ISSN 2215-5910
Investigación

Predicción de la Diabetes Utilizando Técnicas de Minería de Datos: Un Estudio Comparativo

Gary Carballo Barrantes
André Montoya Fernández
Marco Fonseca León
José Ugalde-Rodriguez
César Rodríguez Bravo

Resumen

La diabetes es una enfermedad metabólica crónica que afecta a una gran parte de la población mundial. La detección temprana y la intervención son cruciales para prevenir sus complicaciones. En este trabajo se presenta un enfoque de minería de datos para predecir el riesgo de diabetes utilizando un conjunto de datos de registros médicos de pacientes. El estudio explora dos algoritmos de aprendizaje automático, regresión logística y bosque aleatorio, para construir modelos predictivos para la diabetes. El rendimiento de los modelos se evaluó utilizando varias métricas, incluyendo precisión, especificidad, sensibilidad y el área bajo la curva (AUC). Los resultados sugieren que el modelo de regresión logística tiene un mejor desempeño para predecir los resultados de la diabetes en este conjunto de datos. El estudio demuestra el potencial de las técnicas de minería de datos en la predicción de la diabetes y la importancia de la detección temprana para prevenir complicaciones.

Términos clave

Diabetes, minería de datos, regresión logística, bosque aleatorio, modelado predictivo, aprendizaje automático.

Abstract

Diabetes is a chronic metabolic disease that affects a large portion of the population worldwide. Early detection and intervention are crucial in preventing its complications. This paper presents a data mining approach to predict the risk of diabetes using a dataset of patients' medical records. The study explores two machine learning algorithms, logistic regression, and random forest, to build predictive models for diabetes. The performance of the models was evaluated using several metrics, including accuracy, precision, specificty, sensitivity and auc. The results suggest that the logistic regression model performs better for predicting diabetes outcomes in this dataset. The study demonstrates the potential of data mining techniques in predicting diabetes and the importance of early detection to prevent complications.

Index Terms

Diabetes, data mining, logistic regression, ran-dom forest, predictive modeling, machine learning.

Gary Carballo Barrantes: Estudiante Avanzado en la Carrera ING. Ciencias de Datos en LEAD University.
André Montoya Fernández: Estudiante Avanzado en la Carrera ING. Ciencias de Datos en LEAD University.
Marco Fonseca León: Estudiante Avanzado en la Carrera ING. Ciencias de Datos en LEAD University.
José Ugalde Rodriguez: Estudiante Avanzado en la Carrera ING. Ciencias de Datos en LEAD University.

César Rodrígiez Bravo: Profesor de Lead University, Máster en Ciberseguridad e Inventor con más de 100 aplicaciones a patentes en Estados Unidos, Europa y China.

I. INTRODUCCIÓN

La diabetes es una enfermedad crónica que afecta la forma en que el cuerpo convierte los alimentos en energía. Cuando comemos, nuestro cuerpo descompone los alimentos en moléculas más pequeñas, como la glucosa, que luego se liberan en el torrente sanguíneo. La insulina es una hormona producida por el páncreas que ayuda a que la glucosa entre en nuestras células para ser utilizada como energía. En la diabetes, el cuerpo no produce suficiente insulina o no la utiliza correctamente, lo que provoca un aumento en los niveles de azúcar en la sangre. Esta enfermedad puede causar problemas a largo plazo en el cuerpo, como daño en los ojos, riñones, nervios, piel, corazón y vasos sanguíneos.

La diabetes es una enfermedad compleja y diversa que afecta a millones de personas en todo el mundo. Existen tres tipos principales de diabetes: tipo 1, tipo 2 y diabetes gestacional. La diabetes tipo 1, que generalmente se desarrolla en la infancia o adolescencia, es causada por una reacción auto-inmunitaria en la que el cuerpo ataca a sus propias células productoras de insulina, lo que lleva a una falta de insulina en el cuerpo. Las personas con diabetes tipo 1 requieren inyecciones de insulina o una bomba de insulina para controlar sus niveles de azúcar en la sangre y prevenir complicaciones. La diabetes tipo 2 es el tipo más común de diabetes y generalmente se desarrolla en adultos, aunque también puede ocurrir en niños y adolescentes. La diabetes tipo 2 ocurre cuando el cuerpo se vuelve resistente a la insulina o no produce suficiente insulina para satisfacer sus necesidades. A menudo está relacionada con factores de estilo de vida como la obesidad, la inactividad física y una mala dieta, y puede ser controlada mediante cambios en el estilo de vida, medicamentos o una combinación de ambos.

La diabetes gestacional ocurre durante el embarazo y afecta a aproximadamente del 2 al 10 por ciento de las mujeres embarazadas. Suele ser temporal y se asocia a La diabetes gestacional temporal y desaparece después del parto, pero aumenta el riesgo de desarrollar diabetes tipo 2 más adelante en la vida, tanto para la madre como para el hijo. Por lo general, se aconseja a las mujeres con diabetes gestacional que controlen sus niveles de azúcar en la sangre y realicen cambios en su estilo de vida, como seguir una dieta saludable y hacer ejercicio regularmente, para gestionar su condición.

Sin importar el tipo de diabetes, niveles descontrolados de azúcar en la sangre pueden causar problemas de salud graves como enfermedades cardíacas, accidentes cerebro-vasculares, enfermedades renales y daño en los nervios. Es importante que las personas con diabetes trabajen en estrecha colaboración con sus proveedores de atención médica para desarrollar un plan de tratamiento individualizado que incluya el monitoreo regular de los niveles de azúcar en la sangre, el manejo de medicamentos y cambios en el estilo de vida para prevenir o controlar las complicaciones [3]. Es una enfermedad grave que afecta a millones de personas en todo el mundo. Según la Federación Internacional de Diabetes, aproximadamente 537 millones de adultos viven con diabetes en todo el mundo, y se espera que esa cifra aumente a 643 millones para 2030. La diabetes es una de las principales causas de muerte en el mundo y puede causar complicaciones graves si no se trata adecuadamente.

Mediante modelos predictivos desarrollados por el Instituto Nacional de Diabetes y Enfermedades Digestivas y Renales (NIDDK), se están realizando esfuerzos para detectar a pacientes en riesgo de diabetes antes de que la enfermedad se manifieste. Esto puede ser de gran ayuda para prevenir o retrasar el inicio de la diabetes tipo 2, la forma más común de diabetes. Además, el NIDDK también está trabajando en investigaciones para nuevos tratamientos y terapias para la diabetes, con el objetivo de mejorar la calidad de vida de los pacientes y prevenir complicaciones a largo plazo. [?]. La diabetes es una enfermedad grave y crónica que puede tener consecuencias graves para la salud de las personas. Según la Federación Internacional de Diabetes, más de 537 millones de adultos en todo el mundo viven con diabetes y se espera que esta cifra continúe aumentando. Por lo tanto, es importante seguir investigando y desarrollando nuevas formas de prevenir, diagnosticar y tratar la diabetes, y organizaciones como el NIDDK son esenciales para este esfuerzo.

Los esfuerzos del NIDDK para mejorar la capacidad de los médicos para predecir la diabetes en los pacientes mediante modelos estadísticos van más allá de solo predecir el inicio de la diabetes. El instituto también está trabajando para identificar a los pacientes que están en riesgo de desarrollar complicaciones relacionadas con la diabetes, como enfermedad renal, problemas de visión y daño en los nervios. Mediante el uso de una variedad de modelos estadísticos y de aprendizaje automático, los investigadores del NIDDK pueden analizar grandes cantidades de datos para identificar patrones y factores de riesgo asociados con estas complicaciones.

Además, el NIDDK también está explorando el uso de enfoques de medicina personalizada en el manejo de la diabetes. Al analizar las características individuales de un paciente, como su composición genética y factores de estilo de vida, los investigadores esperan desarrollar tratamientos más específicos y efectivos para la diabetes. Este enfoque tiene el potencial de mejorar los resultados para los pacientes con diabetes y reducir la carga general de la enfermedad.

En general, el trabajo del NIDDK en la investigación de la diabetes destaca la importancia de utilizar modelos estadísticos y de aprendizaje automático para comprender mejor la enfermedad y sus complicaciones. Al identificar factores de riesgo y desarrollar tratamientos específicos, los investigadores del NIDDK están dando pasos significativos para mejorar la vida de aquellos afectados por la diabetes.

II. ANTECEDENTES

La diabetes es una enfermedad crónica que afecta a millones de personas en todo el mundo. Los síntomas iniciales de la enfermedad a menudo son leves o incluso inexistentes, lo que significa que muchas personas no son conscientes de que tienen diabetes hasta que han sufrido un daño significativo en su salud. El diagnóstico temprano y la intervención son cruciales para prevenir complicaciones y asegurar un buen resultado en el tratamiento de la enfermedad. Por lo tanto, la necesidad de un modelo de predicción de la diabetes basado en técnicas avanzadas de minería de datos es vital para mejorar la detección temprana de la enfermedad y proporcionar atención preventiva a los pacientes. El desarrollo de un modelo de predicción de la diabetes utilizando técnicas avanzadas de minería de datos tiene el potencial de revolucionar la detección temprana y la prevención de la diabetes. Al analizar grandes cantidades de datos de pacientes, como antecedentes médicos, factores de estilo de vida y marcadores genéticos, estos modelos pueden identificar con precisión a los pacientes que tienen un mayor riesgo de desarrollar diabetes. Esta identificación temprana permite a los proveedores de atención médica implementar medidas preventivas como cambios en el estilo de vida, medicamentos y revisiones regulares, lo que puede reducir significativamente el riesgo de complicaciones y mejorar los resultados de los pacientes.

El costo del tratamiento y manejo de la diabetes puede ser significativo, y se estima que está entre los más altos entre las enfermedades crónicas. La carga financiera de la diabetes se debe al costo de los medicamentos, hospitalizaciones y complicaciones asociadas con la enfermedad, como enfermedades cardiovasculares, neuropatía, retinopatía y enfermedad renal. Además, la diabetes puede tener un impacto significativo en la calidad de vida del paciente, lo que puede llevar a un mayor uso de los servicios de salud y mayores costos. Al identificar a los pacientes que tienen un mayor riesgo de desarrollar diabetes e intervenir temprano con medidas preventivas, como cambios en el estilo de vida y medicamentos, los proveedores de atención médica pueden reducir el riesgo de complicaciones y mejorar los resultados de salud en general de los pacientes con diabetes. Esto puede llevar a una reducción en los costos de atención médica asociados con el tratamiento y manejo de la diabetes, así como una mejora en la calidad de vida de los pacientes con la enfermedad. El desarrollo de un modelo de predicción de la diabetes utilizando técnicas avanzadas de minería de datos puede ser una herramienta eficaz para lograr estos objetivos, exactamente como puede ayudar a los proveedores de atención médica a identificar a los pacientes en riesgo de desarrollar diabetes de manera más temprana e intervenir con medidas preventivas adecuadas.

El impacto potencial de un modelo de predicción de la diabetes basado en técnicas avanzadas de minería de datos no puede ser subestimado. La diabetes es una enfermedad crónica que afecta a millones de personas en todo el mundo y es una de las principales causas de morbilidad y mortalidad a nivel mundial. Los costos de atención médica asociados con la diabetes son significativos, ya que los pacientes a menudo requieren manejo y tratamiento de por vida para prevenir o controlar complicaciones. Sin embargo, al identificar a los pacientes que tienen un mayor riesgo de desarrollar la enfermedad, los proveedores de atención médica pueden intervenir temprano con medidas preventivas, como cambios en el estilo de vida y medicamentos, para reducir el riesgo de complicaciones y los costos asociados con la atención médica.

Además, el desarrollo de un modelo de predicción de la diabetes puede tener un impacto positivo en la salud pública. La detección temprana e intervención de la enfermedad pueden prevenir complicaciones, como enfermedades cardio-vasculares, accidentes cerebro-vasculares y enfermedad renal, que pueden afectar significativamente la calidad de vida del paciente y llevar a discapacidad o muerte prematura. Al identificar con precisión a los pacientes que tienen un mayor riesgo de desarrollar diabetes, los proveedores de atención médica pueden brindar intervenciones específicas y promover comportamientos saludables para prevenir o retrasar el inicio de la enfermedad.

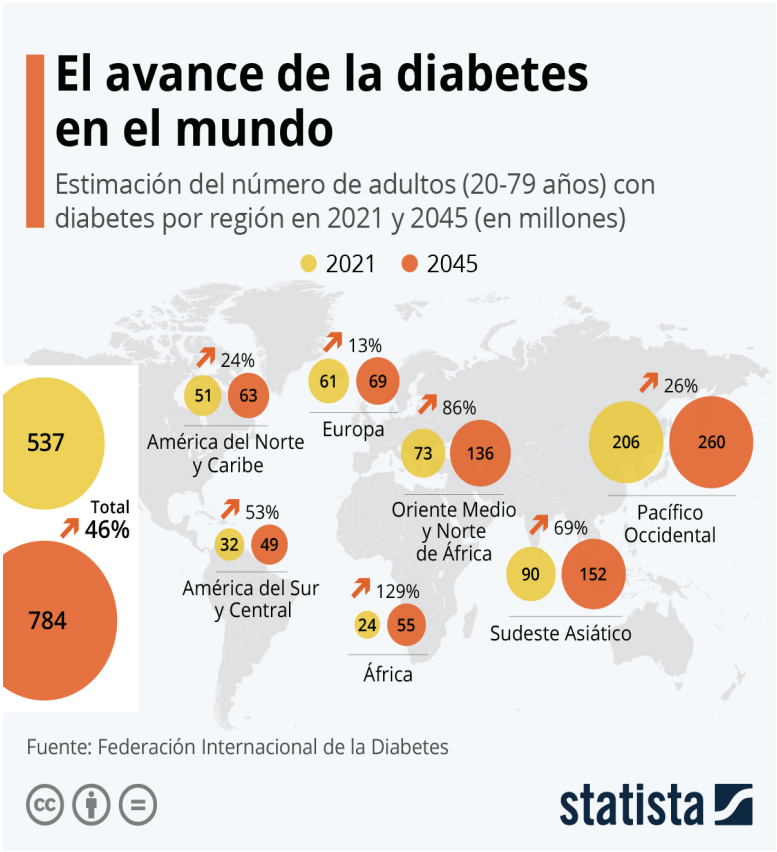
Según la Federación Internacional de Diabetes, la diabetes es una de las principales causas de morbilidad y mortalidad a nivel mundial. La organización estima que en 2019, 463 millones de adultos (de 20 a 79 años) vivían con diabetes, y se espera que esta cifra aumente a 700 millones para 2045 si las tendencias actuales continúan. La diabetes también está asociada con numerosas complicaciones, como enfermedades cardíacas, accidentes cerebro-vasculares, enfermedades renales y ceguera, que pueden afectar significativamente la calidad de vida del paciente y los costos de atención médica.

El uso de técnicas avanzadas de minería de datos, como algoritmos de aprendizaje automático y análisis predictivo, puede ayudar a los proveedores de atención médica a identificar a los pacientes con un mayor riesgo de desarrollar diabetes de manera más temprana. Al analizar diversos datos del paciente, como demografía, antecedentes familiares y factores de estilo de vida, se pueden desarrollar modelos predictivos para identificar con precisión a los pacientes que tienen más probabilidades de desarrollar la enfermedad. Esto, a su vez, permite a los proveedores de atención médica intervenir más temprano con medidas preventivas, como cambios en el estilo de vida y medicamentos, lo que puede reducir significativamente el riesgo de complicaciones relacionadas con la diabetes y los costos asociados con la atención médica.

Finalmente, para desarrollar el modelo de predicción de la diabetes se utilizará R. Este lenguaje es ampliamente utilizado en el análisis de datos y la minería de datos, lo que lo hace ideal para crear modelos de predicción precisos y confiables. Además, R cuenta con una gran cantidad de bibliotecas y paquetes que facilitan la manipulación y análisis de datos, lo que nos permite trabajar de manera eficiente en la creación de modelos de predicción de diabetes. La minería de datos es una herramienta valiosa en la detección temprana y prevención de la diabetes, sin embargo, los desafíos de la privacidad de los datos deben abordarse y asegurarse de que se maneje de manera ética y legal para proteger los derechos de privacidad de los pacientes.

III. TRABAJOS RELACIONADOS

La diabetes es una enfermedad crónica que afecta a millones de personas en todo el mundo, y se estima que su pre-valencia seguirá aumentando en los próximos años. La detección temprana y la intervención son cruciales para prevenir complicaciones y garantizar buenos resultados en el tratamiento de la enfermedad. Por lo tanto, es necesaria con urgencia un modelo predictivo para la diabetes basado en técnicas avanzadas de minería de datos para mejorar la detección temprana de la enfermedad y proporcionar atención preventiva a los pacientes.



El costo del tratamiento y manejo de la diabetes puede ser significativo, y se estima que está entre los más altos entre las enfermedades crónicas. La carga financiera de la diabetes se debe al costo de los medicamentos, hospitalizaciones y complicaciones asociadas con la enfermedad, como enfermedades cardiovasculares, neuropatía, retinopatía y enfermedad renal. Además, la diabetes puede tener un impacto significativo en la calidad de vida de los pacientes, lo que puede llevar a un mayor uso de los servicios de salud y mayores costos. Al identificar a los pacientes que tienen un mayor riesgo de desarrollar diabetes e intervenir temprano con medidas preventivas como cambios en el estilo de vida, medicamentos y revisiones regulares, los proveedores de atención médica pueden reducir el riesgo de complicaciones y mejorar los resultados de salud en general de los pacientes con diabetes. Esto puede llevar a una reducción en los costos de atención médica asociados con el tratamiento y manejo de la diabetes, así como una mejora en la calidad de vida de los pacientes con la enfermedad. El desarrollo de un modelo predictivo de diabetes utilizando técnicas avanzadas de minería de datos puede ser una herramienta eficaz para lograr estos objetivos, ya que puede ayudar a los proveedores de atención médica a identificar a los pacientes en riesgo de desarrollar diabetes e intervenir con medidas preventivas adecuadas.

El impacto potencial de un modelo predictivo de diabetes basado en técnicas avanzadas de minería de datos no puede ser exagerado. La diabetes es una enfermedad crónica que afecta a millones de personas en todo el mundo y es una de las principales causas de morbilidad y mortalidad a nivel mundial. Los costos de atención médica asociados con la diabetes son significativos, ya que los pacientes a menudo requieren tratamiento y manejo de por vida para prevenir o controlar complicaciones. Sin embargo, al identificar a los pacientes que tienen un mayor riesgo de desarrollar la enfermedad, los proveedores de atención médica pueden intervenir temprano con medidas preventivas, como cambios en el estilo de vida y medicamentos, para reducir el riesgo de desarrollar complicaciones y los costos de atención médica. Además, el desarrollo de un modelo predictivo de diabetes puede tener un impacto positivo en la salud pública. La detección temprana e intervención de la enfermedad pueden prevenir complicaciones como enfermedades cardio-vasculares, accidentes cerebro-vasculares y enfermedad renal.

IV. DISEÑO EXPERIMENTAL

Este estudio tiene como objetivo desarrollar un modelo predictivo para la diabetes utilizando técnicas de minería de datos.

A. Descripción general del estudio empírico:

El diseño experimental implica los siguientes pasos:

(1) Recopilación de datos: Se recopiló un conjunto de datos que contiene una cantidad significativa de información sobre pacientes con diabetes, proveniente de fuentes de acceso público.

(2) Exploración y visualización de datos: Se llevaron a cabo estadísticas descriptivas, técnicas de visualización de datos y análisis de correlación para obtener una comprensión de las características del conjunto de datos.

(3) Desarrollo del modelo: Se aplicaron diferentes algoritmos de aprendizaje automático a las características seleccionadas para desarrollar un modelo predictivo para la diabetes.

(4) Evaluación del modelo: El rendimiento de los modelos desarrollados se evaluó utilizando métricas de evaluación apropiadas, como precisión, sensibilidad y especificidad.

B. Selección del conjunto de datos:

El conjunto de datos utilizado en este estudio se obtuvo del Instituto Nacional de Diabetes y Enfermedades Digestivas y Renales. Este conjunto de datos contiene información de 2000 pacientes femeninas que tienen al menos 21 años y son de ascendencia india Pima. Los datos fueron recopilados durante exámenes médicos realizados entre 1988 y 1990. El conjunto de datos incluye diversas variables clínicas y demográficas, como la edad, el índice de masa corporal (IMC), la presión arterial, los niveles de glucosa y el estado de diabetes.

La elección de este conjunto de datos se basó en varios factores. En primer lugar, el conjunto de datos incluye un número significativo de pacientes femeninas, lo cual es importante para estudiar debido a la mayor prevalencia de diabetes en mujeres en comparación con hombres. En segundo lugar, la inclusión de pacientes de ascendencia india Pima también es significativa, ya que esta población tiene una mayor prevalencia de diabetes tipo 2 en comparación con otros grupos étnicos. Finalmente, la disponibilidad de una amplia variedad de variables clínicas y demográficas en el conjunto de datos permite un análisis integral de los factores que contribuyen al desarrollo de la diabetes.

Los datos se recopilaron originalmente como parte de un estudio para determinar los factores asociados con el desarrollo de la diabetes tipo 2 en esta población. Todos los pacientes proporcionaron su consentimiento informado por escrito, y el estudio fue aprobado por el comité de ética del Instituto Nacional de Diabetes y Enfermedades Digestivas y Renales. Los datos se des-identificaron para proteger la privacidad de los pacientes.

En general, la selección del conjunto de datos fue fundamental para el éxito de este estudio, ya que permitió un análisis integral de los factores que contribuyen a la diabetes en una población de alto riesgo. El gran tamaño de la muestra, la diversidad de variables clínicas y demográficas y la recopilación de datos de alta calidad hacen de este conjunto de datos un recurso ideal para los investigadores que estudian la diabetes.

C. Exploración y visualización de datos

La etapa de exploración y visualización de datos en un artículo científico es un paso crítico para comprender los datos e identificar patrones o tendencias que puedan existir. En esta etapa, realizamos un análisis exploratorio de datos (EDA) en el conjunto de datos para obtener una comprensión de la distribución, tendencia central y variabilidad de cada variable.

Una de las formas más comunes de visualizar la distribución de variables es mediante el uso de diagramas de caja (box plots).

A través de los diagramas de caja, podemos identificar la mediana, los cuartiles y los valores atípicos de cada variable, lo que nos ayuda a entender cómo se distribuyen los datos y si existen valores extremos que puedan afectar nuestros análisis.

Además de los diagramas de caja, también podemos utilizar gráficos de dispersión para explorar la relación entre dos variables y detectar posibles correlaciones o patrones. Esto puede ser especialmente útil para identificar posibles factores de riesgo o predictores de la diabetes.

La visualización de datos nos ayuda a hacer inferencias preliminares y a generar hipótesis sobre posibles relaciones entre las variables, lo que nos guiará en el desarrollo de nuestro modelo predictivo. Es una etapa esencial para entender la complejidad de los datos y obtener una visión general antes de aplicar algoritmos de aprendizaje automático para desarrollar el modelo de predicción de diabetes.

TABLE I
VARIABLES ON THE DATASET

Variable name	Description
Pregnancies	To express the Number of pregnancies
Glucose	To express the Glucose level in blood
Blood Pressure	To express the Blood pressure measurement To
Skin Thickness	express the thickness of the skin
Insulin	To express the Insulin level in blood
BMI	To express the Body mass index
Diabetes Pedigree Function	To express the Diabetes percentage
Age	To express the age
Outcome	To express the final result 1 is Yes and 0 is No

Box plots muestran el mínimo, el primer cuartil, la mediana, el tercer cuartil y el máximo de los datos, lo que ayuda a identificar valores atípicos o sesgos en la distribución. Después de explorar la distribución de cada variable, se realizaron pruebas de normalidad utilizando pruebas estadísticas como la prueba de Shapiro-Wilk o la prueba de Kolmogorov-Smirnov. Además, el analista puede querer comparar las medias de dos o más grupos utilizando pruebas como la prueba de Kruskal-Wallis o la prueba de Mann-Whitney. Estas pruebas pueden ayudar a identificar diferencias significativas en las medias de cada grupo y pueden ser utilizadas para respaldar o refutar la hipótesis que se está probando. Finalmente, generamos un gráfico de correlación para examinar la fuerza y dirección de la relación entre pares de variables en el conjunto de datos. Este gráfico puede ayudar a identificar correlaciones significativas entre variables, lo que proporciona información sobre los patrones y relaciones subyacentes en los datos.

D. Desarrollo del Modelo

Antes de comenzar a desarrollar nuestros modelos, necesitamos pre-procesar nuestros datos para extraer las características más informativas. Las características utilizadas en este estudio incluyen la edad, el IMC, la presión arterial, el nivel de glucosa, el nivel de insulina, el número de embarazos y el grosor de la piel. Para seleccionar el mejor modelo, evaluamos el rendimiento de varios algoritmos de aprendizaje automático, incluyendo regresión logística y bosque aleatorio. Dividimos nuestro conjunto de datos en conjuntos de entrenamiento (60%) y prueba (40%). Utilizamos el conjunto de entrenamiento para entrenar nuestros modelos y el conjunto de prueba para evaluar su rendimiento.

Utilizamos validación cruzada de 10 pliegues en el conjunto de entrenamiento para evaluar el rendimiento de cada modelo. Las métricas de evaluación que utilizamos incluyen precisión, valor predictivo positivo, valor predictivo negativo, sensibilidad, especificidad y área bajo la curva de características operativas del receptor (ROC).

Después de entrenar y ajustar nuestros modelos, evaluamos su rendimiento en el conjunto de prueba. Los resultados muestran que el modelo de regresión logística logró la mayor precisión de 0.7872. La precisión, sensibilidad, especificidad y área bajo la curva de características operativas del receptor (ROC) para cada modelo se resumen en la Tabla 3. Basándonos en estas métricas de evaluación, seleccionamos el modelo de Regresión Logística como el modelo de mejor rendimiento para predecir la diabetes.

El diseño experimental se implementó utilizando el lenguaje de programación R y diversas bibliotecas de código abierto.

E. Evaluación del Modelo

Para evaluar el rendimiento de los modelos desarrollados en este estudio, utilizamos una variedad de métricas. En primer lugar, calculamos la precisión de cada modelo en los datos de entrenamiento y en los datos de prueba. La precisión mide la proporción de instancias clasificadas correctamente sobre el total de instancias en el conjunto de datos.

Además de la precisión, también calculamos la precisión, el recall y el puntaje F1 de cada modelo. La precisión mide la proporción de verdaderos positivos (instancias positivas clasificadas correctamente) sobre todas las instancias positivas predichas. El recall mide la proporción de verdaderos positivos sobre todas las instancias positivas reales.

El puntaje F1 es la media armónica de la precisión y el recall, y proporciona un equilibrio entre ambas métricas.

También calculamos el área bajo la curva de características operativas del receptor (AUC-ROC) para cada modelo. El AUC-ROC es una medida de cómo el modelo puede distinguir entre instancias positivas y negativas, y es especialmente útil en problemas de clasificación binaria.

Finalmente, utilizamos la validación cruzada para evaluar el rendimiento de generalización de los modelos. Utilizamos validación cruzada k-fold, donde k se estableció en 10, y evaluamos los modelos en cada pliegue. Esto nos permitió obtener una estimación más robusta del rendimiento del modelo en datos no vistos.

V. ANÁLISIS DE DATOS Y RESULTADOS

Se realizó un análisis exploratorio de datos (EDA) para cada variable dependiente, y se crearon diagramas de caja para identificar posibles valores atípicos o extremos. La normalidad se evaluó utilizando tres pruebas: Shapiro-Wilk, Royston y Henze-Zirkler, todas las cuales arrojaron distribuciones no normales.

Se desarrollaron y evaluaron dos modelos utilizando validación cruzada de 10 pliegues: regresión logística y bosque aleatorio. El modelo de regresión logística tuvo un error de predicción promedio de 15.67843 y una desviación estándar de 0.01079974, mientras que el modelo de bosque aleatorio tuvo un error de predicción promedio de 94.00551 y una desviación estándar de 1.508672. También se calculó el área bajo la curva (AUC) para cada modelo. El modelo de regresión logística tuvo un AUC de 0.8286118, mientras que el modelo de bosque aleatorio tuvo un AUC de 0.93541.

Los resultados finales mostraron que el modelo de regresión logística tuvo un mejor rendimiento que el modelo de bosque aleatorio. La matriz de confusión para el modelo de regresión logística mostró una precisión de 0.7872, un índice kappa de 0.4902 y una sensibilidad de 0.9163. La especificidad fue más baja, con un valor de 0.5385, lo que indica una mayor tasa de falsos positivos. El valor predictivo positivo fue de 0.7928 y el valor predictivo negativo fue de 0.7696. La prevalencia fue de 0.6583 y la tasa de detección fue de 0.6033. La prevalencia de detección fue de 0.7610 y la precisión equilibrada fue de 0.7274.

La matriz de confusión para el modelo de bosque aleatorio, por otro lado, mostró una precisión de 0.3417, mucho más baja que la del modelo de regresión logística. La sensibilidad fue de 0.0000, lo que indica una incapacidad total para detectar la clase positiva. La especificidad fue de 1.0000, lo que significa que el modelo fue capaz de identificar correctamente todos los casos negativos. El valor predictivo positivo fue NaN (no es un número), mientras que el valor predictivo negativo fue 0.9955.

TABLE II
DATASET (DS) INDEPENDENT VARIABLES DESCRIPTIVE STATISTICS

Variable	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Pregnancies	0.0	1.0	3.0	3.7	6.0	17.0
Glucose	0.0	99.0	117.0	121.2	141.0	199.0
Blood Pressure	0.0	63.50	72.00	69.15	80.00	122.00
Skin Thickness	0.0	0.00	23.00	20.93	32.00	110.00
Insulin	0.0	0.00	40.00	80.25	130.00	744.00
BMI	0.0	27.38	32.30	32.19	36.80	80.60
Diabetes Pedigree Function	0.0780	0.2440	0.3760	0.4709	0.6240	2.4200
Age	21.00	24.00	29.00	33.09	40.00	81.00

TABLE III
MODEL PERFORMANCE ON THE TESTING SET

Model	Accuracy	Specificity	Sensitivity	AUC
Logistic Regression	0.7872	0.5385	0.9163	0.8286118
Random Forest	0.3417	1.0000	0.0000	0.93541

El valor predictivo del modelo de bosque aleatorio fue de 0.3417. La prevalencia fue de 0.6583, mientras que la tasa de detección y la prevalencia fueron ambas de 0.0000. La precisión equilibrada fue de 0.5000, lo que indica que el modelo no fue mejor que una predicción aleatoria.

En general, el modelo de regresión logística mostró resultados prometedores en la predicción de los resultados de la diabetes basados en las variables recopiladas, mientras que el modelo de bosque aleatorio tuvo un rendimiento deficiente. Sin embargo, se necesita una investigación adicional para identificar posibles variables de confusión o limitaciones de los modelos.

VI. DISCUSIÓN Y CONCLUSIONES

El uso de técnicas de aprendizaje automático en la predicción de la diabetes es un avance importante en la atención médica. La diabetes es una enfermedad crónica que afecta a millones de personas en todo el mundo y puede tener graves consecuencias para la salud. La detección temprana de la diabetes es crucial para prevenir complicaciones a largo plazo, como enfermedades cardíacas, insuficiencia renal y ceguera.

En conclusión, este estudio demuestra que el aprendizaje automático puede ayudar a los profesionales de la salud a identificar a las personas que tienen un alto riesgo de desarrollar diabetes. Al conocer factores de riesgo específicos, como los niveles de glucosa en sangre, la presión arterial y el índice de masa corporal (IMC), los proveedores de atención médica pueden tomar medidas preventivas, como promover estilos de vida saludables y administrar medicamentos, para reducir la incidencia de la diabetes y sus complicaciones.

Sin embargo, es importante señalar que este estudio tiene limitaciones y que los resultados pueden no ser generalizables a otras poblaciones. También se reconoce que existen otros factores importantes, como el estilo de vida, los antecedentes familiares y la dieta, que pueden afectar el inicio de la diabetes y que no se tuvieron en cuenta en este estudio. Por lo tanto, es crucial que se realicen más investigaciones para validar estos hallazgos y explorar cómo mejorar las técnicas de aprendizaje automático para hacerlas más precisas y aplicables a un grupo más amplio de pacientes.

En resumen, este estudio demuestra que el aprendizaje automático tiene el potencial de revolucionar la forma en que se diagnostica y trata la diabetes en el futuro. La identificación temprana de la diabetes y la prevención de sus complicaciones son fundamentales para mejorar la salud de las personas en todo el mundo. Con el uso de técnicas de aprendizaje automático, los profesionales de la salud pueden tomar decisiones más informadas y personalizadas para el cuidado de los pacientes con diabetes, Con el uso de técnicas de aprendizaje automático, los profesionales de la salud cuentan con una poderosa herramienta para alcanzar este objetivo.

AGRADECIMIENTOS

Los autores desean expresar su más sincero agradecimiento al Dr. Juan Murillo-Morera, Profesor de Minería de Datos Avanzada, por su inestimable orientación, apoyo y estímulo a lo largo del desarrollo de este artículo. Su experiencia en el campo de la minería de datos y su compromiso con la enseñanza han sido fundamentales para moldear nuestra comprensión del tema y han contribuido significativamente al éxito de este proyecto.

Al Profesor César Augusto Rodríguez Bravo por su guía y ayuda invaluable durante todo el proceso de publicación de este artículo. Su experiencia, conocimiento y dedicación en el campo de la investigación y publicación científica han sido fundamentales para el éxito de este trabajo.

El Profesor Rodríguez Bravo ha brindado su apoyo y orientación de manera generosa, proporcionando una dirección clara y valiosos consejos que han enriquecido el contenido y la calidad de este artículo. Su compromiso con el crecimiento académico y su pasión por la enseñanza han sido una fuente de inspiración para nosotros.

Además, quiero agradecerle por su paciencia y disposición para responder nuestras preguntas y resolver nuestras dudas a lo largo de este proceso. Su mentoría ha sido fundamental para nuestro desarrollo como investigadores y autores.

También deseamos extender nuestros agradecimientos a la Universidad por proporcionarnos los recursos e instalaciones necesarios para llevar a cabo esta investigación.

Finalmente, nos gustaría agradecer a todos aquellos que participaron en este estudio, sin quienes este trabajo no hubiera sido posible.

REFERENCIAS

[1] IDF Diabetes Atlas. Brussels, Belgium: International Diabetes Federation.

[2] Standards of medical care in diabetes—2023 abridged for primary care providers.

[3] American Diabetes Association. (2021). Standards of Medical Care in Diabetes–2021 abridged for primary care providers. Clinical diabetes: a publication of the American Diabetes Association,

[4] IDF diabetes atlas (9th ed.). Brussels, Belgium: International Diabetes Federation.

[5] Diabetes Overview. National Institute of Diabetes and Digestive and Kidney Diseases. Retrieved from https://www.niddk.nih.gov/health-information/diabetes/overview.

[6] Diabetes Overview. National Institute of Diabetes and Digestive and Kidney Diseases. Retrieved from https://www.niddk.nih.gov/health-information/diabetes/overview

[7] Type 2 diabetes. The Lancet, 389(10085), 2239-2251. doi: 10.1016/s0140-6736(16)00683-9

[8] Application of machine learning algorithms in predicting diabetes: A review. Diabetes Metabolic Syndrome: Clinical Research Reviews, 14(4), 713-718. doi: 10.1016/j.dsx.2020.04.050.

[9] Predicting diabetes based on longitudinal data: a deep learning approach. BMC Medical Informatics and Decision Making, 19(1), 73.

[10] Impact of social media on academic performance of students: A system-atic review. Education and Information Technologies, 26(3), 2813-2833.

LEAD INSTITUTE
¡Hacia una mejor sociedad!