera del objetivo de la presente investigación. Adicionalmente y, debido a la presencia de una gran cantidad de datos faltantes o falta de congruencia entre la respuesta sobre la presencia o ausencia de alguna alteración de salud (respondida por el paciente directamente) y los niveles o métricas registrados por el personal médico sobre dicho

paciente, se omitieron del *conjunto de datos* variables como *IMC*, *Hipertensión Arterial*, *Diabetes Tipo 2*, *Hemoglobina Glucosilada*, *Nivel de Hemoglobina Glucosilada*, *# de refrescos*, *Triglicéridos alterados*, *Nivel de Triglicéridos*, *Colesterol LDL alterado*, *Nivel Colesterol LDL*, *# de veces que acude a orinar*, *# de kilogramos perdidos* e *Hijos con más de 4 kg al nacer*.

De acuerdo con los niveles de glucosa del total de encuestados, existen 509 personas con características clínicas de DM2 (nivel de glucosa superior a 126 mg/dl), 1,311 pacientes se ubicaron en el grupo de personas saludables con niveles de glucosa inferiores al mismo nivel y 76 personas no contaron con registro de nivel de glucosa en sangre. Si se excluye del análisis a los 76 pacientes de los cuales se desconoce el nivel de glucosa en sangre es posible distinguir dos clases o etiquetas, por un lado, a aquellos pacientes sin características de DM2 (debido a que su glucosa fue menor a 126 mg/dl) y a aquellos pacientes con características de la mencionada enfermedad (glucosa mayor a 126 mg/dl), respectivamente.

Valores faltantes

Por otro lado, se detectaron un total de 922 valores faltantes, la Tabla 1 muestra la distribución de éstos. Para imputar los datos, en el caso de los atributos numéricos se utilizó el Método *k Nearest Neighbors* (kNN), considerando a sus 5 vecinos más cercanos, así como pesos "uniformes", donde todos los puntos en cada vecindario fueron ponderados por igual. La razón de utilizar este imputador es que toma en cuenta a todos los atributos de los vecinos más cercanos (y que más se parecen) para, posteriormente estimar el valor faltante (Ciaburro, 2022), además de ser adecuado cuando existen patrones como aleatoriedad o aleatoriedad completa (MCAR y MAR) en los datos faltantes (Aljrees, 2024). Cabe hacer mención, que existen otros imputadores que toman en cuenta sólo las observaciones del atributo donde se encuentra el mencionado valor perdido, sin embargo, al tratar con variables clínicas que se relacionan fuertemente entre sí, un imputador como kNN es altamente útil. Para los atributos categóricos se empleó la estrategia *"most_frequent"* que reemplaza los valores perdidos usando el valor más frecuente en cada columna. Luego de la imputación se eliminaron los valores faltantes, obteniendo, de manera homogénea 1,820 observaciones en cada atributo del conjunto de datos.

Tabla 1. Registro del total de valores faltantes del conjunto de datos de DM2.

Atributo	Total
Estatura	0
Peso	2
Nivel de presión arterial	38
Genero	6
Edad	0
Obesidad	5
Antecedentes de Diabetes	6
Consumo mayor a 3 refrescos diarios	6
Hambre excesiva	7
Sed excesiva	7
Orina frecuente	7
Pérdida de Peso	6
Actividad física	278
Visión borrosa	239
Presión y estrés	239
Nivel de Glucosa	76
Total	**922**

Balanceo de clases y partición de la base de datos

En la base de datos con 1,820 registros 1,311 fueron identificados como *No diabéticos* y 509 fueron etiquetados como *Diabéticos*. Por tal motivo se trabajó en la aplicación de la técnica de balanceo de datos *Downsampling* (Mîrț *et al*., 2022), la cual llevó a cabo una reducción de muestreo donde se eliminaron registros de la clase mayoritaria, esto es, No diabéticos, creando un conjunto de datos más equilibrado para ser analizado posteriormente mediante los algoritmos de aprendizaje automático. Dicha reducción de muestra aplicada a la clase

mayoritaria consistió en la reducción aleatoria de registros de esa categoría, para concluir con la construcción del denominado *Dataset_balanceado* con una muestra de 1,018 pacientes diabéticos y no diabéticos. Cabe mencionar que para este trabajo de investigación se eligió el enfoque *Downsampling* debido a que a diferencia de otros enfoques como *Upsampling* (Zhang *et al.*, 2022) o SMOTE (Taneja *et al*., 2019), éste no genera datos sintéticos, lo cual se considera adecuado tratándose de una investigación que usa datos provenientes de personas, sin embargo, en futuras investigaciones se explorarán algunos otros enfoques. Adicionalmente, se construyó el *dataset excluded_data* el cual almacena un total de 802 registros con la información excluida durante la etapa de balanceo. La composición de cada *dataset* se aprecia en la Tabla 2.

Tabla 2. Composición de subconjuntos de datos tras el proceso de balanceo de clases.

Nombre del subconjunto de datos	# de Registros	Diabéticos	No diabéticos
Bd_original	1,820	1,311	509
Dataset_balanceado	1,018	509	509
Excluded_data	802	211	591

Creación de subconjuntos de entrenamiento y prueba

Una de las primeras tareas en todo proyecto de aprendizaje automático, es el trabajo en las etapas de entrenamiento y prueba, training y testing, respectivamente. Para efectos de esta investigación se utilizó el *Dataset_balanceado*, aplicando la siguiente partición:

- Entrenamiento de los modelos: 80% (814 filas, 19 columnas).
- Prueba de los modelos: 20% (204 filas, 19 columnas).

Implementación y comparación de modelos de aprendizaje automático

Se trabajó con 14 modelos de aprendizaje automático, cuyo propósito es predecir etiquetas de clase categóricas las cuales son discretas y no poseen un orden específico. Las clases categóricas consideradas son:

- **0 -** Paciente sin DM2
- **1 -** Paciente con DM2

Se realizó una experimentación computacional en tres fases. El primer experimento consta de evaluar el desempeño de los 14 modelos usando los valores por defecto de sus respectivos hiperparámetros. Adicionalmente, para garantizar la reproducibilidad de las pruebas se configuraron desde el inicio parámetros como remoción de valores atípicos, estado de semilla que es utilizado para inicializar un generador de números pseudoaleatorios, y remoción de colinealidad. Asimismo, todas las pruebas se llevaron a cabo usando validación cruzada de 10 subconjuntos para evitar el sobre entrenamiento de los modelos. Para la comparación de éstos se usaron distintas métricas que se enuncian enseguida:

- Exactitud: es una métrica que generalmente describe el rendimiento del modelo en todas las clases. Es útil cuando todas las clases tienen la misma importancia. Se calcula como la relación entre el número de predicciones correctas y el número total de predicciones.
- Sensibilidad: se calcula como la relación entre el número de muestras positivas clasificadas correctamente como positivas y el número total de muestras positivas. Mide la capacidad del modelo para detectar muestras positivas.
- Precisión: se calcula como la relación entre el número de muestras positivas clasificadas correctamente y el número total de muestras clasificadas como positivas, correcta o incorrectamente. La precisión mide la exactitud del modelo al clasificar una muestra como positiva.
- Área Bajo la Curva (AUC): Es la

medida de la capacidad de un clasificador para distinguir entre clases y se utiliza como resumen de la curva ROC (*Receiver Characteristic Operator*). Cuanto mayor sea el AUC, mejor será el rendimiento del modelo para distinguir entre las clases positivas y negativas. Cuando AUC = 1, entonces el clasificador es capaz de distinguir perfectamente entre todos los puntos de clase positivos y negativos correctamente.

Estas métricas son obtenidas mediante la Matriz de Confusión (Bhandari, 2022), la cual permitió observar qué tipos de aciertos y errores se presentaron durante el aprendizaje de los modelos. En este sentido, fue posible distinguir cuatro opciones únicas:

1. Paciente que tiene DM2 y el modelo lo clasificó correctamente. Este caso sería un verdadero positivo o VP.
2. Paciente que no tiene DM2 y el modelo lo clasificó correctamente. Este sería un caso verdadero negativo, es decir, VN.
3. Paciente que tiene DM2 y el modelo lo clasificó como No diabético. Este sería un falso negativo o FN, también conocido en Estadística como Error Tipo II.
4. Paciente que no tiene DM2 y el modelo lo clasificó como Diabético. Este caso sería un falso positivo o FP y se conoce como Error Tipo I.

Bajo este contexto y, en términos de análisis, se enfatiza en detectar aquellos algoritmos que permitan reducir las predicciones de falsos negativos, ya que, bajo este escenario, pacientes con presencia de la enfermedad quedarían alejados de los beneficios de una detección temprana de la DM2 y, además, sacrificarían involuntariamente la utilidad de los medicamentos, así como de sus tratamientos de control. De este primer experimento se obtuvieron los siguientes resultados (ver tabla 3).

Tabla 3. Comparativo de los 14 modelos de Machine Learning.

Modelo	Accuracy	AUC	Recall	Precision
1. k Neighbors	0.6636	0.7011	0.6584	0.6724
2. Naive Bayes	0.6612	0.7041	0.7919	0.6305
3. Linear Discriminant Analysis	0.6586	0.717	0.6841	0.6523
4. Logistic Regression	0.6573	0.7093	0.6866	0.65
5. Ridge	0.656	0	0.6815	0.6505
6. Random Forest	0.6546	0.7027	0.6685	0.6539
7. Quadratic Discriminant Analysis	0.6482	0.7066	0.7766	0.6284
8. Extra Trees	0.6402	0.6926	0.6583	0.6395
9. Gradient Boosting	0.63	0.684	0.6143	0.6366
10. Light Gradient Boosting Machine	0.6287	0.6763	0.6299	0.6336
11. Extreme Gradient Boosting	0.6222	0.665	0.6426	0.6236
12. Ada Boost	0.6144	0.6751	0.6117	0.6214
13. SVM - Linear Kernel	0.6119	0	0.642	0.636
14. Decision Tree	0.6067	0.6068	0.6402	0.6054

Una vez evaluados los modelos, se seleccionaron los tres con mejor desempeño, estos fueron Naive Bayes, Random Forest y k Nearest Neighbors.

Naive Bayes (NB)

NB es un instrumento probabilístico basado en la aplicación de fuertes supuestos de independencia entre las características. Este clasificador proporciona una forma de calcular la probabilidad condicional, $P(c|x)$, a partir de $P(c)$, $P(x)$ y $P(x|c)$. Asume que el efecto del valor de un predictor (x) en una clase dada (c) es independiente de los valores de otros predictores (Rrmoku *et al*., 2020). Esta suposición se llama independencia condicional de clase, como se muestra enseguida:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

Donde,

$P(c|x) = P(x_1|c) \times P(x_2|c) \times \ldots \times P(x_n|c) \times P(c)$

Siendo c y x dos eventos y:

- $P(c)$ la probabilidad de la previa clase
- $P(x)$ $la\ probabilidad\ previa\ del\ predictor$
- $P(c|x)$ es la probabilidad condicional de la clase (objetivo) dado un predictor (atributo)
- $P(x|c)$ es la probabilidad condicional que tiene un predictor dada una clase.

Este clasificador utiliza la teoría de la probabilidad para encontrar la clasificación más probable de una instancia no vista (no clasificada).

Random Forest (RF)

RF es una colección o conjunto de árboles de clasificación y regresión entrenados en conjuntos de datos del mismo tamaño que el conjunto de entrenamiento, llamados *bootstraps*, creados a partir de un remuestreo aleatorio en el propio conjunto de entrenamiento. Una vez que se construye un árbol, se usa como conjunto de prueba un conjunto de *bootstraps*, que no incluye ningún registro en particular del conjunto de datos original. El error de clasificación de los conjuntos de prueba es la estimación del error de generalización de los modelos. En estricto sentido, para clasificar nuevos datos de entrada, cada árbol individual vota por una clase y el bosque predice la clase que obtiene la mayoría de los votos. Algunas de las ventajas de esta técnica incluyen su alta precisión debido al promedio de predicciones de múltiples árboles, así como su poco sobre entrenamiento (Myśliwiec *et al*., 2024).

k Nearest Neighbors (kNN)

kNN es un método de Aprendizaje supervisado basado en distancias como puede ser el caso de la distancia euclidiana. Para implementar kNN primero se dividen los datos de estudio en datos de entrenamiento y de prueba, posteriormente se elige el valor de k y se calculan las distancias de cada punto conocido de entrenamiento con respecto a un punto desconocido de prueba. Las distancias obtenidas se ordenan de forma ascendente y las primeras k distancias son seleccionadas. Para fines de clasificación kNN clasifica basándose en la mayoría de sus vecinos más cercanos. El desempeño del clasificador depende sólo de dos parámetros, k y el tipo de distancia (Chai *et al*., 2023).

Respecto a los resultados del primer experimento se puede mencionar que para el caso del modelo Naive Bayes con afinación de las métricas Exactitud y Sensibilidad, el algoritmo clasificó a 70 pacientes con la presencia de este padecimiento, mismos que, en efecto, respondieron en la encuesta ser diabéticos, estos casos se pueden considerar como verdaderos positivos, VP, según los datos mostrados en su Matriz de Confusión, como se muestra en la Figura 5.

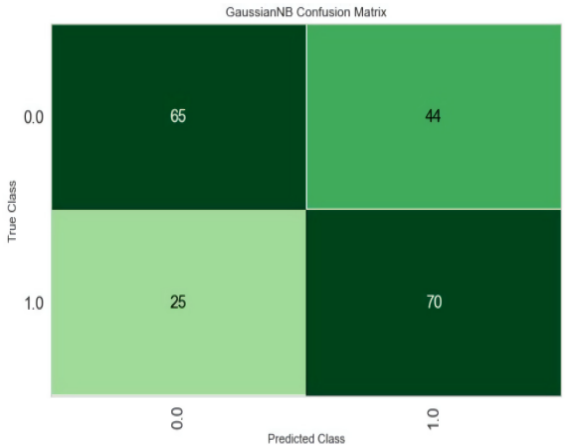

Figura 5. *Matriz de confusión Naive Bayes.*
Fuente: elaboración propia.

En el mismo sentido, logró clasificar a 65 pacientes como no diabéticos, los cuales respondieron no tener Diabetes Mellitus, estos casos se conocen como verdaderos negativos o VN.

Mientras que los casos anteriores representan los aciertos del modelo, la diagonal derecha de la matriz muestra los errores cometidos por éste, en estricto sentido, clasificó a 25 pacientes como no diabéticos cuando en verdad sí tienen la enfermedad (falsos negativos o FN). En el caso contrario, predijo a 44 pacientes con Diabetes, mismos que durante el levantamiento de información declararon no tener la enfermedad (falsos positivos o FP).

Respecto al modelo Random Forest (Figura 6) se pueden destacar los siguientes resultados: Predijo 58 casos de pacientes diabéticos, mismos que confirmaron tener la enfermedad, VP. A un total de 77 pacientes los predijo como no diabéticos, mismos que, a su vez, habían manifestado estar saludables, VN. La misma ejecución sobre datos de prueba arrojó una predicción de 37 casos negativos a Diabetes, cuando en realidad sí tenían dicho padecimiento, FN. Asimismo, obtuvo un total de 32 casos positivos a Diabetes, cuando en realidad los mencionados pacientes manifestaron no ser diabéticos, FP.

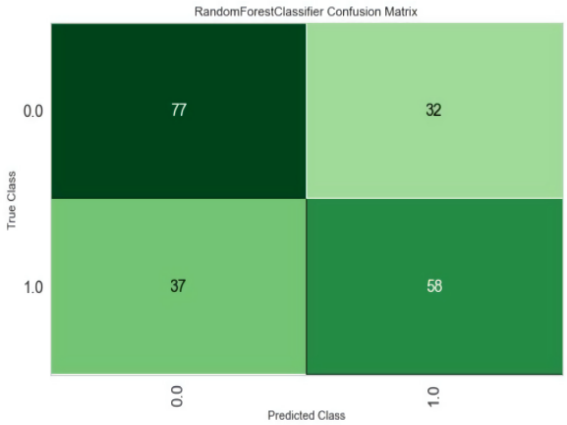

Figura 6. *Matriz de confusión Random Forest.*
Fuente: elaboración propia.

En cuanto al Clasificador kNN (Figura 7), se identificaron 30 VP, que indican el número de pacientes diabéticos que el modelo clasificó correctamente. Por el contrario, 78 casos se corresponden con los VN y se refieren al número de predicciones de pacientes no diabéticos que el modelo clasificó correctamente. A su vez, 65 pacientes fueron etiquetados erróneamente por el modelo como personas sin Diabetes, cuando en realidad sí padecen la enfermedad, FN, mientras que 31 personas fueron clasificadas como diabéticas, cuando en realidad no presentan la enfermedad, FP.

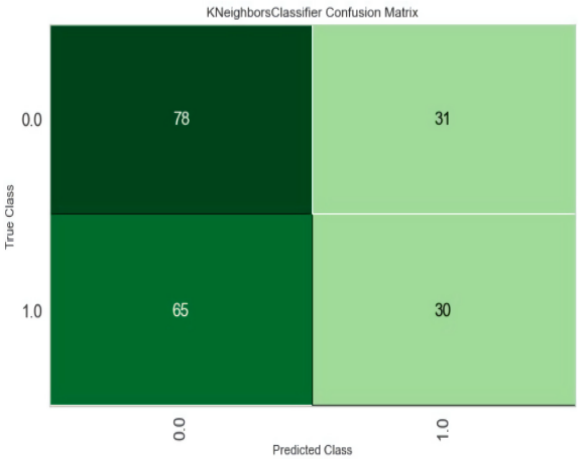

Figura 7. *Matriz de confusión kNN.*
Fuente: elaboración propia.

Posteriormente, como segundo experimento se procedió a implementar la técnica de búsqueda de malla para sintonizar los hiperparámetros de los tres modelos seleccionados con el fin de mejorar su desempeño. Cabe mencionar que algunos de los mejores valores de las métricas de evaluación sobre el conjunto de datos de entrenamiento se obtuvieron mediante la aplicación de dicha técnica de sintonización de hiperparámetros.

La tercera experimentación consistió en probar el desempeño de los modelos sobre el conjunto de datos de prueba (representados por el 20% del total de los datos adquiridos en principio), es decir evaluar la capacidad de clasificación de los modelos sobre un conjunto de datos totalmente desconocido por los mismos. La idea detrás de esta verificación fue observar si dichas predicciones fueron significativamente diferentes a los resultados de validación cruzada durante la etapa de entrenamiento aplicada al 80% de los datos destinados para tal efecto.

Los resultados de estos experimentos computacionales se discuten en la siguiente sección.

IV. Discusión de Resultados

El levantamiento de información permitió el entrenamiento de 14 modelos de Aprendizaje Automático y, las predicciones resultantes fueron puestas a prueba sobre un subconjunto de datos no conocidos por dichos modelos.

En la Tabla 5 se muestran los desempeños obtenidos en la etapa de predicción utilizando el modelo Naive Bayes, una vez concluido el proceso de sintonización de las métricas Exactitud y Sensibilidad. Destaca un comportamiento similar del algoritmo en los resultados sobre el conjunto de prueba, comparado con las cifras obtenidas durante el entrenamiento. Este hecho no mostró indicios de sobreajuste (overfitting), un caso de sobreajuste pudo haber sido detectado con un alto desempeño sobre el conjunto de entrenamiento y bajo en el conjunto de prueba. En esta etapa experimental, la métrica Exactitud obtuvo un valor de 0.6664 durante el entrenamiento y una cifra de 0.6618 en la de prueba. Asimismo, en la métrica Sensibilidad obtuvo un nivel de 0.7996 durante el entrenamiento y un desempeño de 0.7368 durante la etapa de prueba.

A su vez, el algortimo Random Forest arrojó un resultado de 0.7266 en la métrica Área Bajo la Curva ROC (AUC) durante el entrenamiento, contra 0.6772 obtenido en los datos de prueba, lo cual tampoco proporcionó señales de sobreajuste al resultar similares sus cifras en ambas etapas.

Tabla 5. Resultados de Naive Bayes, Random Forest y kNN.

Naive Bayes			
Desempeño	**Entrenamiento**	**Prueba**	**Sobreajuste**
Exactitud	0.6664	0.6618	No
Sensibilidad	0.7996	0.7368	

Random Forest			
Desempeño	**Entrenamiento**	**Prueba**	**Sobreajuste**
Área Bajo la Curva ROC (AUC)	0.7266	0.6772	No

kNN			
Desempeño	**Entrenamiento**	**Prueba**	**Sobreajuste**
Precisión	0.6843	0.4918	Sí

Respecto al modelo kNN, luego de sintonizar la métrica Precisión, fue identificado un caso de sobreajuste sobre los datos de entrenamiento. Este comportamiento se pudo comprobar con la obtención del valor 0.6843 durante la etapa de aprendizaje y un nivel de 0.4918 durante la etapa de prueba, lo cual

sugirió la memorización de patrones en los datos por parte del algoritmo, por tanto, su mal desempeño al momento de llevar a cabo sus predicciones. El sobreajuste indicó, además, que el modelo predictivo no fue lo suficientemente flexible para realizar predicciones sobre datos reales.

Existen distintas estrategias para lidiar con el sobreajuste de los modelos de Aprendizaje Automático, algunas de ellas consisten en utilizar más datos de entrenamiento, utilizar menos variables o atributos, regularizar y optimizar los hiperparámetros, reducir la complejidad del modelo, entre otros. Para efectos de esta investigación, se excluyó al Clasificador kNN, debido a que no cumplió los requerimientos mínimos en términos predictivos para la detección de enfermedades.

El presente estudio de salud enfocado en la detección temprana de la DM2, tuvo como uno de sus objetivos la reducción de los casos de falsos negativos, es decir, se propuso minimizar aquellos casos donde el modelo predice que los pacientes no tienen la enfermedad, cuando en efecto declaran que sí la tienen, es decir, los eventos cuando el modelo no logra detectar a los diabéticos tipo 2.

Tabla 6. Tasa de falsos negativos para Naive Bayes, Random Forest y kNN.

	Naive Bayes	**Random Forest**	**K Neighbors**
Tasa de falsos negativos	0.2632	0.3895	0.6842

De manera complementaria al análisis de predicciones que se presenta en 1a Tabla 6, se muestran los resultados obtenidos por los modelos utilizando la métrica tasa de falsos negativos: el modelo Naive Bayes fue el que registró una tasa menor con 0.2632, seguido del Random Forest con 0.3895 y kNN con un nivel de 0.6842, respectivamente. El modelo Naive Bayes fue el que realizó un mejor trabajo de generalización sobre datos no vistos en la etapa de entrenamiento, el Random Forest presentó un rendimiento medio, mientras que el de kNN presentó el peor desempeño, mostrando su mala capacidad predictiva para detectar a pacientes con la presencia de la enfermedad.

Los resultados obtenidos a través de los modelos seleccionados en esta investigación se limitan a la población de 15 años o más, originaria de la Región Sureste del Estado de Coahuila, particularmente de la ciudad de Saltillo, debido a que la información que sirvió de entrenamiento de dichos modelos pudo no haber sido representativa de cualquier otra población. Para una utilización más generalizada, se tendrían que recopilar muestras más representativas ya que se desconoce la existencia de factores regionales o ambientales que podrían alterar el comportamiento de estos modelos.

El entrenamiento de los clasificadores utilizados en esta investigación, se efectuó con las características (atributos) recomendados por el personal médico del IMSS, relacionadas con los factores de riesgo de la DM2. Su comportamiento predictivo quedó sujeto, entre otros factores, a la calidad de la información proporcionada por el equipo de médicos encuestadores.

V. Conclusiones

Ante la falta de estudios regionales o locales sobre la detección temprana de enfermedades como la DM2, durante esta investigación, se enfocaron esfuerzos en el diseño de un instrumento que permitió identificar los factores de riesgo para el desarrollo de la enfermedad. Dicho instrumento fue materializado en un formulario en línea con el cual un equipo de residentes médicos del IMSS en la ciudad de Saltillo, Coahuila,

recopiló información clínica de pacientes derechohabientes mayores de 15 años con y sin factores de riesgo relacionados con este padecimiento. La información recabada fue posteriormente procesada y empleada en el entrenamiento de distintos modelos de Aprendizaje Automático y sus predicciones puestas a prueba sobre datos no vistos con anterioridad por dichos modelos. En la etapa de validación del desempeño, se utilizaron distintas métricas como Exactitud, Precisión, Sensibilidad, entre otras, para evaluar la capacidad de generalización de las técnicas. Vale la pena precisar que, en los problemas de Aprendizaje Supervisado, particularmente en los de clasificación, el hecho de contar con una gran cantidad de observaciones, o bien, con una cantidad considerada de atributos, no necesariamente garantiza buenos resultados en sus métricas de evaluación. Dicha capacidad predictiva y/o de generalización, estuvo subordinada a la calidad de los datos proporcionada por el equipo médico de encuestadores, mismo que se adecuó a los registros disponibles en los expedientes clínicos de los pacientes entrevistados, así como en las respuestas proporcionadas por ellos mismos. Los resultados demostraron buena capacidad predictiva y de generalización, por parte de los algoritmos Naive Bayes y Random Forest. El clasificador Naive Bayes, por ejemplo, logró una Exactitud de 0.6664 durante el entrenamiento y 0.6618 en la etapa de prueba, no obstante, en la métrica Sensibilidad obtuvo un valor de 0.7996 en la etapa de entrenamiento, mientras que en la prueba obtuvo un valor de 0.7368. Por su parte, el modelo Random Forest obtuvo 0.7266 en la etapa de entrenamiento y 0.6772 en la métrica Área Bajo la Curva ROC (AUC) en la etapa de prueba. Finalmente, el modelo kNN registró en la etapa de entrenamiento un valor de 0.6843 en la métrica Precisión y, un nivel de 0.4918 durante la etapa de predicción, por lo que fue rechazado debido al evidente sobreajuste de datos sobre la etapa previa de entrenamiento.

No obstante, no existe un modelo o clasificador ideal para todos los casos de estudio, por lo que el desempeño de las diversas técnicas de aprendizaje automático está sujeto al problema que se desea atender. Por tal motivo, en este caso de estudio se probó con una variedad de modelos para seleccionar aquellos que satisficieran las necesidades de un tema médico tan relevante como la detección de la DM2. Por otro lado, en este estudio se trabajó con atributos clínicos obtenidos a través de mediciones físicas o declarados directamente por el paciente, excepto el de nivel de glucosa en sangre el cual fue obtenido a través de revisiones capilares a dichos entrevistados. Sin embargo, de acuerdo con la revisión de literatura, se obtuvieron mejores resultados en otros estudios cuando se utilizaron otros atributos sintomáticos o procedentes de métricas de laboratorio, lo cual hace que este trabajo de investigación sea un parteaguas para incitar a los centros de salud a registrar de manera digital la mayor cantidad posible de información acerca de los pacientes. Dicha práctica podría aumentar de manera significativa el desempeño de los modelos presentados y brindar un mejor apoyo a los responsables de la toma de decisiones, no sólo en el área de la detección de diabetes, sino de otras enfermedades y padecimientos.

VI. Agradecimientos

Los autores agradecen al personal directivo de la Unidad de Medicina Familiar No. 82 del Instituto Mexicano del Seguro Social en Saltillo, Coahuila todas las facilidades otorgadas para la realización de esta investigación.

VII. Referencias

1. Encuesta Nacional de Salud y Nutrición (ENSANUT) 2022, Página del

Instituto Nacional de Salud Pública, México, recuperado 21 marzo 2024. https://ensanut.insp.mx/

2. Basto-Abreu, A., López-Olmedo, N., Rojas-Martínez, R., Aguilar-Salinas, C. A., Moreno-Banda, G. L., Carnalla, M., Rivera, J.A., Romero-Martínez, M., Barquera, S., Barrientos-Gutiérrez, T. (2023). Prevalencia de prediabetes y diabetes en México: Ensanut 2022. *Salud Pública de México,* 65, s163-s168.

3. World Health Organization. (2016). *Global report on diabetes: executive summary* (No. WHO/NMH/NVI/16.3). World Health Organization.

4. Basto-Abreu, A., Barrientos-Gutiérrez, T., Rojas-Martínez, R., Aguilar-Salinas, C. A., López-Olmedo, N., De la Cruz-Góngora, V., Rivera-Dommarco, J., Shamah-Levy, T., Romero-Martínez, M., Barquera, S., López-Ridaura, R. Hernández-Ávila, M., Villalpando, S. (2020). Prevalencia de Diabetes y descontrol glucémico en México: Resultados de la ENSANUT 2016. *Salud Pública de México*, 62, 50–59.

5. Escamilla, L. C. (2021). Panorama epidemiológico de la diabetes tipo 2 en la frontera norte de México. El Colegio de Sonora, *Centro de Estudios en Salud y Sociedad,* 1-77.

6. Gobierno del Estado de Coahuila de Zaragoza (2020). www.coahuila.gob.mx, Recuperado el día 04 de Octubre de 2022.

7. Herman, W. H., Ye, W., Griffin, S. J., Simmons, R. K., Davies, M. J., Khunti, K., Rutten, G. E., Sandbaek, A., Lauritzen, T., Borch-Johnsen, K., Brown, M.B. (2015). Early detection and treatment of Type 2 Diabetes reduce cardiovascular morbidity and mortality: A simulation of the results of the Anglo-Danish-Dutch Study of Intensive Treatment in People With Screen-Detected Diabetes in Primary Care (ADDITION-Europe). *Diabetes Care*, 38(8):1449–1455.

8. Villalpando, S., de La Cruz, V., Rojas, R., Shamah-Levy, T., Ávila, M. A., Gaona, B., Rebollar, R., Hernández, L. (2010). Prevalence and distribution of type 2 diabetes mellitus in Mexican adult population: a probabilistic survey. *Salud Pública de México*, 52, S19-S26.

9. Sidey-Gibbons, J. A., Sidey-Gibbons, C. J. (2019). Machine Learning in Medicine: A practical introduction. *BMC Medical Research Methodology,* 19,1–18.

10. Kaur, H., Kumari, V. (2020). Predictive modelling and analytics for Diabetes using a Machine Learning approach. *Applied Computing and Informatics*, 18(1/2), 90-100.

11. Mejía, J. A., Oviedo-Benalcázar, M. A., Ordoñez, J. A., Valencia-Murillo, J. F. (2023). Aprendizaje automático aplicado a la predicción de diabetes mellitus, utilizando información socioeconómica y ambiental de usuarios del sistema de salud. *Revista Facultad Nacional de Salud Pública*, 41(2).

12. Chen, M., Hao, Y., Hwang, K.,Wang, L., andWang, L. (2017). Disease prediction by Machine Learning over Big Data from healthcare communities. *IEEE Access*, 5:8869–8879.

13. Chang, V., Ganatra, M. A., Hall, K., Golightly, L., Xu, Q. A. (2022). An assessment of machine learning models and algorithms for early prediction and diagnosis of diabetes using health indicators. *Healthcare Analytics*, 2, 100118.

14. Oladimeji, O. O., Oladimeji, A., Oladimeji, O. (2021). Classification models for likelihood prediction of Diabetes at early stage using Feature Selection. *Applied Computing and Informatics.*

15. Kumari, S., Kumar, D., Mittal, M. (2021). An ensemble approach for classification and prediction of Diabetes Mellitus using Soft Voting Classifier. *International Journal of Cognitive Computing in Engineering*, 2, 40–46.

16. Das, U., Srizon, A. Y., Islam, M. A., Tonmoy, D. S., Hasan, M. A. M. (2020). Prognostic biomarkers identification for Diabetes prediction by utilizing Machine Learning classifiers. In *2020 2nd*

International Conference on Sustainable Technologies for Industry 4.0 (STI), pp 1–6. IEEE.
17. Muhammad, L., Algehyne, E. A., Usman, S. S. (2020). Predictive Supervised Machine Learning models for Diabetes Mellitus. *SN Computer Science*, 1(5), 240.
18. Reddy, D. J., Mounika, B., Sindhu, S., Reddy, T. P., Reddy, N. S., Sri, G. J., Swaraja, K., Meenakshi, K., Kora, P. (2020). Predictive Machine Learning model for early detection and analysis of Diabetes. Materials Today: Proceedings.
19. Tigga, N. P., Garg, S. (2020). Prediction of Type 2 Diabetes using Machine Learning classification methods. *Procedia Computer Science*, 167:706–716.
20. Islam, M. S., Qaraqe, M. K., Abbas, H. T., Erraguntla, M., Abdul-Ghani, M. (2020). The prediction of Diabetes development: A Machine Learning framework. In *2020 IEEE 5th Middle East and Africa Conference on Biomedical Engineering (MECBME),* pages 1–6. IEEE.
21. Mujumdar, A., Vaidehi, V. (2019). Diabetes prediction using Machine Learning algorithms. *Procedia Computer Science*, 165:292–299.
22. Rawat, V., Suryakant, S. (2019). A classification system for diabetic patients with MachineLearning techniques. *International Journal of Mathematical, Engineering and Management Sciences*, 4(3):729–744.
23. Abed, Mahmood, Ibrıkçı, T. (2019). Comparison between Machine Learning algorithms in the predicting the onset of Diabetes. In *2019 International Artificial Intelligence and Data Processing Symposium (IDAP),* pages 1–5. IEEE.
24. Aada, A., Tiwari, S. (2019). Predicting Diabetes in medical datasets using Machine Learning techniques. *Int. J. Sci. Res. Eng. Trends*, 5:257–267.
25. Faruque, M. F., Sarker, I. H., (2019). Performance analysis of Machine Learning techniques to predict Diabetes Mellitus. In *2019 International Conference on Electrical, Computer and Communication Engineering* (ECCE), pages 1–4. IEEE
26. Kaur, P., Sharma, N., Singh, A., Gill, B. (2018). CI-DPF: A cloud IoT based framework for Diabetes prediction. In *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON),* pp. 654–660. IEEE.
27. Zou, Q., Qu, K., Tang, H. (2018). Predicting Diabetes Mellitus with Machine Learning techniques. *Frontiers in Genetics*, page 515.
28. Wu, H., Yang, S., Huang, Z., He, J., Wang, X., (2018). Type 2 diabetes mellitus prediction model based on data mining. *Informatics in Medicine Unlocked*, 10:100–107.
29. Alehegn, M., Joshi, R., Mulay, P. (2018). Analysis and prediction of Diabetes Mellitus using Machine Learning Algorithm. *International Journal of Pure and Applied Mathematics*, 118(9):871–878.
30. Rrmoku, K., Selimi B., Ahmedi, L. (2022). Application of Trust in Recommender Systems Utilizing Naive Bayes Classifier. *Computation*. 10(1):6.
31. Myśliwiec, P., Kubit, A., Szawara, P. (2024). Optimization of 2024-T3 Aluminum Alloy Friction Stir Welding Using Random Forest, XGBoost, and MLP Machine Learning Techniques. *Materials*. 17(7):1452.
32. Chai, B. et al. (2023). Application of KNN and ANN Metamodeling for RTM Filling Process Prediction. *Materials*. 16(18):6115.
33. Ciaburro G. (2022). Machine fault detection methods based on machine learning algorithms: A review. *Mathematical Biosciences and Engineering*. 19(11): 11453-11490.
34. Zhang, Y., Zhao, W., Sun, B., Zhang, Y., Wen, W.(2022). Point Cloud Upsampling Algorithm: A Systematic Review. *Algorithms*. 15(4):124.
35. Taneja, S., Suri, B., Kothari, C. (2019). Application of Balancing Techniques

with Ensemble Approach for Credit Card Fraud Detection. International Conference on Computing, Power and Communication Technologies.
36. Mîrț, A., Reiche, J., Verbesselt, J., Herold, M. (2022). A Downsampling Method Addressing the Modifiable Areal Unit Problem in Remote Sensing. *Remote Sensing.* 14(21): 5538.
37. Bhandari, A. (2022). AUC-ROC Curve in Machine Learning clearly explained. In Analytics Vidhya. Analytics Vidhya.
38. Aljrees T. Improving prediction of cervical cancer using KNN imputer and multi-model ensemble learning. *PLoS One*. 2024 Jan 3;19(1)

Anexos

Anexo A. ***Recopilación documental de trabajos científicos.***

Autor	Características / Base de Datos	País	Tamaño de la muestra	Algoritmo	Exactitud
Mejía *et al.* (2023)	Datos ambientales, sociales, económicos y sanitarios	Colombia	10889	-kNN - Decision Tree - Random Forest	53.87% 60.02% 59.93%
Chang *et al.* (2022)	Diabetes, hipertensión, colesterol alto, Índice de Masa Corporal, fumador, infarto, enfermedad cardíaca o ataque, actividad física, consumo de frutas, consumo de verduras, consumo excesivo de alcohol, tiene algún tipo de cobertura de atención médica, falta de atención médica debido al costo, estado de salud general, salud mental, salud física, dificultades para caminar o subir escaleras, sexo, edad, nivel de educación, ingreso.	USA	70,692	-Decision Tree -Random Forest -kNN -Logistic Regression -Naive Bayes	81.02% 82.26% 80.55% 72.64% 70.56%
Oladimeji *et al.* (2021).	Edad, sexo, poliuria, polidipsia, pérdida de peso repentino, debilidad, polifagia, candidiasis genital, visión borrosa, hormigueo o picazón, irritabilidad, cicatrización retardada, parálisis parcial, rigidez muscular, pérdida de cabello, obesidad (Diabetes Hospital).	Nigeria	520	-Random Forest + C. Validation -Naive Bayes + C. Validation -J48 + C. Validation -Random Forest + Percentage Split -Naive Bayes + Percentage Split -J48 + Percentage Split -kNN -Random Forest	98.30 % 89.58 % 96.75% 97.92 % 91.67 % 95.56 % 97.92 % 97.92 %
Kumari *et al.* (2021).	Número de embarazos, concentración de glucosa en plasma, presión arterial diastólica, grosor de la piel en tríceps, insulina en plasma, índice de masa corporal, función pedigree de Diabetes, edad (Pima Indians Diabetes *Database*).	India	768	-Logistic Regression -Random Forest -Decision Tree -Naive Bayes -SVM -kNN -Soft Voting Classifier -AdaBoost -Bagging Class -GradientBoost	95.74 % 95.21 % 93.61 % 92.02 % 60.10 % 93.08 % 97.27 % 94.10 % 93.08 % 94.68 %
Das *et al.* (2020).	Edad, sexo, poliuria, pérdida de peso repentino, debilidad, polifagia, candidiasis genital, visión borrosa, hormigueo o picazón, irritabilidad,	Bangl.	520	- Decision Tree -kNN -Logistic Regression -Naive Bayes -Neural Network (Backpropa-	91.35 % 90.38 % 86.54 % 83.65 %

	cicatrización retardada, parálisis parcial, rigidez muscular, pérdida de cabello, obesidad (Sylhet Diabetes Hospital of Sylhet in Bangladesh).			gation) -Random Forest -SVM with Radial Basis Kernel	 89.42 % 94.23 % 98.08 %
Muhammad *et al.* (2020).	Edad, historial familiar, glucosa, colesterol, presión arterial, lipoproteína de alta densidad, triglicéridos, índice de masa corporal (Murtala Mohammed Hospital).	Nigeria	383	-Logistic Regression - SVM - kNN -Random Forest -Naive Bayesi -Gradient boosting	 80.88 % 85.29 % 82.35 % 88.76 % 77.94 % 86.76 %
Reddy *et al.* (2020).	Glucosa, presión arterial, grosor de la piel, insulina, índice de masa corporal, función pedigrí de Diabetes, edad (Pima Indians Diabetes *Database*)	India	2,000	-Logistic Regression - SVM - kNN -Random Forest -Naive Bayes -Gradient boosting	 80.64 % 79.15 % 87.61 % 98.48 % 77.34 % 87.31 %
Tigga, Garg (2020).	Edad, género, historial familiar de Diabetes, hipertensión, actividad física, índice de masa corporal, consumo tabaco, horas sueño, horas sueño profundo, ingesta medicamento, consumo comida chatarra, estrés, presión arterial, número de embarazos, diabetes gestacional, frecuencia de orina (*Dataset* del autor)	India	952	-Logistic Regression - kNN - SVM -Naive Bayes -Decision Tree -Random Forest	 85.70 % 77.30 % 86.50 % 80.60 % 84.00 % 94.10 %
Islam *et al.* (2020).	Edad, índice de masa corporal, glucosa en plasma, insulina en sangre a los 0, 30, 60 y 120 minutos, área bajo la curva de glucosa de 0 a 120 minutos, Índice Matsuda (San Antonio Heart Study SAHS).	Qatar	1,496	-Random Forest -AdaBoots -Bagging -Polynomial SVM -Ensemble of classifiers	90.71 % 88.17 % 90.64 % 66.19 % 81.01 %
Mujumdar, Vaidehi (2019).	Número de embarazos, nivel de glucosa, presión arterial, grosor de la piel, insulina, índice de masa corporal, edad, tipo de empleo (trabajo de oficina, trabajo de campo, trabajo de máquina) (Diabetes *Database*).	India	800	-Decision Tree -Gaussian NB -LDA -SVC - Random Forest -Extra Tree -AdaBoost -Perceptron -Logistic Regression -Gradient Boost Classifier -Bagging -kNN	86.00 % 93.00 % 94.00 % 60.00 % 91.00 % 91.00 % 93.00 % 76.00 % 96.00 % 93.00 % 90.00 % 90.00 %

Rawat, Suryakant (2019).	Número de embarazos, nivel de concentración de glucosa en plasma en una prueba oral de tolerancia a la glucosa, presión arterial (Diastólica), grosor de la piel en el tríceps, cantidad de producción de insulina, índice de masa corporal, función pedigree de Diabetes, edad (Pima Indians *Dataset*).	India	768	-AdaBoost -Logisticboost -Robustboost -Naive Bayes -Bagging	79.68 % 78.64 % 78.64 % 76.04 % 81.77 %
Abed, Ibrikcl (2019).	Número de embarazos, glucosa en plasma, presión Diastólica, grosor de la piel en tríceps, insulina en plasma, índice de masa corporal, función pedigree de Diabetes, edad (Pima Indians Diabetes *Database*).	Turquía	768	-Gradient descent - Gradient descent with Variable Learning Rate -Levenberg-Marquardt Backpropagation -BFGS Quasi-Newton Backpropagation -MLP Bayesain Regularization -CNB -kNN -SVM (RBF Kernel) -LDA	76.30 % 78.60 % 79.90 % 77.10 % 96.00 % 78.80 % 76.60 % 76.60 % 80.50 %
Aada, Tiwari (2019).	Número de embarazos, concentración de glucosa en plasma, presión arterial Diastólica, grosor de la piel en tríceps, insulina en plasma, índice de masa corporal función pedigree de Diabetes, edad (Pima Indians Diabetes *Database*).	India	768	-Decision Tree -SVM -AdaBoost -Linear Regression	74.89 % 94.44 % 91.11 % 93.79 %
Faruque *et al.* (2019).	Edad, sexo, peso, dieta, orina frecuente, consumo de agua, sed excesiva, presión arterial, hipertensión, cansancio, problemas de la vista, problemas de riñón, pérdida de la audición, picazón en la piel, antecedentes genéticos, presencia de Diabetes (Medical Center Bangladesh).	Bangl	200	-SVM -Naive Bayes -kNN -Decision Tree	69.00 % 67.00 % 70.00 % 73.00 %
Kaur *et al.* (2018).	Número de embarazos, concentración de glucosa en plasma, presión arterial Diastólica, grosor de la piel en tríceps, insulina en plasma, índice de masa corporal función pedigree de Diabetes, edad (Pima Indians Diabetes *Database*).	Canadá	768	-N. Bayes + Decision Tree -Decision Tree + Neural Network - Decision Tree / Random Forest - Decision Tree + SVM	90.20 % 94.50 % 76.60 % 78.50 %

				- SVM + Random Forest - R. Forest + Neural Network - R. Forest + Naïve Bayes - Neural Network + N. Bayes - SVM + Neural Network - Naive Bayes + SVM	35.80 % 76.10 % 76.50 % 87.50 % 87.90 % 84.00 %
Zou *et al.* (2018).	Edad, pulso cardíaco, respiración, presión sistólica izquierda, presión sistólica derecha, estatura, peso, índice físico, glucosa en ayuno, cintura lipoproteínas de baja densidad (LDL), lipoproteínas de alta densidad (HDL).	China	289,674	-Decision Tree - Random Forest - Neural Network	78.53 % 80.84 % 78.41 %
Wu *et al.* (2018).	Número de embarazos, concentración de glucosa en plasma en 2 horas en una prueba oral de tolerancia a la glucosa, presión arterial diastólica, grosor de la piel en tríceps, insulina sérica de 2 horas, índice de masa corporal, función pedigree de Diabetes, edad (Pima Indians *Dataset*).	USA	768	-Discrim -MLP -Logdisc -SMART -BayesnNet - Naive Bayes - Random Forest -J48 -SGD -SMO -Backprop -RBF -LMT	77.50 % 73.80 % 78.20 % 76.80 % 71.70 % 74.90 % 76.00 % 76.70 % 76.00 % 77.00 % 75.20 % 75.70 % 76.60 %
Alehegn *et al.* (2018).	Glucosa, presión arterial, grosor de la piel, insulina, índice de masa corporal, función pedigrí de Diabetes, edad (Pima Indians Diabetes *Database*).	India	768	- SVM -Naive Bayes - Decision Stumb -AdaBoost M1	88.80 % 88.54 % 83.72 % 85.68 %

PROYECTO DE INVESTIGACIÓN INSTITUCIONAL IMSS – CIMA 2021

DATOS GENERALES

1) Género
   a) Femenino
   b) Masculino

2) Edad

3) Ciudad de origen

4) Escolaridad
   a) Ninguna
   b) Educación básica
   c) Educación Medio Superior
   d) Educación Superior
   e) Posgrado

DATOS CLÍNICOS

1) Estatura

2) Peso

3) Cintura

4) Índice de Masa Corporal

5) Nivel de glucosa

6) ¿Padece hipertensión arterial?
   a) Sí
   b) No
   c) No sabe

7) Nivel de presión arterial

8) ¿Algún médico le ha dicho que padece Diabetes Tipo II?
   a) Si
   b) No

9) ¿Alguno de sus familiares padece algún tipo de diabetes?
   a) Ninguno
   b) Padre
   c) Madre
   d) Hijo (a)(s)
   e) Hermano (a)(s)
   f) Tío (a)(s)
   g) Abuelo (s) paterno(s)
   h) Abuelo (s) materno(s)
   i) Bisabuelo (s) paterno (s)
   j) Bisabuelo (s) materno (s)

10) ¿Le han tomado su Hemoglobina Glucosilada?
   a) Si
   b) No

11) Hemoglobina Glucosilada

12) ¿Consume más de 3 refrescos al día?
   a) Si
   b) No

13) Indique el # de refrescos que consume al día

14) ¿Le han aparecido alterados sus triglicéridos durante sus últimos análisis de laboratorio?
   a) Si
   b) No

15) Indique el nivel de sus triglicéridos durante sus últimos análisis de laboratorio

16) ¿Le ha aparecido alterado su Colesterol LDL durante sus últimos análisis de laboratorio?
   a) Si
   b) No

17) Indique el nivel de su Colesterol LDL durante sus últimos análisis de laboratorio

18) ¿Recientemente ha presentado episodios de hambre excesiva (Polifagia)?
   a) Sí
   b) No

19) ¿Recientemente ha presentado episodios de sed excesiva (Polidipsia)?
   a) Sí
   b) No

20) ¿Recientemente ha presentado episodios de orina frecuente (Poliuria)?
   a) Sí
   b) No

21) Indique el # de veces que acude al baño a orinar

22) ¿Ha experimentado recientemente pérdida de peso sin detectar la razón o causa aparente?
   a) Si
   b) No

23) En caso de pérdida de peso, indique los kilogramos perdidos recientemente

24) En caso de encuestados del sexo femenino: ¿Durante su(s) embarazo(s), alguna vez le fue diagnosticada Diabetes Gestacional?
   a) Si
   b) No

25) En caso de encuestados del sexo femenino: ¿Durante su(s) embarazo(s), alguno de sus hijos pesó más de 4 kg al nacer?
   a) Si
   b) No

DATOS SOCIOECONÓMICOS

(Llenado **OPCIONAL**, en caso de que el paciente no desee responder la pregunta, favor de seleccionar la opción **"No respondió")**.

1) Seleccione su principal Fuente de Ingresos
   a) Ingreso del trabajo
   b) Renta de propiedad
   c) Transferencias
   d) Estimación del alquiler de la vivienda
   e) Otros ingresos corrientes
   f) No respondió

2) En su familia, ¿a cuánto asciende su nivel de ingreso mensual?
   a) Menos de $ 2,500.00
   b) De $ 2,500.00 a $ 14,999.00
   c) De $ 15,000.00 a $ 99,000.00
   d) Mas de $ 100,000.00
   e) No respondió

NOMBRE DEL ENCUESTADOR