Algoritmos inteligentes para el diagnóstico de Diabetes Mellitus tipo 2

Intelligent algorithms for the diagnosis of Diabetes Mellitus type 2

Isaias Brian Trujillo Rivera, Frank Edmundo Escobedo Bailón

Universidad Nacional Mayor de San Marcos. Lima, Perú

Universidad Nacional Tecnológica de Lima Sur. Lima, Perú

Recibido: 11/02/2023 Revisado: 15/04/2023 Aceptado: 15/05/2023 Publicado: 31/07/2023

Resumen

En los últimos años, a nivel mundial la Diabetes Mellitus tipo 2 es un problema cada vez más común en las personas. En general, el diagnóstico de esta enfermedad es realizado por expertos; pese a ello, algunos de los resultados podrían no ser correctos; además, involucra una inversión del tiempo y dinero por parte del paciente. De este modo, el presente trabajo contribuye a determinar qué algoritmo inteligente es más eficaz para el diagnóstico de Diabetes Mellitus tipo 2 con el objetivo de orientar a futuras investigaciones en el desarrollo de herramientas que faciliten el pronóstico de esta enfermedad en una fase temprana de forma eficaz. Para ello, el presente trabajo emplea el conjunto de datos PIMA en la metodología para la predicción de la diabetes a través de diversos algoritmos de aprendizaje supervisado considerando factores como las ventajas que ofrecen, conjunto de datos, preprocesamiento de datos y la precisión de estos. De lo cual, se concluyó que no hay un algoritmo definitivo que ofrezca los mejores resultados en cualquier escenario; por el contrario, el algoritmo adecuado es el que mejor responde a los factores descritos anteriormente.

Palabras claves: Algoritmos inteligentes, diagnóstico, Diabetes.

Abstract

In recent years, Diabetes Mellitus type 2 has become an increasingly common problem in people worldwide. In general, the diagnosis of this disease is performed by experts; however, some of the results may not be correct; moreover, it involves an investment of time and money by the patient. Thus, the present work contributes to determining which intelligent algorithm is more effective for the diagnosis of Diabetes Mellitus type 2 with the aim of guiding future research in the development of tools that facilitate the prognosis of this disease at an early stage in an effective way. To this end, the present work employs the PIMA dataset in the methodology for the prediction of diabetes through various supervised learning algorithms considering factors such as the advantages they offer, dataset, data preprocessing and the accuracy of these. From which, it was concluded that there is no definitive algorithm that offers the best results in any scenario; on the contrary, the appropriate algorithm is the one that best responds to the factors described above.

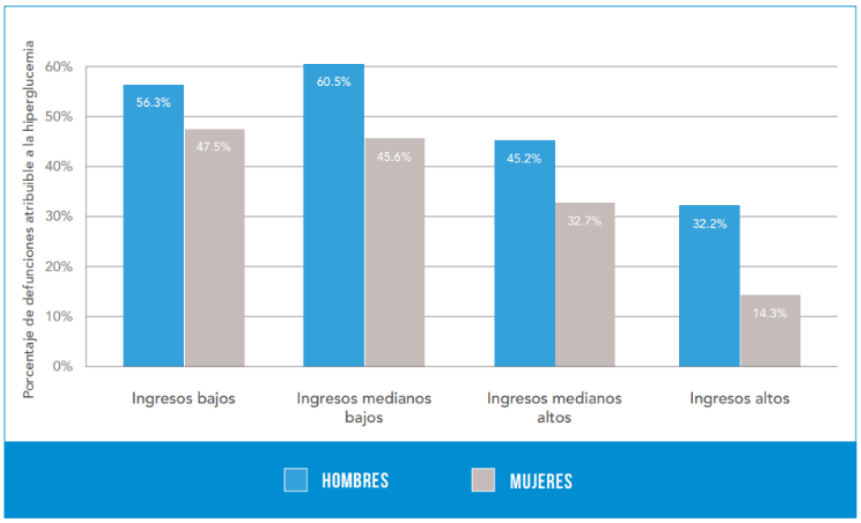
Keywords: Intelligent algorithms, diagnosis, Diabetes.

Introducción

Según la OMS (2016), la diabetes es una enfermedad que se ha duplicado entre 1980, con 108 millones de personas, y 2014 con 422 millones de afectados; debido a, el impulso de los índices de sobrepeso y obesidad. La diabetes se define como una enfermedad que se presenta cuando el organismo no produce suficiente insulina; la cual, regula el azúcar en la sangre. Hay 2 tipos, estas comparten síntomas similares como la poliuria, polidipsia, hambre constante y pérdida de peso. Pero, la diferencia fundamental es que la del tipo 2 puede ser diagnosticada en años posteriores cuando ya han surgido complicaciones; por lo que, la convierte en una enfermedad que opera de forma silenciosa. Según diversos autores (Ismail et al., 2021; Kyrou et al., 2020; Yuan & Larsson, 2020), las principales causas o factores involucrados pueden ser de diferente naturalezas como el estilo de vida, sociales, genéticos, etc.. Uno de ellos es la hiperglucemia, la cual, se define como la distribución de la glucemia por encima de los valores de concentración en sangre, pero, no lo suficiente para considerarse como diabetes, por lo que, dificulta el diagnóstico de esta. Para conocer mejor el impacto que tiene sobre la población mundial y como esta guarda relación con la edad, se presenta la siguiente figura donde se observa que los países con una mayor proporción de defunciones corresponden a aquellos que presentan ingresos bajos y medianos.

Figura 1

Porcentaje de defunciones atribuibles a la hiperglucemia en personas de 20 a 69 años, por sexo y categoría de ingreso de los países, 2012.



Por otro lado, Chang et al. (2022) señalan que la capacidad de predecir el riesgo y susceptibilidad de un individuo a la diabetes u otras enfermedades crónicas de forma temprana permite reducir los costes médicos y el riesgo de presentar complicaciones. Sin embargo, en la mayoría de los casos, como la diabetes tipo 2, pueden pasar desapercibido por la presencia de pocos síntomas o el grado leve en que se manifiestan. Para enfrentar estos problemas se han ido desarrollando soluciones, una de ellas es el empleo de los algoritmos inteligentes; debido a que, estos optimizan el proceso de diagnóstico y ofrecen indicios sobre acciones claves de acuerdo con el resultado del diagnóstico. Por consiguiente, se demuestra que, una tarea clave para enfrentar las enfermedades crónicas como la diabetes de tipo 2 es el diagnóstico durante una fase temprana; debido a que, esta tiene repercusiones en la economía y salud de las personas.

Materiales y métodos

Antecedentes

Ordóñez & Vizcarra (2018) llevaron a cabo una investigación que aborda el pronóstico de la diabetes de tipo 2 es un reto tanto en el contexto mundial como en el caso de países como Perú. Por ello, consideraron el empleo de un modelo predictivo a partir del análisis sintomático de los pacientes; a fin de brindar a entidades prestadoras de servicios de la salud información, resultado del modelo predictivo, que perfile el nivel de riesgo de los clientes.

Ismail et al. (2022) realizaron una investigación donde abordan el problema de la predicción de la diabetes de tipo 2; la cual, es clave para proporcionar un diagnóstico que ayude al personal de salud en el desarrollo de un plan de prevención. Por ello, consideraron desarrollar una alternativa donde se emplean diferentes conjuntos de datos y métricas como los factores de riesgo asociados a la enfermedad; a fin de, construir un modelo predictivo que sea preciso y eficaz considerando el preprocesamiento de datos y la atención en diferenciar la precisión del modelo con la medida de prueba del grado de precisión.

Material

Conjunto de datos de la salud

Para el presente estudio se identificó que se emplearon el conjunto de datos indios PIMA (PIDD); la cual, fue producida por el Instituto Nacional de diabetes y enfermedades digestivas y renales para el modelamiento a través de algoritmos inteligentes. Este conjunto de datos comprende 9 atributos; 8 empleados para la predicción y 1 como etiqueta de clase; es decir, expresa con un valor binario la presencia (1) o ausencia (0) de diabetes.

Figura. 2

Descripción de los atributos del conjunto de datos PIMA.

Sr. no.	Selected Attributes from PIMA Indian dataset	Description of selected attributes	Range
1.	Pregnancy	Number of times a participant is pregnant	0–17
2.	Glucose	Plasma glucose concentration a 2 h in an oral glucose tolerance test	0–199
3.	Diastolic Blood pressure	It consists of Diastolic blood pressure (when blood exerts into arteries between heart)(mm Hg)	0–122
4.	Skin Thickness	Triceps skinfold thickness (mm).It concluded by the collagen content	0–99
5.	Serum Insulin	2-Hour serum insulin (mu U/ml)	0–846
6.	BMI	Body mass index (weight in kg/(height in m)^2)	0–67.1
7.	Diabetes pedigree Function	An appealing attributed used in diabetes prognosis	0.078–2.42
8.	Age	Age of participants	21–81
9.	Outcome	Diabetes class variable, Yes represent the patient is diabetic and no represent patient is not diabetic	Yes/No

Nota.

Adaptado de “Deep learning approach for diabetes prediction using PIMA Indian dataset” (p. 5), por H. Naz et al., 2020, Journal of Diabetes & Metabolic Disorders.

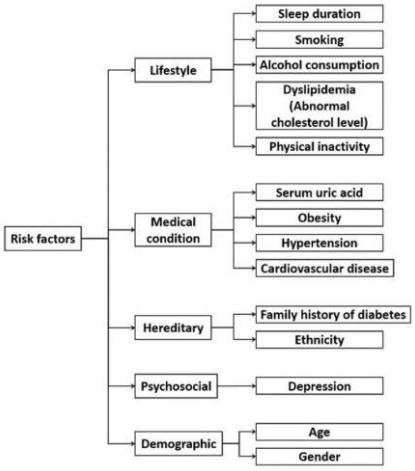
Como se observa en la figura 2, para cada uno de los atributos se presenta una descripción de los atributos seleccionados, así como el rango en el que fluctúa los valores para dichos atributos. De acuerdo con el objetivo del estudio y las herramientas que dispone, se puede considerar omitir o emplear atributos.

Factores de riesgo de la Diabetes tipo 2

Son diversos los factores que contribuyen al riesgo de la diabetes tipo 2. Por ello, Ismail et al. (2022) las presentaron bajo un esquema que las divide en 5 categorías; a fin de, analizar qué categoría de riesgo es significativa en la predicción. Según Altobelli et al. (2020), los factores de estilo de vida están influenciados por el entorno del individuo como consumo de cigarrillos, bebidas alcohólicas y baja actividad física. Los factores basados en la condición médica están relacionados a una dieta poco saludable y de acuerdo con Bellou et al. (2018) involucran la disminución de la actividad física, sedentarismo, entre otros que se traducen en una presión sistólica elevada, diabetes gestacional, síndrome metabólico, parto prematuro. Según Li et al. (2020), los factores de riesgo hereditarios podrían ser transmitidos de una generación a otra en el caso de presentar un gen o grupo de esto en particular. Los factores psicosociales están relacionados con enfermedades de salud mental y estos pueden repercutir en el control glucémico, conducto de autocuidado y calidad de vida. De acuerdo con Kalra et al. (2018), a menudo las necesidades emocionales y psicológicas se comprometen derivando en mayores posibilidades de presentar complicaciones con la diabetes. Los factores demográficos están relacionados a las características del individuo como diferencias étnicas y de género que hacen susceptibles a algunos grupos humanos frente a otros. Además, Pinchevsky et al. (2020) señalan que los niveles de educación, vida urbana, empleo y estado civil contribuyen a los resultados relacionados a la diabetes.

Figura 3

Taxonomía de los factores de riesgo para la diabetes tipo 2.



Nota. Adaptado de "Type 2 Diabetes with Artificial Intelligence Machine Learning: Methods and Evaluation" (p. 2), por L. Ismail et al., 2022, SpringerLink.

Método

Aprendizaje supervisado

Alanis (2018) indica que es un método de análisis de datos de *machine* learning donde un experto etiqueta o clasifica cada patrón dentro de un conjunto de entrenamiento con el fin de obtener el valor asociado a la entrada a partir de los ejemplos proporcionados. Este puede ser categorizado según la tarea que desempeña.

- Clasificación

Etiqueta las entradas en función de sus características.

- Regresión
  Indica que la variable de salida es continua y no categórica a diferencia de la clasificación.
- Recuperación

Aprendizaje no supervisado

Alanis (2018) señala que es un método de análisis de datos de *machine learning* donde éste aprende a partir de ajustar las observaciones. Este se caracteriza por formar agrupaciones o *clustering* para un conjunto determinado de patrones; es decir, no hay conjunto previo de aprendizaje.

Agrupamiento

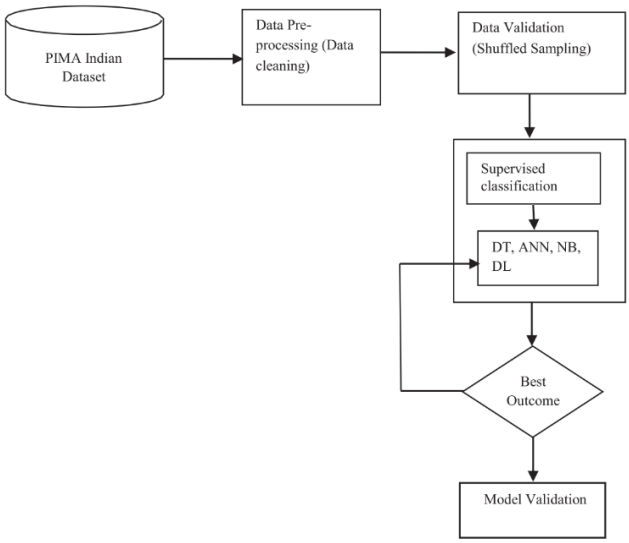
Según Alanis (2018), este proceso consiste en agrupar diferentes vectores o grupos de datos en función de un criterio; es decir, estos grupos comparten características similares; además, involucran el etiquetado de registros según el grupo al que pertenecen y sin disponer de conocimientos previos sobre estos.

Preprocesamiento de datos

Según Fan et al. (2021), conforma una base válida donde se emplean diversas técnicas para la mejora de la calidad de los datos brutos como eliminación de valores atípicos y la imputación de valores perdidos. De este modo, permite a algoritmos de aprendizaje supervisado el procesamiento con un buen rendimiento y representación de datos con distintos niveles de abstracción; además, para Alzubaidi et al. (2021) posibilita la entrada de diferentes tipos de datos como audio, voz, visuales e incluso el lenguaje natural. Para Chaki et al. (2022) la mejora de precisión se suele obtener si se considera como paso previo a este procedimiento y la validación del conjunto de datos resultante. También se pueden emplear diferentes algoritmos para potenciar esta labor.

Figura 4

Diagrama de flujo del modelado para la predicción de la diabetes.



Nota. Adaptado de "Type 2 Diabetes with Artificial Intelligence Machine Learning: Methods and Evaluation" (p. 7), por L. Ismail et al., 2022, SpringerLink.

Árboles de decisión (DT)

Alanis (2018) indica que es un algoritmo empleado en la clasificación de datos no paramétricos a través de secuencias de preguntas con el fin de obtener propiedades, cualidades o características. Esta secuencia se dispone en forma de árbol; en la parte superior, se sitúa el nodo raíz; en la parte inferior, los enlaces o ramas que derivan en nodos posteriores. Según Naz & Ahuja (2020),cada uno de los atributos se considera como un nodo de ramificación y construye una regla que divide los valores según las clases. Según Ismail et al. (2022), en cada uno de los nodos del árbol, el algoritmo selecciona la rama considerando el máximo de información ganada. Según Naz & Ahuja (2020), este algoritmo se presenta como un grafo donde se utiliza el análisis de decisiones y brinda como resultado unas reglas de división para cada atributo específico. En el caso del estudio de enfermedades crónicas como la diabetes tipo 2, se requiere de una analítica a través de modelos predictivos empleado el algoritmo J48. Ordóñez & Vizcarra (2018) indicaron que la modificación de este algoritmo permite un diagnóstico más rápido y eficiente al analizar patrones durante el análisis de clasificación al emplear algoritmos de árboles de decisiones y redes bayesianas.

K-vecinos más cercanos (KNN)

Según Alanis (2018), este algoritmo clasifica las nuevas instancias como la clase mayoritaria de entre los "k" vecinos más cercanos de entre el conjunto de entrenamiento. Además, Uddin et al. (2022) indican que predice la clasificación de datos no etiquetados considerando las características y etiquetas de los datos de entrenamiento. Según Wang (2019), la idea es calcular la distancia entre la muestra actual y la de entrenamiento, luego, encontrar los "k" vecinos más cercanos a fin determinar la categoría de la muestra actual en relación con la categoría de los vecinos. Es decir, el algoritmo es capaz de clasificar conjuntos de datos a través de "consultas" considerando los "k" puntos de datos de entrenamiento más cercanos (vecinos). Posteriormente, se realiza una regla de votación para comprobar que clasificación debe finalizar. Para Uddin et al. (2022), este algoritmo es conocido por su sencillez y diseño adaptable en las tareas de clasificación de conjuntos de datos médicos. Wang (2019) indica que, aunque es necesario considerar la escala de distancia y el conjunto de datos; de lo contrario, se obtendrán resultados distintos.

Redes neuronales artificiales (ANN)

Es un modelo de datos estadísticos no lineales que comparten una similitud con el comportamiento de las neuronas biológicas; debido a que, estas neuronas conforman redes interrelacionadas que se emplean para reconocimiento de patrones a través de *feedforward*. Abiodun et al. (2019) indican que se emplea durante todo el entrenamiento; el cual, inicia con relacionar entradas y salidas. Naz & Ahuja (2020) indican que, mientras se establece esta relación es posible que se revelen patrones, esta información será "comunicada" a lo largo de las capas ocultas, donde se procesan la información y esta a su vez se distribuye en su ciclo de retroalimentación que permite obtener un mejor resultado. Estos modelos de datos han presentado un crecimiento en cuanto a su popularidad debido a su eficacia, eficiencia y tasa de éxito para el reconocimiento de patrones en diferentes problemas. Todas estas cualidades son consecuencia del modelado fácil sobre tareas complejas a diferencia de las técnicas convencionales donde se encontraban con resultados no satisfactorios. Además, Ordóñez & Vizcarra (2018) señalan que este tipo de algoritmo puede ser empleado como un sistema de apoyo de diagnóstico médico a través de modelos predictivos donde se realizó previamente un preprocesamiento a través de distintos algoritmos y la extracción de características para

aumentar el grado de precisión. A continuación, se presentan diversos algoritmos que permitirán apreciar las principales ventajas y aplicaciones en la creación de modelos predictivos empleando el conjunto de datos PIMA para el diagnóstico de diabetes mellitus tipo 2.

Figura 5

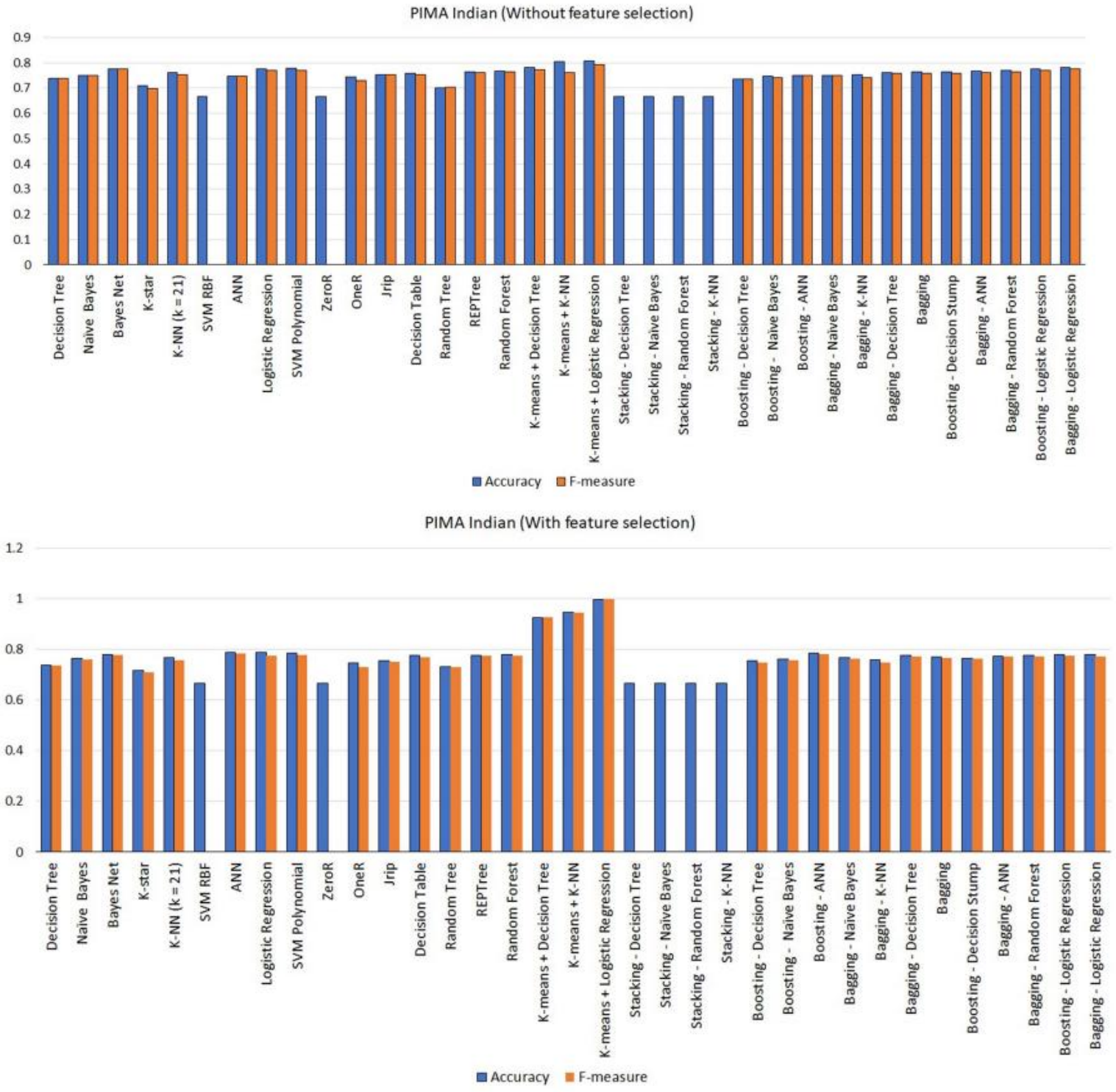
Resumen de las ventajas y desventajas de los algoritmos inteligentes de aprendizaje supervisado.

Algorithm	Advantages	Disadvantages
DT [46]	Suitable for datasets having missing values and data scaling and normalization is not required Easy to implement and interpret	Sensitive to change The probability of overfitting is high
BN [48]	Suitable for datasets having missing values	Not suitable with small datasets Computationally expensive
NB [46]	Suitable for datasets with missing values and is scalable Easy to implement	Suffers from the issue of zero frequency
K-NN [49]	Suitable for datasets having outliers Easy to implement	Determining the value of k is challenging High computation cost
K star [50]	Suitable for datasets having outliers Easy to implement	Not suitable for large datasets High computation cost
LR [51]	Suitable for large datasets Easy to interpret	Not suitable for linear data and datasets having a smaller number of observations than features
SVM [52]	Suitable for high dimensional and non-linear datasets	Feature scaling is required The output is difficult to interpret Selection of kernel is difficult
ANN [53]	Suitable for high dimensional datasets having a greater number of observations and can handle missing values	High computational cost Complex process
ZeroR [54]	Easy to understand Used as a baseline benchmark	No prediction involved
OneR [55]	Easy to understand Used as a baseline benchmark	Only suitable for datasets having categorical features Not suitable for linear data
JRip [56]	Suitable for non-linear data Easy to implement and interpret	Only suitable for datasets having categorical features Suffers from overfitting
DTable [57]	Suitable for dynamic datasets Easy to implement and interpret	Prone to overfitting Complex for high dimensional datasets
RT [58]	Suitable for datasets having missing values and data scaling and normalization is not required Easy to implement and interpret	Sensitive to changes in the dataset The probability of overfitting is high
RF [59]	Suitable for high dimensional datasets and can handle missing values	Difficult to implement Complex algorithm
REPTree [60]	Suitable for large datasets Easy to interpret compared to a decision tree	Sensitive to changes in the dataset Prone to overfitting
K-means [61]	Suitable for large datasets Simple to implement and interpret	Difficult to predict the value of k Sensitive to outliers
Bagging [62]	Suitable for high dimensional datasets and can handle missing values Reduces data overfitting	Model is biased Computationally expensive
Boosting [63]	Reduces data overfitting Easy to interpret	Not suitable for large datasets Sensitive to outliers
Stacking [64]	Reduces overfitting	Memory intensive

Nota. Adaptado de "Type 2 Diabetes with Artificial Intelligence Machine Learning: Methods and Evaluation" (p. 5), por L. Ismail et al., 2022, SpringerLink.

Figura 6

Precisión y F-measure de los algoritmos para el conjunto de datos PIMA en el pronóstico de la diabetes tipo 2 según la ausencia o presencia de la selección de características.



Nota. Adaptado de "Type 2 Diabetes with Artificial Intelligence Machine Learning: Methods and Evaluation" (p. 8), por L. Ismail et al., 2022, SpringerLink.

Figura 7

Resumen de los algoritmos inteligentes para la diabetes tipo 2 empleando el conjunto

Articulo	Técnica	Precisión	Variables	Fuente de data	Pre-procesamiento
Art02 [15]	K-neigborth nest	100%	Ver Tabla 10	Pima Indian Diabetes	-Limpieza de datos -Reducción de datos
Art04 [17]	Árbol de decisión J48	90.04%	Ver Tabla 10	Pima Indian Diabetes	- Reemplazar los valores perdidos y valores imposibles con la media. - Usamos K-means para eliminar las muestras incorrecta-mente clasificadas
Art05 [18]	Levenberg-Marquardt algo-ritmo	0,71	Ver Tabla 10	Pima Indian Diabetes	No se aplico
Art10 [23]	- AdaBoost algo-rithm with deci-sion stump - Machine support vector - Naive Bayes - Árbol de deci-sión	- 80.72% - 79.687% - 79.687% - 77.6%	- Tríceps espesor del pliegue de la piel *(en la data local se obtiene de forma indirecta)* - 2 horas suero de insulina *(en la data local se obtiene de forma indirecta)* - Indice de masa corporal *(en la data para validacion se obtuvo con altura y peso)* - función de la diabetes pedigrí *(en la data local se obtiene de forma indirecta)*	- Pima Indian Diabetes - Data local	Los valores perdidos se sustituyen con los atributos de valor media correspondien-tes en el conjunto de datos global
Art11 [24]	Extreme learning machine(ELM) BackPropagation	- 0.5964 - 0.0575	Ver Tabla 10	Pima Indian Diabetes	Se normaliza para que tengan un cierto rango de valores
Art12 [25]	- Backpropaga-tion - Arbol de desi-cion J48 - Bayes Ingenuo - Vector machines	- 83.11 - 78.26 - 78.97 - 81.69	Ver Tabla 10	Pima Indian Diabetes	- Técnica de normali-zación Min-max - Seleccion de caracte-risticas con chi-cuadrado
Art17 [29]	Árbol de decisión difuso basado en índice de GINI	75.8%	Ver Tabla 10	Pima Indian Diabetes	Se eliminaron los registros con datos perdidos

de datos PIMA.

Nota. Adaptado de "Métodos de aprendizaje supervisado para la predicción de diabetes: una revisión sistemática de la literatura" (p. 21-23), por Y. Aguirre et al., 2019, Repositorio de Tesis Universidad Peruana Unión.

Figura 8

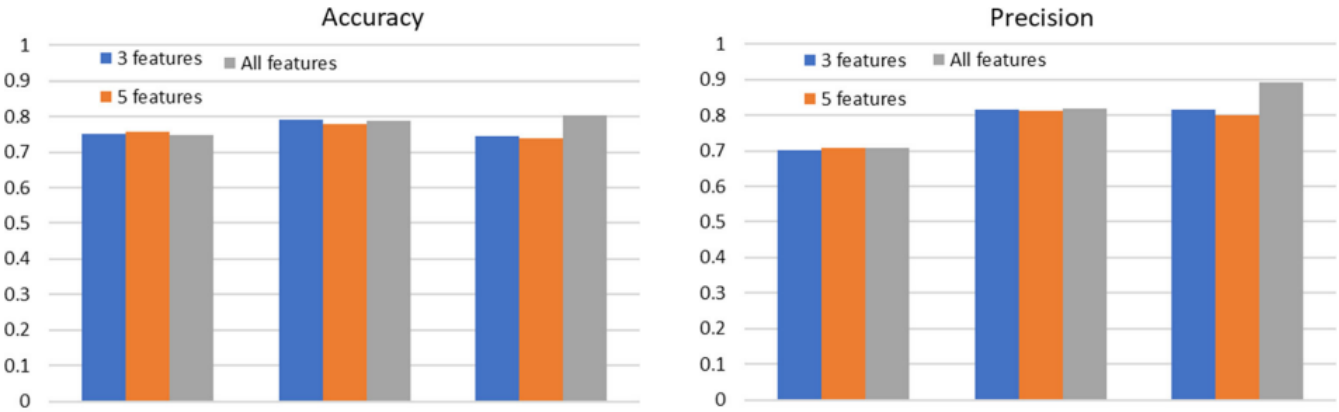
Evaluación de rendimiento de diferentes técnicas de predicción de Diabetes empleando el conjunto de datos PIMA.

Measures	Methods			
	DL	DT	ANN	NB
Accuracy (%)	98.07	96.62	90.34	76.33
Precision (%)	95.22	94.02	88.05	59.07
Recall (%)	98.46	95.45	83.09	64.51
F-Measure (%)	96.81	94.72	85.98	61.67
Specificity (%)	99.29	97.86	91.43	84.29
Sensitivity (%)	95.52	94.03	88.06	59.70

Nota. Adaptado de "Type 2 Diabetes with Artificial Intelligence Machine Learning: Methods and Evaluation" (p. 11), por L. Ismail et al., 2022, SpringerLink.

Figura 9

Comparativa de calidad y precisión de diferentes algoritmos de machine en la predicción de la diabetes según el empleo de selección de carcaterísticas



Nota. Adaptado de "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms" (p. 14), por V. Chang et al., 2022, PubMed Central.

4. Resultados y discusiones

A. Rendimiento de los algoritmos inteligentes según la selección de características

Según Ismail et al. (2022), argumentan que considerar los factores de riesgo de la diabetes tipo 2 como características especiales permite a unos algoritmos destacar frente a otros en su manejo.

- En el caso de que no exista una selección de características del conjunto datos, la adición del algoritmo K-means con los árboles de decisión o regresión logística muestran un grado alto tanto en la precisión como en la prueba de medida de precisión. Por el contrario, en el caso de los algoritmos de árboles de decisión, Naive Bayes, Random Forest presentan los niveles más bajos de precisión.
- En el caso de que exista una selección de características del conjunto datos, la adición del algoritmo K-means con KNN o regresión logística muestran un grado alto tanto en la precisión como en la prueba de medida de precisión. Por el contrario, en el caso de los algoritmos de árboles de decisión, Naive Bayes, Random Forest presentan niveles aún más bajos de precisión en comparación al caso donde se consideran los factores de riesgo (características).

  Según Naz & Ahuja (2020), argumenta que al considerar características especiales dentro del conjunto de datos médicos de PIMA dio como resultado que uno de los algoritmos con mayores grados de precisión es el árbol de decisión combinado con K-means; por el contrario, Naive Bayes y SVM presentaron los peores resultados.

B. Rendimiento de los algoritmos según el preprocesamiento de datos

   De acuerdo con Aguirre, Y. (2019), el preprocesamiento de datos es un factor clave para asegurar la validez de los resultados obtenidos por los algoritmos. De lo contrario, si se tomara como única guía al porcentaje de acierto, dejaría como mejor algoritmo a KNN. Por otro lado, indicó que el árbol de decisión y sus derivados son los más empleados; debido a, el manejo de grandes volúmenes de datos y estabilidad en datos faltantes. Sin embargo, modelos que involucran 2 o 3 algoritmos en conjunto presentan mejores resultados si se acompañan con el preprocesamiento de datos previo.

   De acuerdo con Ismail et al. (2022), presentaron a Bagging-LR, una variante del algoritmo de regresión lineal, como uno de los algoritmos más precisos para procesamiento de datos operando sobre un conjunto de datos balanceados. Por el contrario, para el caso de un conjunto de datos no balanceado, *Random Forest* es más preciso.

5. Respuestas a las preguntas de investigación

A. Análisis sobre el rendimiento de los algoritmos inteligentes según la selección de características

   Algunos de los modelos híbridos que involucran al algoritmo K-means presentan mejores resultados. Por el contrario, Naive Bayes presentó el peor resultado. Además, considerar los factores de riesgo de salud tiene un impacto en la predicción de la diabetes tipo 2.

B. Análisis sobre el rendimiento de los algoritmos según el preprocesamiento de datos

   Los modelos de múltiples algoritmos muestran mejores resultados en comparación a los algoritmos que se evaluaron unilateralmente. Además, el procesamiento de datos permite diferenciar los casos en el cual un algoritmo podría no ser confiable por la forma en cómo ha abordado el tratamiento de datos.

6. Conclusiones

En relación cona la investigación presentada se han identificado las siguientes conclusiones:

No existe un único algoritmo para la detección de la diabetes tipo 2, sino que, de acuerdo con las circunstancias en la que se desarrolla la construcción del modelo predictivo como la selección de características, el conjunto de datos, el preprocesamiento este puede variar. El algoritmo de árbol de decisión ha sido uno de los más empleados y presentados a través de modelos híbridos; por lo que, en un panorama general podría sugerirse como un candidato al mejor algoritmo en el diagnóstico de la diabetes tipo 2.

7. Referencias

Organization, W. H. (2014). Global Report on Diabetes. 2016. *Current Medical Research and Opinion*, *56*(1).

Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Umar, A. M., Linus, O. U., Arshad, H., Kazaure, A. A., Gana, U., & Kiru, M. U. (2019). Comprehensive Review of Artificial Neural Network Applications to Pattern Recognition. *IEEE Access*, *7*(February 2017), 158820–158846. https://doi.org/10.1109/ACCESS.2019.2945545

Aguirre Ascona, Y. D. (2019). *UNIVERSIDAD PERUANA UNIÓN FACULTAD DE INGENIERÍA Y ARQUITECTURA Escuela Profesional de Ingeniería de Sistemas* [Universidad Peruana Union]. http://hdl.handle.net/20.500.12840/2511

Alanis, M. (2018). *Prediagnóstico de enfermedades crónicas mediante algoritmos de cómputo inteligente* [Instituto Politécnico Nacional]. https://tesis.ipn.mx/handle/123456789/26201?show=full

Altobelli, E., Angeletti, P. M., Profeta, V. F., & Petrocelli, R. (2020). Lifestyle Risk Factors for Type 2 Diabetes Mellitus and National Diabetes Care Systems in European Countries. *Nutrients*, *12*, 1–14.

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. In *Journal of Big Data* (Vol. 8, Issue 1). Springer International Publishing. https://doi.org/10.1186/s40537-021-00444-8

Bellou et al. (2018). Risk factors for type 2 diabetes mellitus: An exposure-wide umbrella review of meta-analyses. PLoS ONE [revista en Internet] 2018 [acceso 20 de agosto de 2020]; 13(3): 1-27. *PLoS ONE*, 1–27. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5860745/pdf/pone.0194127.pdf

Chaki, J., Thillai Ganesh, S., Cidham, S. K., & Ananda Theertan, S. (2022). Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review. *Journal of King Saud University - Computer and Information Sciences*, *34*(6), 3204–3225. https://doi.org/10.1016/J.JKSUCI.2020.06.013

Chang, V., Bailey, J., Xu, Q. A., & Sun, Z. (2022). Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. *Neural Computing & Applications*, 1. https://doi.org/10.1007/S00521-022-07049-Z

Fan, C., Chen, M., Wang, X., Wang, J., & Huang, B. (2021). A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data. *Frontiers in Energy Research*, *9*(March), 1–17. https://doi.org/10.3389/fenrg.2021.652801

Ismail, L., Materwala, H., & Al Kaabi, J. (2021). Association of risk factors with type 2 diabetes: A systematic review. *Computational and Structural Biotechnology Journal*, *19*, 1759–1785. https://doi.org/10.1016/J.CSBJ.2021.03.003

Ismail, L., Materwala, H., Tayefi, M., Ngo, P., & Karduck, A. P. (2022). Type 2 Diabetes with Artificial Intelligence Machine Learning: Methods and Evaluation. *Archives of Computational Methods in Engineering*, *29*(1), 313–333. https://doi.org/10.1007/s11831-021-09582-x

Kalra, S., Jena, B. N., & Yeravdekar, R. (2018). Emotional and psychological needs of people with diabetes. *Indian Journal of Endocrinology and Metabolism*, *22*(5), 696–704. https://doi.org/10.4103/ijem.IJEM_579_17

Kyrou, I., Tsigos, C., Mavrogianni, C., Cardon, G., Van Stappen, V., Latomme, J., Kivelä, J., Wikström, K., Tsochev, K., Nanasi, A., Semanova, C., Mateo-Gallego, R., Lamiquiz-Moneo, I., Dafoulas, G., Timpel, P., Schwarz, P. E. H., Iotova, V., Tankova, T., Makrilakis, K., & Manios, Y. (2020). Sociodemographic and lifestyle-related risk factors for identifying vulnerable groups for type 2 diabetes: A narrative review with emphasis on data from Europe. In *BMC Endocrine Disorders* (Vol. 20, Issue 10, pp. 61–63). https://doi.org/10.1186/s12902-019-0463-3

Li, M., Rahman, M. L., Wu, J., DIng, M., Chavarro, J. E., Lin, Y., Ley, S. H., Bao, W., Grunnet, L. G., Hinkle, S. N., Thuesen, A. C. B., Yeung, E., Gore-Langton, R. E., Sherman, S., Hjort, L., Kampmann, F. B., Bjerregaard, A. A., Damm, P., Tekola-Ayele, F., … Zhang, C. (2020). Genetic factors and risk of type 2 diabetes among women with a history of gestational diabetes: Findings from two independent populations. *BMJ Open Diabetes Research and Care*, *8*(1). https://doi.org/10.1136/bmjdrc-2019-000850

Naz, H., & Ahuja, S. (2020). Deep learning approach for diabetes prediction using PIMA Indian dataset. *Journal of Diabetes and Metabolic Disorders*, *19*(1), 391–403. https://doi.org/10.1007/s40200-020-00520-5

Ordóñez, D. A., & Vizcarra, E. R. (2018). *Modelo Predictivo para el diagnóstico de la Diabetes Mellitus Tipo 2 soportado por SAP Predictive Analytics*. https://doi.org/10.19083/tesis/624417

Pinchevsky, Y., Butkow, N., Raal, F. J., Chirwa, T., & Rothberg, A. (2020). Demographic and clinical factors associated with development of type 2 diabetes: A review of the literature. *International Journal of General Medicine*, *13*, 121–129. https://doi.org/10.2147/IJGM.S226010

Uddin, S., Haque, I., Lu, H., Moni, M. A., & Gide, E. (2022). Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. *Scientific Reports 2022 12:1*, *12*(1), 1–11. https://doi.org/10.1038/s41598-022-10358-x

Wang, L. (2019). Research and Implementation of Machine Learning Classifier Based on KNN. *IOP Conference Series: Materials Science and Engineering*, *677*(5). https://doi.org/10.1088/1757-899X/677/5/052038

Yuan, S., & Larsson, S. C. (2020). An atlas on risk factors for type 2 diabetes: a wide-angled Mendelian randomisation study. *Diabetologia*, *63*(11), 2359–2371. https://doi.org/10.1007/S00125-020-05253-X/FIGURES/2