

Clinically Interpretable Risk Stratification for Type 2 Diabetes Using Logistic Regression

Patrick O. Akinwumi

pakinwu@clemson.edu

Clemson University

Meihua Qian

Clemson University

Oyinkansola A. Babatope

Bournemouth University

Research Article

Keywords: Type 2 diabetes mellitus, logistic regression, explainable statistical learning, healthcare analytics, risk stratification, interpretable artificial intelligence

Posted Date: July 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10251506/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Type 2 diabetes mellitus (T2DM) remains a major global public health challenge due to its increasing prevalence, chronic complications, and substantial economic burden. Early risk identification is therefore essential for improving preventive intervention and reducing diabetes-related morbidity and mortality. Although machine learning approaches have increasingly been applied to diabetes prediction, many existing models rely on complex black-box algorithms that lack interpretability and clinical transparency. This study proposes an explainable statistical learning framework for Type 2 diabetes risk stratification using multiple logistic regression.

The study utilized the Pima Indian Diabetes Dataset obtained from Kaggle and originally derived from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). The dataset consisted of 768 records from adult Pima Indian women aged 21 years and above. Exploratory data analysis, correlation analysis, empirical probability assessment, and multiple logistic regression modeling were conducted using JMP statistical software. Model assumptions including binary response structure, independence, linearity of the logit, and multicollinearity were evaluated prior to model development.

The final logistic regression model retained five statistically significant predictors: pregnancies, glucose, blood pressure, body mass index (BMI), and diabetes pedigree function. Glucose emerged as the strongest metabolic predictor of diabetes risk (OR = 1.0353, 95% CI [1.0286, 1.0424], $p < .0001$), while diabetes pedigree function demonstrated the largest relative effect (OR = 2.4859, 95% CI [1.4046, 4.4510], $p = .0020$), indicating substantial hereditary influence on diabetes susceptibility. BMI also significantly increased diabetes odds (OR = 1.0885, 95% CI [1.0595, 1.1199], $p < .0001$), whereas pregnancies moderately elevated diabetes risk (OR = 1.1659, 95% CI [1.1047, 1.2322], $p < .0001$). Blood pressure showed a modest inverse association with diabetes occurrence (OR = 0.9881, 95% CI [0.9783, 0.9978], $p = .0170$).

The model achieved an overall classification accuracy of 77.5%, sensitivity of 57.5%, specificity of 88.2%, F1-score of 0.6403, and Matthews Correlation Coefficient (MCC) of 0.4862, indicating moderate predictive performance with strong identification of non-diabetic individuals. Empirical probability plots further demonstrated clear S-shaped relationships between glucose, BMI, diabetes pedigree function, and diabetes probability.

The findings demonstrate that interpretable statistical learning approaches can provide clinically meaningful diabetes risk stratification while preserving transparency and explanatory clarity. This study contributes to healthcare analytics and explainable artificial intelligence research by reinforcing the practical value of parsimonious and interpretable predictive models for population-level diabetes screening and clinical decision support.

1. Introduction

Type 2 diabetes mellitus (T2DM) has emerged as one of the most significant global public health challenges of the twenty-first century, affecting millions of individuals across diverse demographic and socioeconomic groups. Characterized by chronic hyperglycemia resulting from impaired insulin secretion, insulin resistance, or both, T2DM contributes substantially to premature mortality, long-term disability, and escalating healthcare expenditure worldwide (Magliano et al., 2021). Historically regarded as a disease predominantly affecting older populations in high-income countries, diabetes has increasingly become prevalent among younger populations and low- and middle-income countries due to urbanization, sedentary lifestyles, nutritional transitions, and rising obesity rates (Abdul Basith Khan et al., 2020). Consequently, diabetes has evolved from a localized clinical concern into a major global socioeconomic burden with multidimensional implications for healthcare systems, labor productivity, and quality of life.

The historical evolution of diabetes research reflects continuous attempts to improve disease understanding, diagnosis, and management. Ancient civilizations, including the Egyptians and Greeks, recognized symptoms associated with excessive urination and unexplained weight loss, although the biological mechanisms underlying the disease remained poorly understood for centuries (Bliss, 2021). Significant scientific advancement occurred in the nineteenth and twentieth centuries with the discovery of insulin by Banting and Best in 1921, which transformed diabetes from a fatal disease into a manageable chronic condition (Vecchio et al., 2018). Despite therapeutic progress, the growing prevalence of T2DM continues to outpace healthcare interventions, highlighting the urgent need for improved predictive, preventive, and screening strategies.

Recent epidemiological statistics underscore the alarming scale of the diabetes crisis. According to the International Diabetes Federation (2021), approximately 537 million adults worldwide were living with diabetes in 2021, with projections indicating that this number may rise to 643 million by 2030 and 783 million by 2045 if current trends persist. Studies further identified diabetes as one of the leading causes of blindness, kidney failure, cardiovascular disease, lower limb amputation, and premature mortality globally ("11. Chronic kidney disease and risk management: Standards of Care in Diabetes—2024," 2024; Siam et al., 2024; World Health, 2024). Economically, diabetes imposes substantial direct and indirect costs through hospitalization, medication, productivity losses, and long-term complications. The IDF estimated global diabetes-related healthcare expenditure at approximately USD 966 billion in 2021, representing a significant increase over previous decades. These economic pressures are particularly severe in resource-constrained healthcare systems where early detection and preventive interventions remain limited.

Clinically, the burden of diabetes extends beyond hyperglycemia alone. T2DM is strongly associated with multiple comorbidities, including hypertension, obesity, dyslipidemia, cardiovascular disease, neuropathy, nephropathy, and retinopathy (Ling et al., 2020). Moreover, delayed diagnosis frequently results in irreversible complications before therapeutic intervention begins. Evidence suggests that a substantial proportion of diabetic individuals remain undiagnosed for extended periods, especially in underserved populations lacking regular access to healthcare services (Saeedi et al., 2019). This diagnostic delay increases morbidity and mortality while reducing the effectiveness of subsequent treatment strategies. Consequently, early risk identification and predictive screening models have become critical components of contemporary diabetes management frameworks.

Traditional approaches to diabetes detection and management primarily rely on clinical screening methods such as fasting plasma glucose tests, oral glucose tolerance tests, glycated hemoglobin (HbA1c) measurements, and symptom-based evaluations (Ghazi et al., 2023). While these methods remain

clinically valuable, several limitations constrain their effectiveness at population scale. First, many conventional diagnostic procedures require laboratory infrastructure, trained healthcare personnel, and repeated clinical visits, limiting accessibility in low-resource settings. Second, these approaches are predominantly reactive rather than preventive, often identifying diabetes after significant metabolic dysregulation has already occurred. Third, conventional screening frameworks frequently fail to integrate multidimensional patient-level predictors such as hereditary susceptibility, reproductive history, and obesity-related indicators into unified risk assessment systems.

In response to these limitations, researchers have increasingly explored data-driven predictive analytics and statistical learning approaches to improve diabetes risk stratification (Kolla, 2023; Oikonomou & Khera, 2023). The emergence of healthcare informatics, electronic medical records, and machine learning techniques has facilitated the development of computational models capable of identifying latent disease patterns from clinical data (Topol, 2019). Logistic regression, decision trees, support vector machines, random forests, neural networks, and ensemble learning techniques have all been applied to diabetes prediction tasks with varying degrees of success (Kavakiotis et al., 2017). Among these approaches, machine learning models often demonstrate improved predictive accuracy compared to traditional statistical techniques. However, many advanced machine learning algorithms suffer from significant interpretability limitations that reduce their clinical transparency and practical adoption in healthcare environments.

The growing concern surrounding algorithmic opacity in healthcare has intensified interest in explainable artificial intelligence (XAI) and interpretable statistical learning frameworks. Black-box models such as deep neural networks may achieve high classification performance while providing limited insight into how clinical variables contribute to prediction outcomes (Samek et al., 2021). In clinical decision-making contexts, interpretability remains essential because healthcare practitioners require transparent reasoning pathways to justify diagnostic recommendations, assess patient risk, and maintain ethical accountability (Rudin, 2019). Consequently, interpretable modeling approaches that balance predictive performance with clinical transparency have become increasingly important within healthcare analytics research.

Logistic regression remains one of the most widely adopted interpretable statistical techniques in medical research due to its simplicity, statistical rigor, and explanatory capacity. Unlike highly complex machine learning models, logistic regression provides interpretable coefficients, odds ratios, and statistical significance measures that enable direct understanding of variable contributions to disease risk (Tolles & Meurer, 2016). Furthermore, logistic regression supports clinically meaningful interpretation of associations between predictors and outcomes while maintaining relatively strong predictive performance for binary classification problems. In diabetes research, logistic regression has frequently been used to identify significant predictors such as glucose concentration, body mass index (BMI), blood pressure, age, hereditary predisposition, and pregnancy-related factors.

Despite the continued relevance of logistic regression, several existing studies primarily emphasize predictive accuracy while underexploring interpretability, model transparency, and parsimonious variable selection. Many contemporary studies increasingly favor highly complex machine learning architectures without adequately addressing clinical explainability or the practical challenges associated with deploying opaque systems in real-world healthcare settings (Abdullah et al., 2021). Additionally, some studies fail to critically evaluate statistical assumptions, variable interactions, or empirical probability relationships underlying predictive modeling processes. As a result, there remains a need for interpretable and statistically grounded diabetes prediction frameworks capable of balancing predictive utility with clinical comprehensibility.

The Pima Indian Diabetes Dataset has become one of the most extensively studied benchmark datasets in diabetes prediction research due to its rich representation of metabolic and hereditary risk factors among a high-risk Indigenous female population. Originating from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), the dataset contains diagnostic information for Pima Indian women aged 21 years and older, including variables related to glucose concentration, BMI, insulin levels, blood pressure, pregnancies, and diabetes pedigree function. The dataset is particularly valuable because the Pima Indian population exhibits disproportionately high diabetes prevalence rates linked to genetic susceptibility and environmental influences (Smith et al., 1988). However, despite extensive use of the dataset in predictive modeling research, many studies prioritize algorithmic performance metrics over interpretable clinical insights and transparent risk explanation.

This study addresses these limitations by developing an explainable logistic regression framework for Type 2 diabetes prediction using the Pima Indian women dataset. Unlike black-box machine learning approaches that prioritize predictive optimization at the expense of transparency, the present study emphasizes interpretability, statistical significance analysis, empirical probability assessment, and clinically meaningful risk profiling. Through exploratory data analysis and logistic regression modeling, the study seeks to identify the most influential predictors of diabetes while maintaining model simplicity and explanatory clarity. The proposed framework contributes to ongoing efforts toward transparent healthcare analytics by demonstrating that interpretable statistical models can still provide meaningful predictive insights for diabetes risk stratification.

The findings of this study are expected to contribute to healthcare analytics, explainable artificial intelligence, and public health screening research by reinforcing the value of parsimonious and interpretable predictive models within clinical environments. Furthermore, the study supports the broader objective of developing transparent and accessible screening systems capable of assisting early diabetes identification, particularly in high-risk populations where preventive healthcare interventions remain critically important.

2. Related Work

Diabetes prediction has attracted substantial attention within healthcare analytics, statistical modeling, and machine learning research due to the increasing global prevalence and clinical burden associated with Type 2 diabetes mellitus (T2DM). Over the past two decades, researchers have explored diverse computational approaches aimed at improving early diagnosis, risk stratification, and disease management. Existing studies span traditional statistical methods, machine learning algorithms, deep learning frameworks, and hybrid predictive systems. While these contributions have advanced predictive modeling capabilities, significant gaps remain concerning interpretability, transparency, and clinically meaningful explanation of prediction outcomes.

Early diabetes prediction studies primarily relied on classical statistical approaches, particularly logistic regression analysis. Logistic regression became widely adopted due to its ability to model binary clinical outcomes while producing interpretable coefficients and odds ratios (Hosmer Jr et al., 2013). Numerous studies identified glucose concentration, BMI, age, blood pressure, and hereditary factors as significant predictors of diabetes using logistic regression frameworks. For instance, Colditz et al. (1995) demonstrated that obesity and familial history substantially increased diabetes risk among women, while subsequent epidemiological studies reinforced the association between metabolic dysfunction and T2DM progression. These studies established foundational understanding of diabetes risk factors and highlighted the importance of multivariable statistical analysis in medical diagnostics.

Although traditional logistic regression approaches offered interpretability advantages, several limitations became apparent over time. Many early studies relied on relatively small datasets, linear assumptions, and limited variable interaction analysis, reducing their ability to capture complex nonlinear relationships inherent in metabolic disorders. Additionally, some models exhibited moderate predictive performance when compared with more advanced machine learning approaches. Consequently, researchers increasingly turned toward computational intelligence techniques capable of improving classification accuracy.

Machine learning-based diabetes prediction research expanded rapidly with the availability of healthcare datasets and advances in computational power. Algorithms such as decision trees, support vector machines (SVMs), k-nearest neighbors (KNN), random forests, naïve Bayes classifiers, and artificial neural networks have been extensively applied to diabetes diagnosis tasks (Kavakiotis et al., 2017). Among these approaches, random forest and ensemble learning techniques frequently demonstrated improved predictive accuracy due to their ability to model nonlinear relationships and variable interactions. For example, Sisodia and Sisodia (2018) compared several machine learning algorithms on the Pima Indian Diabetes Dataset and reported strong classification performance using SVM and decision tree models. Similarly, (Ahmad et al., 2018) employed ensemble machine learning techniques to enhance diabetes prediction accuracy and observed improved sensitivity and specificity compared with conventional statistical methods.

Deep learning approaches have also gained prominence in diabetes prediction research. Neural network architectures, including multilayer perceptrons, convolutional neural networks, and recurrent neural networks, have been explored for complex healthcare prediction tasks due to their capacity to learn hierarchical feature representations from large-scale data (Swapna et al., 2018). Deep learning systems often outperform traditional models in terms of predictive optimization, particularly when applied to multimodal healthcare datasets containing imaging, genomic, and electronic health record information. However, despite these performance gains, deep learning models introduce significant interpretability challenges that complicate clinical implementation.

A growing body of literature has criticized black-box predictive systems for their lack of transparency in healthcare contexts. Rudin (2019) argued that opaque machine learning models should not replace interpretable approaches in high-stakes decision-making domains such as medicine. Similarly, Samek et al. (2021) emphasized that explainability remains essential for ensuring accountability, clinician trust, and ethical deployment of artificial intelligence systems in healthcare environments. In diabetes prediction specifically, clinicians require clear understanding of how variables such as glucose levels, obesity, hereditary predisposition, and reproductive history contribute to model predictions. Black-box systems frequently fail to provide such clinically interpretable reasoning pathways.

To address these concerns, recent studies have explored explainable artificial intelligence (XAI) frameworks that integrate interpretability mechanisms into predictive systems. Techniques such as SHAP (Shapley Additive Explanations), LIME (Local Interpretable Model-Agnostic Explanations), and feature importance ranking have been applied to diabetes prediction models to improve transparency (Lundberg & Lee, 2017). For example, several studies combined random forest classifiers with SHAP analysis to identify dominant predictors influencing diabetes outcomes. While these methods partially improve interpretability, they often function as post hoc explanation tools rather than inherently transparent models. Consequently, the underlying predictive systems remain fundamentally complex and computationally opaque.

The Pima Indian Diabetes Dataset remains one of the most widely utilized benchmark datasets in diabetes prediction research. Numerous studies have applied machine learning and statistical techniques to this dataset due to its accessibility and clinical relevance. Researchers have investigated various preprocessing strategies, feature selection approaches, and algorithmic optimization techniques to improve predictive performance. For instance, Chaki et al. (2022) evaluated multiple machine learning algorithms on the Pima dataset and reported favorable performance for ensemble classifiers. Other studies incorporated data balancing techniques, dimensionality reduction, and hyperparameter optimization to address class imbalance and improve sensitivity.

Despite extensive research using the Pima dataset, several important gaps remain insufficiently addressed. First, many studies disproportionately emphasize predictive accuracy metrics such as precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC) while neglecting clinical interpretability and statistical transparency. Second, some machine learning studies fail to provide meaningful explanation of variable significance, limiting their practical value for healthcare practitioners. Third, numerous studies adopt increasingly complex predictive architectures without critically evaluating whether marginal accuracy improvements justify reduced interpretability.

Another significant limitation within existing literature concerns inadequate attention to parsimonious modeling and assumption-driven statistical analysis. Several machine learning studies utilize all available predictors without examining multicollinearity, empirical probability relationships, or variable relevance. Consequently, model complexity often increases unnecessarily while explanatory clarity decreases. Furthermore, some predictive systems lack explicit clinical reasoning regarding why certain variables contribute more strongly to diabetes risk than others. This limitation reduces the translational value of computational findings within real-world healthcare environments.

Existing literature also reveals limited emphasis on population-specific interpretability. While predictive models frequently report classification performance, fewer studies investigate how hereditary, metabolic, and reproductive variables interact within high-risk Indigenous populations such as Pima Indian women. The Pima population presents unique epidemiological importance due to disproportionately elevated diabetes prevalence linked to both genetic and

environmental determinants (Knowler et al., 1978). However, several studies using the dataset treat it merely as a benchmark classification problem rather than contextualizing findings within broader public health and clinical frameworks.

The present study seeks to address these gaps by developing an explainable logistic regression framework emphasizing interpretability, statistical significance assessment, empirical probability analysis, and clinically meaningful predictor evaluation. Unlike studies that prioritize algorithmic complexity and black-box optimization, this research adopts a parsimonious modeling philosophy aimed at balancing predictive utility with transparency and clinical comprehensibility. By focusing on interpretable risk profiling rather than solely maximizing accuracy metrics, the study contributes to ongoing discussions surrounding explainable healthcare analytics and responsible AI deployment in medical decision-making.

Furthermore, this study strengthens methodological rigor through exploratory data analysis and assumption-aware statistical modeling. Rather than treating predictive modeling as a purely computational exercise, the research critically examines variable behavior, significance relationships, and clinical implications associated with diabetes risk prediction. In doing so, the study reinforces the continued relevance of interpretable statistical frameworks within healthcare analytics and demonstrates that meaningful predictive insights can be achieved without relying exclusively on highly complex machine learning systems.

3. Methodology

3.1 Research Design

This study employed a quantitative predictive analytics approach to develop an explainable statistical learning framework for Type 2 diabetes risk stratification using multiple logistic regression. The methodological framework integrated exploratory data analysis (EDA), correlation analysis, empirical probability assessment, and logistic regression modeling to identify statistically significant predictors associated with diabetes occurrence among adult Pima Indian women. Unlike black-box machine learning approaches that prioritize predictive complexity, the present study emphasized interpretability, statistical transparency, and clinically meaningful explanation of predictor effects.

The methodological process was designed to:

1. examine the statistical relationships between health-related variables and diabetes outcome,
2. identify significant predictors of Type 2 diabetes,
3. develop an interpretable logistic regression model for diabetes risk stratification,
4. and evaluate model classification performance using standard predictive metrics.

All statistical analyses and visualizations were conducted using JMP statistical software.

3.2 Data Source and Population

The study utilized the Pima Indian Diabetes Dataset obtained from Kaggle and originally derived from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) (see Table 1 for data information and description). The dataset contains diagnostic measurements collected from adult female members of the Pima Indian population aged 21 years and above.

The Pima Indian population was selected because of its historically high prevalence of Type 2 diabetes and its frequent use in diabetes prediction research. The dataset consisted of 768 observations and included clinical, physiological, and hereditary variables commonly associated with diabetes risk.

The response variable was diabetes outcome, represented as a binary categorical variable, 1 = diabetes diagnosed, 0 = no diabetes diagnosed.

Eight explanatory variables were initially considered including pregnancies, glucose concentration, blood pressure, skin thickness, insulin level, body mass index (BMI), diabetes pedigree function, and age.

After exploratory analysis and model selection procedures, five predictors were retained in the final regression model: pregnancies, glucose, blood pressure, BMI, and diabetes pedigree function.

3.3 Variable Description

Table 1
The variables included in the analysis are summarized below.

Variable	Type	Description
Pregnancies	Quantitative	Number of times pregnant
Glucose	Quantitative	Plasma glucose concentration (mg/dL)
Blood pressure	Quantitative	Diastolic blood pressure (mm Hg)
Skinthickness	Quantitative	Triceps skinfold thickness (mm)
Insulin	Quantitative	2-hour serum insulin (mu U/ml)
BMI	Quantitative	Body Mass Index (weight in kg/(height in m)^2)
DiabetesPedigreeFunc	Quantitative	Genetic predisposition to diabetes (family history index)
Age	Quantitative	Age in years
Outcome	Categorical	1 = Diabetes diagnosed, 0 = No diabetes

3.4 Exploratory Data Analysis

Exploratory data analysis was conducted to evaluate the statistical properties of the dataset and identify potential relationships between predictors and diabetes outcome. Descriptive analysis included the generation of boxplots, scatterplot matrices, correlation matrices, and empirical probability plots.

Boxplot analysis revealed that glucose concentration, BMI, pregnancies, and diabetes pedigree function were generally higher among diabetic individuals than non-diabetic individuals. These patterns suggested meaningful discriminatory relationships between these predictors and diabetes status.

Scatterplot matrices and correlation analysis were additionally used to examine associations among predictors and assess potential multicollinearity. The strongest observed correlations included:

- age and pregnancies ($r = 0.54$),
- BMI and skin thickness ($r = 0.39$),
- and glucose and insulin ($r = 0.33$).

Because all correlation coefficients remained below 0.55, multicollinearity was considered acceptable for logistic regression modeling.

3.5 Empirical Probability Analysis

Empirical probability plots were generated to evaluate the assumption of linearity between predictors and the log-odds of diabetes occurrence. This assumption is fundamental in logistic regression analysis because the model assumes approximately linear relationships between continuous predictors and the logit-transformed outcome probability.

The empirical probability plots demonstrated clear S-shaped relationships for: glucose, BMI, and diabetes pedigree function,

indicating strong linear associations with the log-odds of diabetes and supporting their inclusion in the final model.

Pregnancies and blood pressure exhibited moderate linearity patterns, while insulin and skin thickness demonstrated irregular and non-monotonic probability trends. Consequently, insulin and skin thickness were excluded from the final regression framework due to weaker predictive contribution and less stable empirical relationships.

Age was also excluded because of its moderate correlation with pregnancies and comparatively weaker independent contribution to the final model.

3.6 Logistic Regression Modeling

Multiple logistic regression was employed as the primary predictive modeling technique because the response variable was binary and the study prioritized interpretability and statistical transparency.

The logistic regression model estimates the probability of diabetes occurrence using the logistic function:

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n)}}$$

where:

- $P(Y = 1)$ represents the probability of diabetes occurrence,
- β_0 denotes the intercept,
- β_i represents regression coefficients,
- and X_i represents predictor variables.

The final logistic regression equation retained five predictors:

$$\log\left(\frac{p}{1-p}\right) = -7.95 + 0.153(Pregnancies) + 0.0346(Glucose) - 0.0120(BloodPressure) + 0.0848(BMI) + 0.9106(DiabetesPedigr$$

Maximum likelihood estimation (MLE) was used to estimate regression coefficients, and statistical significance was evaluated using Wald chi-square statistics and corresponding p-values.

3.7 Model Assumptions

Several assumptions associated with logistic regression were evaluated prior to model interpretation.

Binary Response Variable

The response variable satisfied the binary outcome requirement because diabetes status was coded as diabetic or non-diabetic.

Independence of Observations

Each observation represented an independent patient record, satisfying the assumption of independence.

Multicollinearity

Correlation analysis indicated no severe multicollinearity among predictors because all correlation coefficients were below 0.55.

Linearity of the Logit

Empirical probability plots demonstrated approximately linear relationships between several predictors and the log-odds of diabetes occurrence, particularly for glucose, BMI, and diabetes pedigree function.

3.8 Model Evaluation

The predictive performance of the logistic regression model was evaluated using classification metrics derived from the confusion matrix.

The model produced:

- 154 true positives,
- 441 true negatives,
- 59 false positives,
- and 114 false negatives.

Performance metrics included:

- Accuracy = 77.5%
- Sensitivity = 57.5%
- Specificity = 88.2%
- F1-score = 0.6403
- Matthews Correlation Coefficient (MCC) = 0.4862
- Akaike Information Criterion (AIC) = 740.67

The relatively high specificity indicated strong performance in correctly identifying non-diabetic individuals, while the moderate sensitivity suggested that some diabetic cases were misclassified as non-diabetic.

The Matthews Correlation Coefficient indicated moderate agreement between predicted and actual classifications, while the AIC value suggested acceptable model parsimony.

3.9 Ethical Considerations

The study utilized a publicly available anonymized secondary dataset containing no personally identifiable information. Consequently, no direct human participant involvement occurred during the research process.

All analyses were conducted strictly for academic and research purposes in accordance with ethical standards for healthcare data analytics and responsible statistical research.

3.10 Methodological Contribution

The methodological contribution of this study lies in its emphasis on explainable statistical learning and interpretable healthcare analytics. Rather than prioritizing predictive complexity through black-box machine learning systems, the proposed framework demonstrates that parsimonious logistic regression modeling can still provide clinically meaningful diabetes risk stratification with transparent statistical interpretation.

The framework therefore contributes to ongoing efforts toward interpretable artificial intelligence and trustworthy healthcare decision-support systems.

4. Results

4.1 Exploratory Data Analysis

Exploratory data analysis was conducted to examine the distributional characteristics of the predictors and their relationships with diabetes outcome among adult Pima Indian women. Boxplot analysis (Fig. 1) revealed notable differences between diabetic and non-diabetic participants across several health-related variables. Specifically, individuals diagnosed with diabetes generally exhibited higher glucose concentrations, higher body mass index (BMI) values, greater numbers of pregnancies, and elevated diabetes pedigree function scores compared with non-diabetic individuals. Among the predictors, glucose demonstrated the clearest separation between outcome groups, suggesting strong discriminatory capability for diabetes prediction. BMI and pregnancies also displayed substantial upward shifts among diabetic individuals, reinforcing their potential importance in diabetes risk stratification. In contrast, insulin and skin thickness exhibited broader overlap between diabetic and non-diabetic groups, indicating weaker predictive distinction.

The exploratory findings provided preliminary evidence that metabolic, reproductive, and hereditary factors are strongly associated with diabetes occurrence within the study population. These visual patterns supported the inclusion of glucose, BMI, pregnancies, and diabetes pedigree function in the subsequent regression analysis.

4.2 Correlation Analysis

Correlation analysis was performed to evaluate relationships among predictors and assess the potential presence of multicollinearity prior to model development. The scatterplot matrix and correlation matrix (Fig. 2) revealed moderate positive associations among several predictors. The strongest correlation was observed between age and pregnancies ($r = 0.5443$), followed by BMI and skin thickness ($r = 0.3926$), and glucose and insulin ($r = 0.3314$). However, all correlation coefficients remained below 0.55, indicating that severe multicollinearity was not present within the dataset.

The correlation analysis also informed variable selection decisions. Age was excluded from the final model because of its moderate correlation with pregnancies and weaker independent contribution to diabetes prediction. Similarly, insulin and skin thickness were excluded due to weaker predictive influence and irregular empirical probability patterns. These findings supported the development of a more parsimonious and interpretable regression framework.

4.3 Logistic Regression Model Results

Multiple logistic regression analysis identified five statistically significant predictors associated with Type 2 diabetes occurrence: pregnancies, glucose concentration, blood pressure, BMI, and diabetes pedigree function. The final logistic regression model demonstrated that all retained predictors contributed significantly to diabetes risk stratification.

The estimated regression equation was expressed as:

$$\log\left(\frac{p}{1-p}\right) = -7.95 + 0.153 (\text{Pregnancies}) + 0.0346 (\text{Glucose}) - 0.0120 (\text{BloodPressure}) + 0.0848 (\text{BMI}) + 0.9106 (\text{DiabetesPedigree})$$

The multiple logistic regression analysis identified pregnancies, glucose concentration, blood pressure, body mass index (BMI), and diabetes pedigree function as statistically significant predictors of Type 2 diabetes among adult Pima Indian women (see Table 2). Among these variables, glucose emerged as the strongest metabolic predictor of diabetes risk (OR = 1.0353, 95% CI [1.0286, 1.0424], $p < .0001$), indicating that each one-unit increase in glucose concentration increased the odds of diabetes by approximately 3.5%. The narrow confidence interval further demonstrated the stability and precision of the glucose effect estimate, reinforcing its dominant predictive influence within the model.

Body mass index (BMI) also demonstrated a strong positive association with diabetes occurrence (OR = 1.0885, 95% CI [1.0595, 1.1199], $p < .0001$). Specifically, each one-unit increase in BMI increased diabetes odds by approximately 8.9%, suggesting that increasing adiposity substantially elevated diabetes risk. This finding reinforces the established relationship between obesity, insulin resistance, and impaired glucose metabolism in Type 2 diabetes progression.

Pregnancies significantly increased diabetes odds by approximately 16.6% per additional pregnancy (OR = 1.1659, 95% CI [1.1047, 1.2322], $p < .0001$), highlighting the contribution of reproductive history to metabolic risk within the study population. The positive association suggests that cumulative physiological and hormonal stress associated with multiple pregnancies may contribute meaningfully to long-term diabetes susceptibility among women.

Diabetes pedigree function exhibited the largest relative effect among all retained predictors (OR = 2.4859, 95% CI [1.4046, 4.4510], $p = .0020$), indicating that hereditary predisposition more than doubled the odds of diabetes occurrence. This finding highlights the substantial influence of genetic susceptibility and familial risk factors within the Pima Indian population and underscores the importance of hereditary pathways in diabetes development.

Blood pressure demonstrated a modest but statistically significant inverse association with diabetes occurrence (OR = 0.9881, 95% CI [0.9783, 0.9978], $p = .0170$). Although statistically significant, the magnitude of this effect was comparatively small relative to glucose, BMI, and diabetes pedigree function. The finding suggests that blood pressure may function more as a secondary physiological correlate rather than a dominant independent predictor of diabetes risk within the present regression framework.

Table 2
Logistic regression coefficients, odds ratios, confidence intervals, and p-values.

Predictor	Estimate	OR	95% CI	p-value	Interpretation
Pregnancies	0.153	1.166	[1.105, 1.231]	< 0.0001	+ 16.6% odds per pregnancy
Glucose	0.0346	1.035	[1.028, 1.042]	< 0.0001	3.5% odds per unit
Blood Pressure	-0.0120	0.988	[0.979, 0.998]	< 0.0170	Slight negative effect
BMI	0.0848	1.089	[1.059, 1.120]	< 0.0001	+ 8.9% odds per unit
Diabetes Pedigree Function	0.9106	2.486	[1.372, 4.453]	0.0020	Odds more than double

The confidence intervals for all retained predictors excluded 1.0, confirming their statistical significance within the regression framework. Collectively, these findings indicate that diabetes risk within the study population is influenced by a combination of metabolic, hereditary, and reproductive factors.

4.4 Classification Performance

The predictive performance of the logistic regression model was evaluated using multiple classification metrics derived from the confusion matrix. The model achieved an overall classification accuracy of 77.5% (Table 3), indicating that approximately three-quarters of observations were correctly classified. Sensitivity was 57.5%, demonstrating moderate ability to correctly identify diabetic individuals. In contrast, specificity reached 88.2%, indicating strong performance in correctly identifying non-diabetic individuals.

The model additionally achieved an F1-score of 0.6403 and a Matthews Correlation Coefficient (MCC) of 0.4862, reflecting moderate agreement between predicted and actual classifications. The Akaike Information Criterion (AIC) value of 740.67 suggested acceptable model fit while maintaining relative model parsimony.

Table 3
Model performance evaluation metrics.

Metric	Value	Interpretation
AIC	740.67	Lower AIC indicates better model fit compared to alternatives
Accuracy	0.7747	77.5% of outcomes correctly classified
Sensitivity	0.5746	Correctly identifies 57.5% of those with diabetes (moderate)
Specificity	0.8820	Correctly identifies 88.2% of non-diabetic cases
F1 Score	0.6403	Balance between precision and recall
MCC	0.4862	Moderate correlation between predicted and actual classes

These findings indicate that the model performed substantially better in identifying non-diabetic individuals than diabetic cases. Although the relatively lower sensitivity suggests that some diabetic cases were misclassified, the strong specificity demonstrates practical utility for population-level diabetes screening and risk stratification.

4.5 Confusion Matrix Analysis

The confusion matrix provided additional insight into the classification behavior of the logistic regression model. The model correctly classified 154 diabetic individuals as true positives and 441 non-diabetic individuals as true negatives. However, 114 diabetic cases were incorrectly classified as non-diabetic (false negatives), while 59 non-diabetic individuals were incorrectly classified as diabetic (false positives).

The relatively high number of true negatives contributed to the model's strong specificity performance, whereas the false negatives reduced overall sensitivity. These findings suggest that the model is more effective for ruling out non-diabetic individuals than for identifying all positive diabetes cases. Nevertheless, the classification outcomes demonstrate that the logistic regression framework provides acceptable predictive capability while maintaining strong interpretability and statistical transparency.

4.6 Empirical Probability and Profiler Analysis

Empirical probability plots were used to evaluate the relationship between predictor variables and diabetes probability. The plots (Fig. 3) revealed clear S-shaped probability trends for glucose, BMI, and diabetes pedigree function, indicating approximately linear relationships between these predictors and the log-odds of diabetes occurrence. Pregnancies also demonstrated a positive probability trend, although less pronounced than glucose and BMI.

Profiler analysis (Fig. 4) further illustrated how variations in predictor values influenced predicted diabetes probability. Increasing pregnancies from approximately 4 to 10 increased predicted diabetes probability from 29.7% to 52.1%, demonstrating the contribution of reproductive history to diabetes susceptibility. Similarly, elevated glucose and BMI values substantially increased predicted diabetes probability, while higher diabetes pedigree function values consistently reflected increasing hereditary risk.

Among all predictors, glucose exhibited the steepest probability increase, confirming its dominant influence within the predictive framework. BMI similarly demonstrated strong positive effects, particularly above BMI levels near 30. Diabetes pedigree function showed progressive increases in diabetes risk across

higher hereditary predisposition levels, reinforcing the importance of genetic susceptibility in Type 2 diabetes development.

5. Discussion

The present study developed an explainable statistical learning framework for Type 2 diabetes risk stratification among adult Pima Indian women using multiple logistic regression. The findings demonstrated that glucose concentration, body mass index (BMI), diabetes pedigree function, pregnancies, and blood pressure significantly contributed to diabetes risk prediction within the study population. More importantly, the study showed that an interpretable and parsimonious statistical model can provide clinically meaningful diabetes risk stratification while preserving transparency and explanatory clarity. These findings contribute to ongoing discussions within healthcare analytics concerning the balance between predictive performance and model interpretability, particularly in high-stakes clinical environments where transparency and trustworthiness remain essential.

Among all predictors examined, glucose concentration emerged as the strongest predictor of diabetes occurrence. The logistic regression model indicated that each one-unit increase in glucose concentration increased diabetes odds by approximately 3.5% ($OR = 1.0353, p < .0001$). This finding aligns closely with prior epidemiological and predictive modeling studies identifying plasma glucose as the dominant metabolic marker associated with Type 2 diabetes progression (Abdul Basith Khan et al., 2020; Kavakiotis et al., 2017). Elevated glucose levels directly reflect impaired insulin regulation and chronic metabolic dysregulation, which are central pathophysiological mechanisms underlying diabetes development. The strong statistical significance of glucose within the present model therefore reinforces its continued relevance in both clinical screening and computational diabetes prediction frameworks.

The observed dominance of glucose concentration is also consistent with several machine learning studies utilizing the Pima Indian Diabetes Dataset. Sisodia and Sisodia (2018), for instance, reported that glucose consistently ranked among the most influential features across support vector machine, decision tree, and naïve Bayes classifiers. Similarly, Chaki et al. (2022) found that glucose-related variables contributed strongly to predictive accuracy in ensemble-based diabetes detection systems. However, while previous studies frequently emphasized feature importance rankings primarily within black-box predictive systems, the present study advances the field by situating glucose within an interpretable logistic regression framework that provides direct clinical interpretation through regression coefficients, odds ratios, and empirical probability relationships. This distinction is particularly important because healthcare practitioners require transparent reasoning mechanisms rather than solely algorithmic performance metrics.

Body mass index (BMI) also emerged as a statistically significant predictor of diabetes occurrence, increasing diabetes odds by approximately 8.9% per unit increase ($OR = 1.0885, p < .0001$). This finding supports extensive evidence linking obesity to insulin resistance and metabolic dysfunction (Colditz et al., 1995; Ling et al., 2021). Obesity contributes to chronic inflammation, impaired insulin signaling, altered glucose metabolism, and reduced insulin sensitivity, thereby increasing susceptibility to hyperglycemia and long-term metabolic complications. The persistence of BMI significance after controlling for glucose and other predictors suggests that obesity independently contributes to diabetes vulnerability rather than functioning merely as a secondary correlate. Empirical probability analysis further demonstrated progressively increasing diabetes probability at higher BMI values, particularly above BMI levels associated with obesity, reinforcing the clinical significance of adiposity in diabetes progression.

The findings concerning diabetes pedigree function are particularly noteworthy. The diabetes pedigree function demonstrated the largest relative effect among the retained predictors, increasing diabetes odds by approximately 149% ($OR = 2.4859, p = .0020$). This result emphasizes the importance of hereditary predisposition in diabetes susceptibility among the Pima Indian population and supports earlier epidemiological studies reporting disproportionately elevated diabetes prevalence among Pima Indians due to genetic and familial risk factors (Knowler et al., 1978). Furthermore, the present findings align with broader research highlighting the multifactorial nature of Type 2 diabetes, in which hereditary susceptibility interacts with environmental and metabolic influences to shape disease progression (Saeedi et al., 2019).

Importantly, the present study extends existing literature by emphasizing hereditary risk within an explainable modeling framework rather than treating genetic predisposition merely as another predictive variable. Many contemporary machine learning studies prioritize predictive optimization while offering limited explanation regarding why hereditary variables contribute significantly to disease prediction. In contrast, the present analysis explicitly interprets diabetes pedigree function as evidence supporting the clinical relevance of inherited susceptibility pathways. This interpretive emphasis enhances the translational value of the findings for population-level risk assessment, preventive healthcare planning, and early intervention strategies targeting genetically vulnerable individuals.

Pregnancy count also demonstrated statistically significant contribution to diabetes prediction, with each additional pregnancy increasing diabetes odds by approximately 16.6% ($OR = 1.1659, p < .0001$). This finding is consistent with previous studies indicating that reproductive history may influence long-term metabolic regulation and insulin sensitivity among women. Multiple pregnancies may increase cumulative physiological stress and hormonal adaptation demands, potentially contributing to increased diabetes vulnerability later in life. While some prior machine learning studies acknowledged pregnancies as an input variable, relatively few critically discussed its broader reproductive and endocrinological implications. The present study therefore contributes additional interpretive depth by situating pregnancies within women's metabolic health and chronic disease risk contexts.

Blood pressure exhibited a weaker but statistically significant inverse association with diabetes occurrence ($OR = 0.9881, p = .0170$). Although statistically significant, the magnitude of this effect was comparatively modest relative to glucose and BMI. This finding partially aligns with prior literature identifying hypertension as a common comorbidity associated with diabetes and metabolic syndrome (Ling et al., 2021). However, unlike some studies reporting stronger hypertensive effects, the present analysis suggests that blood pressure may function more as a secondary physiological correlate rather than a dominant independent predictor within this specific population. This discrepancy may reflect population-specific metabolic characteristics, differences in variable interactions, or methodological variations across studies. Consequently, the blood pressure findings should be interpreted cautiously and warrant further investigation across more diverse populations.

One of the most important contributions of this study lies in its emphasis on interpretability and parsimonious modeling. Contemporary healthcare analytics research increasingly favors highly complex machine learning architectures such as deep neural networks, ensemble classifiers, and hybrid learning systems (Swapna et al., 2018). Although these approaches often achieve marginal improvements in predictive accuracy, they frequently suffer from limited transparency and poor clinical explainability. Rudin (2019) strongly criticized the deployment of black-box systems in high-stakes domains such as healthcare, arguing that interpretable models should be preferred whenever possible. The present findings directly support this argument by demonstrating that a relatively simple logistic regression framework can still produce meaningful predictive insights while remaining understandable to clinicians, healthcare practitioners, and decision-makers.

The study therefore contributes to the growing body of explainable artificial intelligence (XAI) research within healthcare analytics. Unlike post hoc explanation techniques such as SHAP and LIME, which attempt to interpret inherently opaque systems, logistic regression provides intrinsic interpretability through statistically meaningful coefficients, odds ratios, confidence intervals, and empirical probability relationships (Lundberg & Lee, 2017). Consequently, the present framework offers transparency not as an auxiliary feature but as a foundational characteristic of the predictive model itself. This distinction is particularly important in clinical environments where accountability, trust, interpretive clarity, and ethical decision-making remain essential requirements for practical implementation.

Another important contribution of this study lies in its emphasis on methodological transparency and explainable statistical learning rather than predictive complexity alone. While several previous studies utilizing the Pima Indian Diabetes Dataset primarily focused on maximizing classification accuracy through sophisticated machine learning architectures, many provided limited evaluation of statistical assumptions, empirical probability trends, or variable relevance. In contrast, the present study incorporated exploratory data analysis, correlation analysis, empirical probability assessment, assumption-aware feature selection, odds ratio interpretation, and profiler visualization to support model development. This methodological rigor enhances the credibility, robustness, and interpretability of the resulting predictive framework while demonstrating that meaningful healthcare prediction can be achieved without sacrificing transparency for predictive complexity. Consequently, the study reinforces the continued relevance of logistic regression as a valuable methodological approach for healthcare prediction tasks because it effectively balances statistical rigor, predictive capability, interpretability, and clinical relevance. By integrating transparent statistical reasoning with explainable modeling, the study advances a trustworthy analytical framework that supports clinically interpretable and ethically responsible healthcare decision-making, thereby aligning with growing concerns regarding explainability, accountability, and responsible artificial intelligence deployment in clinical practice.

The model's performance characteristics additionally provide important clinical insights. The model achieved an overall classification accuracy of 77.5%, specificity of 88.2%, sensitivity of 57.5%, F1-score of 0.6403, and Matthews Correlation Coefficient (MCC) of 0.4862. The relatively high specificity suggests that the framework performs effectively in identifying non-diabetic individuals. However, the comparatively moderate sensitivity indicates that some diabetic cases may remain undetected. This tradeoff reflects a common challenge in medical prediction systems, where improving sensitivity may reduce specificity and vice versa. From a public health perspective, the model may therefore function more effectively as an initial population-level screening tool rather than a definitive diagnostic system. This interpretation aligns with previous studies emphasizing the role of predictive analytics in early risk stratification rather than standalone clinical diagnosis (Topol, 2019).

Despite its contributions, the study possesses limitations to be considered. First, the dataset is restricted to adult Pima Indian women, limiting generalizability to broader populations with different demographic, genetic, and environmental characteristics. Second, the study relied on observational cross-sectional data, preventing causal inference regarding relationships between predictors and diabetes outcomes. Consequently, the identified relationships should be interpreted as predictive associations rather than direct causal mechanisms. Third, the dataset lacked several important behavioral and socioeconomic variables such as dietary habits, physical activity levels, smoking behavior, healthcare access, and socioeconomic status, all of which may substantially influence diabetes risk.

Nevertheless, these limitations also reveal important directions for future research. Future studies may integrate interpretable statistical models with explainable machine learning frameworks to balance predictive performance and clinical transparency more effectively. Additionally, expanding predictive analysis to ethnically and geographically diverse populations could improve generalizability and strengthen population-level diabetes screening systems. Longitudinal studies may also provide deeper insight into temporal diabetes progression, causal risk pathways, and evolving metabolic interactions over time. Incorporating lifestyle, behavioral, and socioeconomic indicators may further improve predictive robustness and ecological validity.

Conclusion

The present study demonstrated that glucose concentration, body mass index (BMI), diabetes pedigree function, pregnancies, and blood pressure significantly contribute to Type 2 diabetes risk stratification among adult Pima Indian women, with glucose emerging as the strongest metabolic predictor and diabetes pedigree function exhibiting the largest relative effect, thereby highlighting the combined influence of metabolic, hereditary, and reproductive factors in diabetes susceptibility. The findings further showed that interpretable and parsimonious logistic regression models can provide meaningful predictive capability while preserving transparency, statistical rigor, and clinical interpretability. Rather than prioritizing predictive complexity alone, the study reinforces the importance of explainable statistical learning and transparent healthcare analytics in developing clinically understandable decision-support systems. Overall, the research contributes to ongoing efforts toward trustworthy, interpretable, and ethically responsible artificial intelligence in healthcare by demonstrating that explainable statistical models remain highly valuable for population-level diabetes screening and risk assessment.

Declarations

Author Contribution

PA: Conceptualization, Data curation, Methodology, Investigation, Visualization, Writing – original draft, Writing – review & editing, Formal analysis, Project administration, Software, Validation, Visualization. **MQ:** Writing – review & editing, Supervision. **OB:** Methodology, Visualization, Formal Analysis, Writing – original draft.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Clinical Trial Number: Not applicable.

Ethical Approval: This study used publicly available, de-identified data and was therefore exempt from institutional review board (IRB) oversight.

Conflict of Interest.

No conflict of interest

References

11. Chronic kidney disease and risk management: Standards of Care in Diabetes—2024. (2024). *Diabetes Care*, 47(Supplement_1), S219-S230.
2. Abdul Basith Khan, M., Hashim, M. J., King, J. K., Govender, R. D., Mustafa, H., & Al Kaabi, J. (2020). Epidemiology of type 2 diabetes—global burden of disease and forecasted trends. *Journal of epidemiology and global health*, 10(1), 107-111.
3. Abdullah, T. A. A., Zahid, M. S. M., & Ali, W. (2021). A review of interpretable ML in healthcare: taxonomy, applications, challenges, and future directions. *Symmetry*, 13(12), 2439.
4. Ahmad, M. A., Eckert, C., & Teredesai, A. (2018). Interpretable machine learning in healthcare.
5. Bliss, M. (2021). *The discovery of insulin: special centenary edition*. University of Toronto Press.
6. Chaki, J., Ganesh, S. T., Cidham, S. K., & Theertan, S. A. (2022). Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review. *Journal of King Saud University-Computer and Information Sciences*, 34(6), 3204-3225.
7. Colditz, G. A., Willett, W. C., Rotnitzky, A., & Manson, J. E. (1995). Weight gain as a risk factor for clinical diabetes mellitus in women. *Annals of internal medicine*, 122(7), 481-486.
8. Federation, I. D. (2021). IDF Diabetes Atlas 2021. *International Diabetes Federation*.
9. Ghazi, M. H., Abdulkareem, R., & Abbas, A. A. (2023). JOURNEY FROM SYMPTOM RECOGNITION TO SCREENING APPROACHES: A LITERATURE REVIEW ON DIABETES. *World*, 2(5).
10. Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.
11. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine Learning and Data Mining Methods in Diabetes Research. *Computational and Structural Biotechnology Journal*, 15, 104-116. <https://doi.org/https://doi.org/10.1016/j.csbj.2016.12.005>
12. Knowler, W. C., Bennett, P. H., Hamman, R. F., & Miller, M. (1978). Diabetes incidence and prevalence in Pima Indians: a 19-fold greater incidence than in Rochester, Minnesota. *American journal of epidemiology*, 108(6), 497-505.
13. Kolla, S. K. (2023). Big Data–Driven Machine Learning Frameworks for Clinical Risk Prediction. *International Journal of Medical Toxicology and Legal Medicine*, 26(3), 44-59.
14. Ling, S., Brown, K., Miksza, J. K., Howells, L., Morrison, A., Issa, E.,...Zaccardi, F. (2020). Association of type 2 diabetes with cancer: a meta-analysis with bias analysis for unmeasured confounding in 151 cohorts comprising 32 million people. *Diabetes care*, 43(9), 2313-2322.
15. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
16. Magliano, D. J., Boyko, E. J., & Atlas, I. D. F. D. (2021). COVID-19 and diabetes. In *IDF DIABETES ATLAS [Internet]. 10th edition*. International Diabetes Federation.
17. Oikonomou, E. K., & Khera, R. (2023). Machine learning in precision diabetes care and cardiovascular risk prediction. *Cardiovascular Diabetology*, 22(1), 259.
18. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5), 206-215.
19. Saeedi, P., Petersohn, I., Salpea, P., Malanda, B., Karuranga, S., Unwin, N.,...Ogurtsova, K. (2019). Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas. *Diabetes research and clinical practice*, 157, 107843.
20. Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K.-R. (2021). Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3), 247-278.
21. Siam, N. H., Snigdha, N. N., Tabasumma, N., & Parvin, I. (2024). Diabetes mellitus and cardiovascular disease: exploring epidemiology, pathophysiology, and treatment strategies. *Reviews in Cardiovascular Medicine*, 25(12), 436.
22. Sisodia, D., & Sisodia, D. S. (2018). Prediction of Diabetes using Classification Algorithms. *Procedia Computer Science*, 132, 1578-1585. <https://doi.org/https://doi.org/10.1016/j.procs.2018.05.122>
23. Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus.
24. Swapna, G., Vinayakumar, R., & Soman, K. P. (2018). Diabetes detection using deep learning algorithms. *ICT express*, 4(4), 243-246.

25. Tolles, J., & Meurer, W. J. (2016). Logistic regression: relating patient characteristics to outcomes. *Jama*, 316(5), 533-534.

26. Topol, E. (2019). *Deep medicine: how artificial intelligence can make healthcare human again*. Hachette UK.

27. Vecchio, I., Tornali, C., Bragazzi, N. L., & Martini, M. (2018). The discovery of insulin: an important milestone in the history of medicine. *Frontiers in endocrinology*, 9, 613.

28. World Health, O. (2024). *Guidance on global monitoring for diabetes prevention and control: framework, indicators and application*. World Health Organization.

Figures

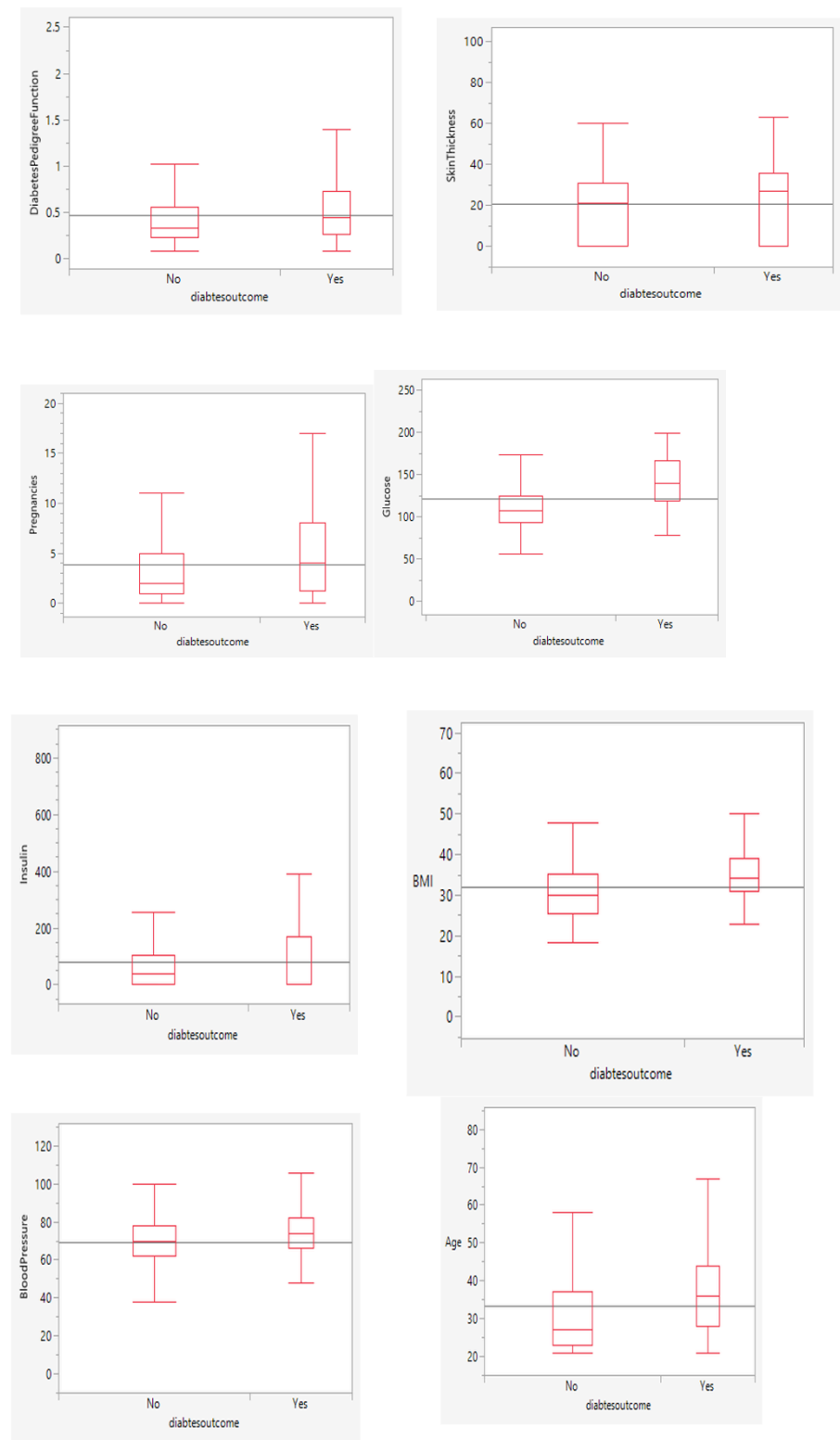


Figure 1

Boxplots comparing predictor distributions between diabetic and non-diabetic participants.

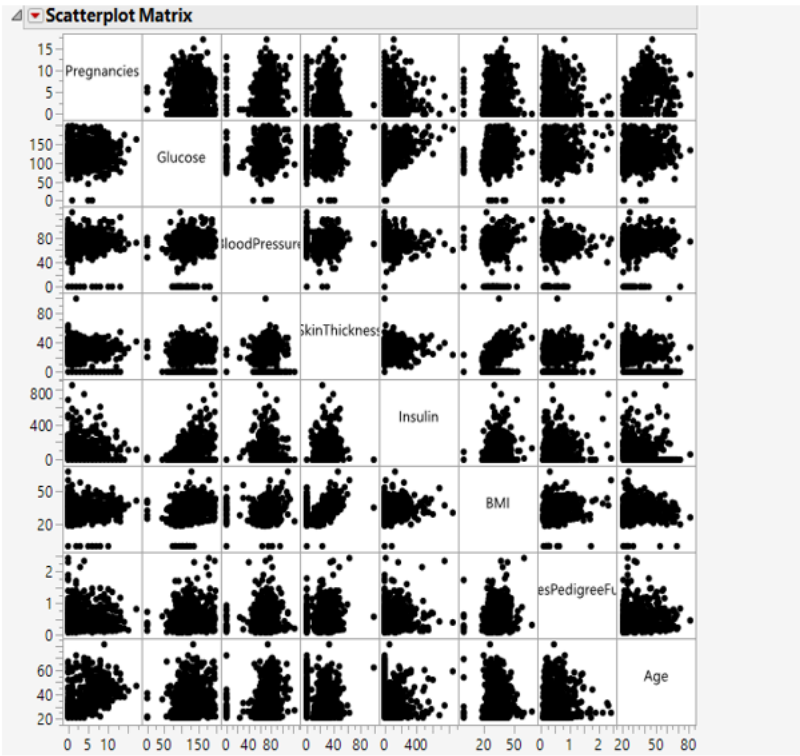


Figure 2

Scatterplot matrix and correlation matrix of predictor variables.

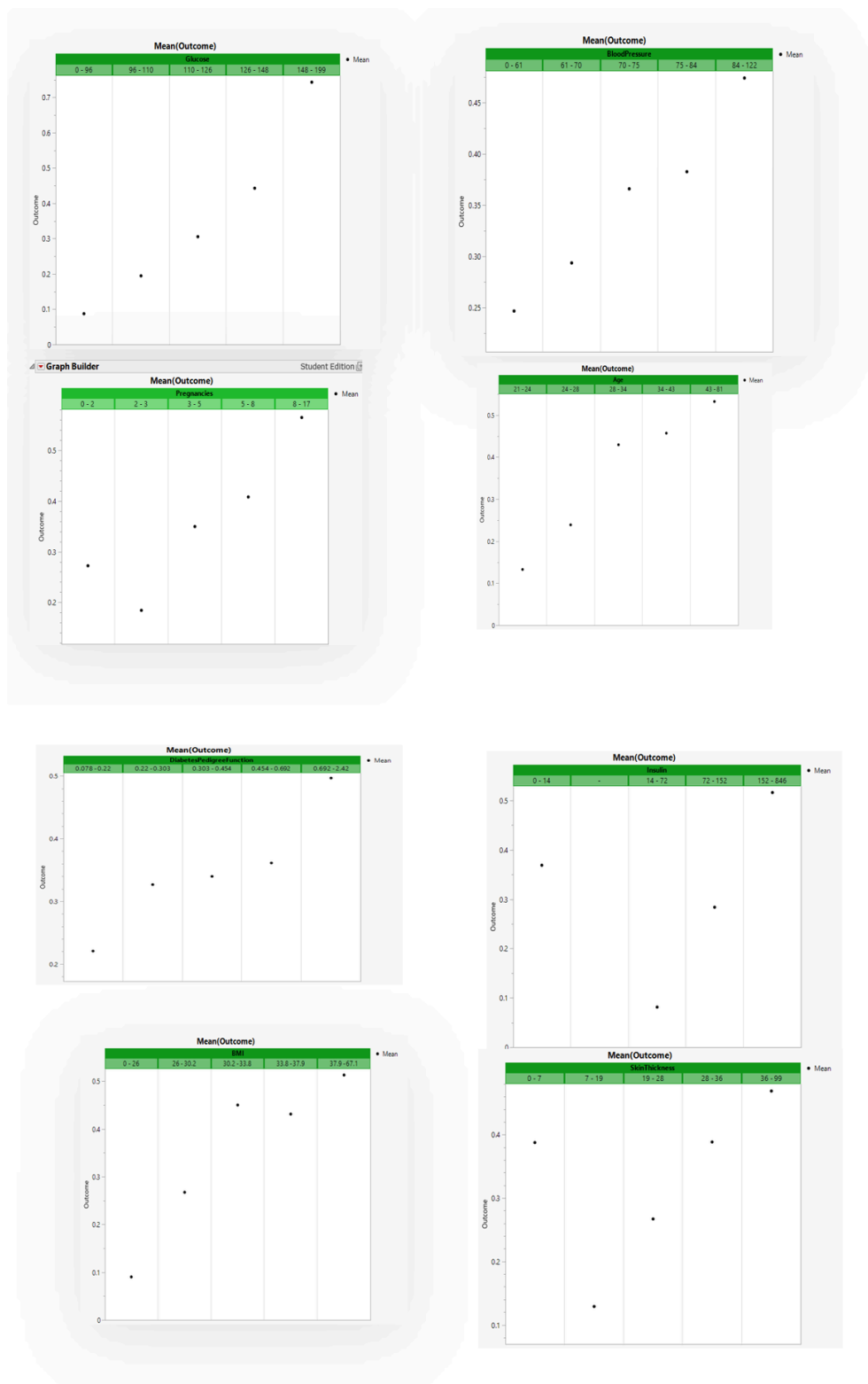


Figure 3

Empirical probability plots for retained predictor variables.

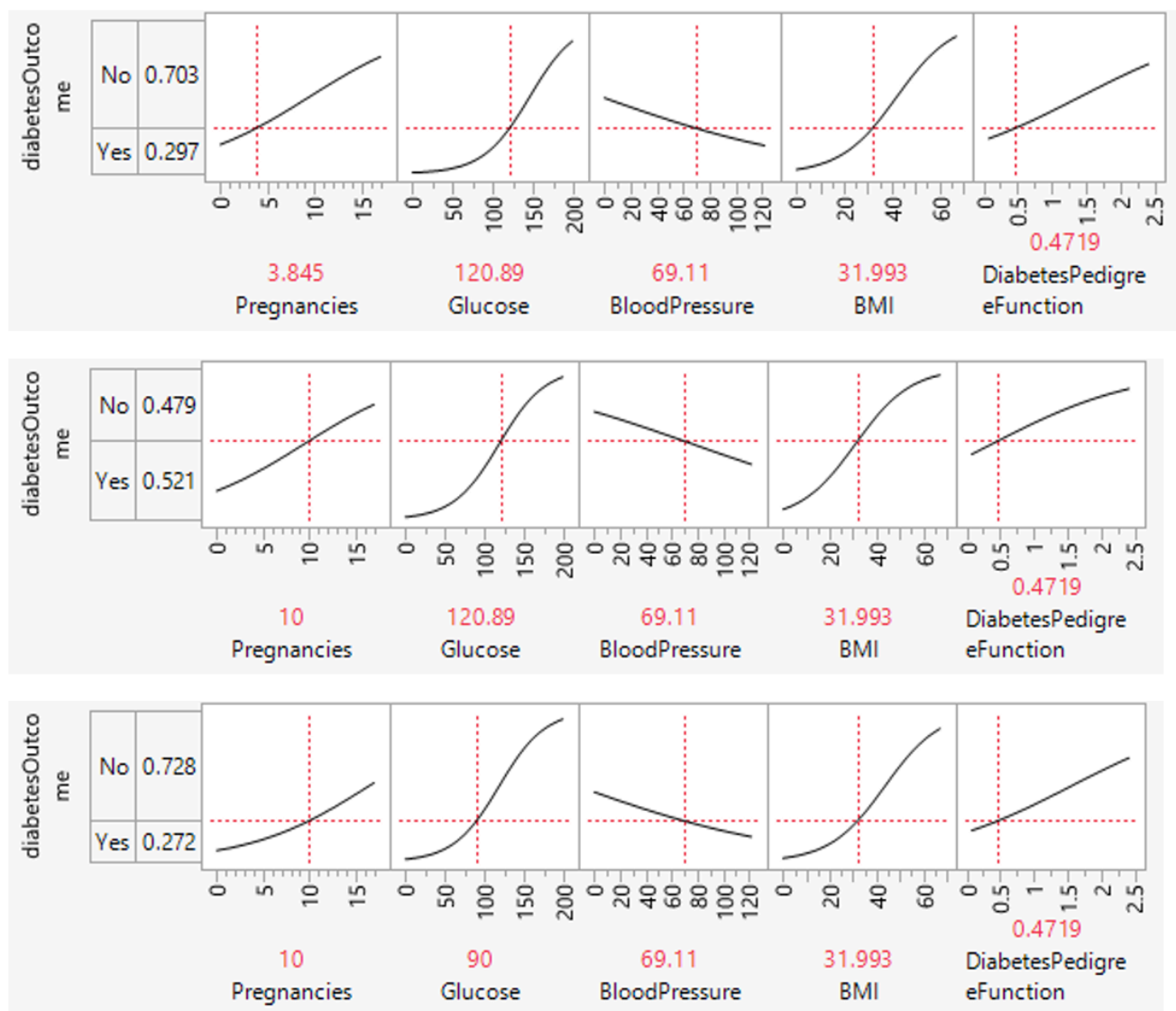


Figure 4

Profiler plots showing changes in diabetes probability across predictor values.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Highlights.docx](#)