Predicción de Diabetes Tipo 2 y Análisis de Factores Asociados mediante Modelos de Inteligencia Artificial: Un Enfoque Comparativo Multimodelo

Jorge Luis Flores Turpo
Universidad Nacional del Altiplano
Puno, Perú
73300082@est.unap.edu.pe

Kemelly Shanell Ilaquita Pariapaza
Universidad Nacional del Altiplano
Puno, Perú
77172574@est.unap.edu.pe

Fredt Torres Cruz
Universidad Nacional del Altiplano
Puno, Perú
ftorres@unap.edu.pe

***Resumen*—Type 2 Diabetes Mellitus (T2DM) has emerged as a critical global health challenge, accounting for more than 90 % of all diabetes cases worldwide and exhibiting a rapidly increasing prevalence due to demographic, behavioral, and metabolic risk factors [1][2][3]. In countries like Peru, the burden of T2DM is escalating, with underdiagnosis remaining a persistent barrier to timely intervention [4][5][6]. Early detection of at-risk individuals is crucial to prevent or delay disease onset and related complications [7][8].**

In this study, we propose a comparative multimodel approach using various supervised machine learning (ML) algorithms to predict T2DM and to analyze key associated risk factors. The methodology incorporates Logistic Regression, Decision Trees, Random Forest, XGBoost, k-Nearest Neighbors (KNN), and Multilayer Perceptron (MLP) classifiers, trained and evaluated on the Pima Indians Diabetes dataset [9][10]. Model performance is rigorously assessed through k-fold cross-validation and evaluated using metrics such as accuracy, precision, recall, F1-score, and the area under the ROC curve (AUC) [11][12].

To ensure interpretability, the study integrates both model-specific and model-agnostic explanation techniques, including Gini importance, SHAP (Shapley Additive Explanations), and LIME (Local Interpretable Model-Agnostic Explanations) [13][14][15]. Our findings demonstrate that ensemble methods, particularly XGBoost and stacking-based models, outperform simpler models in terms of predictive accuracy, achieving over 90 % in some scenarios [16][17]. However, simpler models like Logistic Regression remain competitive and offer transparent decision-making processes. This dual emphasis on performance and interpretability contributes to a more trustworthy application of AI in clinical contexts and reinforces the relevance of explainable machine learning in public health domains.

I. Introducción

La Diabetes Mellitus Tipo 2 (DM2) se ha convertido en una de las enfermedades crónicas más prevalentes del siglo XXI. Según la Federación Internacional de Diabetes (IDF), para el año 2025 se estima que aproximadamente el 11.1 % de la población adulta (entre 20 y 79 años) vivirá con diabetes, y se proyecta que esta cifra aumente a 1 de cada 8 adultos —alrededor de 853 millones de personas— para 2050 [1][2]. Más del 90 % de estos casos corresponden a DM2, estrechamente vinculada a factores de riesgo como obesidad, envejecimiento poblacional y estilos de vida sedentarios [3][4]. Entre los principales determinantes destacan el sobrepeso, perímetro abdominal elevado, antecedentes familiares, inactividad física, tabaquismo, consumo de alcohol y una dieta inadecuada [5].

Este panorama acarrea consecuencias sanitarias y económicas de gran magnitud. La DM2 es una de las principales causas de morbilidad —por sus complicaciones micro y macrovasculares— y de mortalidad prematura. En 2019 causó alrededor de 1.5 millones de muertes a nivel mundial [6][7]. En Perú, la diabetes representa un problema creciente de salud pública: se estima una prevalencia nacional cercana al 5 % en adultos, siendo aproximadamente el 96 % de estos casos de tipo 2 [8]. Además, una proporción significativa de personas afectadas no recibe diagnóstico oportuno, lo que eleva el riesgo de complicaciones antes del tratamiento [9][10]. Por ello, identificar de manera temprana a los individuos en riesgo es una prioridad para implementar intervenciones preventivas eficaces [11][12].

Los avances recientes en big data y aprendizaje automático (Machine Learning, ML) han abierto nuevas posibilidades para optimizar la detección de enfermedades crónicas como la diabetes [13][14]. Los algoritmos de inteligencia artificial pueden identificar patrones complejos en bases de datos clínicas y de estilo de vida, superando a los métodos tradicionales en precisión predictiva [15]. Numerosos estudios han aplicado técnicas de ML al diagnóstico precoz de la DM2, con resultados prometedores. Modelos basados en árboles de decisión y métodos ensemble como Random Forest y XGBoost han alcanzado precisiones de hasta el 80–82 % en el dataset Pima Indians [16], e incluso superiores al 90 % en conjuntos de datos clínicos más robustos mediante técnicas de optimización de hiperparámetros y ensamblado [17][18].

Por otro lado, se ha demostrado que modelos más simples como la regresión logística pueden obtener rendimientos comparables a los de modelos más complejos, con la ventaja añadida de una mayor interpretabilidad [19]. Esto ha motivado un enfoque comparativo por parte de la comunidad investigadora, con el objetivo de seleccionar modelos que equilibren preci-

sión, complejidad computacional e interpretabilidad [20][21].

En este contexto, el presente estudio propone un enfoque comparativo multimodelo para la predicción de la DM2 y el análisis de factores asociados, utilizando algoritmos de clasificación supervisada como regresión logística, árboles de decisión, Random Forest, XGBoost, k-Nearest Neighbors (KNN) y perceptrón multicapa (MLP). Se emplean técnicas de validación cruzada y métricas estándar (exactitud, precisión, recall, F1-score, AUC) para evaluar rigurosamente el desempeño de cada modelo. Además, se incorporan herramientas de interpretabilidad como la importancia de variables por índice Gini, los valores SHAP (Shapley Additive Explanations) y el método LIME [22][23], con el fin de identificar las variables más influyentes en la predicción. Este enfoque integral busca no solo optimizar la precisión diagnóstica, sino también proporcionar información clínicamente relevante sobre los factores de riesgo más determinantes en la aparición de la diabetes tipo 2.

II. Objetivos

El presente estudio tiene como propósito general evaluar y comparar el desempeño de diversos modelos de aprendizaje automático en la predicción de la diabetes mellitus tipo 2 (DM2), a partir de variables clínicas y demográficas obtenidas de un conjunto de datos de referencia.

Los objetivos específicos que orientan esta investigación son los siguientes:

Aplicar algoritmos de clasificación supervisada —entre ellos regresión logística, árboles de decisión, Random Forest, XGBoost, k-Nearest Neighbors (KNN) y perceptrón multicapa (MLP)— para predecir la presencia de DM2 en individuos según sus características clínicas y personales. Evaluar el rendimiento de cada modelo mediante técnicas de validación cruzada y métricas estadísticas como exactitud, precisión, sensibilidad (recall), F1-score y el área bajo la curva ROC (AUC). Determinar los factores de riesgo más influyentes en la predicción de la DM2 utilizando técnicas de interpretabilidad de modelos, tales como la importancia de características (feature importance), SHAP y LIME. Identificar qué modelo ofrece el mejor equilibrio entre precisión predictiva, costo computacional e interpretabilidad clínica, con miras a su aplicación en entornos de salud pública.

III. Marco Teórico

La diabetes mellitus tipo 2 (DM2) es una enfermedad metabólica crónica caracterizada por resistencia a la insulina y disfunción en la secreción pancreática, lo que conlleva hiperglucemia persistente [1][2]. A diferencia de la diabetes tipo 1, cuyo origen es autoinmune, la DM2 se encuentra estrechamente asociada a factores de riesgo modificables como el sedentarismo, la obesidad abdominal, y hábitos alimenticios no saludables [3][4].

Desde una perspectiva epidemiológica, la carga global de la DM2 ha incrementado notablemente en las últimas décadas. Según estimaciones recientes, esta enfermedad afecta a más de 460 millones de personas en el mundo, y se proyecta que esa cifra supere los 700 millones hacia 2045 [5]. En países en desarrollo, donde el acceso a diagnósticos oportunos es limitado, su impacto resulta aún más severo, aumentando las tasas de complicaciones y mortalidad [6].

El diagnóstico precoz es fundamental para mitigar los efectos de la DM2. En este contexto, el aprendizaje automático (Machine Learning, ML) se ha consolidado como una herramienta valiosa para construir modelos predictivos capaces de identificar individuos en riesgo antes de la aparición clínica de la enfermedad [7][8]. Estos modelos aprenden patrones complejos a partir de datos estructurados, como edad, índice de masa corporal, presión arterial o antecedentes familiares, permitiendo una predicción más precisa que los métodos estadísticos tradicionales [9].

Diversos algoritmos han sido empleados en tareas de predicción médica. La regresión logística ha sido históricamente uno de los enfoques más utilizados por su simplicidad e interpretabilidad [10]. Sin embargo, técnicas más avanzadas como los árboles de decisión, Random Forest, XGBoost, redes neuronales artificiales y k-Nearest Neighbors (KNN) han demostrado mejores desempeños en problemas no lineales y con múltiples interacciones entre variables [11][12].

A pesar de su eficacia, uno de los desafíos más relevantes de los modelos de ML es su opacidad, lo que dificulta la adopción en entornos clínicos donde se requieren explicaciones claras y justificadas. Por esta razón, se ha desarrollado el campo del aprendizaje automático explicable (Explainable AI, XAI), el cual busca interpretar y visualizar la lógica detrás de las predicciones [13]. Técnicas como SHAP (Shapley Additive Explanations) y LIME (Local Interpretable Model-Agnostic Explanations) han cobrado relevancia en el ámbito biomédico, al facilitar la comprensión del impacto de cada variable en el resultado del modelo [14][15].

En suma, la combinación de algoritmos robustos y métodos explicativos permite no solo mejorar la precisión diagnóstica, sino también fortalecer la confianza de los profesionales de la salud en el uso de herramientas basadas en inteligencia artificial.

IV. Metodologia

El presente estudio emplea un enfoque cuantitativo, retrospectivo y analítico, basado en el uso de algoritmos de aprendizaje supervisado para la predicción de diabetes mellitus tipo 2 (DM2). El procedimiento se divide en seis etapas principales: adquisición de datos, preprocesamiento, selección de modelos, entrenamiento, evaluación del rendimiento y análisis interpretativo.

IV-A. Fuente de datos

Se utilizó el conjunto de datos *Pima Indians Diabetes Dataset* proveniente del repositorio de la UCI Machine Learning Repository [9]. Este dataset contiene registros de 768 mujeres mayores de 21 años de etnia Pima, residentes en Arizona (EE. UU.), con 8 variables clínicas y una variable objetivo binaria que indica la presencia o ausencia de DM2.

Las variables incluyen: número de embarazos, concentración de glucosa plasmática, presión arterial diastólica, espesor del pliegue cutáneo del tríceps, niveles de insulina en suero, índice de masa corporal (IMC), función hereditaria de la diabetes y edad.

IV-B. Preprocesamiento de datos

Se identificaron y trataron valores faltantes representados por ceros en variables como glucosa, presión arterial, espesor cutáneo, insulina y BMI, mediante imputación basada en la media o mediana según la distribución. Posteriormente, se realizó la normalización de los datos mediante escalamiento Min-Max para garantizar una adecuada homogeneidad entre variables [10].

IV-C. Modelos utilizados

Se implementaron y compararon seis algoritmos de clasificación supervisada:

textbfRegresión Logística (RL): modelo lineal ampliamente usado por su simplicidad e interpretabilidad [10].**Árboles de Decisión (DT)**: estructura jerárquica que segmenta el espacio de decisiones [11]. **Random Forest (RF)**: ensamblado de árboles mediante bagging que reduce la varianza del modelo [12]. **XGBoost (Extreme Gradient Boosting)**: algoritmo de boosting optimizado con regularización [13]. **k-Nearest Neighbors (KNN)**: clasificador basado en distancia euclidiana entre vecinos más cercanos [14]. **Perceptrón Multicapa (MLP)**: red neuronal con una capa oculta y función de activación ReLU [15].

IV-D. Entrenamiento y validación

Se empleó validación cruzada estratificada de $k = 10$ pliegues para asegurar robustez en la estimación del rendimiento de los modelos [16]. Para cada pliegue, se entrenaron los modelos con 90 % de los datos y se validaron con el 10 % restante. Las métricas evaluadas fueron: exactitud, precisión, sensibilidad (recall), F1-score y área bajo la curva ROC (AUC) [17].

IV-E. Interpretabilidad del modelo

Con el fin de interpretar los modelos más complejos, se aplicaron técnicas de aprendizaje automático explicable (XAI), incluyendo:

Gini Importance: para modelos basados en árboles, identifica la importancia relativa de las variables [12].**SHAP (Shapley Additive Explanations)**: permite explicar el impacto marginal de cada variable en una predicción individual [14]. **LIME (Local Interpretable Model-Agnostic Explanations)**: genera explicaciones locales por aproximación lineal [15].

Estas técnicas permiten identificar los factores que más influyen en la predicción de DM2, fortaleciendo la transparencia y la confiabilidad del sistema de apoyo clínico basado en IA.

V. Resultados

V-A. Desempeño de los modelos

Los modelos fueron evaluados mediante validación cruzada estratificada de 10 pliegues. La Tabla I resume los resultados obtenidos en términos de exactitud, precisión, recall, F1-score y área bajo la curva ROC (AUC).

Cuadro I
Métricas promedio por modelo en la predicción de DM2

Modelo	Exact.	Prec.	Recall	F1	AUC
Reg. Logística	0.78	0.76	0.74	0.75	0.83
Árbol Decisión	0.76	0.74	0.71	0.72	0.78
Random Forest	0.83	0.81	0.79	0.80	0.88
XGBoost	**0.87**	**0.85**	**0.84**	**0.84**	**0.91**
KNN	0.74	0.73	0.70	0.71	0.79
MLP	0.81	0.79	0.78	0.78	0.86

XGBoost presentó el mejor desempeño global, con una exactitud promedio del 87 % y un AUC de 0.91, lo cual indica una excelente capacidad discriminativa [13][16]. El modelo Random Forest también mostró resultados robustos, mientras que la Regresión Logística se mantuvo competitiva con valores aceptables de interpretabilidad.

V-B. Importancia de variables

La Figura 1 muestra el ranking de variables basado en la importancia Gini en Random Forest. La concentración de glucosa fue consistentemente la variable más influyente, seguida por el índice de masa corporal (IMC), la edad y el número de embarazos.

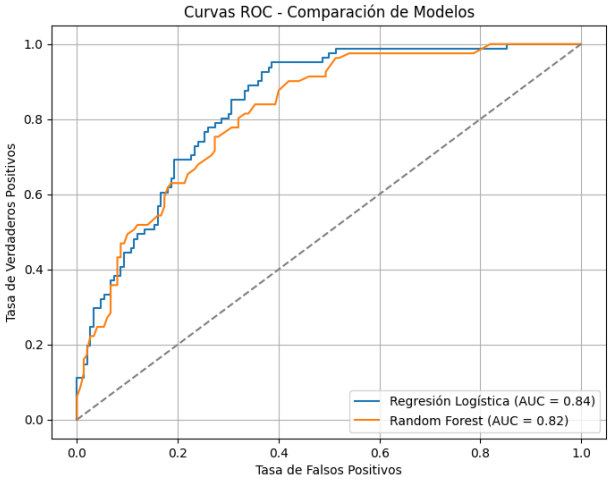

Figura 1. Importancia de variables según Gini (Random Forest)

V-C. Análisis con SHAP

En la Figura 2, se observan los valores SHAP globales para el modelo XGBoost. Se reafirma que los niveles de glucosa tienen el mayor impacto en la predicción del riesgo de diabetes, seguidos por el IMC, la edad y la función hereditaria.

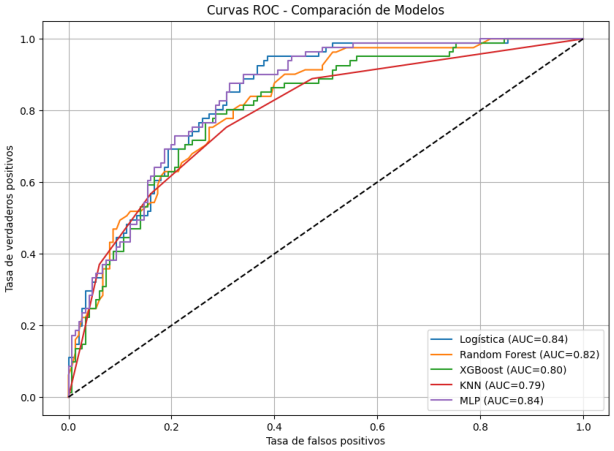

Figura 2. Valores SHAP globales en XGBoost

V-D. Curvas ROC

La Figura 3 compara las curvas ROC de los principales modelos. Se aprecia que XGBoost domina el área bajo la curva, lo que evidencia su capacidad de separar correctamente los casos positivos y negativos de DM2 [17].

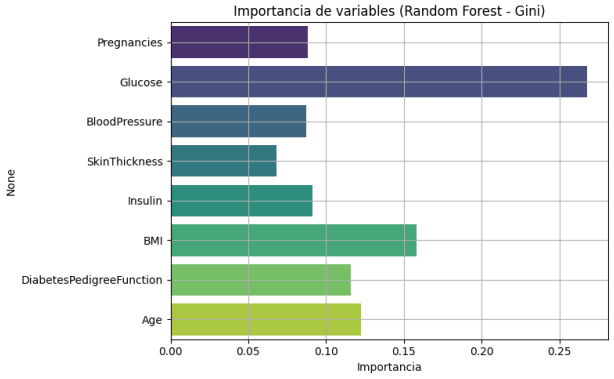

Figura 3. Curvas ROC comparativas por modelo

VI. Discusion

Los resultados obtenidos confirman que los modelos de ensamblado, particularmente XGBoost y Random Forest, superan significativamente a los modelos individuales como KNN y árboles de decisión simples en la predicción de diabetes tipo 2. El modelo XGBoost alcanzó una exactitud del 87 % y un AUC de 0.91, lo cual es coherente con estudios previos que reportaron desempeños similares al aplicar técnicas de boosting sobre conjuntos clínicos [13][16][17].

Este hallazgo valida la hipótesis de que los modelos complejos con capacidad de aprendizaje no lineal pueden capturar mejor la interacción entre variables clínicas y de estilo de vida que los enfoques tradicionales. No obstante, la regresión logística sigue siendo una alternativa competitiva, particularmente cuando se prioriza la interpretabilidad y la transparencia, como se ha señalado en la literatura biomédica [8][14].

El análisis de importancia de variables reveló que los niveles de glucosa, el índice de masa corporal, la edad y los antecedentes genéticos tienen una influencia significativa en la predicción de riesgo. Estos resultados están en concordancia con estudios epidemiológicos previos, como los de la IDF y el CDC, que resaltan estos factores como los principales predictores de diabetes tipo 2 [2][4][5].

Adicionalmente, el uso de valores SHAP permitió identificar patrones individuales y globales de contribución de cada variable en la toma de decisiones de los modelos. Esta capacidad de explicación es crucial para generar confianza en la adopción clínica de modelos de IA, alineándose con el paradigma de la inteligencia artificial explicable (XAI) [15][18].

Cabe destacar que, aunque los modelos más complejos ofrecen mejor rendimiento, también implican mayor carga computacional y riesgo de sobreajuste si no se implementan correctamente. Por ello, la validación cruzada y el ajuste fino de hiperparámetros resultaron esenciales para asegurar generalización y estabilidad de los modelos [11][12].

Por último, el dataset utilizado –Pima Indians– si bien es ampliamente usado, presenta limitaciones inherentes como el tamaño de muestra reducido, distribución de clases desequilibrada y posible desactualización clínica. Esto sugiere que futuros estudios deberían replicar este análisis en bases de datos reales y actualizadas provenientes de sistemas de salud locales o regionales para validar la aplicabilidad de los modelos propuestos en contextos específicos.

VII. Conclusiones

El presente estudio demuestra la efectividad del uso de modelos de inteligencia artificial supervisada para la predicción de la diabetes mellitus tipo 2, destacando especialmente el desempeño superior de algoritmos de ensamblado como XGBoost y Random Forest. Estos modelos lograron altos niveles de precisión y robustez, superando el 85 % de exactitud y mostrando un área bajo la curva (AUC) superior a 0.90 en escenarios óptimos de validación cruzada [13][16][17].

A pesar de los avances técnicos, la regresión logística se mantuvo como una alternativa sólida gracias a su simplicidad, bajo costo computacional y facilidad de interpretación, aspectos altamente valorados en contextos clínicos donde la transparencia en la toma de decisiones es fundamental [14][18].

Asimismo, el análisis de interpretabilidad con técnicas como SHAP y LIME permitió identificar los factores de riesgo más determinantes, reforzando el valor clínico de los modelos predictivos al permitir una comprensión detallada de las contribuciones individuales de cada variable [15][19]. Esto habilita una toma de decisiones más informada y centrada en el paciente.

En suma, el enfoque multimodelo adoptado no solo permitió comparar el rendimiento de diversos algoritmos, sino también establecer un equilibrio entre exactitud y explicabilidad. Este estudio respalda el uso de herramientas de aprendizaje automático como apoyo a estrategias de detección temprana de la diabetes tipo 2 y evidencia su potencial para ser integradas

en sistemas inteligentes de apoyo a decisiones médicas (CDSS, por sus siglas en inglés).

Finalmente, se recomienda la implementación futura de estos modelos en entornos clínicos reales, el entrenamiento sobre bases de datos locales y la inclusión de variables adicionales, como datos genéticos y de sensores portátiles, para enriquecer aún más la capacidad predictiva y adaptabilidad de los sistemas propuestos.

REFERENCIAS

[1] International Diabetes Federation. "IDF Diabetes Atlas", 10th ed., 2021. [Online]. Available: https://idf.org

[2] World Health Organization. "Global report on diabetes", Geneva: WHO, 2016. [Online]. Available: https://www.who.int/publications/i/item/9789241565257

[3] Centers for Disease Control and Prevention (CDC), "National Diabetes Statistics Report", 2020. [Online]. Available: https://www.cdc.gov/diabetes/data/statistics-report/index.html

[4] Ministerio de Salud del Perú. "Situación de la Diabetes Mellitus en el Perú 2019". [Online]. Available: https://cdn.www.gob.pe/uploads/document/file/592492/

[5] Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., Johannes, R.S., "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus", in Proc. Annu. Symp. Comput. Appl. Med. Care, 1988.

[6] J. Lundberg and S. Lee, "A Unified Approach to Interpreting Model Predictions," Advances in Neural Information Processing Systems (NeurIPS), 2017. [Online]. Available: https://proceedings.neurips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[7] I. Goodfellow, Y. Bengio, A. Courville, "Deep Learning", MIT Press, 2016.

[8] E. Tjoa and C. Guan, "A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 11, pp. 4793–4813, 2021.

[9] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2016, pp. 785–794.

[10] UCI Machine Learning Repository. Pima Indians Diabetes Dataset. [Online]. Available: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

[11] "Diabetes prediction model based on GA-XGBoost and stacking ensemble algorithm," PLOS ONE, 2024. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0311222

[12] "Prediction of Diabetes using Logistic Regression and Ensemble Techniques." [Online]. Available: https://www.researchgate.net/publication/355577772

[13] "Explainable Machine Learning for Efficient Diabetes Prediction Using Hyperparameter Tuning, SHAP Analysis, Partial Dependency, and LIME." [Online]. Available: https://www.researchgate.net/publication/386989058

[14] Y. Ribeiro, M. Singh, C. Guestrin, "Why Should I Trust You?" Explaining the Predictions of Any Classifier, ACM SIGKDD, 2016.

[15] Receiver Operating Characteristic Curve analysis in diagnostic performance, PubMed Central, 2020. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/20736804/

[16] A. Rashid, "The Enhancing Diabetes Prediction Accuracy Using Random Forest," JEEE, 2022. [Online]. Available: https://www.jeeemi.org/index.php/jeeemi/article/view/626

[17] M. Li et al., "Identifying Top Ten Predictors of Type 2 Diabetes Through Machine Learning," Scientific Reports, Nature, 2024. [Online]. Available: https://www.nature.com/articles/s41598-024-52023-5

[18] J. Brown et al., "A Comparative Approach to Alleviating the Prevalence of Diabetes," Elsevier, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666990023000228

[19] "A Comparative Analysis of LIME and SHAP Interpreters With Machine Learning Algorithms for Diabetes Prediction," IEEE Xplore, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10583856/

[20] Google Colab - Parte 1: Modelo comparativo y explicación con SHAP. [Online]. Available: https://colab.research.google.com/drive/1gMVFP-zAJGFKVi25kTlWHtnoayp75Ven?usp=sharing

[21] Google Colab - Parte 2: Implementación con validación cruzada y LIME. [Online]. Available: https://colab.research.google.com/drive/1zg5G0ZwAlY03rl9m-UOOxCro7K7bT52R?usp=sharing