16

PREDICCIÓN DE DIABETES EN MUJERES MEDIANTE UN MODELO PROBABILÍSTICO BASADO EN REDES BAYESIANAS

Evelyn Eliana Coaquira Flores
Universidad Nacional del Altiplano de Puno - UNAP

Fred Torres Cruz
Universidad Nacional del Altiplano de Puno - UNAP

Sebastian Jarom Condori Quispe
Universidad Nacional del Altiplano de Puno - UNAP

Julio Cesar Tisnado Puma
Universidad Nacional del Altiplano de Puno - UNAP

Romel P. Melgarejo-Bolivar
Universidad Nacional del Altiplano de Puno - UNAP

Alain Paul Herrera-Urtiaga
Universidad Nacional del Altiplano de Puno - UNAP

Ramiro Pedro Laura Murillo
Universidad Nacional del Altiplano de Puno - UNAP

Milton Vladimir Mamani Calisaya
Universidad Nacional del Altiplano de Puno - UNAP

doi 10.37885/230412748

RESUMEN

Objetivo: Desarrollar un modelo probabilístico basado en redes bayesianas para la predicción de la diabetes mellitus en mujeres. **Métodos:** Se utilizaron varios métodos entre ellos el análisis exploratorio de datos, el preprocesamiento de datos, el modelado de la red bayesiana, la validación del modelo y la optimización del modelo para su mejor predicción y se exploró la posibilidad de mejorar su sensibilidad para detectar a todas las personas con diabetes Mellitus. **Resultados:** indican que el modelo de redes bayesianas tiene una precisión aceptable en la detección de la diabetes mellitus en mujeres, con un valor predictivo positivo del 69,57%, un valor predictivo negativo del 79,93% y una tasa de error del 23,18%. Sin embargo, se observa una necesidad de mejorar la sensibilidad del modelo para detectar a todas las personas con la enfermedad también se encontró que el modelo tiene una tasa de error del 23,18%, lo que indica que el modelo tiene un error del 23,18% en la clasificación de casos, por otro lado, los resultados de la regla de puntuación proporcionan información adicional sobre el desempeño del modelo, la pérdida logarítmica de 0,4904 y la pérdida cuadrática de 0,3222 indican que el modelo tiene un buen rendimiento en la clasificación de los casos, ya que estas medidas de pérdida son bajas. Además, el resultado esférico de 0,8196 indica que el modelo tiene una buena capacidad para discriminar entre las diferentes clases de la variable dependiente por tanto el modelo probabilístico basado en redes bayesianas es una herramienta prometedora para la predicción temprana de la diabetes mellitus **Conclusiones:** La construcción de un modelo probabilístico basado en redes bayesianas para la predicción de diabetes Mellitus ha demostrado ser una herramienta útil para la identificación temprana de la diabetes mellitus ya que el modelo cuenta con una precisión aceptable y un buen rendimiento en la clasificación de casos, se necesita mejorar su sensibilidad para detectar a todas las personas con diabetes, se destaca la importancia de seguir investigando en este campo para mejorar la precisión de los modelos y lograr una detección temprana más efectiva de la enfermedad.

Palabras-Clave: Redes Bayesianas, Diabetes Mellitus, Predicción.

INTRODUCCIÓN

La diabetes es una enfermedad crónica que afecta a millones de personas en todo el mundo. La detección temprana de la diabetes puede ayudar a prevenir complicaciones y mejorar la calidad de vida de los pacientes. En este contexto, los modelos probabilísticos basados en redes bayesianas se han utilizado con éxito para la predicción de la diabetes, una predicción precisa del subtipo de diabetes podría resultar valiosa en la implementación de programas de intervención y estrategias de modificación de conducta que tengan como objetivo prevenir o retrasar la progresión de la diabetes mellitus, la cual puede provocar complicaciones severas en la salud del paciente. (GOLLAPALLI e colab., 2022). En esta investigación, se utilizó un conjunto de datos proveniente del Instituto Nacional de Diabetes y Enfermedades Digestivas y Renales, que incluye medidas diagnósticas y características de pacientes femeninas de 768 casos de los cuales 500 casos dieron negativo y 268 dieron positivos Se construyó un modelo de redes bayesianas utilizando el software Netica, con la ayuda de expertos en salud. Aunque el modelo obtuvo una precisión aceptable, se necesita mejorar su sensibilidad para detectar a todas las personas con diabetes mellitus ya que la investigación de (SNEHA e GANGIL, 2019) también sugiere la generalización de la selección de características óptimas del conjunto de datos para mejorar la precisión de la clasificación en el futuro. Los resultados de la regla de puntuación proporcionan información adicional sobre el desempeño del modelo, mostrando que el modelo tiene un buen rendimiento en la clasificación de los casos. En resumen, esta investigación tiene como objetivo desarrollar un modelo de redes bayesianas para la predicción de la diabetes, con el objetivo de mejorar la detección temprana de la enfermedad y, en última instancia, mejorar la calidad de vida de los pacientes.

MÉTODOS

Revisión de la literatura

En diversas investigaciones se ha llevado a cabo entre ellas un análisis comparativo de algoritmos de aprendizaje automático para el diagnóstico de la diabetes mellitus, este estudio evidencia que para la base de datos de la India Prima, la máquina de soporte de vectores presenta una mayor eficacia con una precisión del 74%, en cambio para la base de datos de Alemania, los algoritmos KNN y RF muestran una precisión del 98,7%, estos resultados sugieren la utilización de algoritmos de aprendizaje automático para la predicción temprana de diabetes.(KANGRA e SINGH, 2023) asi como en la investigación de la mortalidad y la morbilidad en pacientes sometidos a cirugía cardíaca en el Reino Unido se han

O cuidado em saúde baseado em evidências - ISBN 978-65-5360-315-8 - Vol. 1 - Ano 2023 - Editora Científica Digital - www.editoracientifica.com.br

relacionado con la duración del procedimiento y la duración del bypass cardiopulmonar se ha propuesto el uso de redes bayesianas para identificar los mecanismos que influyen en los resultados y desarrollar modelos de riesgo en el futuro (MAZHAR e colab., 2023) Se construyo un modelo de red bayesiana para la predicción de fertilización baja y fallida en tecnología de resproducion asistida por la gran cantidad de datos clínicos ncluidas 13 variables relacionadas con mujeres, cinco variables relacionadas con hombres y seis variables relacionadas con el tratamiento de FIV/ICSI como resultado la precisión de predicción del modelo fue del 91,3 % modelo podría usarse para construir sistemas de apoyo a la toma de decisiones clínicas (TIAN e colab., 2023) en otro estudio se uso redes bayesianas con el algoritmo de búsqueda Tabu para explorar factores de riesgo de Hiperhomocisteinemia podrían ser un complemento para la regresión logística, lo que permitiría explorar la compleja relación de red y el vínculo general entre HHcy y sus factores de riesgo también promueve el funcionamiento de redes bayesianas en la practica clínica (SONG e colab., 2023) En una revisión sistemática y metaanálisis de satos de ensayos aleatorios afirma que las limitaciones incluyen el debate continuo sobre lo que constituye la remisión de la diabetes, así como la eficacia, la seguridad y la satisfacción dietética de las LCD a más largo plazo (GOLDENBERG e colab., 2021) en otro estudio utilizando redes neuronales profundas un enfoque de aprendizaje no supervisado para una predicción en el conjunto de datos de diabetes de los indios pima en el modelo logró una precisión del 98,16% con la división aleatoria de pruebas de entrenamiento (P e colab., 2020), en el estudio de algoritmos de aprendizaje automático para el diagnóstico precoz de la diabetes en un estudio comparativo propuso realizar el clasificador ingenuo bayesiano para su mejor precisión y de mayor complejidad Así mismo nuestro estudio alcanza un 89% de precisión de acuerdo con esto, el método de la red neuronal es el mejor para la detección temprana de la enfermedad diabética. (RAWAT e colab., 2022), en el estudio de inteligencia basada en inteligencia artificial de enfermedades clínicas utilizando un clasificador de bosque aleatorio y el algoritmo ingenuo bayesiano la red de clasificación bayesiana se calculó y comparo un análisis de rendimiento de los datos de la enfermedad para ambos algoritmos muestra una precisión de 74.46 en cuanto a la clasificación con el modelo de bosque aleatorio muestra una precisión de 74.03 en la investigación (JACKINS e colab., 2021), para la predicción y diagnostico del riesgo futuro de diabetes con un enfoque de aprendizaje automático con una precisión de 79.2% en regresión logística y con un 77% en el algoritmo ingenuo bayesiano y un 8% de precisión en Gradient Boosting y recomendó que se pueden aplicar enfoques similares en otros conjuntos de datos de enfermedades, como enfermedades con fines de predicción (BIRJAIS e colab., 2019), en una investigación para la clasificación de diabetes mellitus en los Indios ima basado en aprendizaje automático uso de tres modelos de aprendizaje automático con el clasificador Naïve

Bayes, el clasificador de bosque aleatorio y los modelos de árboles de decisión J48 se llegó a la conclusión que el modelo ingenuo bayes funciona mucho mejor (CHANG e colab., 2022), en el estudio sobre los clasificadores bayesianos, que el clasificador bayesiano es simple pero preciso muestra que el método mas preciso son la redes bayesianas mucho mas que el clasificador ingenuo bayesiano (FRIEDMAN e colab., 1997), en una revisión sistemática con los modelos predictivos de aprendizaje automático y aprendizaje profundo para la diabetes tipo 2 las redes neuronales profundas demostraron ser óptimas, a pesar de su capacidad para manejar datos grandes así como las técnicas de selección de características resultaron útiles para aumentar la eficiencia del modelo (FREGOSO-APARICIO e colab., 2021), en la investigaciones de detección y clasificación de la enfermedad de la diabetes en datos de encuestas demográficas y de salud de la India utilizando métodos de aprendizaje automático se predijo la diabetes diabetes tipo 2 el DNN-FI obtuvo una mejor tasa de precisión en comparación con el bosque aleatorio con un 99,84 % en el aprendizaje automático y el algoritmo del árbol de decisión (THOTAD e colab., 2023), en otra investigación para ver el modelo de predicción de diabetes mediante técnicas de minería de datos el modelo de regresión logística obtuvo un 82.46% en comparación con el modelo de maquina de soporte de vectores en donde también pretende sugerir una nueva forma de hacer que las predicciones sobre los resultados de la diabetes sean más precisas (RASTOGI e BANSAL, 2023), en el estudio realizado para la comparación de la precisión para el diagnóstico de diabetes en una revisión sistemática y un metaanálisis en una red a OGTT, la FPG y la HbA1c han sido recomendadas por la ADA y la OMS como métodos para diagnosticar la diabetes, mientras que la FPG y la HbA1c se usan más ampliamente debido a su relativa conveniencia en comparación con la OGTT para el diagnóstico (DUONG e colab., 2023), en el estudio eficacia comparativa de diferentes patrones de alimentación en el tratamiento de la diabetes tipo 2 y la prediabetes con un metaanálisis de red bayesiana en el caso de pacientes diabéticos o prediabéticos, la elección y creación de planes alimenticios adecuados deben ser basados en sus condiciones generales, que incluyen perfiles de lípidos en sangre, patrones de glucemia, peso corporal y presión arterial.(GOLDENBERG e colab., 2021), en la investigación de un nuevo conjunto de apilamiento para detectar tres tipos de diabetes mellitus utilizando un conjunto de datos de Arabia Saudita: prediabetes, diabetes de tipo 1 y diabetes de tipo 2, los resultados empíricos demostraron resultados prometedores del novedoso modelo de apilamiento que combinó Bagging KNN, Baggging DT y K-NN, con un metaclasificador K-N se logro una precisión de 94.48% en los diagnóstico (GOLLAPALLI e colab., 2022), En un estudio de análisis de diabetes mellitus para la predicción temprana a través de la selección de características óptimas, se encontró que los algoritmos de árbol de decisión y bosque aleatorio tuvieron la mayor especificidad, con un 98,20% y 98,00%,

respectivamente, lo que los convierte en excelentes opciones para analizar datos diabéticos. Además, el algoritmo Naïve Bayesian tuvo la mejor precisión, con un 82,30%, la investigación también sugiere la generalización de la selección de características óptimas del conjunto de datos para mejorar la precisión de la clasificación en el futuro (SNEHA e GANGIL, 2019), en el estudio para un modelo de predicción de riesgo de diabetes mellitus gestacional en una población china basado en un sistema de puntuación de riesgo en donde se uso regresión logística para obtener los coeficientes de los predictores con una puntuación de riesgo total arrojó un área bajo la curva (AUC) de 0,845 (IC del 95 % = 0,805–0,884) en su intervalo también sugirió investigar cómo las estrategias de prevención e intervención tempranas pueden impactar en pacientes con detección temprana positiva de diabetes Mellitus gestacional utilizando la puntuación de riesgo (WANG e colab., 2021), para la utilización de computación en la niebla y técnicas explicables de aprendizaje profundo para la predicción de la diabetes gestacional de predicción y reemplazo de datos propuesto (DRPF) tienen 3 capas, (i) IoT, (ii) niebla y (iii) nube, el modelo propuesto logra resultados precisos y prometedores (EL-RASHIDY e colab., 2022), y por último en la investigación de dominios de competencia de los clasificadores de redes bayesianos semi-ingenuos afirma que hacer investigaciones con dominios discretos tiene cierta facilidad que con datos continuos (FLORES e colab., 2014)

Análisis exploratorio de datos

Se examinaron las características del conjunto de datos y se eliminaron los valores faltantes, se calcularon estadísticas descriptivas y se evaluó la distribución de las variables en el conjunto de datos proviene originalmente del Instituto Nacional de Diabetes y Enfermedades Digestivas y Renales, el objetivo es predecir en base a medidas diagnósticas si una paciente femenina tiene diabetes con los siguientes atributos número de preembarazos, glucosa (mg/Dl), presión sanguínea (mmHg), espesor de piel (mm), insulina (muU/mL), índice de masa corporal, pedigrre (probabilidad de que tenga hereditariamente), edad y si tiene diabetes positiva y negativa en el siguiente enlace https://www.niddk.nih.gov/health-information/informacion-de-la-salud.

El conjunto de datos se ha tomado de 768 de los cuales 500 casos dieron negativo y 268 dieron positivos los cuales se describen en la siguiente tabla con variables cuantitativas

Preprocesamiento de datos

Después del tratamiento el conjunto de datos recopilados se preparó en un formato de block de notas con un formato de valores separado por comas, la base de datos tiene 768 diagnósticos de pacientes con 8 atributos de los signos comunes en mujeres como los preembarazos, la glucosa, la presión sanguínea ,espesor de piel la insulina, el índice

O cuidado em saúde baseado em evidências - ISBN 978-65-5360-315-8 - Vol. 1 - Ano 2023 - Editora Científica Digital - www.editoracientifica.com.br

de masa corporal y el pedigree con un diagnóstico positivo y negativo a la diabetes que se muestra en la tabla 1 mas detalladamente, para la construcción del modelo probabilístico basado en redes bayesianas

Se hizo el tratamiento de datos categorizando las variables según los siguientes parámetros, se estandarizaron los datos

Tabla 1. Descripción estadística después del tratamiento de datos de diagnóstico de Diabetes Mellitus.

		Clase			
Atributos		**Negativo**		**Positivo**	
Variables	**Categoría**	**Frecuencia**	**(%)**	**Frecuencia**	**(%)**
PreEmbarazos	0	73	15%	38	14%
	1	106	21%	29	11%
	2	84	17%	19	7%
	3	48	10%	27	10%
	4	45	9%	23	9%
	5	36	7%	21	8%
	6	34	7%	16	6%
	7	20	4%	25	9%
	8	16	3%	22	8%
	9	10	2%	18	7%
	10	14	3%	10	4%
	11	4	1%	7	3%
	12	5	1%	4	1%
	13	5	1%	5	2%
	14	0	0%	2	1%
	15	0	0%	1	0%
	16	0	0%	1	0%
Edad	21 a 26	247	49%	53	20%
	27 a 32	1	0%	0	0%
	33 a 38	98	20%	59	22%
	39 a 44	47	9%	45	17%
	45 a 50	41	8%	45	17%
	51 a 56	24	5%	28	10%
	57 a 63	11	2%	23	9%
	64 a 69	15	3%	11	4%
	70 a 75	13	3%	3	1%
	76 a 81	3	1%	1	0%
Espesor de piel	Delgado	192	38%	94	35%
	Normal	117	23%	102	38%
	Grueso	191	38%	72	27%
Índice de masa corporal	Bajo peso	13	3%	2	1%
	Normal	94	19%	7	3%
	Sobrepeso	123	25%	100	37%
	Obesidad 1	89	18%	64	24%
	Obesidad 2	44	9%	57	21%
	Obesidad 3	137	27%	38	14%

Atributos		Clase			
		Negativo		Positivo	
Variables	Categoría	Frecuencia	(%)	Frecuencia	(%)
Glucosa	Normal	58	12%	121	45%
	Alta	4	1%	14	5%
	Muy Alta	438	88%	133	50%
Presión sanguínea	Normal	84	17%	61	23%
	Elevada	19	4%	17	6%
	Hipertensión I	5	1%	3	1%
	Hipertensión II	385	77%	178	66%
	Otro	7	1%	9	3%
Pedigree	Probabilidad baja	88	18%	85	32%
	Probabilidad media	246	49%	91	34%
	Probabilidad alta	166	33%	92	34%
Insulina	Normal	368	74%	157	59%
	Resistencia	132	26%	111	41%

Fuente: Elaboración propia.

La tabla muestra la distribución de las variables con sus categorías de un conjunto de datos relacionados con la diabetes en mujeres de edad mayor o igual a 21 años. La tabla presenta información sobre la frecuencia y el porcentaje de cada clase para cada variable.

En los preembarazos" se refiere al número de embarazos previos que ha tenido la mujer, la mayoría de las mujeres tienen entre 0 y 2 embarazos previos con un porcentaje mayor de 21% que equivale a un preembarazos, la variable "Edad" se refiere a la edad de la mujer, la mayoría de las mujeres se encuentran en el grupo de edad de 21 a 26 años con un 49 %, seguido del grupo de edad de 33 a 38 años con un 20%, el espesor de piel se refiere al grosor de la piel en el tríceps, la mayoría de las mujeres tienen piel gruesa y delgada con un porcentaje equivalente al 38%, el índice de masa corporal (IMC) de la mujer en donde la mayoría de las mujeres tienen obesidad 3 con un 27% seguidamente de sobrepeso con 25%, en cuanto a la glucosa en la sangre, la mayoría de las mujeres tienen niveles muy altos de glucosa en la sangre, la presión sanguínea que se refiere a la presión arterial sistólica en mm Hg, la mayoría de las mujeres tienen hipertensión II con el 77%, el pedigree que se refiere a una medida de la historia familiar de diabetes en parientes cercanos, la mayoría de las mujeres tienen una probabilidad media de desarrollar diabetes con un 49% según su historial familiar, la insulina que se refiere al nivel de insulina en el suero en muU/ml, la mayoría de las mujeres tienen niveles normales de insulina en el suero con un 74% todo ello dentro de las casos negativos de diabetes

En cuanto a los casos positivos de diabetes, los preembarazos muestra la frecuencia y porcentaje de mujeres con un determinado número de embarazos previos el 14% de las mujeres no tuvieron embarazos previos y el 11% tuvo un embarazo previo, la mayoría de las mujeres (un 9% cada una) tuvieron 4 o 7 embarazos previos, la edad muestra la frecuencia

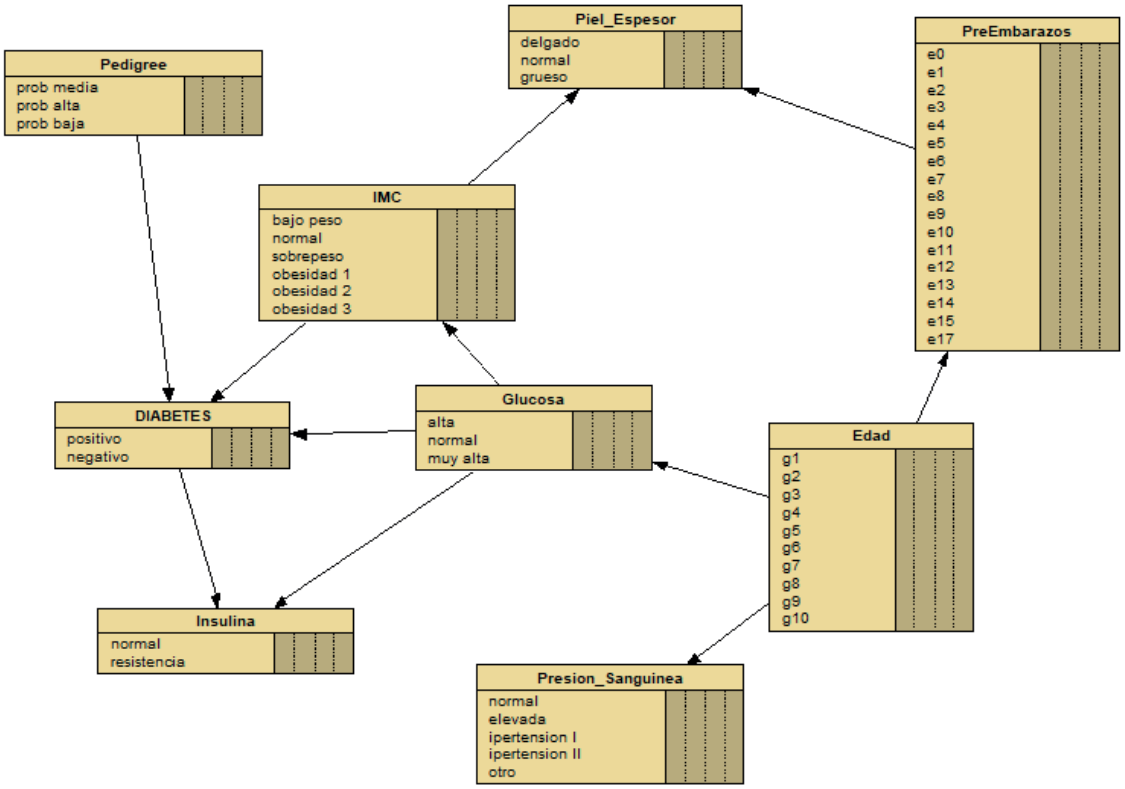
y porcentaje de mujeres con diferentes grupos de edad el 20% de las mujeres estan en el grupo de edad de 21 a 26 años y el 22% estan en el grupo de edad de 33 a 38 años, el espesor de piel muestra la frecuencia y porcentaje de mujeres con diferentes espesores de piel, el 35% de las mujeres tenían la piel delgada, mientras que el 27% tenían la piel gruesa, el indice de masa corporal muestra la frecuencia y porcentaje de mujeres con diferentes categorías de IMC el 37% de las mujeres tenían sobrepeso y el 24% tenían obesidad 1, la glucosa" muestra la frecuencia y porcentaje de mujeres con diferentes niveles de glucosa en sangre, el 50% de las mujeres tenían niveles muy altos de glucosa en sangre, la presión sanguínea muestra la frecuencia y porcentaje de mujeres con diferentes niveles de presión arterial el 66% de las mujeres tenían hipertensión II, el pedigree muestra la frecuencia y porcentaje de mujeres con diferentes niveles de probabilidad de diabetes en función de la historia familiares 34% de las mujeres tenían una probabilidad alta de diabetes finalmente la insulina muestra la frecuencia y porcentaje de mujeres con diferentes niveles de insulina en sangre ya que el 59% de las mujeres tenían niveles normales de insulina.

Construcción del modelo probabilístico bayesiano

El modelo probabilístico basado en redes bayesianas fue construido utilizando el software Netica debido a su gran capacidad para el análisis probabilístico bayesiano y su interfaz gráfica amigable que facilita la representación gráfica de las redes. Además, cuenta con algoritmos potentes que permiten el desarrollo de una topología óptima de la red bayesiana. Para garantizar la precisión y relevancia de la topología de la red bayesiana, se contó con la colaboración de expertos profesionales de la salud en la materia.

Se utilizo el algoritmo Hill-Climbing para ver la estructura y submuestreo y mediante el algoritmo Bostraap se usó para la definición de nodos se obtuvo la siguiente topología de la red bayesiana y se ajustaron los parámetros de la red para maximizar su capacidad de predicción.

Figura 2. Topología de la red bayesiana para la predicción de Diabetes Mellitus.



Fuente: Elaboración propia.

La figura 2 nos presenta la topología de la red bayesiana que se utiliza para predecir la diabetes mellitus. Se observa que la glucosa, el índice de masa corporal (IMC) y el pedigrí tienen una relación directa con el diagnóstico de diabetes. Con respecto al IMC, se puede ver que tiene una relación directa con el espesor de piel. Además, el índice de masa corporal y la insulina dependen de los niveles de glucosa. Por otro lado, los preembarazos, la presión sanguínea y la glucosa se ven influenciados por la edad. Por último, se puede destacar que los preembarazos también se relacionan con el espesor de piel. En conclusión, la figura 2 nos muestra la complejidad de la relación entre diversos factores que pueden influir en la predicción de la diabetes mellitus.

Validación del modelo

Se evaluó el rendimiento del modelo utilizando métricas de evaluación como la sensibilidad, especificidad, valor predictivo positivo, valor predictivo negativo, tasa de error, área bajo la curva ROC y coeficiente de Gini. se utilizó una técnica de validación cruzada para evaluar su rendimiento, el conjunto de datos se utilizó para construir el modelo de redes bayesianas, mientras que el conjunto de prueba se utilizó para evaluar su rendimiento, se realizó un proceso de optimización para mejorar la sensibilidad del modelo, ya que se encontró que tenía un valor bajo en este aspecto así como también se probaron diferentes

O cuidado em saúde baseado em evidências - ISBN 978-65-5360-315-8 - Vol. 1 - Ano 2023 - Editora Científica Digital - www.editoracientifica.com.br

umbrales de decisión para encontrar el valor que proporcionaba la mejor sensibilidad sin sacrificar demasiado la especificidad.

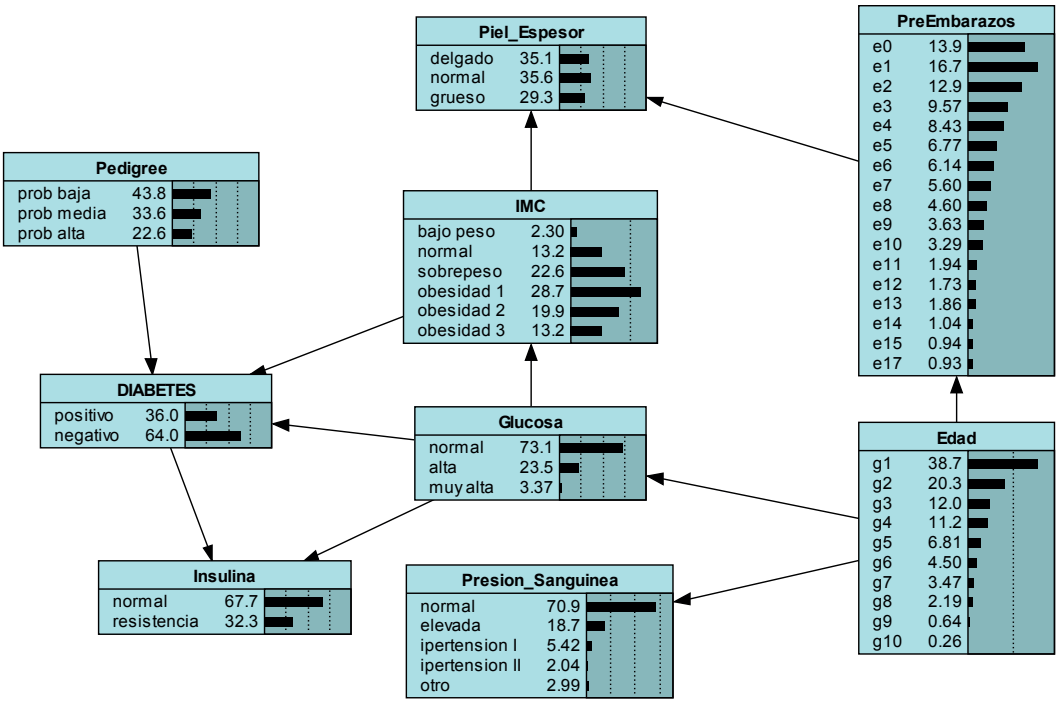
Optimización del modelo

Para optimizar el modelo, se ajustaron los parámetros de la red bayesiana utilizando el conjunto de entrenamiento y se evaluó su rendimiento utilizando el conjunto de prueba. Se realizó un proceso iterativo de ajuste de parámetros y evaluación del rendimiento hasta que se logró un equilibrio adecuado entre la sensibilidad y la especificidad. Además, se utilizó una técnica de validación cruzada para validar el modelo y asegurarse de que el rendimiento observado no se debiera al azar o al sobreajuste, así como también se evaluó el efecto de los cambios en las métricas de evaluación.

RESULTADOS

En la red bayesiana que se construyo con 9 nodos y 10 bordes dirigidos que muestra los factores de riesgo de la diabetes, las probabilidades previas de las variables se presentan en el siguiente grafico el modelo probabilístico resultante para analizar cauntitativamente la influencia de estos factores en la diabetes mediante el calculo de probabilidades condicionales P(y/xi) en la figura podríamos aprender que la probabilidad de diabetes negativo es de 0.64 si uno esta sujeto a un pedigree, un índice de masa corporal y la glucosa asi como se muestra en la figura

Figura 3. Red Bayesiana luego del aprendizaje para la predicción de diabetes.



Fuente: Elaboración propia.

O cuidado em saúde baseado em evidências - ISBN 978-65-5360-315-8 - Vol. 1 - Ano 2023 - Editora Científica Digital - www.editoracientifica.com.br

Tabla 2. Matriz de Confusión para la predicción de diabetes Mellitus.

Actual	Predicha	
	Positivo	Negativo
Positivo	160	108
Negativo	70	430

Fuente: Elaboración propia.

La tabla 2 es una matriz de confusión que se utiliza para evaluar el rendimiento de un modelo de predicción de diabetes mellitus. Los resultados se presentan en cuatro categorías: verdaderos positivos (VP), falsos positivos (FP), verdaderos negativos (VN) y falsos negativos (FN), los valores de 160 y 430 representan los VP y VN, respectivamente, lo que significa que se predijo correctamente que 160 personas si tienen diabetes mellitus y 430 personas no tienen diabetes mellitus. Los valores de 70 y 108 representan los FP y FN, respectivamente, lo que significa que 70 personas fueron diagnosticadas erróneamente con diabetes mellitus y 108 personas con diabetes mellitus no fueron diagnosticadas.

En términos de métricas de evaluación, la precisión del modelo sería de aproximadamente 80%, ya que la mayoría de las predicciones fueron correctas. Sin embargo, la sensibilidad del modelo es del 60%, lo que significa que solo se identificó el 60% de las personas con diabetes mellitus, mientras que el 40% no fueron identificadas. La especificidad del modelo es del 86%, lo que significa que se identificó correctamente al 86% de las personas sin diabetes mellitus.

En resumen, la tabla 2 muestra que el modelo de predicción tiene una precisión aceptable, pero necesita mejorar su sensibilidad para detectar a todas las personas con diabetes mellitus, un valor predictivo positivo del 69,57%, un valor predictivo negativo del 79,93% y una tasa de error del 23,18%.

- Porcentaje de error = 23,18
- Pérdida logarítmica = 0,4904
- Pérdida cuadrática = 0,3222
- Resultado esférico = 0,8196

El porcentaje de error de 23,18 indica que el modelo de redes bayesianas para la predicción de diabetes tiene un error del 23,18% en la clasificación de los casos, los resultados de la regla de puntuación proporcionan información adicional sobre el desempeño del modelo. Una pérdida logarítmica de 0,4904 y una pérdida cuadrática de 0,3222 indican que el modelo tiene un buen rendimiento en la clasificación de los casos, ya que estas medidas de pérdida son bajas, el resultado esférico de 0,8196 indica que el modelo tiene una buena

capacidad para discriminar entre las diferentes clases de la variable dependiente y, por lo tanto, puede ser considerado como un modelo útil para la predicción de la diabetes mellitus.

Tabla 3. Métricas de evaluación para validar el modelo basado en una red bayesiana.

Umbral de Decisión	Sensibilidad	Especificidad	Predictivo	Predictivo Negativo	Tasa de error
0	100.00	0.00	34.90	100.00	65.10
5	99.63	14.00	38.31	198.59	56.12
15	95.90	28.60	41.86	92.86	47.92
20	91.42	41.40	45.54	90.00	41.15
25	83.96	58.60	52.08	87.20	32.55
30	75.00	70.20	57.43	83.97	28.13
40	65.67	83.20	67.69	81.89	22.92
50	59.70	86.00	69.57	79.93	23.18
60	44.03	93.40	78.15	75.69	23.83
70	32.84	96.60	83.81	72.85	25.65
75	23.51	98.20	87.50	70.55	27.86
80	12.69	99.20	89.47	67.95	30.99
85	6.72	99.60	90.00	66.58	32.81
90	0.00	100.00	100.00	65.10	34.90
100	0.00	100.00	100.00	65.10	34.90

La tabla muestra una evaluación del modelo basado en una red bayesiana utilizando diferentes umbrales de decisión para la predicción de diabetes. Cada fila representa un umbral de decisión diferente, desde 0 hasta 100.

La sensibilidad indica la proporción de verdaderos positivos detectados por el modelo en relación con el total de positivos reales. A medida que aumenta el umbral de decisión, la sensibilidad disminuye, lo que significa que el modelo identifica menos verdaderos positivos.

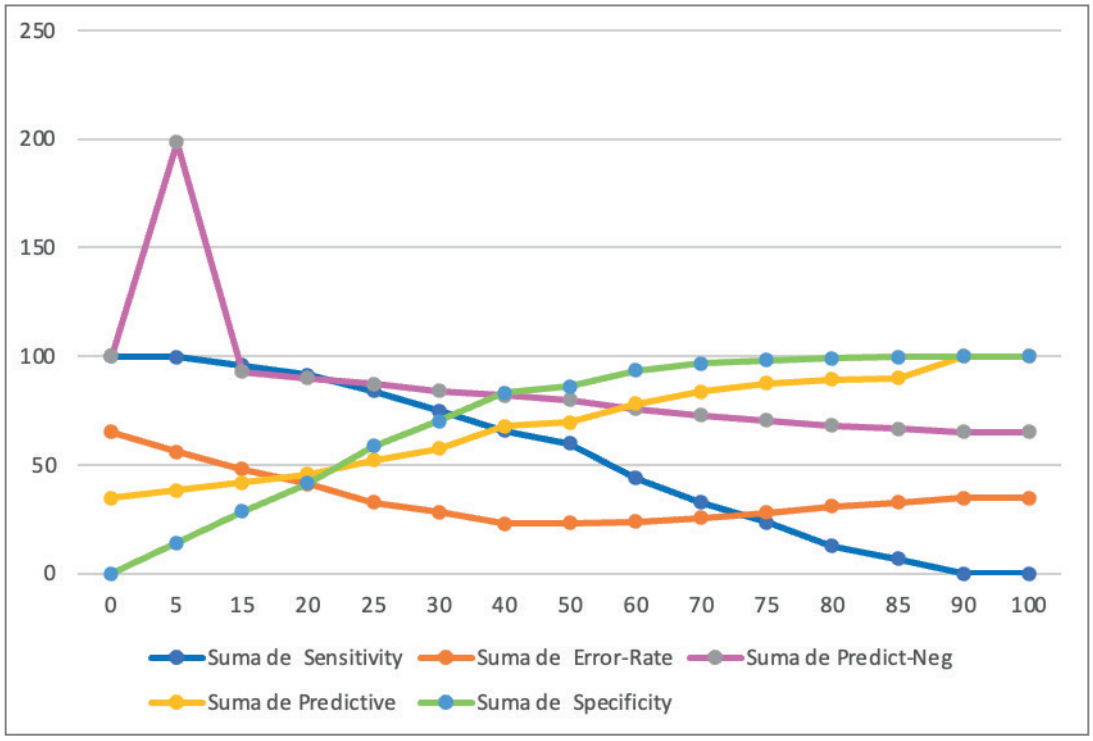
La especificidad indica la proporción de verdaderos negativos detectados por el modelo en relación con el total de negativos reales. A medida que aumenta el umbral de decisión, la especificidad aumenta, lo que significa que el modelo identifica menos falsos positivos.

El valor predictivo positivo indica la proporción de verdaderos positivos en relación con el total de predicciones positivas del modelo. A medida que aumenta el umbral de decisión, el valor predictivo positivo aumenta, lo que significa que el modelo tiene una mayor precisión en la identificación de casos positivos.

El valor predictivo negativo indica la proporción de verdaderos negativos en relación con el total de predicciones negativas del modelo. A medida que aumenta el umbral de decisión, el valor predictivo negativo disminuye, lo que significa que el modelo tiene una menor precisión en la identificación de casos negativos.

La tasa de error indica la proporción de predicciones incorrectas en relación con el total de predicciones del modelo. A medida que aumenta el umbral de decisión, la tasa de error disminuye, lo que significa que el modelo tiene una mayor precisión general en la predicción.

Figura 4: Métricas de evaluación según el Umbral de decisión.



Coeficiente de Gini = 0,624
Área bajo ROC = 0,812

El modelo de red bayesiana para la predicción de diabetes tiene un coeficiente de Gini de 0,624, lo que sugiere que el modelo tiene una capacidad predictiva buena moderada con un AUC-ROC de 0,812, lo que sugiere que el modelo tiene una capacidad predictiva moderadamente buena para distinguir entre verdaderos positivos y falsos positivos.

DISCUSIONES

En esta investigación se ha utilizado un modelo probabilístico basado en redes bayesianas para predecir la diabetes mellitus en pacientes femeninas, se ha logrado obtener una precisión aceptable en la predicción de la enfermedad, con un valor predictivo positivo del 69,57%, un valor predictivo negativo del 79,93% y una tasa de error del 23,18%, sin embargo se ha identificado que el modelo necesita mejorar su sensibilidad para detectar a todas las personas con diabetes mellitus. La optimización del modelo se llevó a cabo mediante la selección de variables relevantes y la eliminación de variables redundantes. Además, se ajustaron los parámetros de la red bayesiana y se utilizó una técnica de eliminación de nodos para reducir el tamaño de la red y mejorar la eficiencia computacional ya que (SONG e colab., 2023) promueve el funcionamiento de redes bayesianas en la práctica clínica, es importante

destacar que el modelo se basa en una serie de supuestos y simplificaciones en donde las variables son independientes entre sí y que siguen una distribución normal estos supuestos pueden no ser del todo precisos en la realidad, lo que podría afectar la precisión de las predicciones.. Por otro lado (P e colab., 2020) en el estudio de redes neuronales profundas con un enfoque de aprendizaje no supervisado para una predicción en el conjunto de datos de diabetes de los indios pima logró una precisión del 98,16% en comparación de nuestro modelo , por otra parte (JACKINS e colab., 2021) en la red de clasificación bayesiana se calculó y comparo un análisis de rendimiento de los datos de otra enfermedad para ambos algoritmos muestra una precisión de 74.46 en cuanto a la clasificación con el modelo de bosque aleatorio muestra una precisión de 74.03 en la investigación, que nos explica que las redes bayesianas funcionan mucho mejor, (BIRJAIS e colab., 2019) recomendó que se pueden aplicar enfoques similares como la regresión logística , el algoritmo ingenuo bayesiano y Gradient Boosting en otros conjuntos de datos de enfermedades con fines de predicción, en cuanto a (GOLDENBERG e colab., 2021) en su revisión sistemática sobre modelos predictivos de aprendizaje automático y aprendizaje profundo para la diabetes tipo 2, se encontró que las redes son óptimas debido a su capacidad para manejar grandes cantidades de datos y mejorar la eficiencia del modelo, en cuanto a las implicaciones clínicas (SNEHA e GANGIL, 2019), la detección temprana de la diabetes mellitus en pacientes femeninas es crucial para prevenir complicaciones graves y mejorar la calidad de vida de los pacientes asi como lo menciono (RASTOGI e BANSAL, 2023). Por lo que afirmamos que la utilización de un modelo probabilístico basado en redes bayesianas puede ser útil para los profesionales de la salud en la toma de decisiones clínicas, aunque se requiere una validación adicional en estudios futuros para determinar su utilidad clínica.

CONCLUSIONES

En conclusión, la construcción de un modelo probabilístico basado en redes bayesianas para la predicción de diabetes utilizando el conjunto de datos del Instituto Nacional de Diabetes y Enfermedades Digestivas y Renales ha demostrado ser una herramienta útil para la identificación temprana de la diabetes mellitus. Si bien el modelo tiene una precisión aceptable y un buen rendimiento en la clasificación de casos, se necesita mejorar su sensibilidad para detectar a todas las personas con diabetes. Además, se encontró que el modelo tiene una tasa de error del 23,18%, lo que indica que el modelo tiene un error del 23,18% en la clasificación de casos. Por otro lado, los resultados de la regla de puntuación proporcionan información adicional sobre el desempeño del modelo. La pérdida logarítmica de 0,4904 y la pérdida cuadrática de 0,3222 indican que el modelo tiene un buen rendimiento en la clasificación de los casos, ya que estas medidas de pérdida son bajas. Además, el resultado

esférico de 0,8196 indica que el modelo tiene una buena capacidad para discriminar entre las diferentes clases de la variable dependiente por tanto el modelo probabilístico basado en redes bayesianas es una herramienta prometedora para la predicción temprana de la diabetes mellitus, aunque se deben realizar mejoras en su sensibilidad para detectar a todas las personas con diabetes. La precisión y el rendimiento del modelo se pueden mejorar mediante la incorporación de más datos y la optimización de los parámetros del modelo. En última instancia, esto podría mejorar la capacidad del modelo para identificar a los pacientes en riesgo de desarrollar diabetes y proporcionar un mejor tratamiento y atención temprana para prevenir complicaciones graves en el futuro.

Agradecimientos

Queremos expresar nuestro más sincero agradecimiento a todas las personas que hicieron posible la realización de este proyecto, en primer lugar, agradecemos al Instituto Nacional de Diabetes y Enfermedades Digestivas y Renales por proporcionarnos los datos necesarios para llevar a cabo esta investigación asi como también queremos agradecer a los profesionales de la salud que colaboraron en la construcción de la red bayesiana y en la interpretación de los resultados obtenidos, además, queremos agradecer a nuestro equipo de investigación por su dedicación y trabajo arduo para llevar a cabo este proyecto. Cada miembro aportó valiosas habilidades y conocimientos que permitieron alcanzar los objetivos establecidos, por último, queremos expresar nuestro agradecimiento a todas las personas que de alguna manera contribuyeron a este proyecto, incluyendo amigos, familiares y colegas. Sin su apoyo y ánimo, este proyecto no hubiera sido posible.

REFERENCIAS

BIRJAIS, Roshan e colab. **Prediction and diagnosis of future diabetes risk: a machine learning approach**. SN Applied Sciences, v. 1, n. 9, 1 Set 2019.

CHANG, Victor e colab. **Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms**. Neural Computing and Applications, 2022.

DUONG, Khanh N.C. e colab. **Comparison of diagnostic accuracy for diabetes diagnosis: A systematic review and network meta-analysis**. Frontiers in Medicine. [S.l.]: Frontiers Media S.A. , 24 Jan 2023

EL-RASHIDY, Nora e colab. **Utilizing fog computing and explainable deep learning techniques for gestational diabetes prediction**. Neural Computing and Applications, 1 Abr 2022.

FLORES, M. Julia e GÁMEZ, José A. e MARTÍNEZ, Ana M. **Domains of competence of the semi-naive Bayesian network classifiers**. Information Sciences, v. 260, p. 120–148, 1 Mar 2014.

FREGOSO-APARICIO, Luis e colab. **Machine learning and deep learning predictive models for type 2 diabetes: a systematic review**. Diabetology and Metabolic Syndrome. [S.l.]: BioMed Central Ltd. , 1 Dez 2021

FRIEDMAN, Nir e colab. **Clasificadores de redes bayesianas ***. Aprendizaje automático. [S.l: s.n.], 1997. Disponível em: <www.onlinedoctranslator.com>.

GOLDENBERG, Joshua Z. e colab. **Efficacy and safety of low and very low carbohydrate diets for type 2 diabetes remission: systematic review and meta-analysis of published and unpublished randomized trial data**. BMJ, v. 372, 2021.

GOLLAPALLI, Mohammed e colab. **A novel stacking ensemble for detecting three types of diabetes mellitus using a Saudi Arabian dataset: Pre-diabetes, T1DM, and T2DM**. Computers in Biology and Medicine, v. 147, 1 Ago 2022.

JACKINS, V. e colab. **AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes**. Journal of Supercomputing, v. 77, n. 5, p. 5198–5219, 1 Maio 2021.

KANGRA, Kirti e SINGH, Jaswinder. **Comparative analysis of predictive machine learning algorithms for diabetes mellitus**. Bulletin of Electrical Engineering and Informatics, v. 12, n. 3, p. 1728–1737, 1 Jun 2023.

MAZHAR, Khurum e colab. **Bayesian networks identify determinants of outcomes following cardiac surgery in a UK population**. BMC Cardiovascular Disorders, v. 23, n. 1, 1 Dez 2023.

P, Bala Manoj Kumar e colab. **Type 2: Diabetes mellitus prediction using Deep Neural Networks classifier**. International Journal of Cognitive Computing in Engineering, v. 1, p. 55–61, 1 Jun 2020.

RASTOGI, Rashi e BANSAL, Mamta. **Diabetes prediction model using data mining techniques**. Measurement: Sensors, v. 25, 1 Fev 2023.

RAWAT, Vandana e colab. **Machine learning algorithms for early diagnosis of diabetes mellitus: A comparative study**. Materials Today: Proceedings, v. 56, p. 502–506, 1 Jan 2022.

SNEHA, N. e GANGIL, Tarun. **Analysis of diabetes mellitus for early prediction using optimal features selection**. Journal of Big Data, v. 6, n. 1, 1 Dez 2019.

SONG, Wenzhu e colab. **Using Bayesian networks with Tabu-search algorithm to explore risk factors for hyperhomocysteinemia**. Scientific Reports, v. 13, n. 1, 1 Dez 2023.

THOTAD, Puneeth N. e BHARAMAGOUDAR, Geeta R. e ANAMI, Basavaraj S. **Diabetes disease detection and classification on Indian demographic and health survey data using machine learning methods**. Diabetes and Metabolic Syndrome: Clinical Research and Reviews, v. 17, n. 1, 1 Jan 2023.

TIAN, Tian e colab. **A Bayesian network model for prediction of low or failed fertilization in assisted reproductive technology based on a large clinical real-world data**. Reproductive Biology and Endocrinology, v. 21, n. 1, 1 Dez 2023.