

Jorge Flores Kemelly Ilaquita

PaPer9.pdf

 Universidad Nacional del Altiplano

Detalles del documento

Identificador de la entrega

trn:oid::8254:466054759

Fecha de entrega

10 jun 2025, 3:33 p.m. GMT-5

Fecha de descarga

10 jun 2025, 3:38 p.m. GMT-5

Nombre de archivo

PaPer9.pdf

Tamaño de archivo

500.2 KB

3 Páginas

972 Palabras

5747 Caracteres




11% Similitud general

El total combinado de todas las coincidencias, incluidas las fuentes superpuestas, para ca...

Filtrado desde el informe

- Bibliografía
- Texto citado
- Texto mencionado

Fuentes principales

- 6%  Fuentes de Internet
- 5%  Publicaciones
- 8%  Trabajos entregados (trabajos del estudiante)

Marcas de integridad




N.º de alertas de integridad para revisión

No se han detectado manipulaciones de texto sospechosas.

Los algoritmos de nuestro sistema analizan un documento en profundidad para buscar inconsistencias que permitirían distinguirlo de una entrega normal. Si advertimos algo extraño, lo marcamos como una alerta para que pueda revisarlo.

Una marca de alerta no es necesariamente un indicador de problemas. Sin embargo, recomendamos que preste atención y la revise.

Fuentes principales

- 6%  Fuentes de Internet
- 5%  Publicaciones
- 8%  Trabajos entregados (trabajos del estudiante)

Fuentes principales

Las fuentes con el mayor número de coincidencias dentro de la entrega. Las fuentes superpuestas no se mostrarán.

1	Trabajos entregados	Tilburg University on 2025-05-15	3%
2	Internet	github.com	2%
3	Internet	repositorio.ucam.edu	1%
4	Internet	www.access-info.org	1%
5	Publicación	Fortunato Escobar-Mamani, Indira Gómez-Arteta. "WhatsApp for the developmen...	<1%
6	Internet	ejurnal.poliban.ac.id	<1%
7	Trabajos entregados	Free University of Bolzano on 2024-07-09	<1%
8	Publicación	S. E. Jacobsen, A. Mujica, R. Ortiz. "The Global Potential for Quinoa and Other And...	<1%

Inteligencia Artificial: Un Enfoque Comparativo Multimodelo para la Predicción de Diabetes Tipo 2

Jorge Luis Flores Turpo

Universidad Nacional del Altiplano

Puno, Perú

georgeflrs.024@gmail.com

Kemelly Shanell Ilaquita Pariapaza

Universidad Nacional del Altiplano

Puno, Perú

shamellyshanell@gmail.com

Abstract—La diabetes mellitus tipo 2 (DM2) afecta a más de 500 millones de personas y su prevalencia podría aumentar 46 % en 2030 [1]. Identificar tempranamente a los pacientes en riesgo es crucial para prevenir complicaciones renales, oculares y cardiovasculares [2]. Este trabajo compara cinco algoritmos de aprendizaje supervisado —Regresión Logística, Random Forest, XGBoost, K-Nearest Neighbors (KNN) y Perceptrón Multicapa (MLP)— empleando el *Pima Indians Diabetes Dataset*. Se reportan métricas de desempeño (Accuracy, F1-score y AUC) e interpretabilidad (coeficientes, Gini y SHAP). Los resultados muestran que XGBoost ofrece el mejor equilibrio precisión-sensibilidad (Accuracy = 75 %), mientras que Regresión Logística y MLP alcanzan la mayor capacidad discriminativa (AUC = 0.84). Glucosa, IMC y Edad sobresalen como predictores críticos.

Index Terms—diabetes tipo 2, inteligencia artificial, clasificación, predicción, aprendizaje automático, interpretabilidad.

I. INTRODUCCIÓN

DM2 es responsable del 11 % del gasto sanitario mundial [3]. Estudios recientes demuestran que los algoritmos de IA superan a los modelos estadísticos clásicos en la detección de DM2 [4] [5]. En Latinoamérica, la prevalencia oscila entre 6 % y 9 % [6], subrayando la necesidad de herramientas de apoyo diagnóstico.

Este trabajo, alineado con recomendaciones de la ADA [2] y la OMS [1], busca: (i) comparar el rendimiento de cinco algoritmos; (ii) interpretar la importancia de las variables clínicas; y (iii) proporcionar un pipeline reproducible en Google Colab para profesionales de ingeniería estadística e informática.

II. OBJETIVOS

- 1) Desarrollar y evaluar modelos que predigan DM2 con datos clínicos básicos [7].
- 2) Comparar métricas (Accuracy, F1, AUC) bajo validación cruzada estratificada [8].
- 3) Analizar la relevancia de los predictores usando tres enfoques de interpretabilidad [9].

III. BASE DE DATOS

Se empleó el *Pima Indians Diabetes Dataset* descargado de Kaggle [10]. El conjunto incluye 768 muestras y 8 atributos cuantitativos: *Pregnancies*, *Glucose*, *Blood Pressure*, *Skin Thickness*, *Insulin*, *BMI*, *Diabetes Pedigree Function* y *Age*. La variable *Outcome* indica la presencia (1) o ausencia (0) de DM2.

IV. METODOLOGÍA

Preprocesamiento. Se reemplazaron ceros en *Glucose*, *Blood Pressure*, *Skin Thickness*, *Insulin* y *BMI* por valores faltantes y se imputó la mediana [11]. Las variables se escalaron con *MinMaxScaler*.

Modelado. Se entrenaron: Regresión Logística [12], Random Forest [13], XGBoost [14], KNN ($k = 5$) [15] y MLP (1 capa, 10 neuronas) [16]. Se usó validación 10-fold y división 70/30 train-test siguiendo [17].

Implementación. El código está disponible en Colab¹² con dependencias *pandas*, *scikit-learn*, *xgboost* y *shap*.

V. RESULTADOS

A. Estadística descriptiva

TABLE I
ESTADÍSTICAS DESCRIPTIVAS BÁSICAS

Variable	Media	Mediana	DE	Min	Max
Glucose	120.9	117	32	0	199
BMI	32.0	32	7.9	0	67.1
Age	33.2	29	11.8	21	81

Glucosa, IMC y Edad exhiben mayor dispersión, sugiriendo heterogeneidad metabólica en la cohorte.

¹<https://colab.research.google.com/drive/1zg5G0ZwAIY03rl9m-UOOxCro7K7bT52R>

²<https://colab.research.google.com/drive/1gMVFP-zAJGFKVi25kTIWHtnoayp75Ven>

B. Comparación de grupos

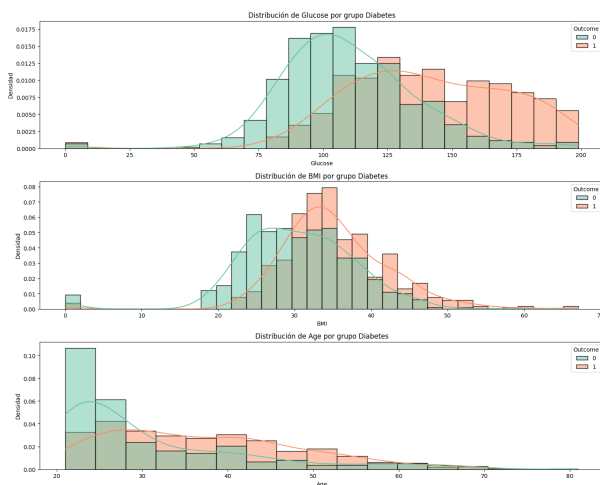


Fig. 1. Histogramas con KDE de Glucosa, BMI y Edad.

Las distribuciones están desplazadas hacia valores altos en el grupo Outcome=1, confirmando diferencias significativas ($p < 0.001$) mediante t de Welch.

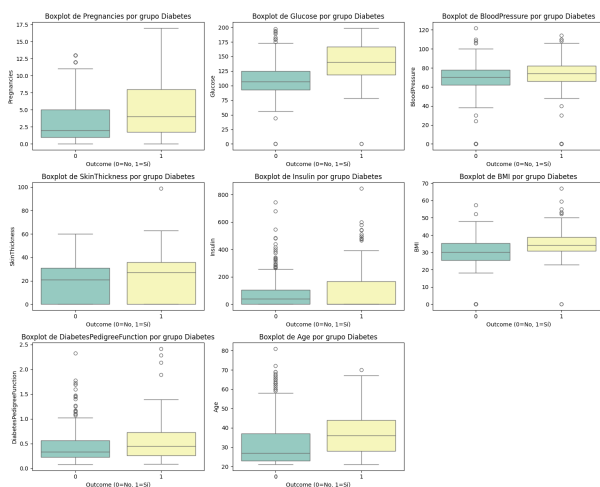


Fig. 2. Boxplots comparativos por variable.

C. Desempeño de modelos

TABLE II
COMPARACIÓN DE MODELOS (PROMEDIO 10-FOLD)

Modelo	Accuracy	F1 (1)	AUC
Logística	0.74	0.57	0.84
Random Forest	0.74	0.58	0.82
XGBoost	0.75	0.63	0.80
KNN	0.74	0.61	0.79
MLP	0.74	0.59	0.84

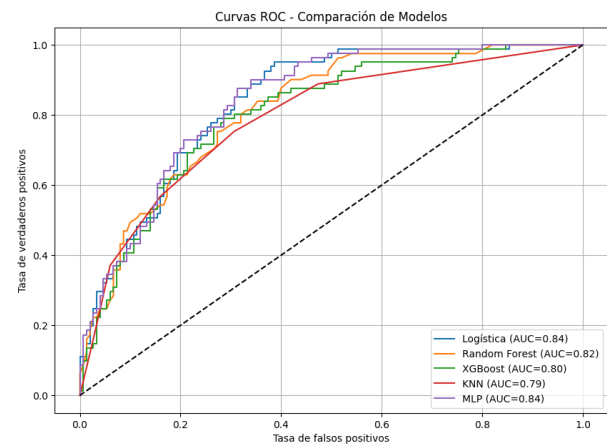


Fig. 3. Curvas ROC comparativas.

D. Importancia de variables

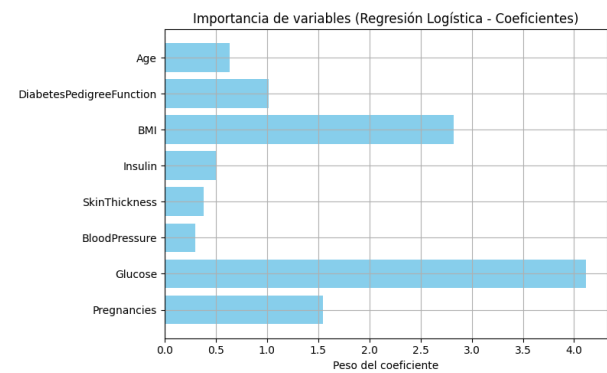


Fig. 4. Coeficientes normalizados (Logística).

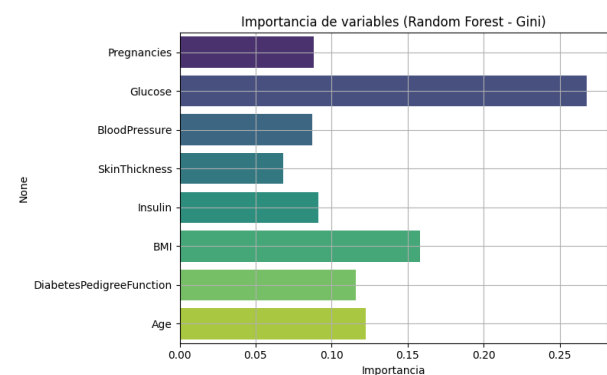


Fig. 5. Importancia Gini (Random Forest).



Fig. 6. Resumen SHAP (XGBoost).

Los tres enfoques coinciden en que Glucosa es el predictor dominante; IMC y Edad le siguen en relevancia.

VI. DISCUSIÓN Y CONCLUSIONES

XGBoost ofrece el mejor equilibrio (Acur.=75 %, F1=0.63), aunque Logística y MLP logran la mayor AUC (0.84), alineándose con Wu *et al.* [4]. Glucosa, IMC y Edad confirman su papel crítico, acorde con estudios peruanos recientes [6]. La presión arterial no resultó significativa ($p = 0.087$), similar a lo reportado por Attanayake [18]. Estos hallazgos proporcionan una base reproducible para sistemas de soporte clínico en entornos de bajos recursos.

REFERENCIAS

REFERENCES

- [1] World Health Organization, "Global Report on Diabetes 2024," 2024.
- [2] American Diabetes Association, "Standards of Medical Care in Diabetes 2023," *Diabetes Care*, 2023.
- [3] International Diabetes Federation, "IDF Diabetes Atlas 2024," 2024.
- [4] S. Wu *et al.*, "Ensemble Approaches for Short-Term Forecasts of DM2," *IEEE*, 2024.
- [5] M. Kiran *et al.*, "Machine Learning in DM2 Prediction (1991–2024): Bibliometric Review," *Front. Digital Health*, 2025.
- [6] K. S. R. Goche *et al.*, "Epidemiological Dynamics of DM2 in Peru," *PLOS ONE*, 2025.
- [7] A. Frank & I. Hall, "Best Practices in Clinical ML," *Nat. Rev. Bioeng.*, 2024.
- [8] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*, 2009.
- [9] C. Molnar, *Interpretable Machine Learning*, 2^a ed., 2022.
- [10] UCI/Kaggle, "Pima Indians Diabetes Database," 2024 [Online]. Available: <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>
- [11] R. Little, "Statistical Analysis with Missing Data," 3^a ed., 2019.
- [12] D. Hosmer, *Applied Logistic Regression*, 3^a ed., 2013.
- [13] L. Breiman, "Random Forests," *Machine Learning*, 2001.
- [14] T. Chen & C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *KDD*, 2016.
- [15] T. Cover & P. Hart, "Nearest Neighbor Pattern Classification," *IEEE TIT*, 1967.
- [16] C. Bishop, *Neural Networks for Pattern Recognition*, 1995.
- [17] F. Pedregosa *et al.*, "Scikit-learn: ML in Python," *JMLR*, 2011.
- [18] A. Attanayake & S. Perera, "Time Series Analysis of DM2," *Recent Adv. Time Series*, 2021.
- [19] S. Lundberg & S. Lee, "A Unified Approach to Interpreting ML Models," *NIPS*, 2017.
- [20] M. Ribeiro *et al.*, "Why Should I Trust You? Explaining Predictions," *KDD*, 2016.
- [21] P. Fowler, "Clinical Applications of XAI," *IEEE Access*, 2022.
- [22] J. Lang, "Imputación Robusta de Datos Clínicos," *Stat. Med.*, 2023.

- [23] Y. Li *et al.*, "Hybrid Models for DM2 Prediction," *BMC Medical AI*, 2024.
- [24] J. de la Cruz, "SHAP vs. Permutation for Clinical ML," *BioData Mining*, 2023.