RESEARCH ARTICLE
OPEN ACCESS

REVISTA RIICS Investigación e innovación en ciencias de la salud

Explainable Machine Learning to Understand the Social and Biological Determinants of Type 2 Diabetes in Mexico

Aprendizaje de máquina explicativo para entender los determinantes sociales y clínicos de la diabetes de tipo 2 en México

Mario Daniel Cervantes-Guerrero[1], **Carlos E. Galván-Tejada**[1], **Miguel Cruz**[2], **Jorge I. Galván-Tejada**[1], **Rodrigo C. Barros**[3], **Lucas Kupssinskü**[3]

[1] *Unidad Académica de Ingeniería Eléctrica, Universidad Autónoma de Zacatecas, Zacatecas, Mexico*

[2] *Instituto Mexicano del Seguro Social, Mexico City, Mexico*

[3] *Machine Learning Theory and Applications Lab, School of Technology, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, Brazil*

Correspondence
Mario Daniel Cervantes-Guerrero.
Email:
danielcervantesguerrero@uaz.edu.mx

Editor
Fraidy-Alonso Alzate-Pamplona, MSc.

Declaration of interests
The authors declare no conflicts of interest.

Funding
This research received no external funding.

Ethics statement
This study was approved by the Ethics Committee of the Instituto Mexicano del Seguro Social (IMSS) (approval no. R-2011-785-018). All participants provided written informed consent prior to participation.

Abstract

Introduction. Type 2 diabetes mellitus (T2DM) is a multifactorial disease associated not only with clinical factors but also with social determinants such as income, education, and access to healthcare. Its global prevalence continues to rise, particularly in low- and middle-income countries, making early detection and equitable management a public health priority.

Objective. To evaluate and interpret the performance of machine learning models that incorporate both clinical and socioeconomic data to predict the presence of type 2 diabetes in a Mexican population.

Methods. A dataset including 898 diabetic patients and 889 non-diabetic individuals randomly selected from the Hospital de Especialidades Siglo XXI (IMSS, Mexico City) was analyzed. Ten supervised algorithms (logistic regression, SVM, decision tree, random forest, KNN, naive bayes, AdaBoost, LightGBM, CatBoost, and XGBoost) were trained and validated using stratified 10-fold cross-validation. Data preprocessing included multiple imputation by chained equations (MICE), normalization, and encoding of categorical variables. Model interpretability was evaluated using SHapley Additive exPlanations (SHAP) to identify the most influential predictors.

Results. Ensemble methods, particularly XGBoost and LightGBM, achieved the best performance (accuracy = 0.94, AUC > 0.95). Glucose, diastolic blood pressure, and age were the strongest predictors, while socioeconomic variables such as income and education contributed additional predictive value. The integration of contextual variables improved model interpretability without reducing accuracy.

Data availability
All data supporting the findings of this study are available within the article. For additional details, please contact the corresponding author.

Author Contributions
Mario Daniel Cervantes-Guerrero: conceptualization, formal analysis, investigation, software, validation, visualization, writing – original draft.
Carlos E. Galván-Tejada: funding acquisition, methodology, project administration, resources, supervision, validation, writing – review & editing.
Miguel Cruz: data curation, validation, visualization.
Jorge I. Galván-Tejada: funding acquisition, resources, writing – original draft, writing – review & editing.
Rodrigo C. Barros: methodology, supervision, validation.
Lucas Kupssinskü: methodology, supervision, validation.

Generative AI declaration
The authors declare that no generative artificial intelligence tools were used in the writing of this manuscript, data analysis, or interpretation of the results presented. Grammarly (Premium version, 2025) was used solely as a language support tool for grammatical and linguistic editing to improve clarity and readability. The authors reviewed and edited the manuscript and take full responsibility for the accuracy and integrity of the final content.

Cite this article
Cervantes-Guerrero MD, Galván-Tejada CE, Cruz M, Galván-Tejada JI, Barros RC, Kupssinskü L. Explainable machine learning to understand the social and biological determinants of type 2 diabetes in Mexico. Revista de Investigación e Innovación en Ciencias de la Salud. 2026;8(2):XX-XX. e-v8n2a535. https://doi.org/10.46634/riics.535

Received: 11/11/2025
Revised: 01/03/2026
Accepted: 04/01/2026
Published: 04/16/2026

Conclusions. Explainable machine learning models integrating both clinical and socioeconomic data can enhance diagnostic precision and promote equity in diabetes detection. These tools offer potential applications in public health programs, facilitating the early identification of at-risk populations and supporting more equitable prevention and care strategies.

Keywords

Machine learning; explainable AI; diabetes; socioeconomic determinants; SHAP; Mexico.

Resumen

Introducción. La diabetes mellitus tipo 2 (DM2) es una enfermedad multifactorial asociada no solo a factores clínicos, sino también a determinantes sociales como el ingreso, la educación y el acceso a los servicios de salud. Su prevalencia global continúa en aumento, especialmente en países de ingresos bajos y medios, lo que convierte su detección temprana y su manejo equitativo en una prioridad de salud pública.

Objetivo. Evaluar e interpretar el desempeño de modelos de aprendizaje automático que integran datos clínicos y socioeconómicos para predecir la presencia de diabetes tipo 2 en una población mexicana.

Método. Se analizó un conjunto de datos conformado por 898 pacientes con diabetes y 889 individuos no diabéticos seleccionados aleatoriamente del Hospital de Especialidades Siglo XXI (IMSS, Ciudad de México). Se entrenaron y validaron diez algoritmos supervisados (regresión logística, SVM, árbol de decisión, bosque aleatorio, KNN, naive Bayes, AdaBoost, LightGBM, CatBoost y XGBoost) mediante validación cruzada estratificada de 10 pliegues. El preprocesamiento incluyó imputación múltiple por ecuaciones encadenadas (MICE), normalización y codificación de variables categóricas. La interpretabilidad del modelo se analizó mediante SHapley Additive exPlanations (SHAP) para identificar los predictores más influyentes.

Resultados. Los métodos de ensamble, particularmente XGBoost y LightGBM, mostraron el mejor desempeño (exactitud = 0.94, AUC > 0.95). La glucosa, la presión arterial diastólica y la edad fueron los predictores más relevantes, mientras que las variables socioeconómicas como ingreso y educación aportaron valor predictivo complementario. La integración de variables contextuales mejoró la interpretabilidad sin comprometer el rendimiento predictivo.

Conclusiones. Los modelos explicables de aprendizaje automático que integran información clínica y socioeconómica pueden mejorar la precisión diagnóstica y promover la equidad en la detección de la diabetes. Estas herramientas ofrecen un alto potencial para su aplicación en programas de salud pública, facilitando la identificación temprana de poblaciones en riesgo y fortaleciendo estrategias de prevención y atención más equitativas.

Palabras clave

Aprendizaje automático; inteligencia artificial explicable; diabetes; determinantes socioeconómicos; SHAP; México.

Introduction

According to the World Health Organization (WHO), diabetes mellitus is a chronic condition that occurs when blood glucose levels increase because the body cannot produce, or does not produce enough insulin, or cannot efficiently use the insulin it produces. Insulin allows glucose in the bloodstream to pass into the body's cells, where it is converted into energy or stored. Insulin is essential for metabolism and for the synthesis of proteins and fats; a deficiency of insulin or the inability of cells to react to it results in increased blood glucose levels, a condition known as hyperglycemia. When hyperglycemia persists for long periods of time, it can cause damage to the body, resulting in more serious complications such as kidney, nerve, and eye damage, cardiovascular diseases, or limb amputation, which may cause loss of vision or blindness. It is also linked to other complications such as loss of cognitive abilities, liver damage, or cancer. The most common type of diabetes is type 2. It occurs when the body becomes resistant to insulin or when it does not produce enough of it; over 90% of all diabetes worldwide corresponds to this kind of diabetes. In contrast, type 1 diabetes is caused by an autoimmune process, where the body's immune system detects the beta cells of the pancreas as a threat, attacking them and causing an insulin deficiency. It corresponds to around 5% to 10% of the world's cases.

The International Diabetes Federation (IDF) Diabetes Atlas provides estimates for diabetes in 2024. It is estimated that 589 million adults are living with diabetes, which represents 11.1% of the world's adult population, and this figure is projected to rise to 853 million by 2050. An estimated 252 million people are unaware that they have this condition, meaning that worldwide, four in ten adults living with diabetes are undiagnosed. Globally, 87% of all undiagnosed people with diabetes live in low and middle-income countries [1].

Considering the growing global burden of type 2 diabetes, data-driven and AI-based approaches have emerged as powerful tools for early detection and management. Recent advances in machine learning and explainable artificial intelligence (XAI) have enabled highly accurate predictive models for diabetes. Studies using both private and public datasets have reported strong performance from ensemble methods such as XGBoost, often enhanced through data balancing techniques like ADASYN. These works have also demonstrated the value of interpretability tools, particularly SHAP and LIME in identifying influential clinical predictors such as glucose levels and blood pressure, thus supporting transparent and clinically relevant decision-making [2,3]. Beyond traditional clinical variables, other research has explored the integration of lifestyle and nutritional factors into predictive frameworks. For example, incorporating dietary antioxidants into an XGBoost model for cardiovascular risk prediction in diabetic populations has achieved outstanding discrimination [4], while interpretable deep learning models have been applied to glucose forecasting in type 1 diabetes to ensure alignment with physiological principles [4]. Such developments demonstrate the growing convergence of advanced modeling techniques, interpretability methods, and multidimensional health data.

Despite these advances, most predictive models overlook socioeconomic determinants, or treat them as secondary variables, even though evidence shows they significantly influence the risk and detection of type 2 diabetes particularly in low- and middle-income countries [5,6]. Previous work has linked factors such as income, education, insurance status, ethnicity, and geographic location to disparities in disease prevalence, diagnosis, and treatment adherence [6,7]. A recent global systematic review confirmed that insurance status and ethnicity are the most consistent predictors of adherence to antidiabetic medication, while income and

education also play important but context-dependent roles [7]. These findings underscore the importance of integrating socioeconomic data into predictive modelling to create more equitable and context-aware tools. In this context, integrating socioeconomic and clinical variables in predictive modelling helps create more equitable and context-aware diagnostic tools. The present study evaluates ten supervised learning algorithms using combined clinical and socioeconomic data from a Mexican cohort, and applies SHAP to interpret the models' outputs. This approach aims not only to improve diagnostic accuracy but also to provide insight into the social dimensions of type 2 diabetes risk.

This study addresses an important limitation in existing diabetes prediction research: the frequent exclusion of socioeconomic determinants in machine learning models. While previous predictive models for type 2 diabetes have primarily relied on biomedical variables such as glucose, BMI, or blood pressure, recent evidence suggests that social determinants such as education level, income, and access to healthcare significantly influence both disease risk and the probability of diagnosis. In contrast to most previous studies, the present work explicitly integrates socioeconomic indicators with clinical biomarkers within multiple machine learning models and interprets their joint contribution using SHAP (SHapley Additive exPlanations). This integration provides a more context-aware predictive framework, particularly relevant for middle-income countries such as Mexico, where social inequality strongly shapes health outcomes.

Method

The original dataset was obtained through the systematic collection of data from 898 diabetic patients and 889 non-diabetic individuals at the Hospital de Especialidades Siglo XXI of the Mexican Institute of Social Security (IMSS) in Mexico City. Data on patients and non-diabetic individuals were collected in the Hospital's Biochemistry Research Unit laboratory. Clinical assessments were carried out, including measurements of weight, height, and blood pressure by trained clinical personnel following standard clinical protocols. Blood pressure was measured using calibrated sphygmomanometers with the patient in a seated position after a short resting period. Blood samples were collected after a 12-hour fasting period and analyzed in the hospital clinical laboratory using standard biochemical methods for glucose and lipid determination. Diabetic patients were identified through clinical diagnosis confirmed by laboratory measurements according to institutional protocols, while non-diabetic controls were randomly selected from individuals attending routine medical evaluations at the same hospital during the data collection period. This approach ensured comparable clinical assessment conditions across groups. Individuals with incomplete core clinical records, missing outcome information, or inconsistencies between laboratory and questionnaire data were excluded from the final analytic dataset. All patients and members of the control group signed an informed consent form, and the protocol met the criteria of the Declaration of Helsinki. The Ethics Committee of the IMSS approved it under the number R-2011-785-018. Because participants were recruited from a tertiary-care hospital, the sample should not be considered fully representative of the general Mexican population, and the findings should therefore be interpreted within the context of a hospital-based cohort. To evaluate the social and biological determinants of type 2 diabetes, this study included demographic variables (age and sex), clinical variables (body mass index, glucose, systolic blood pressure, and diastolic blood pressure), and socioeconomic variables (income level and education level). Diabetes status was used as the outcome variable. The general workflow of this study is presented in Figure 1.

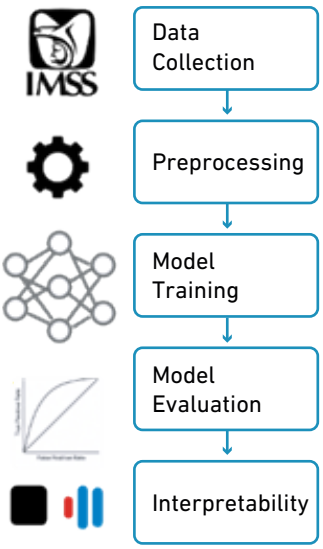

Figure 1. General workflow of the machine learning pipeline applied in this study, including data collection, preprocessing, model training, evaluation, and interpretability.

Data preprocessing and normalization are essential steps in the development of predictive models for the diagnosis and management of diseases such as diabetes mellitus. Preprocessing procedures included the detection and correction of errors, the transformation and coding of variables, and the imputation of missing values. Categorical features were standardized into binary numerical values to facilitate modeling: the sex variable, originally encoded as "F" (female) and "M" (male), was recoded into 0 and 1, respectively, while the diabetes status variable ("Status"), originally labeled as "control" (non-diabetic) and "case" (diabetic), was also recoded as 0 and 1. Missing values appeared in heterogeneous formats, including empty cells, extreme outliers, the placeholder "-9", and the string "null". These were first homogenized into a single format and then imputed using the Multivariate Imputation by Chained Equations (MICE) [8] method, an iterative approach that estimates missing entries through regression models based on the observed values of other features until stable estimates are obtained. The multivariate imputation by chained equations procedure was implemented using iterative regression models with ten iterations and predictive mean matching. Convergence was assessed by monitoring the stability in the imputed distributions across iterations. Features with direct or redundant associations to the outcome, such as patient ID, presence of complications, or time since diagnosis, were excluded to avoid data leakage and improve model generalizability. Finally, normalization was applied to numerical features to transform them to a common scale, using min–max rescaling or standardization based on the mean and standard deviation. This step was particularly important for algorithms sensitive to data scale, such as distance-based and gradient-based models, and contributed to convergence stability and improved predictive accuracy.

Before model training, categorical variables were encoded and continuous variables were normalized when required by the corresponding algorithm. Monthly income and educational level were treated as ordinal variables because both represent naturally ordered categories and are commonly modeled in epidemiological analyses as rank-based socioeconomic indicators. Missing data were handled using Multiple Imputation by Chained Equations (MICE) with

ten iterations and predictive mean matching. To prevent data leakage, all preprocessing steps, including imputation, encoding, and normalization, were performed exclusively within each training fold and then applied to the corresponding validation fold during cross-validation.

In this research, various supervised classification methods were used with the aim of developing predictive models for the identification of patients with diabetes. Classification algorithms are machine learning techniques that allow instances to be categorized into predefined classes, based on patterns learned from a set of labeled data. These methods analyze relevant features of the data set to build a model that can generalize and make accurate predictions about new observations.

The ten supervised learning algorithms were selected to cover a diverse range of modeling paradigms, from interpretable linear models to highly flexible ensemble methods. Logistic Regression [9] was included as a baseline due to its interpretability and well-established use in clinical research, making it useful for comparing the effect of clinical and socioeconomic variables such as glucose, BMI, income, and education with findings from previous epidemiological studies. Support Vector Machines (SVMs) [10] were chosen for their effectiveness in high-dimensional spaces and their ability to handle non-linear relationships between mixed data types, which is relevant for the present dataset that combines continuous clinical measures and ordinal socioeconomic indicators. Decision Trees [11] offer an intuitive, rule-based representation of decision-making, facilitating clinical interpretability and allowing direct visualization of how thresholds in glucose, blood pressure, or income can segment the population. Random Forests extend this approach through bagging and random feature selection, providing robustness to overfitting when dealing with correlated predictors such as systolic and diastolic blood pressure.

K-Nearest Neighbors (KNN) [12] was included as a distance-based method that can capture local patterns in the data without assuming a specific functional relationship, potentially revealing clusters of patients with similar socioeconomic and clinical profiles. Gaussian Naive Bayes [13] was selected for its simplicity, low computational cost, and ability to handle noisy data, serving as a probabilistic benchmark that can perform well even when variables are not strongly correlated. AdaBoost [14] was incorporated as a boosting algorithm that sequentially focuses on misclassified instances, improving the performance of weak learners such as shallow decision trees. LightGBM [15], XGBoost [16] and CatBoost [17] were chosen as gradient boosting frameworks capable of handling heterogeneous features and missing values efficiently, while offering high predictive accuracy and built-in regularization to mitigate overfitting. These three methods have consistently demonstrated superior performance in structured diabetic medical datasets [18-21], making them strong candidates for the present application. Hyperparameters for ensemble models were optimized through nested cross-validation combined with randomized search. Key parameters explored included learning rate (0.01–0.3), maximum tree depth (3–10), number of estimators (100–500), and subsampling ratios (0.6–1.0).

To evaluate the performance of the classification models, standard metrics for supervised learning problems were calculated, including accuracy, precision, sensitivity (recall), and F1-score [22], and analyzed independently for each class: patients without diabetes (class 0) and patients with diabetes (class 1). Accuracy was defined as the proportion of correct predictions out of the total samples evaluated. Precision reflects the proportion of true positives among the total number of cases predicted as positive, while sensitivity, also known as recall, indicates the model's ability to correctly identify true positive cases. The F1 score, as the harmonic mean between precision and recall, provides a balance between both metrics, which is especially useful in contexts with uneven class distribution or when seeking to minimize both false

positives and false negatives. Additionally, the Receiver Operating Characteristic (ROC) [23] curve and the area under the ROC curve (AUC) were used to evaluate the overall discriminatory capacity of the models. The ROC curve plots the true positive rate (sensitivity) versus the false positive rate (1 - specificity) for different classification thresholds, allowing the model's sensitivity-specificity trade-off to be visualized. The AUC quantifies this discriminatory ability in a single value between 0 and 1, where values close to 1 indicate excellent performance in distinguishing between patients with and without diabetes.

To ensure robust estimation of model performance and reduce the risk of overfitting, all models were evaluated using stratified 10-fold cross-validation [24]. The dataset was randomly partitioned into ten equally sized folds while preserving the proportion of diabetic and non-diabetic cases in each fold. In which nine folds were used for training and one of them was used for testing at each iteration; the process was repeated until every fold had served as a test set. To further reduce potential overfitting, additional validation was performed by comparing training and test performance across folds. Regularization parameters of ensemble methods (such as learning rate and maximum depth) were optimized using nested cross-validation. This procedure ensured model stability and generalization. Reported performance metrics (accuracy, precision, recall, F1-score, and AUC) correspond to the average across all folds.

Results

The database contains 1,787 complete observations. The average age of the subjects is 52.77 years (SD = 10.12), with a range of 30 to 93 years. The body mass index (BMI) has an average of 28.64 kg/m² (SD = 4.88), which is consistent with the overweight range. Blood glucose levels (GLU) show an average of 120.05 mg/dl (SD = 59.26), with a marked dispersion and extreme values (max. = 538 mg/dl), which suggests the presence of severe or atypical cases. Systolic blood pressure (SBP) and diastolic blood pressure (DBP) have averages of 122.53 mmHg (SD = 15.67) and 78.06 mmHg (SD = 11.11), respectively, with a wide range that includes clinically hypertensive values. Regarding sociodemographic variables, income (Sal), on an ordinal scale, has a mean of 2.00 (SD = 0.89), while educational level (Edu) averages 3.10 (SD = 1.57) on a scale from 0 to 6. The sex variable was nearly evenly distributed (mean = 0.499), and the outcome variable (STATUS) was also approximately balanced (mean = 0.503), indicating a similar proportion of diabetic and non-diabetic participants in the sample. These results are presented in Table 1.

The Correlation Matrix in Figure 2 shows how the STATUS variable, which indicates the presence of diabetes in the individuals in the dataset, presents positive correlations with various biometric variables. A positive correlation with blood glucose levels (GLU mg/dl) stands out, which is expected, since high glucose levels are a key diagnostic criterion for diabetes. Positive correlations are also observed with systolic blood pressure (SBP), diastolic blood pressure (DBP), body mass index (BMI), and age, suggesting that people with diabetes tend to be older, have higher body weight, and exhibit elevated blood pressure, patterns commonly associated with this disease. On the other hand, when analyzing the relationship between glucose and sociodemographic variables, a negative correlation is observed with both educational level (Edu) and salary (Sal). This indicates that people with a lower educational level or lower income tend to have higher glucose levels. This relationship can be explained by structural limitations associated with fewer socioeconomic resources, such as reduced access to healthy foods, less time available for physical activity due to longer or more demanding work hours, and limited access to preventive health information or services. These conditions can encourage the adoption of unhealthy habits that, over time, increase the risk of developing diseases such as diabetes.

Table 1. Descriptive Statistics of the Dataset for every Feature.

Variable	Unit	Mean	SD	Min	25%	Median	75%	Max
Age	Years	52.78	10.12	30.00	45.00	52.00	60.00	93
BMI	kg/m^2	28.64	4.88	15.34	25.28	27.96	31.14	60
Glucose	mg/Dl	120.05	59.26	38.00	85.00	96.00	137.00	538
Systolic BP	MmHg	122.53	15.67	70.00	110.00	120.00	130.00	210
Diastolic BP	mmHg	78.06	11.11	30.00	70.00	80.00	85.00	125
Income Level	Ordinal scale (0–5)	2.00	0.89	0.00	1.00	2.00	3.00	5.00
Education Level	Ordinal scale (0–6)	3.10	1.57	0.00	2.00	3.00	5.00	6.00
Sex	Binary (0 = female, 1 = male)	0.50	0.50	0.00	0.00	0.00	1.00	1.00
Diabetes Status	Binary (0 = no diabetes, 1 = diabetic)	0.50	0.50	0.00	0.00	1.00	1.00	1.00

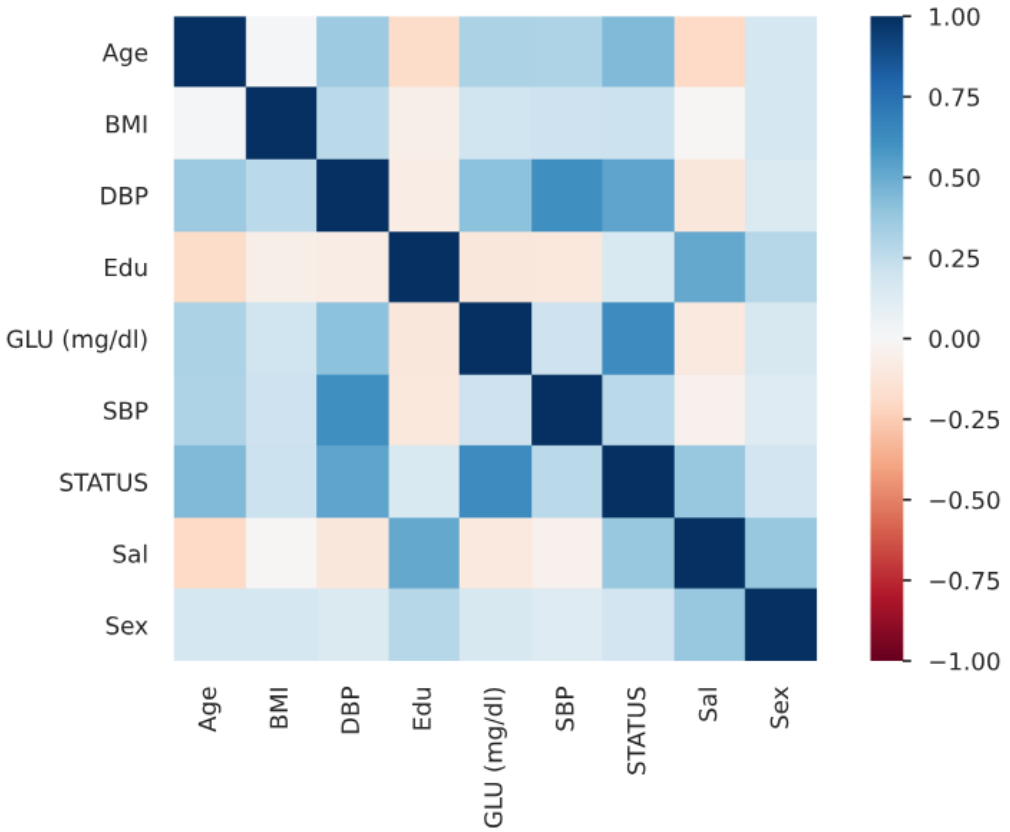

Figure 2. Correlation Matrix.

The performance of ten classification algorithms for predicting the presence of diabetes in patients was evaluated using accuracy, precision, recall, and F1-score metrics, differentiated for classes 0 (without diabetes) and 1 (with diabetes). The models that demonstrated the best overall performance were XGBoost, LightGBM, AdaBoost, and Random Forest, achieving accuracy values between 0.93 and 0.94, with balanced F1-scores across both classes (0.92–0.94). XGBoost and LightGBM showed outstanding performance both in detecting patients with diabetes and in correctly identifying those without the disease, with high recall (≥0.91) in both classes. Catboost stood out for its very high accuracy (0.95), although its recall was slightly lower than that of the previous models. Random Forest, on the other hand, offered a solid balance across all metrics, with consistently high accuracy and F1-score values. At an intermediate performance level were models such as Decision Tree, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Gaussian Naive Bayes, which achieved an accuracy close to 90% and satisfactory results in precision and sensitivity. It is worth noting that SVM presented a particularly high recall for class 0 (0.95), suggesting a tendency to correctly classify non-diabetic patients, although with less effectiveness in detecting positive cases. Finally, Logistic Regression was the model with the lowest performance (accuracy of 0.83), although its metrics were relatively balanced, which can be useful in contexts where model interpretability is a priority. In general, models based on ensemble techniques, especially boosting algorithms, demonstrated greater predictive capacity and better discrimination between classes, which is essential in clinical contexts where minimizing false negatives is crucial to ensure timely and adequate detection of diabetes. These results are summarized in Table 2.

Table 2. Performance metrics of ten supervised learning models for type 2 diabetes prediction using clinical and socioeconomic variables.

Model	Accuracy	Precision (0)	Recall (0)	F1-score (0)	Precision (1)	Recall (1)	F1-score (1)
Logistic Regression	0.83	0.83	0.82	0.83	0.84	0.84	0.84
Decision Tree	0.91	0.92	0.88	0.9	0.9	0.92	0.91
SVM	0.9	0.85	0.95	0.9	0.95	0.84	0.89
XGBoost	0.94	0.92	0.96	0.94	0.96	0.92	0.94
Random Forest	0.93	0.96	0.89	0.93	0.96	0.89	0.93
K-Nearest Neighbors	0.9	0.94	0.86	0.9	0.94	0.86	0.9
Gaussian Naive Bayes	0.9	0.93	0.87	0.9	0.93	0.87	0.9
AdaBoost	0.93	0.98	0.88	0.92	0.98	0.88	0.92
LightGBM	0.93	0.96	0.91	0.94	0.96	0.91	0.94
Catboost	0.95	0.96	0.93	0.95	0.96	0.93	0.95

Evaluating the area under the ROC curve (AUC-ROC) for classification models complements traditional metrics and provides a robust measure of each model's discriminatory power, which is particularly relevant in clinical settings where misclassification costs can be asymmetric. The presented ROC curve in Figure 3 shows that algorithms based on boosting techniques,

such as XGBoost, LightGBM, AdaBoost, and CatBoost, exhibit outstanding performance, with curves approaching the upper left corner, indicating a high true positive rate with a low false positive rate. This is consistent with their previously observed high precision, recall, and F1-score values. Random Forest also demonstrates high performance, with a similarly close-to-optimal ROC curve. On the other hand, models such as Logistic Regression, Naive Bayes, and Decision Tree present ROC curves that, while acceptable, are lower than those obtained by ensemble models, reflecting a lower ability to discriminate between patients with and without diabetes. The performance of SVM and K-Nearest Neighbors (KNN) is also notable, showing curves above the reference diagonal and close to the leading models, although with slight variations in the false positive rate. Overall, the ROC curve results reinforce the superiority of boosting-based models, not only in terms of overall accuracy but also in their ability to achieve an optimal balance between sensitivity and specificity, which is essential for medical applications where the goal is to minimize both false negatives and false positives.

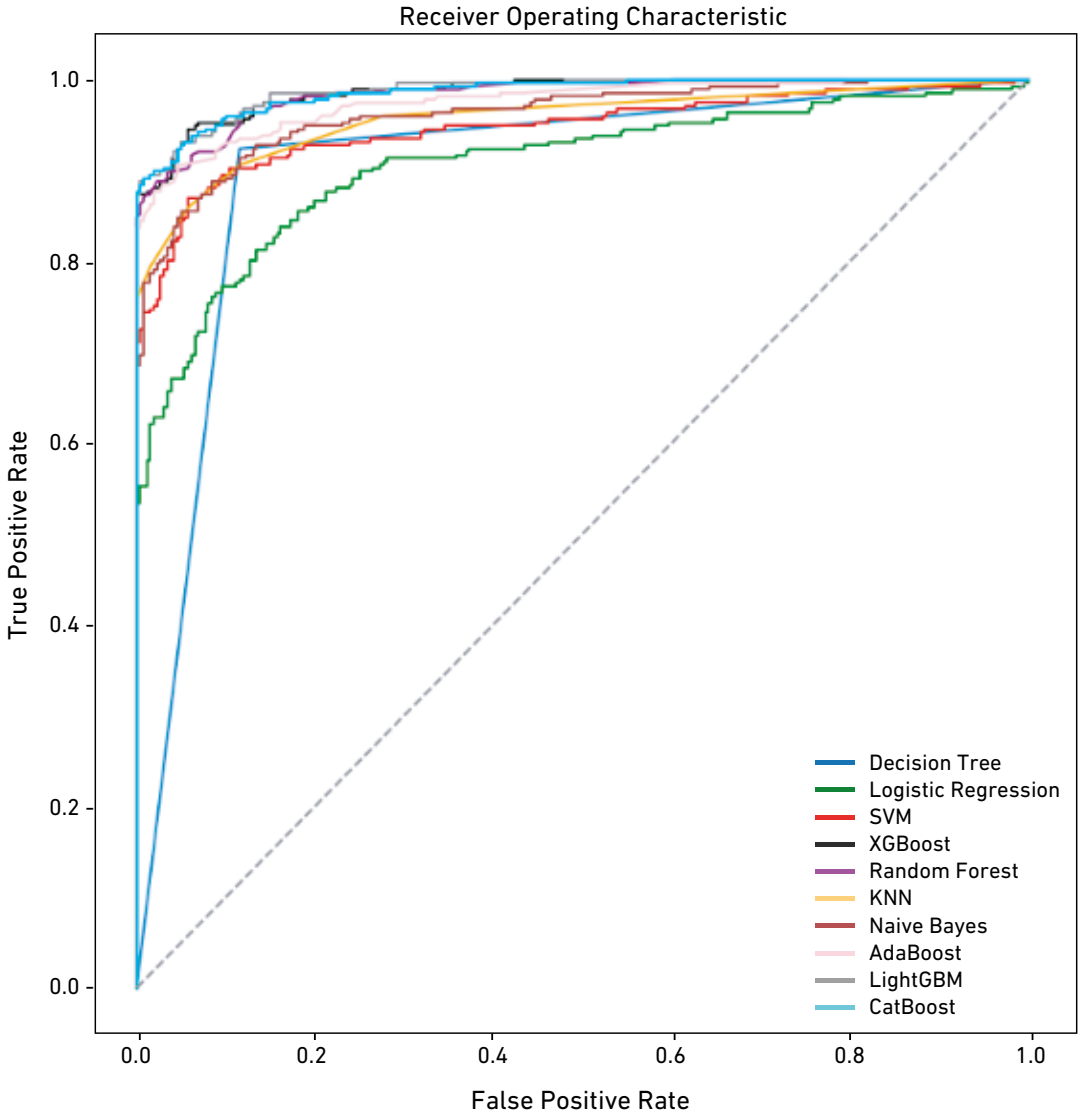

Figure 3. ROC Curve of ten supervised learning models for type 2 diabetes prediction using clinical and socioeconomic variables.

Figure 4 compares the mean absolute SHAP values of the top eight predictors across the three best-performing tree-based models: XGBoost, LightGBM, and CatBoost. Glucose concentration consistently emerged as the most influential feature, with LightGBM assigning the highest relative importance (mean |SHAP| ≈ 5.3), followed by XGBoost (≈ 4.2) and CatBoost (≈ 3.2). Diastolic blood pressure (DBP) and age ranked second and third across all models, although their contributions varied slightly between algorithms. Socioeconomic indicators, particularly monthly salary (Sal) and educational level (Edu), exhibited lower overall influence but remained consistently relevant, underscoring the added value of integrating contextual variables into the prediction task. Notably, LightGBM and XGBoost allocated greater weight to clinical measures such as BMI and systolic blood pressure (SBP), whereas CatBoost distributed relatively more importance to demographic variables such as sex. These differences highlight how model architecture and feature interactions can shape variable importance, reinforcing the need to interpret results in light of both predictive performance and domain relevance.

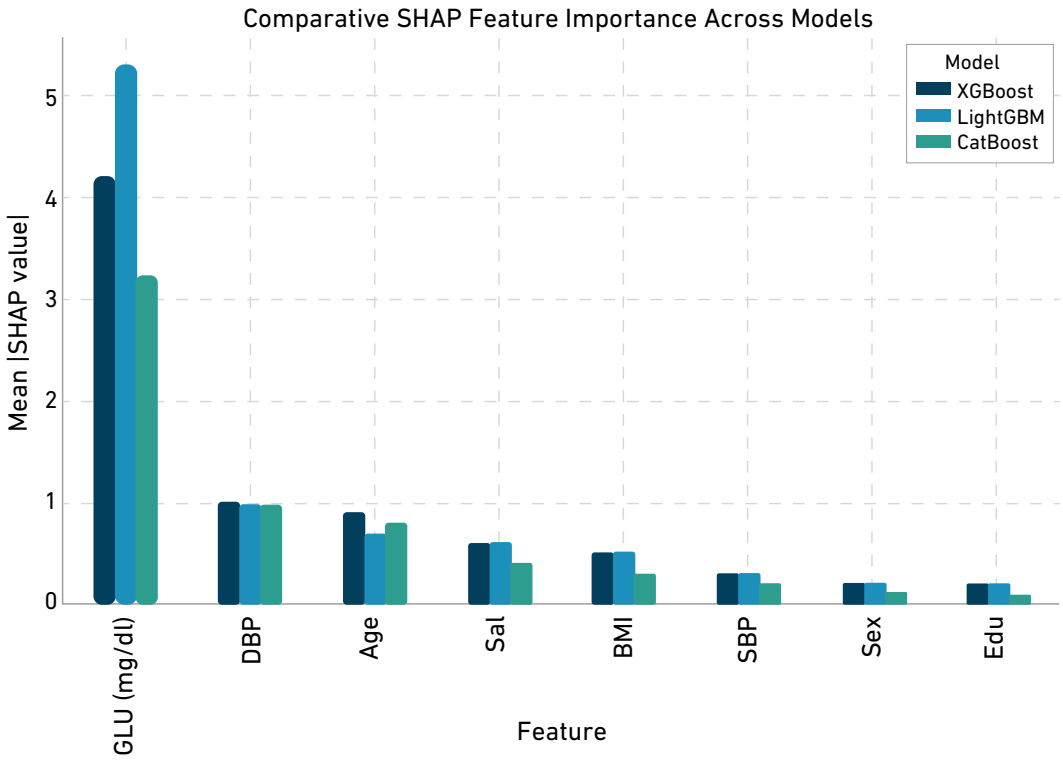

Figure 4. SHAP Feature Importance Across the 3 best-performing models.

Discussion

The SHAP analysis for XGBoost, LightGBM, and CatBoost confirmed that blood glucose (GLU) was the strongest predictor, in line with its central diagnostic role (World Health Organization). Diastolic blood pressure (DBP), age, and income also ranked highly, showing that both physiological and socioeconomic variables, such as income, were consistently associated

with predictions. Body mass index (BMI) showed a moderate impact, consistent with its recognition as a metabolic risk factor [24], whereas systolic blood pressure (SBP), sex, and education had a lower but still measurable influence. The variation in feature importance between algorithms, such as the higher weighting of demographic factors by CatBoost and the greater emphasis on BMI and SBP by LightGBM and XGBoost highlights how model architecture can shape interpretability outputs.

These results highlight the importance of integrating biomedical and contextual variables into predictive models to enhance both accuracy and real-world relevance. While a clinical perspective is essential for identifying physiological alterations such as insulin resistance or beta cell dysfunction, and for guiding evidence-based interventions like hypoglycemic therapy and continuous glucose monitoring, it often operates only after disease onset. A complementary social determinants approach addresses upstream risk factors, including poverty, food insecurity, unplanned urbanization, and limited access to healthy diets, enabling primary prevention strategies that can reduce incidence at the population level. Integrating these two perspectives can improve quality of life for individuals already living with diabetes and support the design of more equitable and sustainable public health policies.

In the correlation matrix, education and age presented moderate negative coefficients, indicating that higher educational attainment and younger age were associated with a lower probability of type 2 diabetes. This is consistent with previous findings that link education to better health knowledge and preventive behaviors [5], while age reflects the cumulative exposure to metabolic and lifestyle risk factors over time [23].

Unlike most existing machine learning studies that rely exclusively on biomedical variables, this work demonstrates that adding socioeconomic indicators such as income and educational level can consistently contribute to predictive performance without reducing accuracy. In fact, the best-performing ensemble models maintained AUC values above 0.90, comparable to those achieved in purely clinical datasets [2,3]. This aligns with epidemiological evidence that socioeconomic disadvantage is linked to both higher risk and poorer management of diabetes [6,7].

From a practical perspective, incorporating socioeconomic determinants into explainable AI models offers tangible opportunities for implementation in health systems, particularly in low- and middle-income countries. Such models could be embedded into primary care screening workflows, mobile health units, or community-based programs, allowing earlier risk detection even where laboratory resources are limited. Providing both a risk score and an interpretable breakdown of contributing factors would help healthcare providers tailor advice, prioritize follow-up, and allocate resources more effectively to underserved populations. These findings also emphasize the importance of developing context-sensitive predictive models tailored to populations with high social inequality. Strengthening early detection policies and access to preventive care in low-income communities could significantly reduce the prevalence of undiagnosed diabetes. Explainable machine learning models integrating socioeconomic information could be implemented in decision-support systems within primary care environments. For example, screening tools embedded in electronic health records or community health programs could estimate diabetes risk using both biomedical measurements and contextual indicators, allowing healthcare professionals to identify vulnerable populations earlier and prioritize preventive interventions.

Limitations and Recommendations

This study has limitations that must be acknowledged. First, the study design was cross-sectional, which precludes causal inference and limits the ability to evaluate risk trajectories over time; the associations observed here should be interpreted as correlational rather than causal. Second, because the sample was obtained from a single tertiary-care hospital, the findings may not be fully generalizable to other healthcare settings or to the broader Mexican population. Third, while data preprocessing included imputation and normalization, the specific methods applied may influence replicability, and the ordinal treatment of socioeconomic variables may not fully capture their complexity. In addition, key socioeconomic indicators such as education and monthly income were self-reported and may therefore be subject to reporting error or category misclassification. Finally, although the models showed strong discriminatory performance, calibration and subgroup fairness metrics were not assessed in the present analysis. Future work should evaluate these models in external cohorts, incorporate richer contextual and behavioral variables, and complement discrimination-based evaluation with calibration analysis and subgroup performance assessment to strengthen generalizability, clinical usefulness, and equitable real-world implementation.

Conclusions

This study concludes that while blood glucose and blood pressure remain the dominant predictors of type 2 diabetes across machine learning models, socioeconomic variables such as income and education provide consistent complementary value. By integrating biomedical and contextual factors, explainable AI models can improve both predictive accuracy and practical relevance, supporting strategies that address not only clinical management but also upstream social determinants of health. Despite limitations related to dataset scope, cross-sectional design, and variable representation, the findings highlight the potential of combining clinical and socioeconomic information to inform earlier detection, guide resource allocation, and promote more equitable public health interventions. Although the present findings are encouraging, they should be interpreted in light of the study's hospital-based and cross-sectional design. Future research should evaluate these models in external cohorts, incorporate richer contextual and behavioral variables, and assess calibration and subgroup performance to strengthen generalizability, clinical usefulness, and equitable real-world implementation.

References

1. Magliano DJ, Boyko EJ. IDF Diabetes Atlas [Internet]. 11th ed. Brussels: International Diabetes Federation (IDF); 2025. 125 p. Available from: https://diabetesatlas.org/resources/idf-diabetes-atlas-2025/

2. Tasin I, Nabil TU, Islam S, Khan R. Diabetes prediction using machine learning and explainable AI techniques. Healthc Technol Lett [Internet]. 2023;10(1-2):1-10. doi: https://doi.org/10.1049/htl2.12039

3. Zhang X, Lin S, Zeng Q, Peng L, Yan C. Machine learning and SHAP value interpretation for predicting cardiovascular disease risk in patients with diabetes using dietary antioxidants. Front Nutr [Internet]. 2025;12:1612369. doi: https://doi.org/10.3389/fnut.2025.1612369

4. Ejiyi CJ, Qin Z, Amos J, Ejiyi MB, Nnani A, Ejiyi TU, et al. A robust predictive diagnosis model for diabetes mellitus using Shapley-incorporated machine learning algorithms. Healthc Anal [Internet]. 2023;3:100166. doi: https://doi.org/10.1016/j.health.2023.100166

5. Agardh E, Allebeck P, Hallqvist J, Moradi T, Sidorchuk A. Type 2 diabetes incidence and socio-economic position: a systematic review and meta-analysis. Int J Epidemiol [Internet]. 2011;40(3):804-18. doi: https://doi.org/10.1093/ije/dyr029

6. Walker RJ, Smalls BL, Campbell JA, Strom Williams JL, Egede LE. Impact of social determinants of health on outcomes for type 2 diabetes: a systematic review. Endocrine [Internet]. 2014;47(1):29-48. doi: https://doi.org/10.1007/s12020-014-0195-0

7. Studer CM, Linder M, Pazzagli L. A global systematic overview of socioeconomic factors associated with antidiabetic medication adherence in individuals with type 2 diabetes. J Health Popul Nutr [Internet]. 2023;42(1):122. doi: https://doi.org/10.1186/s41043-023-00459-2

8. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res [Internet]. 2011;20(1):40-9. doi: https://doi.org/10.1002/mpr.329

9. LaValley MP. Logistic regression. Circulation [Internet]. 2008;117(18):2395-9. doi: https://doi.org/10.1161/circulationaha.106.682658

10. Suthaharan S. Support vector machine. In: Sharda R, Vob S, editors. Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning, vol 36 [Internet]. New York: Springer; 2016. p. 207-35. doi: https://doi.org/10.1007/978-1-4899-7641-3_9

11. De Ville B. Decision trees. WIREs Comput Stat [Internet]. 2013;5(6):448-55. doi: https://doi.org/10.1002/wics.1278

12. Guo G, Wang H, Bell D, Bi Y, Greer K. KNN model-based approach in classification. In: Meersman R, Tari Z, Schmidt DC, editors. On the Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE. OTM Confederated International Conferences [Internet]. Berlin: Springer; 2003. p. 986-96. doi: https://doi.org/10.1007/978-3-540-39964-3_62

13. Rish I. An empirical study of the naive Bayes classifier. In: Hoos HH, Stutzle T, editors. IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence [Internet]. Seattle: IJCAI; 2001. p. 41-6. Available from: https://faculty.cc.gatech.edu/~isbell/reading/papers/Rish.pdf

14. Schapire RE. Explaining AdaBoost. In: Scholkopf B, Luo Z, Vovk V, editors. Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik [Internet]. Berlin: Springer; 2013. p. 37-52. doi: https://doi.org/10.1007/978-3-642-41136-6_5

15. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: A highly efficient gradient boosting decision tree. In: Luxburg U, Guyon I, editors. NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems [Internet]; 2017 Dec 4 – Dec 9; Long Beach, USA. New York: Curran Associates Inc; 2017. p 3149-57. Available from: https://dl.acm.org/doi/10.5555/3294996.3295074

16. Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, et al. XGBoost: Extreme Gradient Boosting. R Package. Version 0.4-2 [software]. 2015 [cited 2025 Sep 12]. doi: https://doi.org/10.32614/CRAN.package.xgboost

17. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. Arxiv [Internet]. 2018. doi: https://doi.org/10.48550/arXiv.1706.09516

18. Mazhar F, Akbar W, Sajid M, Aslam N, Imran M, Ahmad H. Boosting early diabetes detection: an ensemble learning approach with XGBoost and LightGBM. JCBI [Internet]. 2024;6(2):127-38. Available from: https://www.jcbi.org/index.php/Main/article/view/347

19. Jaiswal S, Gupta P. Ensemble approach: XGBoost, CatBoost, and LightGBM for diabetes mellitus risk prediction. In: 2022 Second International Conference on Computer Science, Engineering and Applications (ICCSEA) [Internet]; 2022 Sep 8. GIET University, India. New York: IEEE; 2022. p. 1-6. doi: https://doi.org/10.1109/ICCSEA54677.2022.9936130

20. Rufo DD, Debelee TG, Ibenthal A, Negera WG. Diagnosis of diabetes mellitus using gradient boosting machine (LightGBM). Diagnostics [Internet]. 2021;11(9):1714. doi: https://doi.org/10.3390/diagnostics11091714

21. Donepudi S, Nakka R, Thota KK, Ajmeera M, Praveen SP, Sindhura S. Enhancing Person-Centric Health Care for Diabetes Prediction: A Comparative Study of LightGBM, XGBoost, and Hybrid LIGB Model. In: Barsocchi P, Srinivasu PN, Bhoi AK, Palumbo F, editors. Enabling Person-Centric Healthcare Using Ambient Assistive Technology, Volume 2: Personalized and Patient-Centric Healthcare Services in AAT [Internet]. Cham: Springer; 2025. p. 127-55. Available from: https://scite.ai/reports/enhancing-person-centric-health-care-for-ejrE2DQm

22. Vujovic Z. Classification model evaluation metrics. IJACSA [Internet]. 2021;12(6):599-606. doi: https://doi.org/10.14569/IJACSA.2021.0120670

23. Hoo ZH, Candlish J, Teare D. What is an ROC curve? Emerg Med J [Internet]. 2017;34(6):357-9. doi: https://doi.org/10.1136/emermed-2017-206735

24. Wong TT, Yeh PY. Reliable accuracy estimates from k-fold cross validation. IEEE transactions on knowledge and data engineering [Internet]. 2020;32(8):1586-94. doi: https://doi.org/10.1109/TKDE.2019.2912815