JOURNAL OF INTERDISCIPLINARY PERSPECTIVES

ISSN Print: 2984-8288, ISSN Online: 2984-8385
Vol. 3 No. 8, 689-698, August 2025

Predictive Modeling of Diabetes Classification using Binomial Logistic Regression on Biomedical Indicators

Jojie E. Campugan*, Melani G. Aguaras
College of Development Management, University of Southeastern Philippines, Davao City, Philippines

*Corresponding Author Email: jecampugan02202400400@usep.edu.ph

Date received: May 29, 2025
Date revised: June 28, 2025
Date accepted: July 18, 2025

Originality: 95%
Grammarly Score: 99%
Similarity: 5%

Recommended citation:
Campugan, J., & Aguaras, M. (2025). Predictive modeling of diabetes classification using binomial logistic regression on biomedical indicators. *Journal of Interdisciplinary Perspectives*, *3*(8), 689-698.
https://doi.org/10.69569/jip.2025.465

Abstract. Diabetes mellitus remains a significant health and economic burden globally, with early detection frequently delayed in resource-limited settings such as the Philippines. Addressing this gap, the present study aimed to develop a predictive model for classifying individuals as diabetic or non-diabetic using biomedical indicators, including Body Mass Index (BMI), Low-Density Lipoprotein (LDL), Glycated Hemoglobin (HbA1c), and Triglycerides. Guided by Roy Baumeister's Self-Regulation Theory of Illness Behavior, the study employed a multi-method classification approach involving binomial logistic regression, K-means clustering, and decision tree analysis. A total of 947 participants aged 24 to 79 years were included. K-means clustering categorized participants into two distinct groups based on biomarker profiles, differentiating those at higher and lower risk of diabetes. Logistic regression identified BMI as the most significant predictor ($\chi^2(1) = 104.44$, $p < .001$), followed by HbA1c ($\chi^2(1) = 51.80$, $p < .001$), Triglycerides ($\chi^2(1) = 12.44$, $p < .001$), and LDL ($\chi^2(1) = 9.15$, $p = .002$). The model demonstrated excellent predictive performance, with an McFadden R^2 of 0.80 and a Nagelkerke R^2 of 0.85. Decision tree analysis confirmed BMI as the primary classifier, with HbA1c enhancing classification accuracy, thereby highlighting the combined diagnostic utility of both. These findings suggest that incorporating BMI and HbA1c thresholds as accessible, cost-effective screening tools within barangay health systems could improve early identification of individuals at risk for diabetes. Integrating predictive analytics with behavior modification programs based on self-regulation theory may empower communities to adopt preventive health measures. The study recommends prioritizing risk-based screening protocols, subsidizing access to essential biomarker testing, and integrating predictive modeling frameworks into primary healthcare. This multi-method model presents a robust, scalable tool to enhance diabetes risk prediction and support targeted health interventions in underserved Philippine communities.

Keywords: Biomedical indicators; Diabetes classification; Predictive analytics; Self-regulation theory.

1.0 Introduction

Diabetes mellitus is a chronic non-communicable disease characterized by sustained hyperglycemia resulting from impaired insulin secretion, action, or both (Hossain et al., 2024). Its global prevalence has risen dramatically, with cases increasing from approximately 200 million in 1990 to over 830 million by 2022, doubling adult prevalence rates to 14% (WHO, 2024). This growing burden disproportionately affects low- and middle-income countries, where healthcare systems often struggle with limited diagnostic resources, high out-of-pocket costs, and delayed interventions. Early detection remains a critical challenge, and without timely diagnosis, individuals

This work is licensed under a Creative Commons Attribution-Noncommercial 4.0 International License (CC BY-NC 4.0).

face elevated risks of cardiovascular disease, neuropathy, and premature mortality (Sun et al., 2022).

In the Philippines, diabetes affects an estimated 6.3% of adults, translating to over 4.3 million diagnosed cases as of 2021, with an additional 2.8 million remaining undiagnosed (Cando et al., 2024). While national guidelines conform to global management protocols, socioeconomic disparities hinder equitable access to care, particularly in rural communities. The reliance on costly laboratory diagnostics such as oral glucose tolerance tests or fasting blood glucose screening limits widespread early detection. Consequently, there is a pressing need for cost-effective and scalable screening tools that utilize easily obtainable biomedical indicators.

Recent studies have highlighted the predictive value of biomedical markers, including Body Mass Index (BMI), Low-Density Lipoprotein (LDL) cholesterol, Glycated Hemoglobin (HbA1c), and Triglycerides, in identifying diabetes risk (Weng et al., 2021; Abdel-Razek et al., 2020). However, many existing predictive models are constrained by binary classification approaches or univariate analyses that neglect potential interaction effects among biomarkers. Moreover, multi-method classification frameworks combining traditional regression with data-driven techniques remain underexplored, particularly in Southeast Asian contexts (Zhu et al., 2023). To address this gap, the present study developed a predictive model for classifying individuals as diabetic or non-diabetic based on the aforementioned biomedical markers. The study employed binomial logistic regression to determine the individual and interactive predictive strength of these indicators, supplemented by K-means clustering and decision tree analysis to enhance classification accuracy. Grounded in Roy Baumeister's Self-Regulation Theory of Illness Behavior, which posits that health risk awareness prompts behavior modification, this study aimed not only to develop a robust, multi-method classification model but also to inform community-based preventive health initiatives. By integrating predictive analytics with behavior-oriented health promotion strategies, the research aims to contribute to more accessible and efficient diabetes risk screening practices within the Philippine primary healthcare system.

The primary objective of this study was to determine the predictive strength of selected biomedical indicators using a binomial logistic regression model. Specifically, it aimed to (1) assess the individual predictive value of BMI, LDL cholesterol, HbA1c, and triglycerides on diabetes status; (2) evaluate their interaction effects in predicting diabetes classification; and (3) explore how the predicted probability of being diabetic varies across different levels of these biomarkers. To strengthen the model's predictive accuracy, the study also employed K-means clustering and decision tree analysis. Anchored in Roy Baumeister's Self-Regulation Theory of Illness Behavior, the study viewed risk awareness—such as identifying elevated biomarker levels—as a catalyst for behavior change. Through this lens, the research aimed not only to develop a statistically sound classification model but also to contribute to community-level preventive health strategies.

2.0 Methodology

This study employed a combination of statistical techniques to develop a predictive model for classifying individuals as diabetic or non-diabetic based on clinical indicators. A total of 947 individuals participated in the study, comprising 435 females and 512 males, aged 24 to 79 years. The study utilized three statistical methods. Binomial Logistic Regression, an estimation method for predicting binary outcomes(Harris, 2021), was used to assess the likelihood of diabetes occurrence based on Body Mass Index (BMI), LDL cholesterol, HbA1c, and triglyceride levels. K-means Clustering was applied to identify homogeneous groups within the dataset(Sinaga & Yang, 2020), enabling the identification of clusters of individuals with similar clinical profiles. Decision Tree Analysis, a statistical tool that graphically represents the decision-making process under specified conditions, (Bansal et al., 2022)was utilized to develop interpretable classification rules. These methods were integrated to improve the accuracy, robustness, and interpretability of the predictive model for diabetes classification.

2.1 Research Design

This research utilized a quantitative predictive research design aimed at classifying individuals as diabetic or non-diabetic based on selected biomedical indicators. The study aimed to identify statistically significant relationships among biomedical variables and to develop a model that could accurately predict diabetes status.

2.2 Research Instrument

This study utilized secondary data sourced from Kaggle, last updated in April 2025, which includes variables such as Gender, Age, HbA1c (Glycated Hemoglobin), TG (Triglycerides), LDL (Low-Density Lipoprotein Cholesterol Level), BMI (Body Mass Index), and the class variable (1-Diabetic or 0-Non-diabetic). To analyze the data, JAMOVI

software version 2.3.28 was used, which provided core functions such as data entry and manipulation, rule-based data filtering, variable transformation, and support for advanced statistical techniques, including binomial logistic regression, K-means clustering, and decision tree analysis. Binomial logistic regression predicted diabetes risk based on these variables, K-means clustering identified natural groupings within the data, and decision tree analysis provided an interpretable classification model. These methods enabled a comprehensive understanding of the key factors influencing diabetes classification.

2.3 Data Gathering Procedure

This study utilized secondary data obtained from the Kaggle online repository (Patel, 2025), which was last updated in April 2025. The dataset was downloaded in CSV format and contained anonymized clinical records relevant to diabetes prediction, including variables such as Body Mass Index (BMI), LDL cholesterol, HbA1c, and triglyceride levels. Prior to analysis, the dataset underwent an initial screening to identify missing values, duplicates, and inconsistencies. Necessary data cleaning procedures, such as handling null values and standardizing variable formats, were applied to ensure data quality and integrity. The cleaned dataset was then prepared for statistical modeling using logistic regression, K-means clustering, and decision tree analysis.

2.4 Data Analysis Procedure

Data analysis was conducted using JAMOVI software version 2.3.28, which facilitated the execution of multiple statistical techniques to develop a predictive model for diabetes classification. Initially, data cleaning and preparation were performed, including handling missing values and verifying data consistency. Binomial logistic regression was employed to estimate the probability of diabetes occurrence based on selected clinical indicators, assessing the significance and strength of predictors such as BMI, LDL cholesterol, HbA1c, and triglycerides. Subsequently, K-means clustering was applied to identify natural groupings within the dataset, revealing clusters of individuals with similar biomedical profiles. Finally, decision tree analysis was used to generate clear, interpretable classification rules, aiding in the visualization and understanding of the decision-making process behind diabetes status prediction. The integration of these methods allowed for a robust, multi-faceted analysis that improved the accuracy and interpretability of the predictive model.

2.5 Ethical Considerations

This study used anonymized secondary data from Kaggle, which contained no personally identifiable information. Data handling followed standard ethical guidelines, including secure storage and honest reporting. The software and tools used in the analysis included components licensed under the MIT License. In compliance with the license, the original copyright notices from Mark Otto (2013) and Andrew Fong (2017) were acknowledged. The license permits unrestricted use, distribution, and modification of the software, provided that proper attribution is maintained. The software is provided "as is" without warranties or liabilities.

3.0 Results and Discussion

This section presents the results of the binomial logistic regression, K-means clustering, and decision tree analysis applied to biomedical indicators (BMI, HbA1c, LDL, and triglycerides) to classify diabetes status. The findings were clearly illustrated through tables and figures, demonstrating the interactions between these markers and their impact on diabetes classification.

3.1 Binomial Logistic Regression

Table 1 presents the results of the binomial logistic regression model, which demonstrated strong performance in predicting diabetes status. The model significantly outperformed the null model, as indicated by the omnibus chi-square test, $\chi^2(5) = 523$, $p < .001$, and showed a low deviance value of 128, suggesting it explained a large portion of the variability. Model fit was further supported by the Akaike Information Criterion (AIC = 140) and Bayesian Information Criterion (BIC = 169), indicating model efficiency. The high pseudo R-squared values (McFadden's $R^2 = .80$, Nagelkerke's $R^2 = .85$, and Tjur's $R^2 = .78$) confirmed a strong relationship between the predictors and the outcome, while the Cox & Snell R^2 (.42) reflected a moderate fit. These results indicated that the selected predictors, including BMI, HbA1c, and lipid profiles, provided an accurate classification of diabetes status, highlighting the model's clinical utility for risk assessment.

Table 1. *Overall Model Fit Statistics for Diabetes Classification*

Model	Deviance	AIC	BIC	R^2_{McF}	R^2_{CS}	R^2_N	R^2_T	Overall Model Test		
								χ^2	df	P-value
1	128	140	169	0.80	0.42	0.85	0.78	523	5	<.001

These findings align with recent studies emphasizing the importance of integrating anthropometric and biochemical markers in predictive models to improve diabetes risk identification.Ojulari et al., (2023) demonstrated that combining BMI, HbA1c, and lipid profiles enhances early detection, particularly in resource-constrained settings. Consistent with Baumeister's Self-Regulation Theory of Illness Behavior, Khan et al. (2022) emphasized that biomarker-based risk awareness fosters engagement in preventive behaviors, such as healthier dietary habits and increased physical activity. Integrating predictive models based on accessible indicators like BMI and HbA1c into community-level screening programs could improve early detection rates, promote proactive health management, and reduce the growing diabetes burden, particularly in underserved populations (Bandi, M., et. al., 2024).

Table 2 presents the omnibus likelihood ratio tests, highlighting the contribution of each predictor in the binomial logistic regression model. The results showed that BMI was the strongest predictor of diabetes status, $\chi^2(1) = 104.44$, $p < .001$, accounting for a significant portion of the variance. HbA1c also contributed significantly, $\chi^2(1) = 51.80$, $p < .001$, reinforcing its clinical importance. Triglycerides (TG), $\chi^2(1) = 12.44$, $p < .001$, and LDL, $\chi^2(1) = 9.15$, $p = .002$, also showed significant effects, though weaker than BMI and HbA1c. The interaction term combining BMI, LDL, HbA1c, and TG was marginally significant, $\chi^2(1) = 3.83$, $p = .050$, suggesting a minor effect of their combined influence. These findings underscored BMI and HbA1c as key predictors, with lipid profile indicators offering additional diagnostic value.

Table 2. *Omnibus Likelihood Ratio Tests for Predictors in the Logistic Regression Model*

Predictor	χ^2	df	P-value
BMI	104.44	1	< .001
LDL	9.15	1	0.002
HbA1c	51.80	1	< .001
TG	12.44	1	< .001
BMI ✻ LDL ✻ HbA1c ✻ TG	3.83	1	0.050

This study's findings have practical implications for early diabetes detection, as predictive models based on BMI, HbA1c, and lipid profiles can identify individuals at risk for timely intervention. According to Baumeister's Self-Regulation Theory of Illness Behavior, individuals who are informed about their health risks are more likely to adopt health-promoting behaviors (Leary & Tangney, 2003). Supporting this, research has shown that awareness of diabetes risk encourages individuals to adopt healthier lifestyle choices, such as an improved diet and increased physical activity. Thus, integrating predictive models into screenings can enhance both early intervention and self-regulation of health behaviors.

As shown in Table 3, logistic regression analysis revealed that HbA1c (B = 1.85, SE = 0.31, z = 6.07, $p < .001$), triglycerides (TG; B = 1.61, SE = 0.35, z = 4.56, $p < .001$), LDL cholesterol (B = 0.96, SE = 0.28, z = 3.45, $p < .001$), and BMI (B = 0.83, SE = 0.14, z = 5.97, $p < .001$) significantly increased the odds of being classified as diabetic. The corresponding odds ratios were 6.37 for HbA1c, 5.02 for TG, 2.60 for LDL, and 2.30 for BMI. Additionally, a significant four-way interaction among these variables was observed (B = −0.0021, SE = 0.0007, z = −3.06, p = .002), suggesting a modest moderating effect. Collectively, these results indicated that elevated levels of these biomarkers substantially increased the likelihood of diabetes classification.

Table 3. *Model Coefficients for Predicting Diabetes Status (Diabetic vs non-diabetic)*

Predictor	95% Confidence Interval			SE	Z	P-value	95% Confidence Interval		
	Estimate	Lower	Upper				Odds ratio	Lower	Upper
Intercept	-33.45	-43.01	-23.89	4.88	-6.86	< .001	2.97e-15	2.10e-19	4.20e-11
BMI	0.83	0.56	1.11	0.14	5.97	< .001	2.301	1.75	3.02
LDL	0.96	0.41	1.50	0.28	3.45	< .001	2.604	1.51	4.49
HbA1c	1.85	1.25	2.45	0.30	6.07	<.001	6.372	3.51	11.58
TG	1.61	0.92	2.31	0.35	4.56	< .001	5.021	2.51	10.05
BMI ✻ LDL ✻ HbA1c ✻ TG	-0.002	-0.003	-7.69e−4	6.96e-4	-3.06	0.002	0.998	0.997	0.999

These findings confirm the predictive value of biomedical indicators for diabetes risk. (Han et al., 2023) Found

that elevated HbA1c levels strongly predict diabetes, often outperforming other single biomarkers, while Khan et al. (2022) demonstrated that triglycerides and LDL cholesterol significantly enhance risk stratification when combined with BMI. Wang et al., (2024) Likewise, they emphasized that multi-biomarker models integrating HbA1c, BMI, and lipid profiles offer superior predictive accuracy compared to isolated measures. From a behavioral perspective, Aldubayan et al., (2022) researchers reported that biomarker-based risk awareness improves engagement in health-promoting behaviors, supporting Baumeister's Self-Regulation Theory of Illness Behavior and updated frameworks Greene et al., (2023), which highlight how risk perception fosters proactive health management and positions health as a personal, behavior-driven goal.

Table 4 presents the multicollinearity diagnostics for the predictors used in the binomial logistic regression model. Variance Inflation Factor (VIF) values for BMI (1.50), LDL (1.84), HbA1c (2.59), and Triglycerides (TG; 2.88) were all well below the conventional cutoff of 10, indicating no significant multicollinearity concerns. Tolerance values similarly exceeded the 0.10 threshold, confirming acceptable levels of collinearity among predictors. Although the interaction term (BMI ✻ LDL ✻ HbA1c ✻ TG) displayed a moderately higher VIF of 4.12, it remained within acceptable limits, ensuring stable and interpretable regression estimates. These results affirm the independence of each biomedical marker, enhancing the reliability of their contributions to diabetes risk prediction.

Table 4. *Collinearity Statistics for Logistic Regression Predicting Diabetes Status*

Predictor	VIF	Tolerance
BMI	1.50	0.67
LDL	1.84	0.54
HbA1c	2.59	0.39
TG	2.88	0.35
BMI ✻ LDL ✻ HbA1c ✻ TG	4.12	0.24

This statistical clarity has important clinical and public health implications, as it enables healthcare providers to interpret individual risk factors without confounding overlaps, improving the precision of patient education and intervention strategies. (Bayman & Dexter, 2021) Noted that high multicollinearity can obscure risk assessments, while Beaudart et al., (2021) emphasizing that clear, individualized risk communication enhances patient engagement, self-management, and adherence to preventive behaviors. Maintaining low multicollinearity not only sharpens clinical decision-making but also strengthens patient-centered health messaging for proactive diabetes risk management.

Table 5 presented the estimated probabilities of being diagnosed with diabetes based on three important health measures: glycated hemoglobin (HbA1c), low density lipoprotein cholesterol levels (LDL), and body weight (BMI). The data showed that as any of these measures increase, so does the likelihood of diabetes. Specifically, across BMI levels from −1 standard deviation (25.0) to +1 standard deviation (34.8) and LDL levels from −1 SD (1.49) to +1 SD (10.95), the probability of diabetes rose with higher HbA1c values. Even at lower BMI and LDL levels, the chance of being diabetic was already high, ranging from about 92% to nearly 100%. At higher BMI and LDL levels, this probability approached certainty, regardless of the HbA1c level, indicating that these factors work together to increase the risk of diabetes significantly. These findings highlight that when blood sugar, cholesterol, and body weight are all unfavorable, the chance of remaining non-diabetic is almost zero.

These findings have important implications for diabetes prevention in the Philippines. Since the chance of having diabetes is already high when even one or two health markers are elevated, early and combined screening is essential. However, many Filipinos delay seeking care due to a lack of health knowledge, money, or trust in the healthcare system. According to Tolabing et al. (2022), many people struggle to understand health information, while Pagaddu (2021) noted that poor health and poverty often go hand in hand, which is furthered by Baumeister's Self-Regulation Theory of Illness Behavior, suggesting that people may neglect their health when cognitive, emotional, or environmental resources are limited. To address these challenges, community-based programs—such as free barangay-level screenings, trained health workers, and mobile health tools—can help individuals better understand their risk and take timely action to manage it. These simple yet targeted efforts can significantly reduce the burden of diabetes, especially in poor and rural areas.

Table 5. *Estimated Marginal Means of the Probability of Being Diabetic by Levels of HbA1c, LDL, and BMI*

BMI	LDL	HbA1c	Probability	SE	95% Confidence Interval	
					Lower	Upper
25.0-	1.49-	5.86-	0.92	0.04	0.78	0.97
		8.41 μ	0.10	0.00	0.99	1.00
		10.95+	1.00	3.36e-5	0.10	1.00
	2.62 μ	5.86-	0.93	0.03	0.85	0.97
		8.41 μ	0.10	0.00	0.99	1.00
		10.95+	1.00	4.88e-5	0.10	1.00
	3.74+	5.86-	0.95	0.02	0.89	0.98
		8.41 μ	0.10	0.00	0.99	1.00
		10.95+	1.00	7.38e-5	0.10	1.00
29.9 μ	1.49-	5.86-	0.10	0.00210	0.98	1.00
		8.41 μ	1.00	4.89e-5	0.10	1.00
		10.95+	1.00	1.06e-6	1.00	1.00
	2.62 μ	5.86-	0.10	0.00	0.99	1.00
		8.41 μ	1.00	6.34e-5	0.10	1.00
		10.95+	1.00	2.10e-6	1.00	1.00
	3.74+	5.86-	0.10	0.00	0.99	1.00
		8.41 μ	1.00	8.64e-5	0.10	1.00
		10.95+	1.00	4.33e-6	1.00	1.00
34.8+	1.49-	5.86-	1.00	7.08e-5	0.10	1.00
		8.41 μ	1.00	1.55e-6	1.00	1.00
		10.95+	1.00	3.36e-8	1.00	1.00
	2.62 μ	5.86-	1.00	7.36e-5	0.10	1.00
		8.41 μ	1.00	2.58e-6	1.00	1.00
		10.95+	1.00	9.09e-8	1.00	1.00
	3.74+	5.86-	1.00	7.88e-5	0.10	1.00
		8.41 μ	1.00	4.46e-6	1.00	1.00
		10.95+	1.00	2.56e-7	1.00	1.00

Table 6 presents the classification results of the predictive model for diabetes status, using a 0.5 probability threshold. The model correctly identified 828 out of 844 actual diabetic cases, yielding a sensitivity (true positive rate) of 98.10%. Meanwhile, it accurately classified 94 out of 103 non-diabetic cases, resulting in a specificity (actual negative rate) of 91.30%. These findings demonstrate strong discriminatory power of the model, particularly in detecting diabetic individuals, with only a small number of false negatives (n = 16) and false positives (n = 9). The high overall correct classification rates suggest that the model is well-calibrated at the 0.5 cut-off and capable of performing with excellent predictive accuracy in clinical or population screening contexts. Such performance indicates its potential utility in early identification strategies, where correctly identifying at-risk individuals is critical to timely intervention and management.

Table 6. *Classification Table for Predicting Diabetes Status at a 0.5 Cut-Off*

Observed	Predicted		% Correct
	Diabetic	Non-Diabetic	
Diabetic	828	16	98.10
Non-Diabetic	9	94	91.30

Note. The cut-off value is set to 0.5

With only a small number of false results, the model demonstrated a strong potential for early screening and intervention, which is especially important in the Philippine context. Many Filipinos face cultural (Pillai, 2021), financial (Yilmaz, 2024), and informational barriers that delay recognition of diabetes symptoms and seeking care. This predictive tool can provide objective, data-driven assessments even before symptoms appear, helping individuals recognize their risk earlier. Such early identification can encourage timely self-regulation—prompting people to seek medical advice, adopt healthier lifestyles, or adhere to treatment—thus improving outcomes. Given the rising diabetes prevalence in the Philippines, which reached 8.2 percent among adults in 2021 according to the Daily Tribune in 2024, integrating this model into barangay-level health programs and primary care could offer a scalable, evidence-based approach to curb the growing burden of diabetes nationwide.

As shown in Table 7, the diabetes classification model demonstrated strong predictive performance, with a sensitivity of 0.98, a specificity of 0.91, and an Area Under the Curve (AUC) of 0.99. The near-perfect AUC suggests the model has excellent discriminative ability, accurately distinguishing between diabetic and non-diabetic individuals across varying thresholds. High sensitivity indicates the model's capacity to effectively detect

individuals with diabetes, while strong specificity reflects its reliability in minimizing false-positive results. These metrics collectively confirm the robustness of the model in both identifying at-risk individuals and preventing misclassification, critical characteristics for population-based screening and early detection initiatives.

Table 7. *Predictive Performance Metrics for the Diabetes Classification Model*

Accuracy	Specificity	Sensitivity	AUC
0.97	0.98	0.91	0.99

Note. The cut-off value is set to 0.5

In the real world, this means the model can be confidently used in health screenings—such as in community clinics or mobile health programs—to identify people at risk early on. This targeted approach helps doctors and health workers focus care on those who need it most, avoiding unnecessary worry or tests for others, and making diabetes prevention and management more effective and efficient. Such applications highlight machine learning's potential as a valuable and cost-effective tool for early diabetes detection and monitoring, especially in resource-limited settings, providing important guidance for healthcare policy and diabetes management in countries facing similar challenges (Adua et al., 2021), such as the Philippines.

Figure 1 illustrates the Receiver Operating Characteristic (ROC) curve for the diabetes classification model, plotting sensitivity (actual positive rate) against 1 – specificity (false positive rate). The steep ascent of the red line toward the top-left corner demonstrates the model's high sensitivity and specificity across a range of classification thresholds. This visually confirms the model's excellent discriminative power, correctly identifying a high proportion of actual positive cases (diabetic individuals) with relatively few false positives. The ROC curve's proximity to the upper-left boundary aligns with the reported AUC of 0.99, reflecting near-perfect classification accuracy. The early, steep rise along the y-axis suggests that the model achieves high sensitivity even with low false-positive rates, making it highly suitable for clinical screening contexts where early detection and minimizing false alarms are paramount.

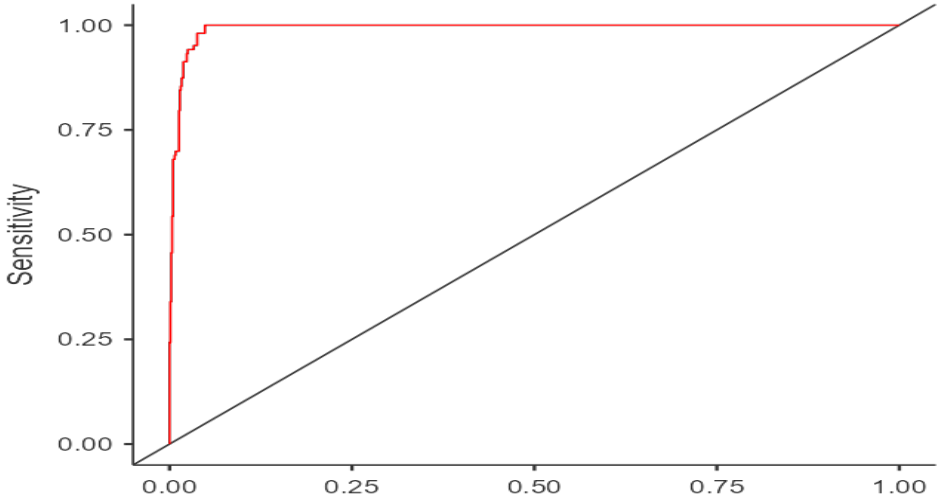

Figure 1. *Receiver Operating Characteristic (ROC) Curve for the Diabetes Classification Model*

The value of using predictive models like this lies in the decisions they help inform. Just as in quality checks, where measurements indicate whether something meets a standard, in healthcare, the model helps determine if a person should undergo further tests or treatment. The ROC curve is important because it shows how well the model tells apart people with and without diabetes (Pendrill et al., 2023). Grounded in Self-Regulation Theory, such data-driven feedback can also motivate individuals to take proactive steps—such as seeking medical advice or adjusting their lifestyle—when informed that they may be at risk.

3.2 K-means Clustering

This section presents the K-means clustering results, which grouped participants into two clusters based on biomarker values. The centroids revealed distinct patterns in HbA1c, triglycerides, LDL cholesterol, and BMI, offering insights into diabetes risk prediction.

Table 8. *Centroids of clusters Table*

	Cluster No	HbA1c	TG	LDL	BMI	CLASS
1	1.00	0.19	0.06	-0.00	0.12	0.35
2	2.00	-1.51	-0.52	0.01	-1.54	-2.86

Table 8 presents the K-means clustering results, which identify two distinct clusters based on key biomarkers associated with diabetes risk. Cluster 1 showed higher levels of HbA1c, triglycerides, and BMI, indicating a higher risk of diabetes, while Cluster 2 displayed lower levels of these biomarkers, suggesting a lower risk. These findings align with Roy Baumeister's Self-Regulation Theory of Illness Behavior, which suggests that individuals' cognitive and emotional responses to health risks influence their actions. In Cluster 1, individuals may adopt health-promoting behaviors due to their awareness of elevated biomarkers, while Cluster 2 individuals may delay intervention, perceiving themselves as low-risk.

In the Philippines, where healthcare resources are limited and diabetes remains a significant health burden, diabetes mellitus (DM) has consistently ranked among the top five causes of death from 2018 to 2022. In 2021 alone, 55,636 deaths were attributed to DM (Baron, 2024). Using K-means clustering to group individuals by risk—based on key biomarkers such as HbA1c, triglycerides, and BMI—provides a practical strategy to improve care delivery. This approach enables healthcare providers to prioritize high-risk individuals for immediate medical intervention and lifestyle counseling, while directing low-risk individuals toward preventive care, ultimately improving outcomes and resource efficiency.

3.3 Decision Tree

The statistical result, illustrated in the figure below, was derived from the decision tree analysis. This analysis visually demonstrates how BMI, HbA1c, LDL, and triglyceride levels influence diabetes classification. The decision tree highlights key thresholds and decision nodes, providing clear and actionable rules for identifying individuals at risk of diabetes, thereby supporting early intervention and effective risk management strategies.

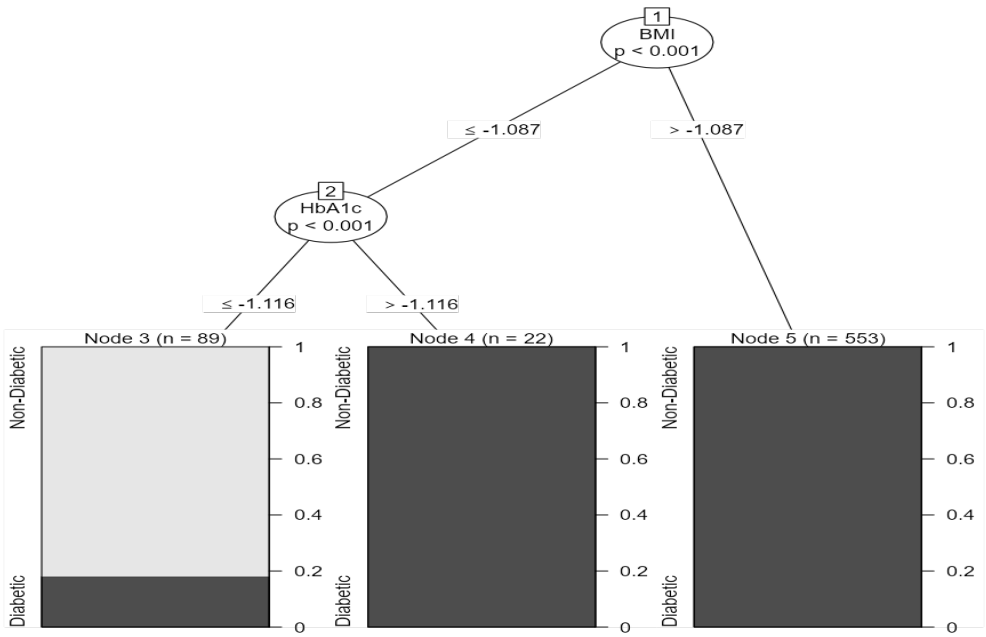

Figure 2. *Decision Tree Plot*

Figure 2 illustrates a classification decision tree that predicts diabetes status based on BMI and HbA1c. The tree first splits on BMI, with those above -1.087 predominantly classified as diabetic, indicating a strong association between higher BMI and the risk of Type 2 diabetes. For individuals with lower BMI, the tree further splits on HbA1c, where those with elevated HbA1c are classified as diabetic, highlighting the importance of glycemic control in diabetes prediction. This aligns with established research linking obesity and poor glycemic control to diabetes.

In resource-limited settings, such as the Philippines, utilizing decision tree models based on BMI and HbA1c levels can significantly enhance diabetes risk stratification and management. The Philippine Practice Guidelines on the Diagnosis and Management of Diabetes Mellitus recommend laboratory testing for individuals with risk factors such as a BMI over 23 kg/m² or elevated triglyceride levels, highlighting the importance of targeted screening. A study Cando et al., (2024) reported that an estimated 4.3 million Filipinos were diagnosed with diabetes in 2021, with an additional 2.8 million remaining undiagnosed, underscoring the urgent need for practical and accessible screening tools. Decision tree models can support healthcare providers in identifying high-risk individuals for immediate intervention and directing low-risk individuals toward preventive care, thus improving health outcomes and optimizing limited healthcare resources. This strategy also reflects Roy Baumeister's Self-Regulation Theory of Illness Behavior, which posits that individuals' awareness of their health risks—such as having elevated BMI or HbA1c—can influence them to adopt healthier behaviors and seek timely medical attention.

4.0 Conclusion

This study evaluated the predictive value of BMI, LDL, HbA1c, and triglycerides in classifying diabetes status using binomial logistic regression, decision tree analysis, and K-means clustering. Addressing a key gap in Philippine research, it introduced a multi-method classification framework that integrates biomedical markers with behavioral insights—an approach rarely used in local studies. In achieving its objectives, the study first confirmed that all four biomarkers were significant predictors of diabetes, with HbA1c and BMI emerging as the most influential. Second, a marginally significant interaction effect revealed that analyzing biomarkers in combination provides greater predictive accuracy than treating them independently—an area that has been previously underexplored in localized models. Third, probability modeling and K-means clustering identified a high-risk subgroup with poor biomarker profiles. At the same time, the decision tree produced simple, interpretable rules suitable for use by community health workers. Collectively, these findings enhance diabetes risk stratification and provide practical tools for early detection at the community level.

The study also addressed behavioral knowledge gaps by applying Baumeister's Self-Regulation Theory of Illness Behavior, demonstrating how awareness of objective risk factors, such as BMI and HbA1c, can support preventive action in low-access healthcare settings. These results reinforce the importance of risk visibility in promoting self-monitoring and early intervention, particularly in high-prevalence, underserved populations.

It is therefore recommended that the Department of Health (DOH) and local government units adopt simplified, risk-based screening protocols at the barangay level. Community health workers should be trained to measure BMI and refer individuals at high risk for HbA1c testing. These actions should be supported by public education campaigns that align with self-regulation principles to improve health-seeking behavior. Additionally, the inclusion of HbA1c and lipid profile testing in PhilHealth coverage under the Universal Health Care (UHC) program is advised to ensure more equitable access to preventive diagnostics.

For future research, the model should be expanded to include behavioral and psychosocial predictors such as dietary habits, physical activity, stress, and health literacy to capture a more holistic risk profile. Longitudinal studies are recommended to validate the model over time and in varied population groups. Finally, integrating mobile health (mHealth) tools for real-time data collection and follow-up can enhance patient engagement and monitoring. These directions will help build a more responsive, data-driven, and inclusive public health strategy for diabetes prevention and management in the Philippines.

5.0 Contribution of Authors

Author 1: Data preparation and cleaning, dataset execution, statistical analysis, literature review, proofreading.
Author 2: Dataset execution, statistical analysis, results description, literature review.

6.0 Funding

This research received no specific grant or financial support from any funding agency, public or private. The researchers personally shouldered all costs related to the completion and publication of this paper.

7.0 Conflict of Interest

The authors declare that they have no known financial, personal, or professional conflicts of interest that could have influenced the conduct of the research, the analysis of the data, or the preparation and publication of this manuscript.

8.0 Acknowledgment

The researchers would like to express their sincere gratitude to Dr. Joeteddy Bugarin for his invaluable mentorship and expertise in data analytics, which significantly guided the analytical aspect of this study. Special thanks are also extended to Dr. John Vianne Murcia for his brilliance in teaching machine learning, which greatly enriched the researchers' methodological approach. The researchers are beyond grateful to have undertaken this study under the umbrella of the College of Development Management, University of Southeastern Philippines. The experience was never easy, but it has been genuinely fulfilling. To God be all the glory.

9.0 References

Aldubayan, M. A., Pigsborg, K., Gormsen, S. M. O., Serra, F., Palou, M., Mena, P., Wetzels, M., Calleja, A., Caimari, A., Del Bas, J., Gutierrez, B., Magkos, F., & Hjorth, M. F. (2022). Empowering consumers to PREVENT diet-related diseases through OMICS sciences (PREVENTOMICS): Protocol for a parallel double-blinded randomised intervention trial to investigate biomarker-based nutrition plans for weight loss. BMJ Open, 12(3), e051285. https://doi.org/10.1136/bmjopen-2021-051285

Adua, et al., (2021). Predictive model and feature importance for early detection of type II diabetes mellitus. Translational Medicine Communications, 6(1), 17. https://doi.org/10.1186/s41231-021-00096-z

Bandi, M., Masimukku, A. K., Vemula, R., & Vallu, S. (2024). Predictive analytics in healthcare: Enhancing patient outcomes through data-driven forecasting and decision making. International Numeric Journal of Machine Learning and Robots, 8(8), 1–20. https://injmr.com/index.php/fewfewf/article/view/144

Bansal, M., Goyal, A., & Choudhary, A. (2022). A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning. Decision Analytics Journal, 3, 100071. https://doi.org/10.1016/j.dajour.2022.100071

Baron, G. (2024). Diabetes a growing concern among Filipinos—DOH. Daily Tribune. https://tinyurl.com/ys7van7t

Bayman, E. O., & Dexter, F. (2021). Multicollinearity in Logistic Regression Models. Anesthesia & Analgesia, 133(2), 362–365. https://doi.org/10.1213/ANE.0000000000005593

Beaudart, C., Li, N., Hiligsmann, M., & Silverman, S. (2021). Effective risk communication and improving adherence. In E. Dennison (Ed.), Osteoporosis Treatment (pp. 115–143). Springer International Publishing. https://doi.org/10.1007/978-3-030-78128-6_8

Cando, L. F., et al (2024). Current status of diabetes mellitus care and management in the Philippines. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 18(2), 102951. https://doi.org/10.1016/j.dsx.2024.102951

Daily Tribune. (2024). 4.3M Filipinos living with diabetes — DOH. https://tribune.net.ph/2024/11/23/43m-filipinos-living-with-diabetes-doh

Diabetes, U. K. (2006). What is diabetes? Accessed March 21.

Greene, D., Palmer, M. J., & Relman, D. A. (2023). Motivating proactive biorisk management. Health Security, 21(1), 46–60. https://doi.org/10.1089/hs.2022.0101

Han, F., Shi, X., Pan, J., Wu, K., Zhu, Q., Yuan, C., Xiao, W., Ding, Y., Yu, X., Jiao, X., Hu, L., Lu, G., & Li, W. (2023). Elevated serum HbA1c levels, rather than a previous history of diabetes, predict disease severity and clinical outcomes in acute pancreatitis. BMJ Open Diabetes Research & Care, 11(1), e003070. https://doi.org/10.1136/bmjdrc-2022-003070

Harris, J. K. (2021). Primer on binary logistic regression. Family Medicine and Community Health, 9(Suppl 1), e001290. https://doi.org/10.1136/fmch-2021-001290

Hello Doctor. (2023). Diabetes statistics in the Philippines – Where do we stand? Hello Doctor Philippines. Retrieved May 29, 2025, from https://tinyurl.com/45k4nsy4

Hossain, Md. J., Al-Mamun, Md., & Islam, Md. R. (2024). Diabetes mellitus, the fastest growing global public health concern: Early detection should be focused. Health Science Reports, 7(3), e2004. https://doi.org/10.1002/hsr2.2004

Johnson, J. E. (1999). Self-regulation theory and coping with physical illness. Research in Nursing & Health, 22(6), 435–448. https://doi.org/10.1002/(SICI)1098-240X(199912)22:6<435::AID-NUR2>3.0.CO;2-Q

Leary, M. R., & Tangney, J. P. (2003). The self as an organizing construct in the behavioral and social sciences. Handbook of self and identity, 15, 3–14.MIT License. (2013, 2017). Released by Mark Otto and Andrew Fong. https://opensource.org/licenses/MIT

Okosun, I. S., Davis-Smith, M., & Seale, J. P. (2012). Awareness of diabetes risks is associated with healthy lifestyle behavior in diabetes free American adults: Evidence from a nationally representative sample. Primary Care Diabetes, 6(2), 87–94. https://doi.org/10.1016/j.pcd.2011.12.005

Ojulari, L. S., Sulaiman, S. E., Ayinde, T. O., & Kadir, E. R. (2023). Handgrip strength as a screening tool for diabetes in resource-constrained settings: A potential solution to overcome barriers to diagnosis. Endocrinology (including Diabetes Mellitus and Metabolic Disease). https://doi.org/10.1101/2023.10.19.23297260

Pagaddu, J. V. A. (2021). An insight on poverty reduction and health inequalities in the Philippines: Social determinants, constraints, and opportunities. Book of Life Publication. ISBN 978-621-419-265-6.

Patel, M. (2025). Diabetes prediction dataset [Data set]. Kaggle. https://www.kaggle.com/datasets/marshalpatel3558/diabetes-prediction-dataset-legit-dataset

Pendrill, L. R., Melin, J., Stavelin, A., & Nordin, G. (2023). Modernising Receiver Operating Characteristic (ROC) Curves. Algorithms, 16(5), 253. https://doi.org/10.3390/a16050253

Pillai, B. (2021). Cultural Barriers. In H. Tohid & H. Maibach (Eds.), International Medical Graduates in the United States (pp. 117–124). Springer International Publishing. https://doi.org/10.1007/978-3-030-62249-7_7

Ryan, P. (2009). Integrated theory of health behavior change: Background and intervention development. Clinical Nurse Specialist, 23(3), 161–170. https://doi.org/10.1097/NUR.0b013e3181a42373

ŞahiN, M., & Aybek, E. (2020). Jamovi: An easy-to-use statistical software for social scientists. International Journal of Assessment Tools in Education, 6(4), 670–692. https://doi.org/10.21449/ijate.661803

Sinaga, K. P., & Yang, M.-S. (2020). Unsupervised K-Means Clustering Algorithm. IEEE Access, 8, 80716–80727. https://doi.org/10.1109/ACCESS.2020.2988796

Thoolen, B. J., Ridder, D. D., Bensing, J., Gorter, K., & Rutten, G. (2009). Beyond good intentions: The role of proactive coping in achieving sustained behavioural change in the context of diabetes management. Psychology & Health, 24(3), 237–254. https://doi.org/10.1080/08870440701864504

Tolabing, MA. C. C., et al,. (2022). Prevalence of limited health literacy in the Philippines: First national survey. HLRP: Health Literacy Research and Practice, 6(2). https://doi.org/10.3928/24748307-20220419-01

Yilmaz, G. S. (2024). Does a financing scheme matter for access to healthcare service? In N. Rezaei (Ed.), Integrated Science for Sustainable Development Goal 3 (Vol. 24, pp. 49–74). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-64292-0_3

World Health Organization. (2024). Diabetes. https://www.who.int/en/news-room/fact-sheets/detail/diabetes