Aprendizaje automático aplicado a la predicción de diabetes mellitus, utilizando información socioeconómica y ambiental de usuarios del sistema de salud*

Machine learning applied to the prediction of diabetes mellitus, using socioeconomic and environmental information from health system users

Aprendizagem automática aplicada à predição de diabetes mellitus, utilizando informação socioeconômica e ambiental de usuários do sistema de saúde

Jessner Alexander Mejía[1]; Mario Andrés Oviedo-Benalcázar[2]; José Armando Ordoñez[3]; José Fernando Valencia[4]

1 Maestría en Ingeniería, Ingeniero, Lumen Innovations, Colombia. jessner@lumeninnovations.org, ORCID: https://orcid.org/0000-0002-0729-4924

2 Maestría en Ingeniería, Ingeniero, ASMET Salud EPS SAS, Colombia. marioandres7@hotmail.com, ORCID: https://orcid.org/0000-0002-4350-5346

3 Doctorado, Universidad ICESI, Colombia. jaordonez@icesi.edu.co, ORCID: https://orcid.org/0000-0001-6544-0283

4 Doctorado, Universidad de San Buenaventura, Colombia. jfvalenc@usbcali.edu.co, ORCID: https://orcid.org/0000-0003-2997-2121

Recibido: 07/10/2022. Aprobado: 15/02/2023. Publicado: 27/03/2023

Mejía JA, Oviedo-Benálcazar MA, Ordoñez JA, Valencia JF. Aprendizaje automático aplicado a la predicción de diabetes mellitus, utilizando información socioeconómica y ambiental de usuarios del sistema de salud. Rev. Fac. Nac. Salud Pública. 2023;41(2):e351168. DOI: https://doi.org/10.17533/udea.rfnsp.e351168

Resumen

Objetivo: Se propuso aplicar modelos basados en técnicas de aprendizaje automático como apoyo para el diagnóstico temprano de la diabetes mellitus, utilizando variables de datos ambientales, sociales, económicos y sanitarios, sin la dependencia de la toma de muestras clínicas. **Metodología:** Se utilizaron datos de 10 889 usuarios afiliados al régimen subsidiado de salud de la zona suroccidental en Colombia, diagnosticados con hipertensión y agrupados en usuarios sin (74,3 %) y con (25,7 %) diabetes mellitus. Se entrenaron modelos supervisados utilizando *k* vecinos más cercanos, árboles de decisión y bosques aleatorios, así como modelos basados en ensambles, aplicados a la base de datos antes y después de balancear el número de casos en cada grupo de diagnóstico. Se evalúo el rendimiento de los algoritmos mediante la división de la base de datos en datos de entreno y de prueba (70/30, respectivamente), y se utilizaron métricas de exactitud, sensibilidad, especificidad y área bajo la curva. **Resultados:** Los valores de sensibilidad aumentaron considerablemente al utilizar datos balanceados, pasando de

* Este texto es producto de la investigación: "Predicción de diabetes mellitus tipo 2 a partir de variables ambientales, sociales, económicas y sanitarias de la población del régimen subsidiado en Colombia", Tesis de la Maestría en Tecnologías de la Información para la Analítica de Datos, Universidad de San Buenaventura Cali (Valle del Cauca). Fecha de inicio: 13 de octubre de 2021; fecha de terminación: 7 de septiembre de 2022.

valores máximos del 17,1 % (datos sin balancear) a valores de hasta 57,4 % (datos balanceados). El valor más alto de área bajo la curva (0,61) fue obtenido con los modelos de ensambles, al aplicar un balance en el número de datos por cada grupo y al codificar las variables categóricas. Las variables de mayor peso estuvieron asociadas con aspectos hereditarios (24,65 %) y con el grupo étnico (5.59 %), además de la dificultad visual, el bajo consumo de agua, una dieta baja en frutas y verduras, y el consumo de sal y azúcar. **Conclusiones:** Aunque los modelos predictivos, utilizando información socioeconómica y ambiental de las personas, surgen como una herramienta para el diagnóstico temprano de la diabetes mellitus, estos aún deben ser mejorados en su capacidad predictiva.
----------*Palabras clave:* aprendizaje automático, diabetes mellitus, factores ambientales, factores socioeconómicos, modelo predictivo.

Abstract

Objective: The objective was to apply models based on machine learning techniques to support the early diagnosis of diabetes mellitus, using environmental, social, economic and health data variables, without dependence on clinical sample collection. **Methodology:** Data from 10,889 users affiliated with the subsidized health system in the southwestern area of Colombia, diagnosed with hypertension and grouped into users without (74.3%) and with (25.7%) diabetes mellitus, were used. Supervised models were trained using k-nearest neighbors, decision trees, and random forests, as well as ensemble-based models, applied to the database before and after balancing the number of cases in each diagnostic group. The performance of the algorithms was evaluated by dividing the database into training and test data (70/30, respectively), and metrics of accuracy, sensitivity, specificity, and area under the curve were used. **Results:** Sensitivity values increased significantly when using balanced data, going from maximum values of 17.1% (unbalanced data) to values as high as 57.4% (balanced data). The highest value of area under the curve (0.61) was obtained with the ensemble models, by applying a balance in the amount of data for each group and by coding the categorical variables. The variables with the greatest weight were associated with hereditary aspects (24.65%) and with the ethnic group (5.59%), in addition to visual difficulty, low water consumption, a diet low in fruits and vegetables, and the consumption of salt and sugar. **Conclusions:** Although predictive models, using people's socioeconomic and environmental information, emerge as a tool for the early diagnosis of diabetes mellitus, their predictive capacity still needs to be improved.
---------*Keywords:* machine learning, diabetes mellitus, environmental factors, socioeconomic factors, predictive model

Resumo

Objetivo: Propôs-se aplicar modelos baseados em técnicas de aprendizagem automática como apoio para o diagnóstico precoce da diabetes mellitus, utilizando variáveis de dados ambientais, sociais, econômicos e sanitários, sem a dependência da coleta de amostras clínicas. **Metodologia:** Usaram-se dados de 10.889 usuários filiados ao regime subsidiado de saúde da zona sudoeste da Colômbia, diagnosticados com hipertensão e agrupados em usuários sem (74,3%) e com (25,7%) diabetes mellitus. Foram treinados modelos supervisionados utilizando k vizinhos mais próximos, árvores de decisão e florestas aleatórias, assim como modelos baseados em montagens, aplicados à base de dados antes de depois de equilibrar o número de casos em cada grupo de diagnóstico. Avaliou-se o desempenho dos algoritmos por meio da divisão da base de dados de treino e teste (70/30, respectivamente), e utilizaram-se métricas de exatidão, sensibilidade, especificidade e área sob a curva. **Resultados:** Os valores de sensibilidade aumentaram de maneira significativa ao utilizar dados equilibrados, passando de valores máximos de 17,1% (dados sem equilibrar) a valores de até 57,4% (dados equilibrados). O valor mais elevado de área sob a curva (0,61) foi obtido com os modelos de montagens, ao aplicar um balanço no número de dados por cada grupo e codificar as variáveis categóricas. As variáveis de maior peso estiveram associadas com fatores hereditários (24,65%) e com o grupo étnico (5,59%), além da dificuldade visual, o baixo consumo de água, um regime baixo em frutas e vegetais e o consumo de sal e açúcar. **Conclusões:** Embora os modelos preditivos, utilizando informação socioeconômica e ambiental das pessoas, surgem como uma ferramenta para o diagnóstico precoce da diabetes mellitus, ainda devem ser melhorados em sua capacidade preditiva.
---------*Palavras-chave:* aprendizagem automática, diabetes mellitus, fatores ambientais, fatores socioeconômicos, modelo preditivo

Introducción

La diabetes mellitus (DM) es una de las 10 enfermedades más graves en el mundo y se caracteriza por complicaciones progresivas que incluyen enfermedades cardiovasculares, retinopatía, enfermedades cerebrovasculares, amputación de miembros del cuerpo y en ocasiones la muerte [1]. De acuerdo con Bernardini [2], entre el 6 y el 8 % de la población mundial está afectada por la DM, con alrededor de 400 millones de personas diagnosticadas que reciben tratamiento. Este mismo estudio estima que los costos en la atención de la DM para el año 2030 ascenderán a los 490 miles de millones de dólares, lo que equivale al 12 % de los gastos médicos de todas las enfermedades [2]. Para América, se prevé que existan 109 millones de personas con diagnóstico de DM para el 2040, con una mayor concentración en países con ingresos bajos y medios [3].

En Colombia, de acuerdo con el estudio realizado por la Cuenta de Alto Costo [4], frente a la situación de la enfermedad renal crónica, la hipertensión arterial (HTA) y la DM en el país para 2020, se observa que, en los últimos 6 años, la prevalencia de la HTA y la DM se ha incrementado, dando como resultado que para ese año se presentaron 4 527 098 de casos prevalentes de HTA y 1 426 574 de casos de DM. Ese mismo estudio indica que para el periodo entre julio de 2019 y junio de 2020 fallecieron 31 316 personas con diagnóstico de DM, lo cual significa una tasa de mortalidad general de 62,78 casos por cada 100 000 habitantes [4]. Dado que, según la Organización Mundial de la Salud [5], alrededor del 45 % de la población con DM no sabe que la padece, una tarea crucial para los sistemas de salud es poder tener un diagnóstico oportuno de los pacientes que la sufren o están en mayor riesgo de desarrollarla, con el objetivo de poder promover tratamientos que permitan tener un manejo terapéutico del paciente más eficiente.

El uso de modelos basados en técnicas de aprendizaje automático (*Machine Learning*, ML) ha demostrado excelentes resultados en diferentes áreas y especialidades de la medicina, apoyando el diagnóstico de enfermedades de manera efectiva y de forma temprana, a fin de iniciar los tratamientos oportunamente [6-9]. En particular, se han llevado a cabo diversos esfuerzos para estudiar y proponer avances en la prevención de la DM [10,11], incluyendo la creación de sistemas de recomendación para la promoción de estilos de vida saludable [12,13], la construcción de redes de áreas corporal para el monitoreo de la glucosa en sangre [14] y la creación de modelos que permitan predecir la DM [1,15-20]. En estos estudios, los modelos basados en ML utilizaron para su entrenamiento entradas derivadas de parámetros tomados tanto de variables fisiológicas como de variables sociodemográficas, ambientales y de estilo de vida, registradas en poblaciones de India [1,19], México [15], China [16] y Reino Unido [18].

Aunque los mejores predictores de DM están asociados a variables fisiológicas extraídas de muestras clínicas como son las tomas de sangre para análisis de glucosa, estás pruebas son costosas y su estudio puede tomar un tiempo considerable, por el manejo de muestras y la logística que se requiere en el traslado a los laboratorios de análisis. Acceder a estos datos puede llegar a ser muy complejo en países en desarrollo como Colombia, especialmente en las zonas rurales. Para estos casos, es mucho más factible disponer de información de los pacientes a través de variables sociodemográficas, ambientales, económicas y de estilo de vida. Por lo tanto, desde el análisis de datos ambientales, sociales, económicos y sanitarios, sin la dependencia de la toma de muestras clínicas, el presente estudio propone el desarrollo de modelos basados en técnicas de ML para apoyar el diagnóstico temprano de la DM o la predicción de la misma, a fin de permitir a los profesionales de la salud establecer estrategias de prevención o tratamiento oportunos de la DM.

Metodología

En esta sección se describe la base de datos, las técnicas y los procedimientos utilizados para el entrenamiento, validación y prueba de los modelos que se proponen en el presente estudio para el diagnóstico temprano de la DM.

Base de datos

Los datos utilizados en el presente estudio fueron tomados de la empresa ASMET Salud, una entidad promotora de salud (EPS) del régimen subsidiado de salud en Colombia con una amplia cobertura en zonas rurales y de difícil acceso, principalmente el suroccidente y nororiente del país.

Dentro de sus procesos de identificación de riesgos en salud, ASMET Salud realiza una encuesta, de donde se obtienen variables asociadas al modo de vida, los hábitos alimenticios, los antecedentes familiares, las condiciones de vida, el nivel económico y las condiciones ambientales para sus afiliados, los cuales, a su vez, están relacionados con el grupo de riesgo al que pertenece cada afiliado (cáncer, virus de la inmunodeficiencia humana, HTA, diabetes, enfermedad renal crónica, hemofilia, etc.).

La base de datos cuenta con 128 501 registros de usuarios. De estos, inicialmente se seleccionaron 11 423, correspondientes a los usuarios que previamente han sido identificados en el grupo de riesgo de HTA. Entre estos, 2883 son usuarios que también están diagnosticados con DM.

Posteriormente, y considerando que estas dos patologías (HTA y DM) se presentan principalmente en el gru-

po de personas mayores de 40 años [21], se excluyeron del estudio aquellas menores de 40 años, quedando con un total de 10 889 registros con HTA, de los cuales 2808 tiene un padecimiento de DM (25,7 %).

De cada uno de los registros se eliminaron los datos que pudieran individualizar al afiliado, como documento de identificación, nombres, apellidos y aquella información que no aporta para este estudio, como foto de la fachada de la vivienda, datos del gestor que realiza la encuesta, teléfonos, correo electrónico, etc.

Se excluyeron 6 registros en los que todas sus variables estaban vacías, quedando 10 883 registros, con un 80 % de pertenencia a los estratos económicos 1 y 2. Cada registro cuenta con 69 variables, de las cuales 7 son continuas y 62 son categóricas.

Se realizó una primera selección de variables teniendo en cuenta el criterio médico según el aporte de la variable al objetivo del estudio. De esta manera, solo 18 variables fueron seleccionadas por cada registro: 1 variable continua (edad), 3 variables categóricas nominales (tipo de sangre, Rh y grupo étnico) y 14 variables categóricas ordinales (estrato, nivel educativo, discapacidad, actividad física, dieta frutas y verduras, consumo agua, sal en comidas, azúcar en comidas, dificultad visual, lesiones en piel, debilidad en cuerpo, pérdida de fuerza, familiar con diabetes y diabetes). Para el presente estudio, la variable objetivo a predecir es diabetes, categorizada en dos grupos: pacientes con DM y sin DM. La categorización está basada en el diagnóstico médico de diabetes registrado en la base de datos de ASMET Salud.

Por cada variable se identificaron los valores nulos, encontrando que las variables "estrato" y "dificultad visual" tuvieron el mayor número de registros con ausencia de dato. Los valores nulos fueron reemplazados por la moda, dado que todas las variables con valores nulos eran categóricas. De acuerdo con las categorías registradas en la base de datos, las variables ordinales fueron codificadas por escalas. Por ejemplo, para las variables "dieta frutas y verduras" y "consumo agua" se utilizó la siguiente escala: 0-Nunca; 0,5-Menos de una vez por semana; 1-Una vez por semana; 2-Entre dos y tres veces por semana; 4-Entre cuatro y seis veces por semana; 5-Todos los días.

Para la variable "familiar con diabetes" se utilizó: 0-Ninguno; 2-Sí: Otros parientes; 3-Sí: Abuelos, tíos, primos hermanos; 5-Sí: Padres, hermanos o hijos.

Las variables dicotómicas (discapacidad, sal en comidas, azúcar en comidas, dificultad visual, lesiones en piel, debilidad en cuerpo, pérdida de fuerza, diabetes) se codificaron con "0" y "1", donde "1" indica la presencia de enfermedad o de exposición a algún factor, y "0", su ausencia.

Por otra parte, para la codificación de las variables nominales se usó el método de codificación efectiva de un bit (*one-hot Encoding*), que consiste en crear un vector de *N* columnas para codificar las *N* clases de la variable nominal, y, para cada registro, marcar con un 1 la columna a la que pertenezca dicho registro y dejar las demás con 0.

Modelos predictivos

Los modelos utilizados se seleccionan teniendo en cuenta lo siguiente: 1) se trata de un estudio de diagnóstico retrospectivo aplicado a un problema de clasificación de dos clases, una clase con personas que tienen HTA, pero no presentan DM (No-DM), y otra clase que incluye personas con HTA y también manifiestan DM (Sí-DM); 2) se pueden aplicar métodos basados en aprendizaje supervisado, considerando que la base de datos contiene la variable objetivo a predecir (diabetes); 3) las clases a clasificar pertenecen a un conjunto de datos no balanceados, es decir, el porcentaje de usuarios No-DM es mucho mayor que el porcentaje de usuarios Sí-DM (relación de 3 a 1); y 4) son los modelos más comunes empleados o evaluados en las soluciones encontradas en el estado del arte [11-16], tanto individualmente como en algoritmos compuestos.

Para el preprocesamiento de la base de datos, entreno, validación y evaluación de los modelos propuestos se utilizó Python® 3.8, con las librerías Scikit-Learn, Pandas, NumPy y Plotly, y como servicios de centralización de datos se usó Google® BigQuery.

A continuación se describe cada uno de los modelos empleados.

K vecinos más cercanos

Es un algoritmo que asigna a una observación la clase más común entre las clases más cercanas (*K* vecinos más cercanos, KNN) a dicha observación.

Para evaluar la cercanía, se determina la distancia entre el punto de prueba y cada una de las observaciones de entrenamiento, utilizando métricas como la distancia euclidiana, de Manhattan, de Minkowski y de Chebyshev.

Una vez obtenidas las distancias, se toman los *K* vecinos más cercanos y se identifica la categoría a la que pertenece cada vecino. La categoría con más vecinos cercanos será la que se le asigne a la observación que se está clasificando.

Normalmente, se toma un valor de *K* impar, para facilitar el desempate entre cuál es la clase más cercana al punto de prueba, y un valor de *K* pequeño para reducir el tiempo de cómputo, esto debido al número de comparaciones que debe de efectuar el algoritmo.

Este algoritmo no genera un modelo explícitamente y, por el contrario, debe comparar cada instancia u observación de prueba con todas las observaciones de entrenamiento [22].

Árbol de decisión

Son algoritmos versátiles de aprendizaje automático que pueden llevar a cabo tanto tareas de clasificación como

de regresión, e incluso tareas de salida múltiple. Este algoritmo es reconocido por ser de fácil lectura al ojo humano, ya que define sus caminos respondiendo preguntas, lo que permite decidir o crear vías para llegar a una decisión final.

Los árboles de decisión (*Decision Tree*, DT) se ajustan realizando numerosas iteraciones y mediante la creación umbrales de decisión, a partir de los valores que toman las características de un conjunto de datos [22].

Las preguntas producen un esquema de entrada, pregunta y salida, que técnicamente se ven como nodos y ramas.

Existen diferentes algoritmos propuestos en la literatura [23]. Entre los más populares se destacan dos: C4.5 y Classification and Regression Trees (CART). En el caso puntual del presente trabajo se usó este último, provisto por la librería Scikit-Learn [22].

Para el proceso de ajuste, se utilizan diversos hiperparámetros, como la profundidad del árbol (cantidad máxima de niveles de nodos internos que deben crearse) y la impureza de un nodo, la cual se determina por medio de la entropía o el índice de Gini.

Para el caso del índice de Gini, que es una medida utilizada por defecto para el proceso de entrenamiento en la librería Scikit-Learn, se dice que un nodo es puro si Gini = 0, mientras que entre más cercano a 1 sea el valor de Gini, más probable es que el algoritmo equivoque su predicción (nodo impuro).

En los DT y demás algoritmos derivados de ellos es posible determinar el nivel de importancia de las variables de entrada al modelo, que se determina calculando la media y la desviación estándar de la acumulación de la disminución de impurezas dentro de cada árbol.

Bosque aleatorio

Es un algoritmo robusto a manera de ensamble, que se compone de diversas versiones del mismo algoritmo DT, y donde cada versión es entrenada con subconjuntos aleatorios de la misma base de datos (*bagging*), la cual se divide en secciones que se distribuyen entre los DT que conforman al bosque aleatorio (Random Forest, RF).

La asignación de las muestras a cada DT se hace de manera aleatoria tanto en observaciones como en características (*bootstrapping*), de forma que cada DT se entrena con datos ligeramente diferentes [22].

Los RF permiten controlar los hiperparámetros propios de un DT, como el número de hojas, y agrega otros hiperparámetros, como el número de árboles o la cantidad de unidades centrales de procesamiento utilizadas para el entrenamiento.

Optimización de hiperparámetros

Para la optimización de hiperparámetros, se utilizó la técnica de búsqueda por grilla (*Grid Search*), un método controlado que permite iterar sobre un número finito de valores previamente definidos [22].

Esta optimización se basa en seleccionar un conjunto de *N* valores para cada parámetro de un total de *M* parámetros, evaluando cada posible combinación de hiperparámetros.

Para el presente estudio se estudiaron las siguientes combinaciones: 1) para KNN, n_neighbors {3,5,7,9,11,13,15} y weights {'uniform','distance'}; 2) para DT, max_depth {5,10,15,20,30} y criterion {'gini','entropy'}; 3) para RF, n_estimators {20,30,60,100,150,180,200,300}, criterion {'gini','entropy'} y max_depth {5,10}.

Ensamble de algoritmos

Consiste en la creación de un algoritmo a partir de la unión de múltiples modelos, con el objetivo de mejorar la generalización de las predicciones. Con los ensambles se busca minimizar el sesgo de predicción (promedio de la diferencia entre el valor predicho y el valor real) y minimizar la varianza (capacidad de responder a nuevos datos) del modelo [22].

Entre las técnicas para crear los ensambles se tiene:

Ensamble por votación: donde se unen diversos algoritmos y se entrenan con subconjunto de datos. Una vez entrenado, se predice un nuevo valor de cada uno de los algoritmos y se selecciona la moda.

Método de embolsado o bagging: se entrena el mismo algoritmo con diferentes subconjuntos de datos producidos a partir del conjunto de entrenamiento. La predicción final será la moda de las predicciones obtenidas para cada uno de los subconjuntos.

Método de boosting: es la unión en secuencia de modelos simples que transfieren su tasa de aprendizaje el uno al otro; esto significa que dado un modelo M, el algoritmo siguiente M1 aprenderá de los errores de entrenamiento de M.

El aumento de gradiente (*Gradient Boosting*) se refiere al algoritmo de optimización de descenso de gradiente que se utiliza para ajustar la función de pérdida cuando se entrena un modelo. El *eXtreme Gradient Boosting* (XGB) es una implementación eficiente de código abierto del algoritmo de aumento de gradiente (biblioteca de Python), diseñado para ser computacionalmente eficiente.

Evaluación de algoritmos

Debido a que se trata de una base de datos no balanceada, con relación 3 a 1 en los grupos de usuarios diagnosticados sin diabetes (No-DM) con respecto a los diagnosticados con diabetes (Sí-DM), se realizó un balanceo en el número de registros por cada clase. Para ello, se aplicó una técnica de submuestreo, con la cual se eliminaron de forma aleatoria registros en el grupo con más datos

(No-DM), hasta alcanzar un tamaño similar al grupo con menos datos (Sí-DM).

Para la evaluación del rendimiento de los algoritmos, la base de datos se dividió en dos subconjuntos: *datos de entreno* (*Train*) y *datos de prueba* (*Test*), en proporción 70/30 respectivamente.

Las *métricas de rendimiento* para cada modelo se determinaron a partir de la matriz de confusión, identificando los verdaderos negativos (TN), los falsos negativos (FN), los verdaderos positivos (TP) y los falsos positivos (FP):

1. *Exactitud (Acc):* permite conocer la proporción de elementos clasificados correctamente. Para esto, se considera la relación entre las predicciones correctas (TP + TN) con respecto al total de las observaciones realizadas (TP + TN + FP + FN).
2. *Sensibilidad (Sen):* brinda información sobre la cantidad de casos positivos que el modelo puede predecir correctamente. Esta métrica es de gran importancia en el presente trabajo, por cuanto el modelo predictivo se utilizará por la empresa ASMET Salud para determinar qué pacientes tiene mayor probabilidad de desarrollar DM y promover así la captación temprana de estos pacientes.
3. *Especificidad (Esp):* entrega información sobre la cantidad de casos negativos que el modelo puede predecir de forma correcta.
4. *Área bajo la curva (AUC-ROC):* permite conocer la capacidad que tiene un modelo predictivo para clasificar correctamente las instancias que se le presentan. Puede verse también como una medida de separación entre las clases de un modelo. Se calcula a partir de la curva característica operativa del receptor (*receiver operating characteristic*, ROC), la cual refleja la relación entre la tasa de TP y la tasa de FP para diferentes niveles del umbral.
5. *F1 Score:* combina la precisión y la sensibilidad en un solo valor, lo que permite poder observar el comportamiento de estos dos valores en una sola medida.

A continuación se presentan las ecuaciones utilizadas para calcular las métricas Acc, Sen, Esp y F1 Score:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad \text{Ecuación 1}$$

$$Sen = \frac{TP}{TP + FN} \quad \text{Ecuación 2}$$

$$Esp = \frac{TN}{TN + FP} \quad \text{Ecuación 3}$$

$$F1\,Score = 2 * \frac{TP/(TP + FP) * Sen}{TP/(TP + FP) + Sen} \quad \text{Ecuación 4}$$

Resultados

A continuación se presentan los resultados obtenidos durante el entrenamiento, la validación y la prueba de los modelos que se proponen en el presente estudio para el diagnóstico temprano de la DM, aplicados a la base de datos de ASMET Salud.

Distribución de los afiliados por rangos de edad

En la Figura 1 se muestra la pirámide poblacional de los afiliados, de la base de datos de ASMET Salud, con HTA. Se observa que la mayoría (el 95 %) de los afiliados que padecen HTA tiene una edad mayor o igual que 40 años. Asimismo, se encontró que el 93 % de los pacientes que padecen DM tienen una edad mayor o igual que 40 años. Teniendo en cuenta esta situación, se decide considerar, en el presente estudio, solamente los registros de los afiliados con edades iguales o mayores que 40 años.

Con el propósito de conocer si la pirámide poblacional obtenida con la base de datos de ASMET Salud sigue la misma distribución reportada para la población colombiana, en la Tabla 1 se comparan la distribución de los afiliados por edad y por patología (HTA y DM) en el presente estudio, con respecto a los presentados por la Cuenta de Alto Costo [4].

Aunque la base de datos de ASMET Salud contiene usuarios principalmente de las zonas rurales del suroccidente y nororiente del país, de la Tabla 1 se observa que esta población sigue una distribución similar a la reportada en toda Colombia por la Cuenta de Alto Costo [4]. La similitud es más evidente cuando se compara el grupo de afiliados diagnosticados con HTA, para el cual las diferencias entre ambas poblaciones son de menos del 0,5 % para los rangos de edades que van de los 50 a los 74 años (62,58 vs. 61,24 %) y de los 60 a los 64 años (14,16 vs. 14,26 %). Para el grupo de los afiliados diagnosticados con DM, las diferencias también son menores al 0,5 % para el rango de edades que va de los 60 a los 64 años (15,27 vs. 15,41 %), pero para edades entre los 50 y los 74 años la diferencia es de aproximadamente del 10 % (65,73 vs. 55,80 %).

Clasificación de los grupos de riesgo

Los resultados, obtenidos en la clasificación de los sujetos según el grupo de riesgo, se presentan agrupados para los casos en los que se utilizan datos sin y con codificación, y aplicando o no técnicas de balanceo en el número de registros por cada clase.

Clasificación utilizando datos sin codificación

La Tabla 2 contiene los resultados del desempeño de los modelos de clasificación, aplicando técnicas de ML, para clasificar los pacientes HTA que presentan o no DM. Estos resultados corresponden a los obtenidos cuando se utiliza el conjunto de datos en estado puro, es decir, sin

realizar ninguna transformación a las variables categóricas y continuas. Además, la primera parte de la tabla muestra los resultados obtenidos sin efectuar un balance en el número de registros por cada clase (afiliados HTA con y sin DM). En la segunda parte, se encuentran los resultados después de haber llevado a cabo un balanceo en el número de registros por cada clase, aplicando la técnica de submuestreo.

De la Tabla 2, para los datos no balanceados, se observa que: 1) RF obtuvo el valor más alto de Acc en el grupo de entreno, pero, a su vez, uno de los valores más bajos en el grupo de Test, sugiriendo un posible sobreentrenamiento del modelo y un mal ajuste para la generalización; 2) los valores de Sen fueron muy bajos (entre el 1,43 y el 15,3 %), indicando que la capacidad para detectar afiliados HTA con DM es muy baja; 3) los valores de Esp fueron muy altos (entre el 88,37 y el 99,44 %), señalando que los modelos aprenden a detectar en su mayoría las muestras sanas (afiliados HTA sin DM); 4) los valores de Sen y F1 Score fueron bajos, siendo los peores resultados los obtenidos con el modelo DT; y 5) el valor del parámetro AUC-ROC está entre 0,5043 y 0,5207, lo que significa que los modelos no son capaces de discriminar correctamente las instancias en cada clase posible.

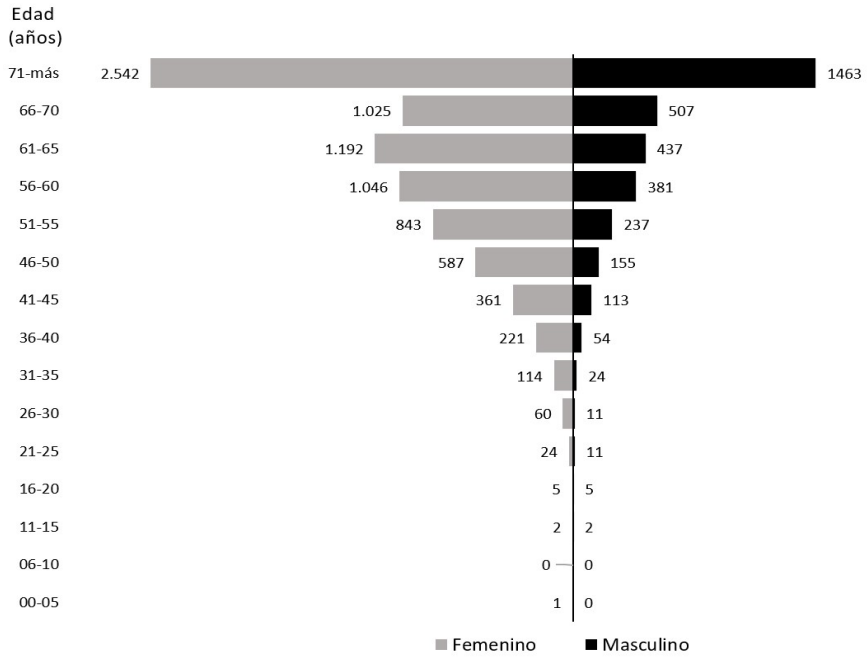

Figura 1. Pirámide poblacional: afiliados diagnosticados con hipertensión (HTA).

Tabla 1. Comparativo: base de datos de ASMET Salud vs. Cuenta de Alto Costo [4]

Enfermedad	Rango de edad (años)	Cuenta de Alto Costo [4] (%)	Base de datos (ASMET Salud) (%)
Diabetes mellitus (DM)	Entre 50-74	65,73	55,80
	Entre 60-64	15,27	15,41
	Menores 35	3,46	3,32
Hipertensión (HTA)	Entre 50-74	62,58	61,24
	Entre 60-64	14,16	14,26
	Menores 35	3,89	2,66

Tabla 2. Desempeño de los modelos de clasificación utilizando datos sin codificar

Modelo	Train	Test				
	Acc(%)	Acc(%)	Sen(%)	Esp(%)	F1 Score	AUC-ROC
Datos no balanceados						
KNN-5	78,72	69,96	15,1	88,99	0,2063	0,5207
DT	74,29	74,18	1,43	99,44	0,0277	0,5043
RF	98,56	69,55	15,3	88,37	0,2060	0,5185
Datos balanceados						
KNN-5	70,27	51,47	51,56	51,38	0,5223	0,5147
DT	60,63	56,28	47,06	66,06	0,5256	0,5656
RF	98,86	54,05	54,15	53,94	0,5482	0,5405

Train: datos de entreno; Test: datos de prueba; KNN-5: 5 vecinos más cercanos; DT: árbol de decisión; RF: bosque aleatorio; Acc: exactitud; Sen: sensibilidad; Esp: especificidad; AUC-ROC: área bajo la curva.

De igual manera, de la Tabla 2, para los datos balanceados, se observa que en comparación con los resultados alcanzados para datos no balanceados: 1) los valores de Acc disminuyeron tanto en el grupo de *Train* como en el grupo de *Test*; 2) el valor de Sen aumentó, llegando a valores entre 47,01 y 54,15 %, es decir, se mejora la capacidad para detectar afiliados HTA con DM; 3) el parámetro Esp disminuyó a valores entre 51,38 y 66,06 %, indicando que se reducen las detecciones de muestras sanas (afiliados HTA sin DM); 4) los parámetros de Sen y F1 Score aumentaron su valor, obteniendo todos los modelos valores similares; y 5) el valor del parámetro AUC-ROC aumentó ligeramente, entre 0,5147 y 0,5656, indicando de nuevo que los modelos no pueden discriminar correctamente las instancias en cada clase posible.

Clasificación utilizando datos codificados

Como paso siguiente, se evalúan otra vez los algoritmos de clasificación, pero en esta ocasión se efectúa una codificación de las variables categóricas. La primera parte de la Tabla 3 presenta los resultados al aplicar la codificación, pero sin realizar un balance en el número de registros por cada clase (afiliados HTA con y sin DM). En la segunda parte, se encuentran los resultados obtenidos después de haber llevado a cabo un balanceo en el número de registros por cada clase, mediante la técnica de submuestreo.

De la Tabla 3, para los datos no balanceados, se observa que, en general, los valores de las métricas Acc, Sen, Esp, F1 Score y AUC-ROC fueron muy similares a los recabados con los datos sin codificar (véase Tabla 2), aunque se origina un ligero aumento de dichas métricas cuando se utilizan datos codificados, especialmente para el modelo DT. Este último alcanzó un incremento importante en el valor de la sensibilidad, pasando de 1,43 a 10,16 %. Sin embargo, el comportamiento a nivel general de los tres algoritmos es deficiente, con un rango de 10,16 a 17,11 % en la sensibilidad. En este sentido, puede inferirse que los modelos: 1) tuvieron una muy baja capacidad para detectar afiliados HTA con DM (Sen baja); y 2) aprenden a detectar, en su mayoría, a los afiliados HTA sin DM (Esp alta).

De igual manera, al analizar los resultados obtenidos con los datos balanceados, se contempla un comportamiento similar a los alcanzados con los datos sin codificar y balanceados (véase Tabla 2), mostrando de nuevo un ligero aumento de dichas métricas cuando se utilizan datos codificados, especialmente para el modelo DT. Es decir: 1) el valor de Sen aumentó, llegando a valores entre 46,19 y 57,44 %, con lo que se mejora la capacidad para detectar afiliados HTA con DM; 2) la métrica Esp disminuyó a valores entre 53,76 y 74,31 %, lo que indica que se reduce la detección de muestras sanas (afiliados HTA sin DM); y 3) el valor del parámetro AUC-ROC se incrementó ligeramente, entre 0,5304 y 0,6025, siendo el modelo DT el que mejor desempeño tuvo.

Para los siguientes análisis se utilizan los registros con los datos codificados y balanceados (aplicando submuestreo).

Ajuste de hiperparámetros

La optimización de hiperparámetros arrojó los siguientes resultados: 1) KNN: KNeighborsClassifier(n_neighbors = 15); 2) Decision Tree: DecisionTreeClassifier(criterion = 'entropy', max_depth = 5); y 3) Random Forest: RandomForestClassifier(max_depth = 10, n_estimators = 100).

En el caso puntual del presente estudio, no se observó una diferencia notable en el rendimiento del algoritmo optimizado por la entropía o con optimización basado en Gini. En la Tabla 4 se muestran los resultados obtenidos en cada uno de los modelos de clasificación con los hiperparámetros optimizados (registros con los datos codificados y balanceados)

Tabla 3. Desempeño de los modelos de clasificación utilizando datos codificados

Modelo	Train	Test			
	Acc (%)	Acc (%)	Sen (%)	Esp (%)	F1 Score	AUC-ROC
Datos no balanceados						
KNN-5	78,26	70,60	17,11	89,17	0,2308	0,5314
DT	78,07	72,67	10,16	94,37	0,1608	0,5226
RF	98,40	70,28	15,69	89,23	0,2139	0,5246
Datos balanceados						
KNN-5	70,07	52,98	51,21	54,86	0,5286	0,5304
DT	61,03	59,84	46,19	74,31	0,5421	0,6025
RF	98,80	55,65	57,44	53,76	0,5714	0,5560

Train: datos de entreno; Test: datos de prueba; KNN-5: 5 vecinos más cercanos; DT: árbol de decisión; RF: bosque aleatorio; Acc: exactitud; Sen: sensibilidad; Esp: especificidad; AUC-ROC: área bajo la curva.

En la Tabla 4 se observa que la optimización de hiperparámetros generó modelos con mejor generalización, lo que se deduce de los valores muy similares obtenidos en Acc durante el entreno y la validación. Los modelos que presentaron el valor más alto en la métrica AUC-ROC fueron el DT-adj (60,45 %) y el RF-adj (60,37 %). Particularmente, el DT-adj presentó una Sen del 45,67 % y una capacidad de detección de pacientes sanos del 75,23 %.

Ensamble de algoritmos

Los resultados de los clasificadores, utilizando modelos basados en ensambles, se muestran en la Tabla 5. En general, los tres métodos de ensamble obtuvieron valores muy similares en las métricas utilizadas para medir el desempeño de los modelos, con valores de AUC-ROC de aproximadamente 0,60, siendo el método *eXtreme Gradient Boosting* (XGB) el que obtuvo el valor ligeramente más alto (0,6092). Este ensamble presentó una sensibilidad de 48,44 % y una especificidad del 73,39 %.

Variables de mayor importancia en los modelos predictivos

La Figura 2 presenta las variables que obtuvieron un nivel de importancia mayor que el 2 % para el clasificador basado en ensamble con el método XGB. El grado de consanguinidad es la variable más relevante como factor decisivo (22,6 %, considerado como valor acumulado entre familiar_diabetes_x), seguido por el grupo étnico mestizo (5,6 %). Otras variables están relacionadas con la dificultad visual, el bajo consumo de agua, una dieta baja en frutas y verduras, y el consumo de sal y azúcar.

Tabla 4. Desempeño de los modelos de clasificación con los hiperparámetros optimizados (registros con los datos codificados y balanceados)

Modelo	Train	Test			
	Acc(%)	Acc(%)	Sen(%)	Esp(%)	F1 Score	AUC-ROC
KNN-adj	53,87	53,87	52,77	55,05	0,5408	0,5391
DT-adj	60,02	60,02	45,67	75,23	0,5404	0,6045
RF-adj	59,93	59,93	45,33	75,41	0,5380	0,6037

KNN-adj (vecinos más cercanos), DT-adj (árbol de decisión) y RF-adj (bosque aleatorio) son los modelos de clasificación ajustados en sus hiperparámetros. Train: datos de entreno; Test: datos de prueba; Acc: exactitud; Sen: sensibilidad; Esp: especificidad; AUC-ROC: área bajo la curva.

Tabla 5. Desempeño de los modelos de clasificación utilizando ensambles

Modelo	Train	Test			
	Acc(%)	Acc(%)	Sen(%)	Esp(%)	F1 Score	AUC-ROC
EVM	63,95	59,84	45,67	74,86	0,5393	0,6027
GBM	60,11	60,11	48,96	71,93	0,5582	0,6044
XGB	60,55	60,55	48,44	73,39	0,5583	0,6092

EVM: ensamble por votación; GBM: aumento de gradiente; XGB: *eXtreme Gradient Boosting*; Train: datos de entreno; Test: datos de prueba; Acc: exactitud; Sen: sensibilidad; Esp: especificidad; AUC-ROC: área bajo la curva.

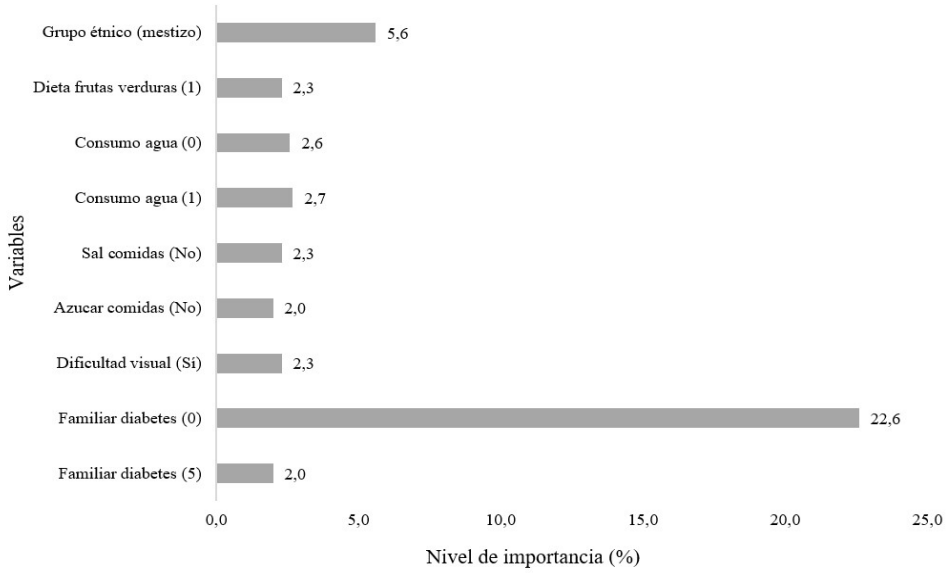

Figura 2. Variables de mayor importancia para el clasificador basado en ensamble con el método *eXtreme Gradient Boosting* (XGB). Dieta frutas verduras (1): Una vez por semana; consumo agua (0): Nunca; consumo agua (1): Una vez por semana; familiar diabetes (0): Ninguno; familiar diabetes (5): Sí - Padres, hermanos o hijos.

Discusión

En el presente trabajo se aplicaron modelos basados en técnicas de ML para predecir los grupos de afiliados que presentan HTA con y sin DM, a partir de los datos registrados en la base de datos de la empresa prestadora de servicios de salud ASMET Salud.

Las variables seleccionadas e incluidas dentro de la base de datos analizada se enfocaron en variables sociodemográficas, antropométricas y de estilo de vida, excluyendo las variables metabólicas, para enfocarse en características que describen el estilo de vida y el contexto del paciente, las cuales son variables que pueden incluso ser tomadas mediante instrumentos de medición a distancia.

De esta población de sujetos se observa que los porcentajes de afiliados distribuidos por rangos de edades (pirámide poblacional en la Figura 1) se corresponden en un alto grado con lo reportado en el estudio de la Cuenta de Alto Costo [4], donde quizás la mayor diferencia se encuentra en el rango de usuarios entre los 50 y 74 años que padecen diabetes, con el 65,73 %, que es más alto que el 55,80 % obtenido con los datos de ASMET Salud. Esta diferencia puede estar relacionada con el tipo de población que está incluida en la base de datos de ASMET Salud, la cual corresponde principalmente a usuarios de las zonas rurales del suroccidente y nororiente del país, mientras que la Cuenta de Alto Costo incluye información de usuarios de entidades de salud de toda Colombia.

De los resultados del presente estudio y reportados en la Tabla 1, se puede afirmar que la HTA y la DM son enfermedades que, para la población bajo estudio, tienen mayor prevalencia en las personas con edades mayores que los 40 años, sin querer decir con esto que la población menor de 40 años está exenta de sufrir estas enfermedades.

El desempeño de los clasificadores utilizados en el presente trabajo (KNN, DT y RF) se ve afectado en gran proporción por el desbalance en la distribución de las dos clases modeladas: afiliados que tienen HTA, pero sin DM (No-DM), y afiliados que tienen HTA y también presentan DM (Sí-DM). Esto se refleja principalmente en la métrica de la sensibilidad (Sen), la cual da información sobre el porcentaje de pacientes positivos (grupo Sí-DM) que el modelo puede predecir. En particular, se pasa de valores que están alrededor del 15 % en la base de datos sin balancear, a valores de hasta el 54 % en la base de datos balanceada.

Lo anterior tiene su explicación en el hecho de que para el entrenamiento del modelo se consideró el Acc como métrica para la optimización de parámetros, la cual permite conocer la relación global de usuarios clasificados correctamente, contando tanto los usuarios clasificados en el grupo Sí-DM, como en el grupo No-DM. Es por esto por lo que el valor de la métrica de especificidad es tan alto cuando se utiliza la base de datos sin balancear (valores entre 88 y 99 %), en comparación con el valor obtenido utilizando la base de datos balanceada

(valores entre 51 y 66 %). En general, al tener los datos balanceados, se tuvo un mejor equilibrio entre las medidas de especificidad y sensibilidad de los tres modelos propuestos.

Para la evaluación de los modelos propuestos se propusieron métricas como la Acc, la Sen, la Esp, el F1 Score y el AUC-ROC. Para el caso del presente estudio se decidió dar prioridad a la métrica de Sen, teniendo en cuenta que el mejor modelo se va a utilizar por parte de los colaboradores de la EPS ASMET Salud, de manera que es de suma importancia conocer los pacientes que pueden llegar a desarrollar DM (positivos verdaderos), para promover la captación temprana de estos pacientes, lo cual permitirá realizar acciones en esquemas de prevención y promoción que podrán evitar o ralentizar el desarrollo de la DM.

Lo anterior, además de poderse ver reflejado en un beneficio en la calidad de vida de las personas que sean tratadas adecuadamente y a tiempo, se puede convertir en un beneficio económico para la EPS, al permitir la mejora de los indicadores que mide la Cuenta de Alto Costo [4] para determinar el porcentaje de compensación que las EPS deben realizar en el manejo de los pacientes que sufren DM.

Aunque estudios como los reportados en Abbas *et al.* [15] han logrado valores de sensibilidad del 81,1 %, es importante mencionar que estos valores surgieron al incluir cuatro variables en los modelos: dos variables fisiológicas, que se midieron en forma directa o se derivaron de una prueba oral de tolerancia a la glucosa, y dos variables sociodemográficas (edad y etnia). El mismo estudio [15] reporta que utilizando solo las dos variables fisiológicas (el área bajo la curva de glucosa en 2 h y el nivel de glucosa en plasma después de 120 min) se alcanza una sensibilidad del 72 % aproximadamente, resaltando el poder predictivo de este tipo de variables fisiológicas. Sin embargo, dicho estudio no reporta valores de sensibilidad, incluyendo solamente variables sociodemográficas.

Gracias a la naturaleza de los modelos empleados (modelos basados en árboles) es factible analizar la importancia de cada variable incluida en el proceso de construcción del modelo. Esto, alineado con un deseo empresarial, que es detectar la importancia de las variables que inciden en el desarrollo de la DM, permitiría iniciar una investigación a detalle de un instrumento de medición que posibilite definir lineamientos o programas internos para la prevención de la enfermedad y la detección temprana.

A nivel general, los resultados demuestran que la consanguinidad es la variable más significativa con más del 20 % de importancia, seguido de la edad y el grupo étnico, dependiendo del modelo. Por su parte, la frecuencia del consumo de fruta y agua son factores decisivos en más del 2 %, siendo este un valor considerable. Es decir, lo anterior se puede unir en cuatro grandes grupos de importancia: 1) nivel de consanguinidad de familiares con diabetes; 2) edad; 3) dieta saludable; y 4) grupo étnico. La ausencia o presencia de cada uno de estos factores determinan más del 2 % el desarrollo de la DM dentro de la población bajo estudio (usuarios del régimen subsidiado de salud ubicados en zona suroccidental de Colombia), lo que confirma la importancia de la herencia genética dentro del desarrollo de diabetes, potenciado por el estilo de alimentación, la edad y la raza. Sin embargo, para el grupo étnico es necesario desarrollar un estudio de la distribución de grupos étnicos en Colombia y de los afiliados a la EPS, que busque evitar el sesgo sobre la predicción y que confirme que la distribución de afiliados no afecta la predicción.

Aunque el mayor valor AUC-ROC alcanzado en el presente trabajo fue del 0,61, se debe considerar que este valor se logró utilizando solamente variables sociodemográficas, antropométricas y de estilo de vida, excluyendo variables metabólicas, lo cual, aunque es el objetivo del estudio, también puede verse como una limitación.

Otros estudios reportados en la literatura [15-19] han alcanzado valores de AUC-ROC superiores al 80 %, lo cual ha sido posible especialmente porque los modelos han introducido como variables de entrada parámetros tomados a partir de variables fisiológicas, incluyendo analíticas de muestras de sangre. Por lo tanto, para trabajos futuros, se sugiere realizar un análisis más detallado de las variables registradas en la base de datos, aplicando transformaciones en dichas variables o incluyendo nuevas variables, como podrían ser las medidas antropométricas de los usuarios u otras de fácil adquisición, y que puedan estar asociadas al padecimiento de HTA y DM.

Conclusión

Se desarrollaron modelos, basados en técnicas de aprendizaje automático (ML), para apoyar en el diagnóstico temprano de la DM o en la predicción de esta, a fin de permitir a los profesionales de la salud establecer estrategias de prevención o tratamiento oportunos de la DM. Los modelos utilizan, como entradas variables, derivadas de datos ambientales, sociales, económicos y sanitarios, sin la dependencia de la toma de muestras clínicas, registradas en usuarios de ASMET Salud, institución del régimen subsidiado de salud en Colombia que cubre principalmente la zona suroccidental del país.

Los mejores valores de AUC-ROC fueron obtenidos con los modelos basados en ensambles, que integran modelos supervisados utilizando KNN, DT y RF. El ensamble utilizando la técnica XGB obtuvo el valor más alto de AUC-ROC (0,61), identificando como variables de mayor peso a las asociadas con aspectos hereditarios (24,65 %) y con el grupo étnico (5,59 %), además de

la dificultad visual, el bajo consumo de agua, una dieta baja en frutas y verduras y el consumo de sal y azúcar.

Agradecimientos

A las directivas de la entidad ASMET Salud EPS SAS por facilitar la base de datos y asesoría para la interpretación de la información registrada en ella.

Declaración financiación

No se relaciona fuente de financiación.

Declaración conflictos de interés

Los autores declaran que no tienen ningún tipo de conflicto de interés.

Declaración de responsabilidad.

Los autores declaran que todo lo escrito y los diferentes puntos de vista es responsabilidad de todos los autores, quienes revisaron el manuscrito final y lo aprobaron. Las instituciones de afiliación no son responsables sobre lo que se describe en este artículo.

Declaración de contribución por autores

Jessner Alexander Mejía, Mario Andrés Oviedo Benalcázar, José Armando Ordoñez y *José Fernando Valencia Murillo* tuvieron una contribución sustancial en el diseño de la investigación, el análisis e interpretación de los datos, en la revisión crítica de su contenido intelectual, la aprobación de la versión final del manuscrito enviado, y están en capacidad de responder por las cuestiones relacionadas con la exactitud o integridad de cualquier parte del trabajo

Referencias

1. Howlader KC, Satu MS, Awal MA, et al. Machine learning models for classification and identification of significant attributes to detect type 2 diabetes. Health Inf Sci Syst 2022;10(2). DOI: https://doi.org/10.1007/s13755-021-00168-2
2. Bernardini D. Sobre los aspectos económicos de la diabetes mellitus. Rev Cubana Aliment Nutr. [internet]. 2022 [citado 2022 ago. 26]; 30(Supl. 2):255-61. Disponible en: http://revalnutricion.sld.cu/index.php/rcan/article/view/1226/1701
3. Organización Mundial de la Salud. Informe mundial sobre la diabetes. Geneva, Switzerland: WHO [internet]; 2016 [citado 2022 ago. 26]. Disponible en: https://apps.who.int/iris/bitstream/handle/10665/254649/9789243565255-spa.pdf
4. Cuenta de Alto Costo, Fondo Colombiano de Enfermedades de Alto Costo. Situación de la enfermedad renal crónica, la hipertensión arterial y la diabetes mellitus en Colombia 2020. Bogotá [internet]; 2021 [citado 2022 ago. 26]. Disponible en: https://cuentadealtocosto.org/site/publicaciones/situacion-de-la-enfermedad-renal-cronica-la-hipertension-arterial-y-diabetes-mellitus-en-colombia-2020/
5. Colombia, Ministerio de Salud y Protección Social. Prevenir la diabetes, clave desde los hábitos saludables. [internet]; 2021 [citado 2022 ago. 26]. Disponible en: https://www.minsalud.gov.co/Paginas/Prevenir-la-diabetes-clave-desde-los-habitos-saludables.aspx
6. Kruczkowski M, Drabik-Kruczkowska A, Marciniak A, et al. Predictions of cervical cancer identification by photonic method combined with machine learning. Sci Rep. 2022;12(1):3762. DOI: https://doi.org/10.1038/s41598-022-07723-1
7. Hameed Z, Zahia S, Garcia-Zapirain B, et al. Breast cancer histopathology image classification using an ensemble of deep learning models. Sensors. 2020;20(16):4373. DOI: https://doi.org/10.3390/s20164373
8. Konnaris MA, Brendel M, Fontana MA, et al. Computational pathology for musculoskeletal conditions using machine learning: Advances, trends, and challenges. Arthritis Res Ther. 2022;24(1):68. DOI: https://doi.org/10.1186/s13075-021-02716-3
9. Lee LS, Chan PK, Wen C, et al. Artificial intelligence in diagnosis of knee osteoarthritis and prediction of arthroplasty outcomes: A review. Arthroplasty. 2022;4(1):16. DOI: https://doi.org/10.1186/s42836-022-00118-7
10. Lazzarini PA, Raspovic A, Prentice J, et al. Guidelines development protocol and findings: Part of the 2021 Australian evidence-based guidelines for diabetes-related foot disease. J Foot Ankle Res. 2022;28:15. DOI: https://doi.org/10.1186/s13047-022-00533-8
11. Patel D, Msosa YJ, Wang T, et al. An implementation framework and a feasibility evaluation of a clinical decision support system for diabetes management in secondary mental healthcare using CogStack. BMC Med Inform Decis Mak. 2022;100(1):22. DOI: https://doi.org/10.1186/s12911-022-01842-5
12. Cerón-Rios GM, Lopez-Gutierrez DM, et al. Recommendation System based on CBR algorithm for the Promotion of Healthier Habits. Sanchez-Ruiz AA, Kofod-Petersen A, editors. Proceedings of ICCBR 2017 Workshops (CAW, CBRDL, PO-CBR), Doctoral Consortium, and Competitions co-located with the 25th International Conference on Case-Based Reasoning (ICCBR 2017). Trondheim, Norway, June 26-28, 2017. CEUR Workshop Proceedings [internet]; 2017. pp. 167-76 [citado 2022 ago. 26]. Disponible en: https://ceur-ws.org/Vol-2028/paper16.pdf
13. Li J, Huang J, et al. Application of artificial intelligence in diabetes education and management: Present status and promising prospect. Front Public Health. 2020;8:173. DOI: https://doi.org/10.3389/fpubh.2020.00173
14. Rohokale V, Rashmi Neeli, Prassad Ramjee. A cooperative internet of things (IoT) for rural healthcare monitoring and control. 2011 2nd International Conference on Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology (Wireless VITAE). 2011; 1-6. DOI: https://doi.org/10.1109/WIRELESSVITAE.2011.5940920
15. Abbas H, Alic L, Rios M, et al. Predicting diabetes in healthy population through machine learning. In: Proceedings - IEEE Symposium on Computer-Based Medical Systems. Institute of Electrical and Electronics Engineers Inc. [internet]; 2019. pp. 567-70

[citado 2022 ago. 26]. Disponible en: https://ieeexplore.ieee.org/document/8787404

16. Zhang L, Wang Y, Niu M, et al. Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: The Henan Rural Cohort Study. Sci Rep. 2020;4406(1):10. DOI: https://doi.org/10.1038/s41598-020-61123-x
17. Dinh A, Miertschin S, et al. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC Med Inform Decis Mak. 2019; 211(1):19. DOI: https://doi.org/10.1186/s12911-019-0918-5
18. Fazakis N, Kocsis O, Dritsas E, et al. Machine learning tools for long-term type 2 diabetes risk prediction. IEEE Access. 2021;9:103737-57. DOI: https://doi.org/10.1109/ACCESS.2021.3098691
19. Shetty G, Katkar V. Type-II diabetes detection using decision-tree based ensemble of classifiers. In: 2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA); 2019. pp. 1-5. DOI: https://doi.org/10.1109/ICCUBEA47591.2019.9129348
20 Haq AU, Li JP, Khan J, et al. Intelligent machine learning approach for effective recognition of diabetes in e-healthcare using clinical data. Sensors. 2020;20(9):2649. DOI: https://doi.org/10.3390/s20092649
21. Leiva AM, Martínez MA, Petermann F, et al. Factores asociados al desarrollo de diabetes mellitus tipo 2 en Chile. Nutr Hosp. 2018;35(2):400-7. DOI: https://doi.org/10.20960/nh.1434
22. Géron A. Hands-on machine learning with Scikit-Learn and TensorFlow. CA: O'Reilly Media; 2017. https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/
23. Priyam A, Abhijeeta, Gupta R, et al. Comparative analysis of decision tree classification algorithms. Int. J. Curr. Eng. Technol. 2013;3(2):334-7. https://inpressco.com/comparative-analysis-of-decision-tree-classification-algorithms/