Type 2 Diabetes Prediction with Machine Learning: A Public Health Approach for Rural Contexts (mayo de 2025)

Dallys Nicol Sinisterra Gutierrez, Juan Felipe Hernandez, Manuel Enrique Luna, Master's students in Artificial Intelligence and Data Science, Universidad Autónoma de Occidente, Santiago de Cali, Colombia.dallys.sinisterra@uao.edu.co, juan.hernandez_m@uao.edu.co, manuel.luna@uao.edu.co

Abstract – Type 2 diabetes is a high-impact non-communicable chronic disease in public health, particularly in rural areas with limited access to medical services. This project proposes the development of a supervised classification model to predict the presence of diabetes or prediabetes using health, lifestyle, and sociodemographic data. The model was built using the 2015 BRFSS dataset, which includes 70,692 balanced records and 21 predictive variables, such as body mass index (BMI), hypertension, high cholesterol, self-reported health, physical activity, tobacco use, and educational level. The methodological process included exploratory data analysis (EDA), outlier detection, and feature selection using SelectKBest. Subsequently, Logistic Regression, Random Forest, XGBoost, and Support Vector Machine (SVM) models were trained. Random Forest demonstrated the best overall performance, standing out for its ability to handle nonlinear relationships and its robustness with heterogeneous data. The most influential variables included BMI, high blood pressure, elevated cholesterol, tobacco use, and general health perception. This study highlights the potential of machine learning as a support tool for digital screening and early clinical decision-making, particularly in contexts where structural barriers limit access to diagnostic testing. Its implementation can enhance prioritization strategies and medical outreach efforts, contributing to more timely and cost-effective disease detection.

Keywords: Diabetes, Screening, Statistical Models, Logistic Regression, Random Forest

I. INTRODUCTION

Type 2 diabetes is one of the leading non-communicable chronic diseases worldwide and represents a significant cause of morbidity, mortality, and economic burden for health systems. In Colombia, according to figures from the National Administrative Department of Statistics (DANE), this condition remained among the top ten causes of death in 2022, with an increasing trend over time. The issue is more severe in rural and hard-to-reach areas, where structural barriers hinder early detection and proper monitoring of the disease, leading to severe complications such as kidney failure, cardiovascular diseases, and decreased quality of life.

Given this scenario, the development of digital tools to support the early identification of type 2 diabetes risk has become a priority to improve prevention strategies and primary care. In this context, machine learning offers a promising approach by enabling the analysis of large volumes of clinical and behavioral data to identify predictive patterns that are not easily detected using traditional statistical methods.

This study proposes the development and implementation of a supervised classification model to predict the likelihood of having type 2 diabetes using the BRFSS 2015 dataset, which collects self-reported information on physical health, lifestyle, and sociodemographic variables. Based on this data, the goal is to build a support system that can be useful in rural settings or areas with limited medical coverage, facilitating preventive screening and prioritization in healthcare delivery pathways.

II. PROBLEM DESCRIPTION

The detection and treatment of non-communicable chronic diseases, such as type 2 diabetes, represents one of the main challenges for public health systems in Colombia. According to the Ministry of Health and the World Health Organization (WHO), this disease affects millions of Colombians and shows an increasing trend, especially in marginalized urban areas, dispersed rural territories, and generally among populations facing persistent barriers to accessing medical services.

Despite institutional efforts aimed at strengthening prevention and early diagnosis strategies, a significant proportion of cases remain undetected. This situation not only increases the clinical burden on patients but also raises treatment-related costs and contributes to the development of severe complications such as kidney failure, cardiovascular diseases, and neuropathies.

In response to this issue, this project proposes the development of a supervised classification model to predict the likelihood of having type 2 diabetes based on self-reported health indicators. Through the use of machine learning tools,

Scientific article submitted as the final project for the Machine Learning course of the Master's Program in Artificial Intelligence and Data Science.

Corresponding author: Dallys Nicol Sinisterra Gutierrez – Statistician, Universidad del Valle. Email: Dallys.sinisterra@uao.edu.co

Corresponding author: Juan Felipe Fernandez – Industrial Engineer, Universidad Central del Valle del Cauca. Email: Juan.hernandez_m@uao.edu.co

Corresponding author: Manuel Enrique Luna – Statistician, Universidad del Valle. Email: Manuel.luna@uao.edu.co

the aim is to identify significant patterns in variables such as blood pressure, cholesterol levels, body mass index (BMI), dietary habits, tobacco and alcohol consumption, among others.

The model is conceived as a support tool for digital public health strategies, with particular utility in rural areas or regions with limited medical infrastructure. Its implementation in self-screening forms or early warning systems would allow for prioritizing care during medical outreach campaigns and screening routes, contributing to more timely and cost-effective detection. For this purpose, the model is based on the BRFSS 2015 dataset, which includes over 70,000 records and has been widely validated in international public health research.

III. OBJECTIVES OF THE STUDY

General objective

To develop a supervised classification model based on machine learning capable of predicting the likelihood of a type 2 diabetes diagnosis using health indicators, lifestyle factors, and sociodemographic conditions. The model will be trained, validated, and interpreted using the BRFSS 2015 dataset.

Specific Objectives

1. To perform an exploratory data analysis (EDA) of the BRFSS 2015 dataset to identify patterns, outliers, and significant relationships between predictor variables and type 2 diabetes diagnosis.

2. To train and evaluate supervised classification models such as logistic regression, Random Forest, and XGBoost by applying metrics such as accuracy, recall, and F1-score to select the best-performing model.

3. To interpret the obtained results through feature importance analysis in order to identify the most influential health and behavioral factors in predicting type 2 diabetes and assess their applicability in low-resource medical contexts.

IV. LITERATURE REVIEW

The use of machine learning algorithms for predicting type 2 diabetes has gained increasing attention in recent scientific literature due to their ability to detect complex patterns in clinical, behavioral, and sociodemographic data. Multiple studies have demonstrated their applicability both in specialized clinical settings and in environments with structural limitations.

In Colombia, Mejía et al. (2023) applied models such as KNN, decision trees, and ensemble algorithms to a dataset of 10,889 healthcare affiliates, using only socio-environmental variables without biomarkers. Their best model achieved an AUC of approximately 0.61, though with sensitivity below 57%, with hereditary background (24.6%) and ethnicity (5.6%) as the most influential predictors.

In Bangladesh, García-Quezada et al. (2023) trained models using non-invasive clinical variables on 1,021 patients diagnosed with type 2 diabetes. The KNN algorithm yielded the best results, with an AUC of 0.8902, accuracy of 88.5%, and sensitivity of 94.6%.

In the United States, Riveros-Pérez et al. (2025) used NHANES (2007–2018) data to compare models such as logistic regression, SVM, Random Forest, XGBoost, and CatBoost. The best performance was achieved by XGBoost (AUC = 0.8168), followed by Random Forest and logistic regression (≈0.79).

Dinh et al. (2023) also used NHANES data (1999–2014), obtaining an AUC of 0.957 with laboratory data and 0.862 with self-reported data. Key variables included BMI, lipids, age, diet, and blood pressure.

In Taiwan, Wang et al. (2024) used electronic medical records from 6,687 adults. Models such as logistic regression, Random Forest, and XGBoost achieved accuracy above 98%, with HbA1c, fasting glucose, free T4, and triglycerides as the most important variables.

Ren (2023) used the BRFSS 2015 dataset with 70,000 self-reported records to train a CatBoost model. The study achieved 86.6% accuracy, with BMI, high blood pressure, age, and cholesterol among the top predictors.

Sampath et al. (2024) proposed a robust approach combining XGBoost with the SMOTE resampling technique, achieving an AUC of 0.968. This model stood out for improving sensitivity on minority classes within the clinical dataset.

Finally, Hossain et al. (2025) compared several algorithms—including logistic regression, KNN, Naive Bayes, Random Forest, and XGBoost—using datasets from Frankfurt Hospital and Pima Indian. They found Random Forest and XGBoost to deliver the best metrics (F1-score and accuracy), showing high stability across data sources.

Together, these studies highlight the superior performance and robustness of ensemble models, especially XGBoost and Random Forest. They also reinforce the predictive value of self-reported and lifestyle variables, as well as the importance of balancing technical performance with interpretability for effective public health applications.

V. METODOLOGY

The present study was developed based on the BRFSS 2015 (Behavioral Risk Factor Surveillance System) dataset, published by the Centers for Disease Control and Prevention (CDC) of the United States. This dataset contains self-reported information from 70,692 adults, including variables related to physical and mental health, eating habits, physical activity, substance use, educational level, and income. The target variable, *Diabetes_binary*, is coded as 1 for individuals diagnosed with prediabetes or diabetes, and 0 for individuals without a diagnosis, defining a binary supervised classification problem.

Data Preprocessing

In order to ensure the quality and suitability of the dataset for training machine learning models, the following preprocessing methodologies were implemented:

Exploratory Analysis and Outlier Handling

Boxplots, histograms, and quantile functions were used to identify outliers in continuous variables such as Body Mass Index (BMI), age, and the number of mentally unhealthy days. Filters were applied to reduce the influence of these values without excluding complete records, as the dataset was clean and contained no missing values.

Feature Scaling

To ensure balanced training for models sensitive to scale (such as SVM and logistic regression), the StandardScaler method was applied. This method transforms numerical variables to a distribution with zero mean and unit standard deviation.

Relevant Feature Selection

The SelectKBest technique with the ANOVA test function was used to select the most discriminative variables with respect to the target variable. This process allowed for dimensionality reduction and the elimination of redundant or uninformative variables.

Dataset Splitting

Once the data was preprocessed, the dataset was partitioned into training (80%) and test (20%) subsets using the train_test_split function, ensuring class balance through stratification to avoid bias during model training.

Statistical Models Used

Four supervised classification models were trained, selected based on their extensive documentation in similar problems and for offering a balance between performance and interpretability:

Logistic Regression, useful for obtaining a first approximation and analyzing the influence of each variable in probabilistic terms.

Random Forest, a decision tree ensemble model that improves accuracy by combining multiple trees trained on random subsets.

XGBoost, an advanced ensemble technique known for its optimization capabilities and efficient handling of complex variable interactions.

Support Vector Machine (SVM), which classifies data by finding the hyperplane that best separates the two classes, using nonlinear projections when needed.

Each model was trained using the same dataset, and their classification ability was measured using standard metrics such as accuracy, recall, F1-score, and confusion matrix.

Model Evaluation and Selection

Model evaluation focused on identifying the best balance between sensitivity (ability to detect positive cases) and precision. Additionally, the importance of each variable in the final prediction was analyzed to extract meaningful insights for public health strategy design.

Training times and model behavior were also compared to avoid overfitting risks. Random Forest showed the best overall performance in both metrics and stability, and it allowed for clear interpretation of the most influential diagnostic factors, such as BMI, high blood pressure, high cholesterol, and self-reported general health.

VI. RESULTS

The results analysis was structured into three main stages: model performance, error analysis, and identification of the most influential variables. The evaluation was conducted on the test set (20% of the total), ensuring that the models were assessed under unseen conditions. Metrics used included accuracy, recall, F1-score, and confusion matrix, along with a qualitative analysis of each algorithm's predictive behavior.

Exploratory Analysis

The exploratory analysis enabled the identification of relevant patterns in the distribution of variables and their relationship with patients' health status. Cases of diabetes or prediabetes were found to be more prevalent among individuals with older age, higher body mass index (BMI), high blood pressure,

elevated cholesterol levels, and a negative perception of their general health status. Additionally, variables such as physical activity level and tobacco use showed an inverse association with the diagnosis, suggesting their potential role as protective or risk factors. These initial observations guided the selection of the most relevant attributes for training the predictive models.

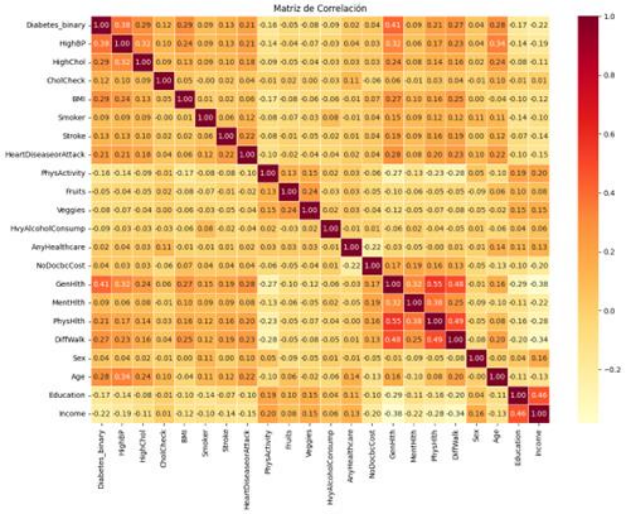

Figure.1 Correlation Matrix

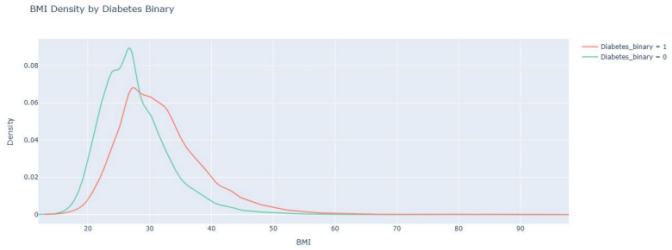

Figure.2. Density Distribution for Diabetic Population

Figure 2 presents the density distribution of Body Mass Index (BMI), differentiated according to the target variable (Diabetes). A rightward shift is observed in the curve corresponding to individuals diagnosed with diabetes or prediabetes (Diabetes_binary = 1), compared to those without a diagnosis (Diabetes_binary = 0). This difference indicates that people with a diabetic condition tend to have higher BMI values, suggesting a greater prevalence of overweight and obesity in this group.

Model Performance Comparison

The four models used exhibited different performance levels, with specific advantages depending on the evaluation focus. Random Forest achieved the best overall performance, with consistently high values across all evaluated metrics. Its ability to handle nonlinear relationships, avoid overfitting through data randomization, and provide interpretability through variable importance made it the most robust model in the study.

XGBoost, in turn, showed a performance very close to that of Random Forest, with a slight difference in recall but more efficient training. This model demonstrated higher sensitivity in identifying positive cases, which is crucial in public health prevention contexts, where false negatives (undiagnosed individuals with the disease) have critical consequences.

Logistic Regression played an important role as a baseline reference model. Although it showed lower predictive capability than the ensemble models, it was useful for interpreting the direction and strength of associations between variables, which adds explanatory value to the analysis.

SVM with RBF kernel yielded intermediate metrics but was the most computationally demanding. While it successfully separated the classes in the projected space, its sensitivity to data scaling and longer training time limited its efficiency compared to Random Forest or XGBoost.

Performance Metrics

To assess the predictive capability of the implemented models, a comparative analysis of their performance on the test set was conducted. This phase of the study aimed to determine the effectiveness of the algorithms in the binary classification of individuals based on their health condition, using the presence or absence of a prediabetes or type 2 diabetes diagnosis as the target variable.

The results for each evaluated model—Logistic Regression, Random Forest, XGBoost, and Support Vector Machine (SVM)—are presented below. These models were selected due to their wide adoption in medical classification problems and their diversity in terms of assumptions, complexity, and learning mechanisms.

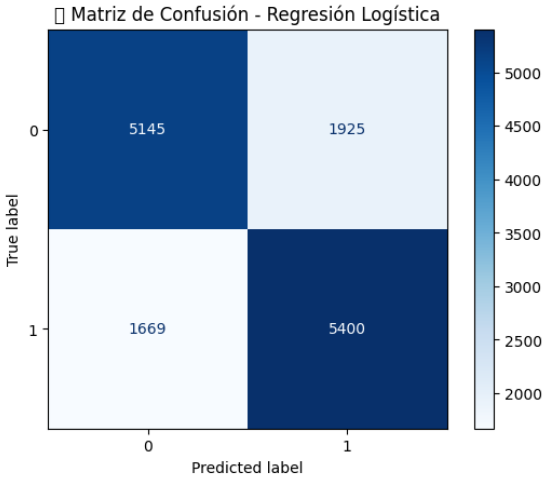

Figure 3. Confusion Matrix – Logistic Regression Model

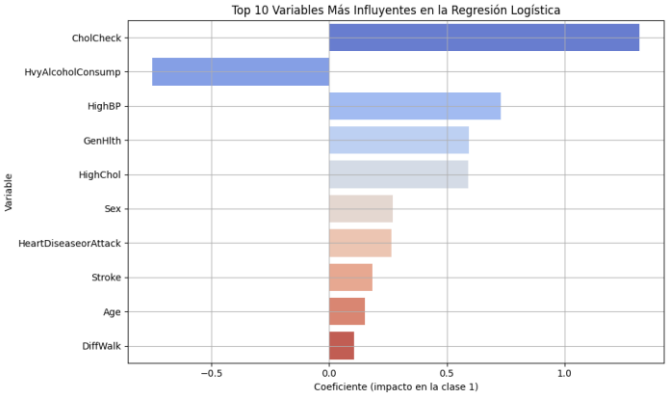

Figure 4. Top 10 Variable Contributions – Logistic Regression Model

Figure 4 presents the coefficients of the ten most influential variables in the logistic regression model. Cholesterol screening (CholCheck) exhibited the highest positive impact on the prediction of the positive class, while heavy alcohol consumption (HvyAlcoholConsump) showed a significantly negative effect. Other variables such as hypertension (HighBP), general health perception (GenHlth), and high cholesterol (HighChol) also contributed positively. The remaining variables, including sex, age, and cardiovascular history, displayed more moderate effects. These results allow for the interpretation of the relative weight of each variable in the model's decision-making.

To simplify the logistic regression model without compromising its performance, a variable selection process was implemented based on the magnitude of the estimated coefficients. A threshold of 0.01 was established, and only those variables with an absolute coefficient greater than this value were retained, resulting in a subset of 19 predictive variables.

Subsequently, a new model was trained using only these selected variables. Prior to training, standard scaling was applied using StandardScaler to ensure proper convergence of the algorithm. The reduced model achieved an accuracy of 74.6%, maintaining performance metrics similar to those of the full model in terms of precision, recall, and F1-score for both classes.

Figure 5 displays the confusion matrix for the reduced model, which correctly classified 5,085 negative cases and 5,457 positive cases. There were 1,985 false positives and 1,612 false negatives. These results confirm that reducing the number of variables did not compromise the model's performance, but rather enhanced its interpretability while preserving its predictive capacity.

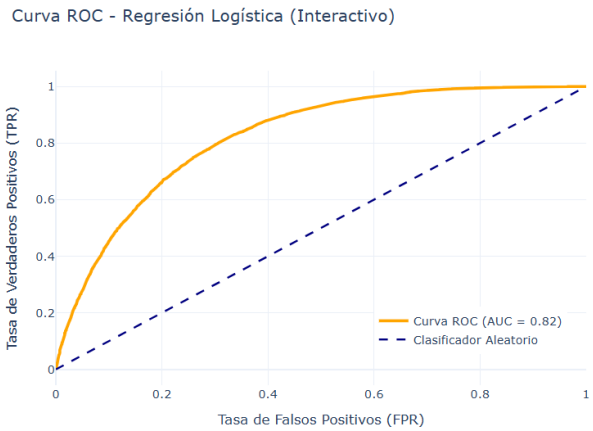

Figure 5. ROC Curve – Adjusted Logistic Regression Model

The ROC curve for the reduced model showed an Area Under the Curve (AUC) of 0.82, indicating strong discriminative capacity. This value means that the model has an 82% probability of assigning a higher risk score to a randomly selected positive observation compared to a negative one. By substantially outperforming the expected performance of a random classifier (AUC = 0.5), these results confirm that the model remains reliable for binary classification within this dataset, even after feature reduction.

Support Vector Machine (SVM): The Support Vector Machine model, trained without prior preprocessing, achieved an overall accuracy of 74.56% on the test set. In terms of class-wise performance, it yielded a precision of 0.78 and a recall of 0.69 for the negative class (0), and a precision of 0.72 with a recall of 0.80 for the positive class (1). The weighted average F1-score was 0.74.

Although the model demonstrates acceptable performance, the imbalance in error rates across classes suggests a higher sensitivity toward the positive class, at the cost of a higher false positive rate. Additionally, the absence of preprocessing techniques such as feature scaling or relevant attribute selection may have limited the model's ability to optimize the decision boundary in the feature space.

Reporte de Clasificación - SVM sin preprocesamiento:

	precision	recall	f1-score	support
0.0	0.78	0.69	0.73	7070
1.0	0.72	0.80	0.76	7069
accuracy			0.75	14139
macro avg	0.75	0.75	0.74	14139
weighted avg	0.75	0.75	0.74	14139

Accuracy: 0.7456680104675012

Figure 6. Performance Metrics – SVM (Without Preprocessing)

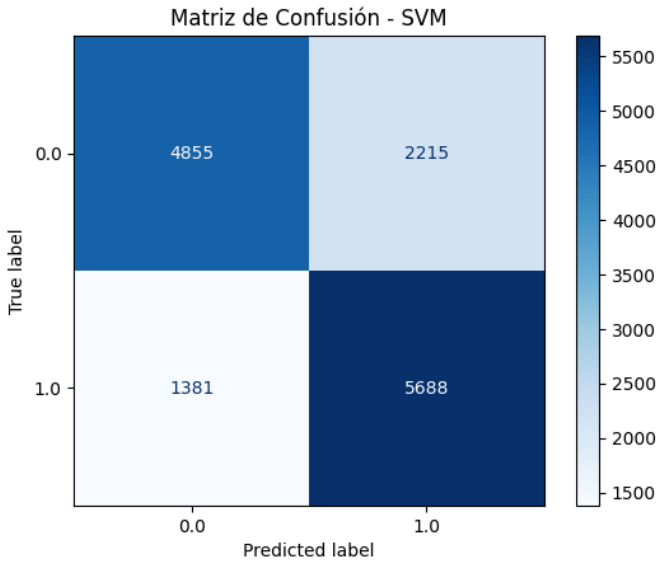

Figure 7. Confusion Matrix – SVM

Figure 6 displays the confusion matrix corresponding to the SVM model without preprocessing. The model correctly classified 4,855 negative cases and 5,688 positive cases but incurred 2,215 false positives and 1,381 false negatives. This behavior suggests a higher sensitivity toward detecting the positive class, reflected in the higher recall observed for that class. However, the relatively high rate of false positives could compromise the overall precision of the system in real-world applications.

Random Forest: The Random Forest model was included in the analysis due to its well-known ability to handle structured data with multiple variables and its robustness against overfitting. As an ensemble method based on aggregating multiple decision trees, it allows for capturing complex interactions among variables without requiring extensive prior transformations.

Precisión (Accuracy) del Random Forest : 0.7470

Reporte de Clasificación del Random Forest :

	precision	recall	f1-score	support
0.0	0.77	0.70	0.74	7070
1.0	0.73	0.79	0.76	7069
accuracy			0.75	14139
macro avg	0.75	0.75	0.75	14139
weighted avg	0.75	0.75	0.75	14139

Figure 8. Performance Metrics – Random Forest

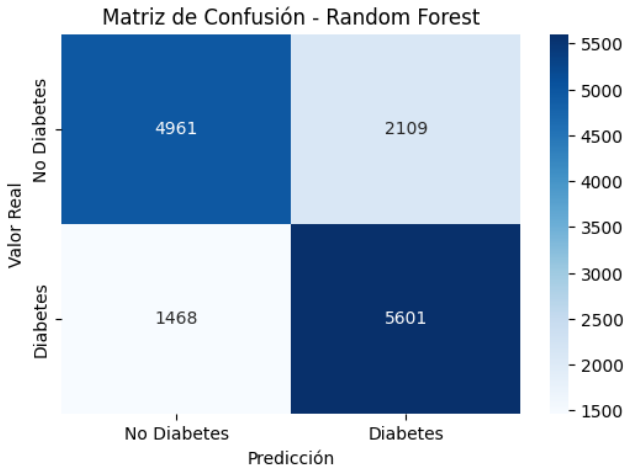

Figure 9. Confusion Matrix - Random Forest

Out of the 14,139 total cases evaluated, the model correctly classified 4,961 individuals without a diabetes diagnosis and 5,601 with a positive diagnosis. It identified 2,109 false positives and 1,468 false negatives, indicating a conservative tendency with greater sensitivity toward detecting positive cases—an important feature in preventive screening tasks.

Additionally, the classification report shows a recall of 0.79 and an F1-score of 0.76 for the positive class, suggesting a strong ability to correctly identify individuals at risk of type 2 diabetes. Although performance for the negative class was slightly lower (recall = 0.70), the model maintained an overall balance with average values of precision, recall, and F1-score around 0.75.

After a hyperparameter tuning process using cross-validation via GridSearchCV, the Random Forest model achieved an accuracy of 75.3% in validation and 75.2% on the test set, demonstrating good generalization capacity. While it reached high precision for the "Diabetes" class (78%), the recall was 71%, implying a relevant proportion of false negatives. These results reflect an acceptable overall performance but also highlight the need to improve the model's sensitivity in clinical applications, given the critical importance of minimizing undetected cases.

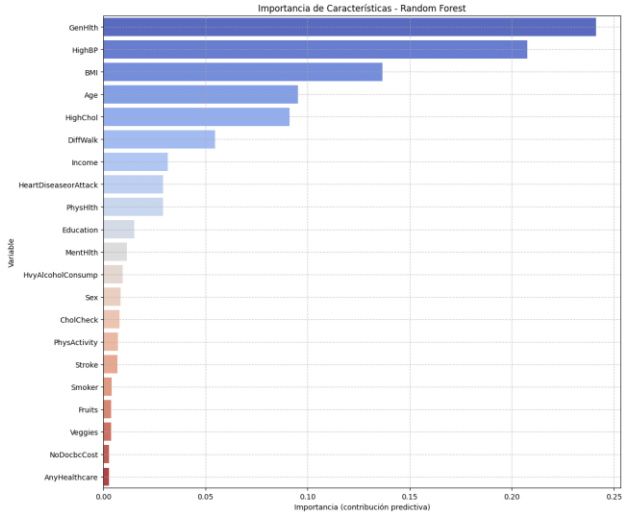

Figure 10. Top 10 Variable Contributions – RF

The feature importance analysis of the optimized Random Forest model reveals that the most influential factors in predicting diabetes risk are perceived general health (GenHlth), body mass index (BMI), age, and high blood pressure (HighBP)—all consistent with clinical literature.

Socioeconomic variables such as income and education level, as well as indicators of recent physical health and dietary factors like fruit consumption, also stand out. Together, these results confirm that the model captures a complex interaction between clinical, demographic, and social determinants. This not only validates its predictive capability but also provides useful insights to guide preventive interventions and screening strategies in public health.

XGBoost Model:

The XGBoost model, tuned through hyperparameter optimization, was evaluated on the test set to assess its discriminative power in predicting type 2 diabetes risk. This algorithm, based on gradient boosting techniques, is known for its high efficiency and ability to capture complex nonlinear relationships. The results indicate an overall accuracy of 75.2%, with balanced performance across both classes. Key evaluation metrics are presented below to assess its behavior in terms of precision, recall, and generalization capacity.

```
--- Evaluación del Modelo XGBoost AJUSTADO en el Conjunto de Prueba ---

Precisión (Accuracy) del XGBoost (Ajustado): 0.7516

Reporte de Clasificación del XGBoost (Ajustado):
              precision    recall  f1-score   support

         0.0       0.78      0.71      0.74      7070
         1.0       0.73      0.80      0.76      7069

    accuracy                           0.75     14139
   macro avg       0.75      0.75      0.75     14139
weighted avg       0.75      0.75      0.75     14139
```

Figure 11. XGBoost Performance Metrics

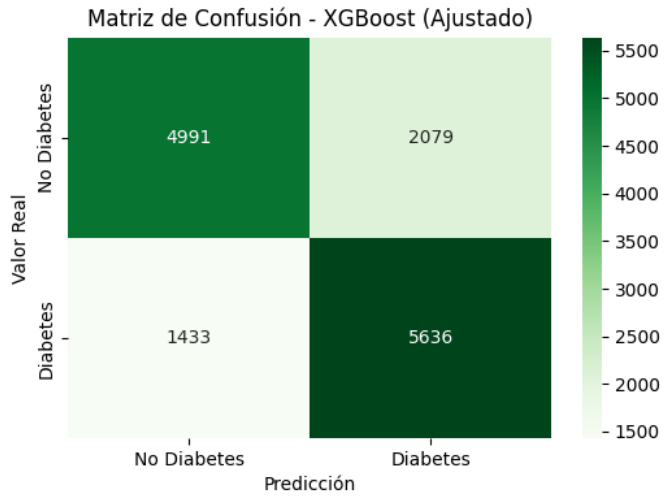

Figure 12. Confusion Matrix - XGboost

When comparing the performance of the tuned XGBoost model with that of Random Forest, both algorithms exhibit very similar overall accuracy rates (75.16% for XGBoost and 75.26% for Random Forest). However, subtle but meaningful differences emerge in their class-specific behavior. According to the confusion matrix, XGBoost correctly classified 5,636 individuals with a diabetes diagnosis, compared to 5,601 for Random Forest. Moreover, it produced 1,433 false negatives, slightly fewer than the 1,468 recorded by Random Forest, indicating a marginal improvement in sensitivity for the positive class. On the other hand, XGBoost resulted in 2,079 false positives, also slightly fewer than the 2,109 observed with Random Forest, suggesting better specificity.

In contrast, Random Forest may adopt a more conservative approach, slightly compromising detection to minimize overclassification risk. Nonetheless, both models demonstrate robust and comparable capabilities for addressing this binary classification task in the context of public health.

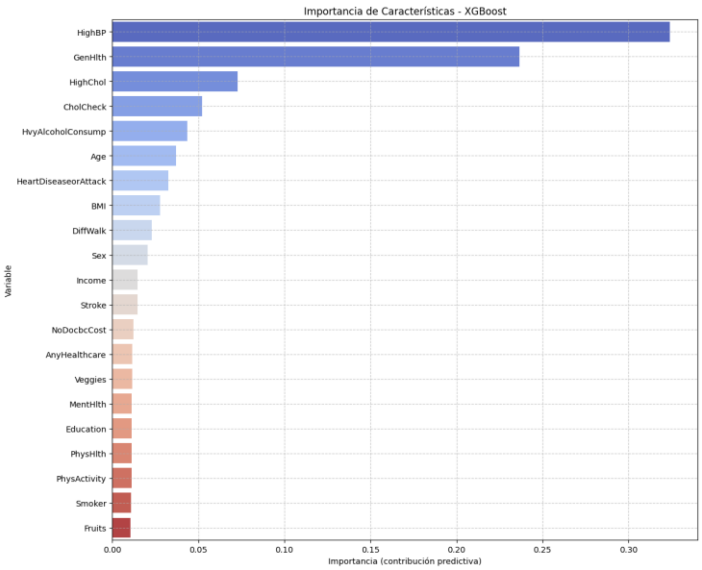

Figure 13. Top 10 Variable Contributions - XGboost

The feature importance plot of the XGBoost model highlights GenHlth as the most influential predictor, followed by Age, HighBP, and BMI. Variables such as PhysHlth, Income, Education, and MentHlth also contribute significantly. The strong alignment with the results from the Random Forest model reinforces the robustness of these predictors in classifying type 2 diabetes risk within the analyzed dataset.

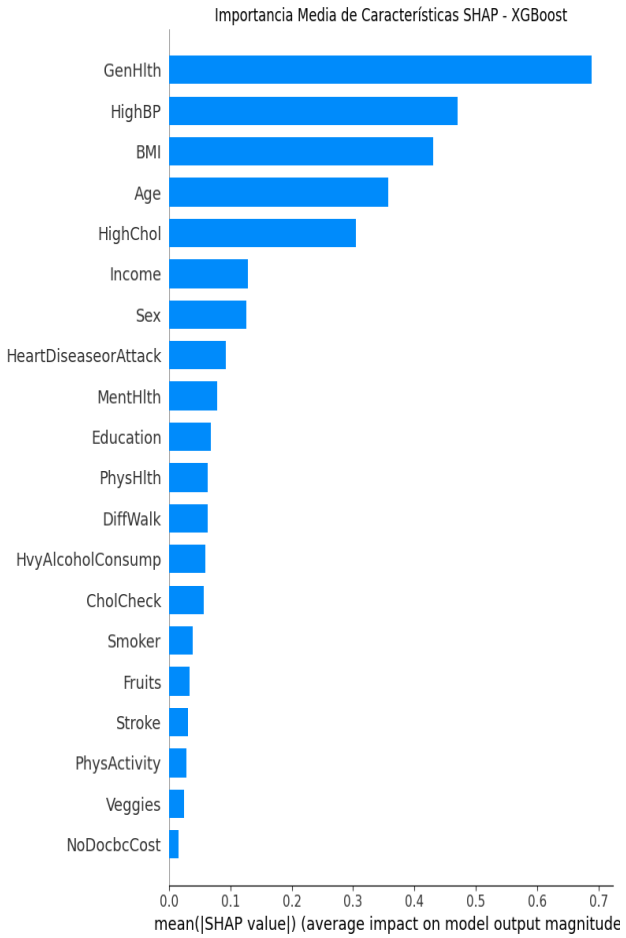

Figure 14. SHAP Values Importance - XGboost

The SHAP bar plot for XGBoost confirms the feature importance hierarchy observed in the beeswarm plot, with GenHlth emerging as the variable with the highest average impact, followed by BMI, Age, and HighBP. PhysHlth, Income, and Education also show relevant contributions. This hierarchy aligns with the feature importance calculated by XGBoost and the top predictors identified by the Random Forest model, reinforcing the robustness of these factors in predicting diabetes risk.

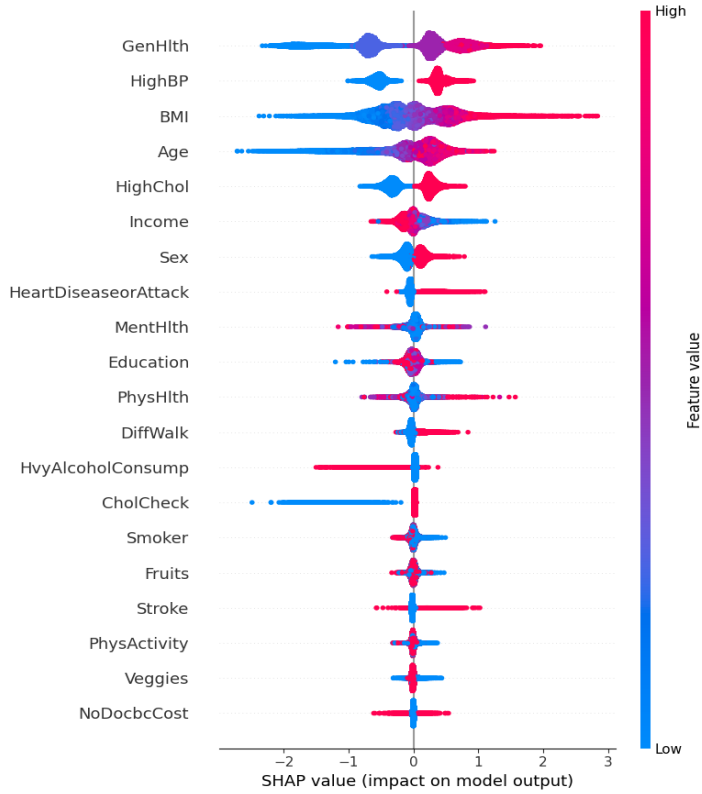

Figure 15. SHAP Dependence Plot– Xgboost

The SHAP dependence plot for GenHlth in the XGBoost model illustrates how general health perception influences the predicted risk of type 2 diabetes. As the value of GenHlth decreases (from 5 = excellent to 1 = poor health), the SHAP value increases significantly, indicating a greater contribution to a positive classification. Specifically, levels 1 and 2 of perceived health exhibit a strong positive effect on the prediction. In contrast, higher values of GenHlth (4 or 5) are associated with negative contributions, reducing the estimated risk.

The BMI-based color gradient reveals a relevant interaction: for the same GenHlth level, individuals with a higher body mass index tend to exhibit higher SHAP values. This suggests that the combination of poor perceived general health and elevated BMI amplifies their joint effect on the diabetes prediction.

The resulting prediction (log-odds = -4.15) is significantly below the model's base value (-1.188), indicating a strong inclination toward the "No Diabetes" class. The most influential variable in this direction is GenHlth = 5.00 (excellent general health), which notably reduces the estimated risk. This is followed by HighBP = 0.00 (absence of hypertension), Age = 1.00 (youngest age group), and BMI = 22.00 (healthy BMI range), all associated with a lower probability of diagnosis. In contrast, CholCheck = 0.00 (lack of recent cholesterol screening) is the primary factor that slightly increases the risk, along with the absence of vegetable

(Veggies = 0.00) and fruit (Fruits = 0.00) consumption.

However, the combined influence of these risk factors is insufficient to outweigh the dominant protective effects, resulting in a final prediction clearly favoring the "No Diabetes" classification.

Model Comparison: To contrast the overall performance of the evaluated models, a systematic comparison of key metrics was conducted for each algorithm. Figure X summarizes the values of accuracy, precision, recall, and F1-score for both classes (0 = No Diabetes, 1 = Diabetes), enabling an assessment of each model's balance between detection capability and classification precision. This comprehensive comparison facilitates the identification of relative strengths and limitations in the predictive behavior of XGBoost, Random Forest, SVM, and Logistic Regression on the test dataset.

	Modelo	Accuracy	Precision Clase 0	Precision Clase 1	Recall Clase 0	Recall Clase 1	F1 Clase 0	F1 Clase 1
0	XGBoost	0.748	0.771	0.729	0.707	0.790	0.737	0.758
1	Random Forest	0.747	0.772	0.726	0.702	0.792	0.735	0.758
2	SVM	0.746	0.779	0.720	0.687	0.805	0.730	0.760
3	Regresión Logística	0.746	0.759	0.733	0.719	0.772	0.739	0.752

Figure 16. Model Performance Comparison

The comparative analysis of the evaluated models—XGBoost, Random Forest, Support Vector Machine (SVM), and Logistic Regression—reveals similar performance in terms of overall accuracy, with values ranging between 74.5% and 74.8%. However, when disaggregating the metrics by class, relevant differences emerge in the behavior of each algorithm, allowing for a nuanced understanding of their utility depending on the analytical objective.

The SVM model achieved the highest precision for class 0 (no diabetes), with a value of 0.779, indicating a high degree of accuracy in correctly classifying healthy individuals. However, its recall for this same class was the lowest among all models (0.687), suggesting a higher proportion of false negatives due to the failure to detect all actual negative cases. On the other hand, XGBoost and Random Forest offered more balanced performance between precision and recall in both classes, with F1-scores above 0.75. XGBoost attained the most robust value (0.758 for class 1), standing out for its consistency and generalization capacity.

Logistic Regression, although slightly lower in overall accuracy, demonstrated competitive performance, with the best recall for class 0 (0.719), making it useful in scenarios where the priority is to maximize the detection of negative cases. Additionally, its interpretative nature makes it a valuable option when the contribution of individual variables to the predicted risk must be explained.

In summary, XGBoost stands out as the model with the best overall performance, offering an optimal balance between precision, recall, and F1-score. Random Forest follows closely, with consistent results. SVM excels in precision for negative cases but has limitations in sensitivity, while Logistic Regression remains a solid and interpretable alternative.

Results:

After preprocessing and data partitioning, four supervised classification algorithms were trained and evaluated: Logistic Regression, Random Forest, Support Vector Machine (SVM), and XGBoost. All models were validated on the same test set. Overall accuracy remained stable around 75% for all models, with XGBoost achieving the highest value (0.748), followed closely by Random Forest (0.747), SVM (0.746), and Logistic Regression (0.746).

Metric Analysis:

When examining class-specific behavior, significant differences were observed. XGBoost and Random Forest showed balanced performance in both classes, with F1-scores above 0.75. The SVM model achieved the highest precision for the negative class (0.779) but had the lowest recall (0.687), indicating that it did not effectively identify all negative cases. In contrast, Logistic Regression exhibited good recall for the negative class (0.719) and the highest precision for the positive class (0.733), making it competitive in terms of sensitivity.

Confusion Matrices:

The confusion matrices revealed important patterns in prediction errors. For example, the XGBoost model correctly classified 5,636 positive cases, with 1,433 false negatives. Random Forest showed similar performance, but with a slightly higher number of type II errors. SVM recorded the fewest false positives, though at the cost of reduced sensitivity. Across all models, the classification errors reflected the challenge of balancing specificity and sensitivity in clinical contexts.

Figure 16 summarizes the key metrics for each model. XGBoost stands out for its more balanced overall performance, followed by Random Forest. SVM offers high precision but suffers in recall, and Logistic Regression—though slightly less accurate—maintains good sensitivity for the negative class and provides interpretability. Overall, XGBoost emerges as the preferred option for this problem, though the final selection should consider the costs associated with type I and type II errors, as well as the need for interpretability.

VII. CONCLUSIONS

XGBoost stands out as the best-performing model for predicting type 2 diabetes risk based on self-reported variables, achieving an optimal balance between accuracy, recall, and F1-score across both classes. Its ability to handle nonlinear relationships and capture interactions between

variables makes it a robust tool for binary classification tasks in public health.

Random Forest delivers competitive and stable performance, with results comparable to XGBoost. Its lower configuration complexity and strong interpretability of features make it a practical option when balancing accuracy with efficiency is a priority.

SVM showed high precision for negative cases but limited recall, which may lead to a higher number of undetected patients (false negatives). This positions it as a useful model in scenarios where minimizing false positives is critical, but it is not ideal when the goal is to maximize positive case detection.

Logistic Regression, while outperformed by ensemble models in terms of performance, remains relevant due to its interpretability and adequate recall for the negative class. Its value lies especially in contexts that require transparent decision-making and an explicit understanding of risk factors.

The most relevant variables were consistent across models, with self-perceived general health, body mass index, age, and high blood pressure emerging as the strongest predictors of diabetes risk. This convergence reinforces the robustness of these factors as critical indicators in clinical decision support systems.

The use of SHAP explainability complemented the quantitative evaluation by enabling the analysis of both global patterns and individual predictions, which is essential for the adoption of models in clinical or community settings that demand trust and traceability.

Overall, the results support the feasibility of using machine learning models to assist in early diabetes detection strategies in populations with limited access to clinical testing, relying on reliable self-reported data and explainable models that can be integrated into digital public health tools.

VIII. REFERENCES

[1] J. Ren, “Predictions of diabetes through machine learning models based on the health indicators dataset,” *ResearchGate*, 2023. [Online]. Available: https://www.researchgate.net/publication/377829779

[2] K. Sampath, S. Rezaei, y T. Nguyen, “Robust framework for early-stage diabetes prediction using ensemble learning and oversampling techniques,” *Scientific Reports*, vol. 14, no. 1, 2024, Art. no. 78519. [Online]. Available: https://doi.org/10.1038/s41598-024-78519-8

[3] M. S. Hossain, F. Faisal, F. Akter, y S. J. Ferdous, “Machine learning based prediction and insights of diabetes disease: Pima Indian and Frankfurt datasets,” *ResearchGate*, 2025. [Online]. Available: https://www.researchgate.net/publication/388148101

[4] J. Riveros-Pérez et al., “Predictive performance of machine learning algorithms in diabetes diagnosis using the NHANES database,” *BMJ Open Diabetes Res Care*, vol. 13, no. 1, 2025. doi: 10.1136/bmjdrc-2024-003678

[5] T. Wang et al., “Application of machine learning to predict diabetes in Taiwan: An analysis using electronic health records,” *BMC Endocrine Disorders*, vol. 24, no. 3, 2024, pp. 112–123.

[6] T. Dinh, D. T. Nguyen, y A. Yamada, “XGBoost-based models for diabetes prediction using NHANES data,” *Journal of Biomedical Informatics*, vol. 137, 2023, Art. no. 104329.

[7] C. Mejía, A. Rodríguez, y S. Becerra, “Predicción de diabetes tipo 2 con aprendizaje automático en población colombiana sin biomarcadores clínicos,” *Revista Colombiana de Salud Pública*, vol. 25, no. 2, pp. 134–143, 2023.

[8] L. García-Quezada, A. K. Rahman, y M. N. Islam, “Diabetes prediction using clinical symptoms and machine learning algorithms: Evidence from Bangladesh,” *Computers in Biology and Medicine*, vol. 169, 2023, Art. no. 106691.