Diabetes Statistical Report

DSA1101: Introduction to Data Science

Bryan Kee Tze Ren | A0192401R | e0363342@u.nus.edu

1. Overview

Abstract:

This report explores various classification methods for diabetes prediction using medical and demographic data. The dataset, comprising 100,000 patient records, was analyzed to identify key features associated with diabetes. Four classification models – K-Nearest Neighbors, Decision Tree, Naïve Bayes, and Logistic Regression – were evaluated based on their Sensitivity (True Positive Rate), Type 2 Error, Receiver Operating Characteristic (ROC) curve and Area Under Curve (AUC). The Decision Tree model emerged as the best model for predicting diabetes, exhibiting the highest Sensitivity (97.0%), the lowest Type 2 Error (28.5%), and the highest AUC value (0.966) among all models. This report concludes that the Decision Tree model is the most effective for diabetes screening, facilitating early intervention and reducing the risk of diabetes-related complications.

Introduction:

Diabetes Mellitus is a chronic metabolic disorder of increasing prevalence worldwide, causing significant health burdens on societies and individuals. One of the critical challenges posed by diabetes is its insidious onset and asymptomatic nature in the early stages, leading many individuals to be unaware of their condition until it progresses to advanced stages. Early detection is crucial to allow timely intervention and mitigate the onset of complications. However, traditional diagnostic methods rely on symptomatic presentation or invasive tests, which may not be feasible for large-scale population screening. Therefore, this report aims to develop a classification method using medical and demographic data to predict diabetes. This method can be used to screen populations and identify those at risk for further confirmatory tests and early intervention.

The dataset, provided by Mohammed Mustafa contains medical and demographic data of patients along with their diabetes status as either positive or negative. It consists of features such as age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, and blood glucose level.

2. Methods

Data Preparation and Preliminary Insights

The dataset comprises 100,000 patient records, featuring 4 quantitative variables and 5 categorical variables. In this report, diabetes status serves as the categorical response variable. Numeric categorical variables were declared as factors in R. The data was randomly split into 80% for training and 20% for testing. Given the 8.5% prevalence of diabetics in the dataset, this proportion was maintained in both the training and testing datasets to mitigate the impact of class imbalance on model performance. Additionally, the quantitative data was standardized solely for the KNN model, while remaining unchanged for other models. Visualizations illustrating the distribution of each variable are presented in Figure 1, including pie charts for categorical data and histograms for quantitative data. Notably, there are slightly more females than males, both heart disease and hypertension patients make up less than 10% of the dataset, and none of the quantitative variables appear to follow a normal distribution.

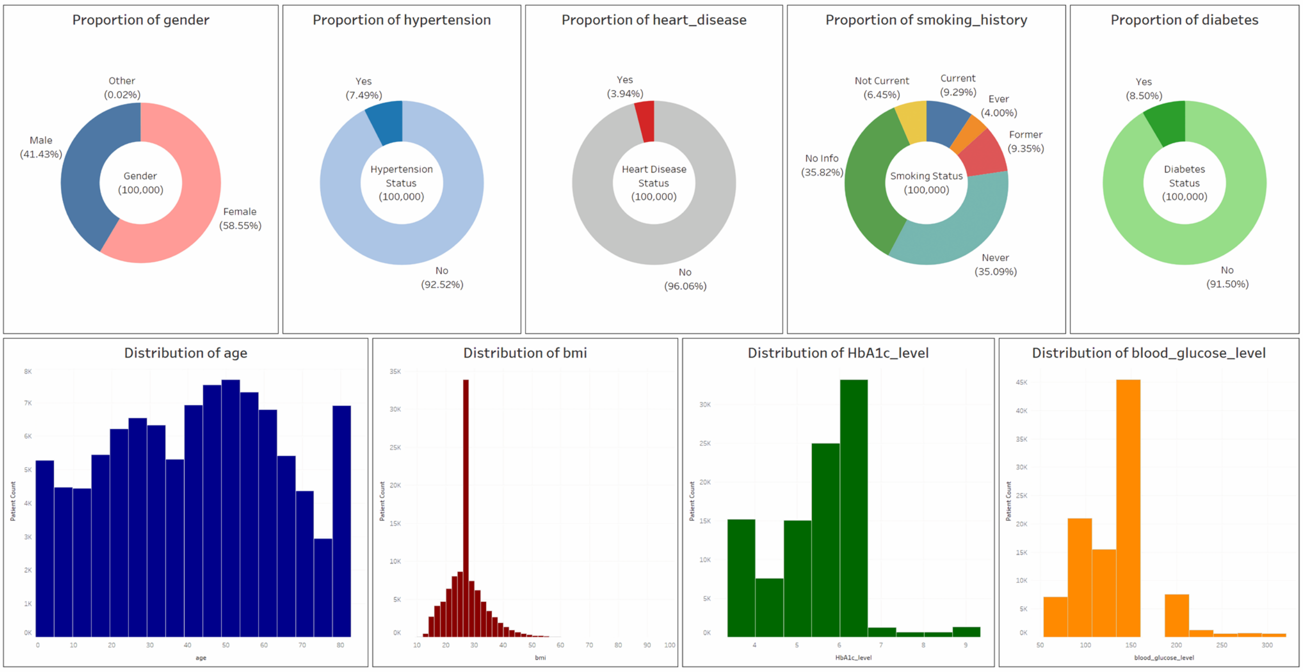

Figure 1: Distribution of Variables in Dataset

Exploratory Data Analysis:

This section explores the relationship between the input variables and diabetes status to determine their inclusion in the model. For categorical variables (gender, hypertension, heart disease, and smoking history), contingency tables were generated to observe the proportion of diabetics within each category. Subsequently, risk ratios were calculated to gauge the impact of each variable on diabetic risk. Visual representations of these proportions were illustrated using bar graphs (refer to Figure 2). Likewise, quantitative variables (age, BMI, HbA1c level, and blood glucose level) were analyzed using box plots to compare their distribution between diabetic and non-diabetic individuals (refer to Figure 3). Due to the non-normal distribution of the quantitative variables, median values were compared to assess their association with diabetes.

The following associations support the inclusion of all input variables into subsequent models:

- **Gender:** Males (9.75%) exhibit a 28% higher diabetic risk than females (7.62%). Comparisons were not made with the "Other" category due to its negligible representation of 18 entries.
- **Hypertension:** Hypertensive patients (27.90%) display a 4.02 times higher diabetic risk than non-hypertensive patients (6.93%).
- **Heart Disease:** Patients with heart disease (32.14%) display a 4.27 times higher diabetic risk compared to patients without heart disease (7.53%).
- **Smoking History:** Individuals with a positive smoking history show a notable elevation in diabetes risk across all categories compared to those who have never smoked, with risk ratios ranging from 1.07 to 1.78. For those with unknown smoking history, despite having a risk ratio smaller than 1 (0.43), no meaningful associations can be made due to the lack of data for this class.
- **Quantitative Variables:** In all cases of age, BMI, HbA1c, and glucose, the boxplots for diabetics consistently exhibit higher median values compared to non-diabetics. However, there is notable overlap between the tails and heads of the boxplots, suggesting some degree of variability. Specifically, the medical variables (HbA1c, glucose) display a smaller overlap in their interquartile ranges between diabetics and non-diabetics compared to the demographic variables (age, BMI).

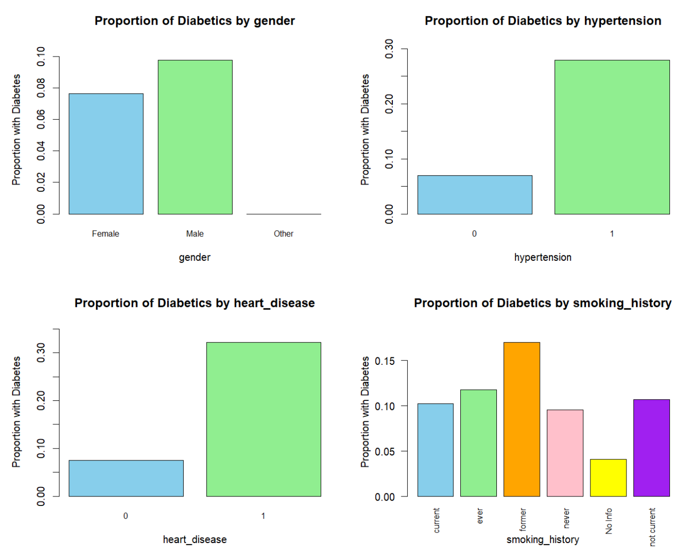

Figure 2: Proportion of Diabetics Across Categorical Variables

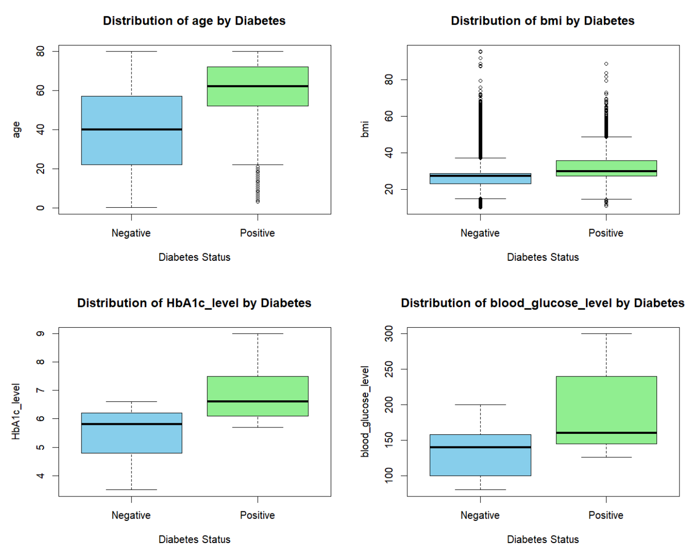

Figure 3: Distribution of Quantitative Variables by Diabetes Status

Model Building:

Model 1: K-Nearest Neighbors

The K-Nearest Neighbors (KNN) model utilized all standardized quantitative features, including age, BMI, HbA1c level, and blood glucose level. A range of k values from 1 to 30 was assessed to determine the most suitable k value for the model. Performance metrics such as Accuracy, Sensitivity, Precision, Type 1 Error and Type 2 Error were computed for each k value, and the results were graphed to visualize the model's performance across different k values (refer to Figure 4). The k value of 3 was chosen, prioritizing Sensitivity and Type 2 Error to align with the report's objective (see Discussion section for detailed rationale). Larger values of k worsened Sensitivity and Type 2 Error without significant gains in the other metrics. Subsequently, the KNN model was trained using the selected k value, and predictions were made on the test dataset.

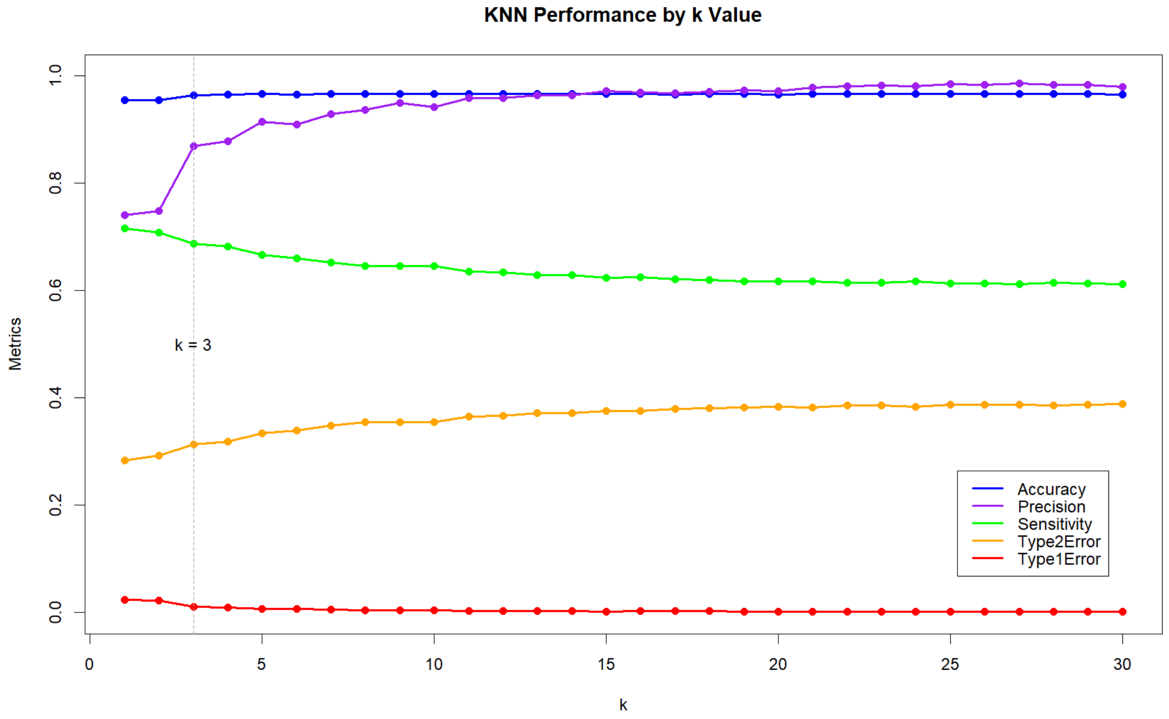

Figure 4: KNN Performance by k Value

Model 2: Decision Tree

The Decision Tree model utilized every input feature. A range of complexity parameter (cp) values from 10^-1 to 10^-10 was assessed to determine the most suitable cp value for the model. Performance metrics such as Accuracy, Sensitivity, Precision, Type 1 Error and Type 2 Error were computed for each cp value, and they were graphed to visualize the model's performance across different cp values (refer to Figure 5). The cp value of 10^-4 was chosen, prioritizing Sensitivity and Type 2 Error to align with the report's objective (see Discussion section for detailed rationale). Larger cp values worsened Sensitivity and Type 2 Error while smaller cp values reduced Precision without affecting other metrics. Finally, the Decision Tree model was trained using the chosen cp value, and predictions were made on the test dataset. Due to the complexity of the final Decision Tree model, visualizing the tree diagram was not feasible. However, the focus was prioritized on optimizing model performance rather than visual representation.

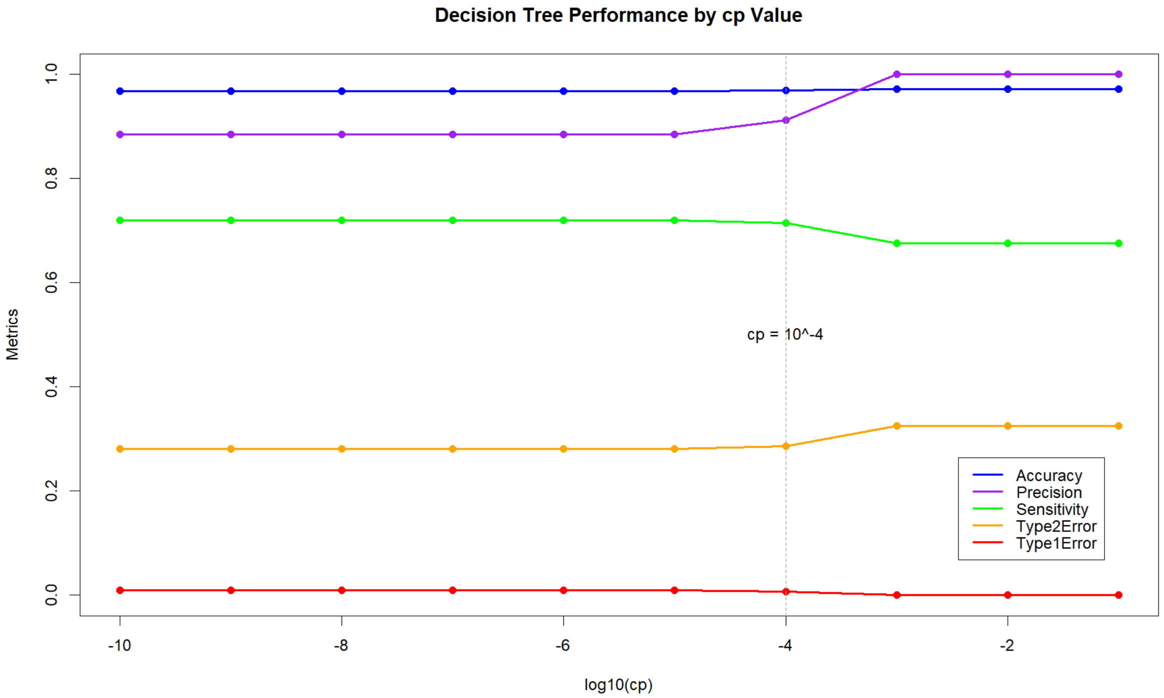

Figure 5: Decision Tree Performance by cp Value

Model 3: Naïve Bayes

The Naïve Bayes model utilized all categorical features, including gender, hypertension, heart disease and smoking history. A significant assumption was made regarding the independence of categorical variables when forming the model. This assumption implies that the presence of one categorical variable does not influence the presence of another, thereby simplifying the modeling process.

Model 4: Logistic Regression

The Logistic Regression model utilized all input features. It was observed that all coefficients for the quantitative variables were found to be statistically significant, with p-values < 0.001. Similarly, at least one class from each categorical variable showed statistical significance. Therefore, all variables were retained in the final logistic regression model. The final equation is shown below in Figure 6.

$$\boxed{\begin{aligned}\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = {} & -27.085 + 0.0462(\text{age}) + 0.0890(\text{bmi}) + 2.340(\text{HbA1c\_level}) + 0.0334(\text{blood\_glucose\_level}) \\ & + 0.272[I(\text{Gender} = \text{Male})] - 9.475[I(\text{Gender} = \text{Other})] \\ & + 0.741[I(\text{hypertension} = 1)] + 0.735[I(\text{heart\_disease} = 1)] \\ & - 0.0510[I(\text{smoking\_history} = \text{ever})] - 0.108[I(\text{smoking\_history} = \text{former})] \\ & - 0.157[I(\text{smoking\_history} = \text{never})] - 0.211[I(\text{smoking\_history} = \text{not current})] \\ & - 0.730[I(\text{smoking\_history} = \text{No Info})]\end{aligned}}$$

Figure 6: Logistic Regression Model Equation

3. Results & Discussion:

Performance Metrics and Rationale

In this report, Sensitivity (True Positive Rate) and Type 2 Error were the primary metrics used to evaluate a model's ability to screen for diabetes. Prioritizing Sensitivity maximizes diabetes detection while minimizing Type 2 Error reduces the risk of overlooking diabetics. The decision to place less emphasis on Accuracy and Precision was due to the dataset's class imbalance, which has the potential to skew these metrics. Despite the importance of Type 1 Error in reducing false positives, medical screening emphasizes Sensitivity over Specificity. This approach can be complemented by confirmatory tests administered by healthcare professionals to verify the model's positives. Given the non-urgent nature of a diabetes diagnosis, some false positives are acceptable to ensure the identification of at-risk individuals.

This rationale guided the selection of k and cp values during model building and will inform the subsequent comparison between models. Each model was used to predict diabetes status using the test data, and the resulting metrics were compiled in a table for comparison (refer to Table 1). For the logistic regression model, a threshold of 0.5 was chosen to align with the default thresholds used in other models.

Model	Accuracy	Sensitivity	Type 1 Error	Type 2 Error	Precision
K Nearest Neighbours	0.965	0.687	0.00967	0.313	0.868
Decision Tree	0.970	0.715	0.00639	0.285	0.912
Naïve Bayes	0.913	0.034	0.00574	0.966	0.356
Logistical Regression	0.959	0.624	0.00984	0.376	0.855

Table 1: Model Performance Metrics

Receiver Operating Characteristic (ROC) Curves

ROC curves were used to illustrate the trade-off between Sensitivity and Specificity across various classification thresholds, with the Area Under Curve (AUC) serving as a performance metric. Figure 7 presents these curves, where steeper curves indicate better discrimination ability and curves closer to the top-left corner exhibit higher Sensitivity and lower False Positive Rate. The legend in Figure 7 contains the AUC values, where higher values (closer to 1) indicate better model performance in distinguishing between diabetic and non-diabetic cases.

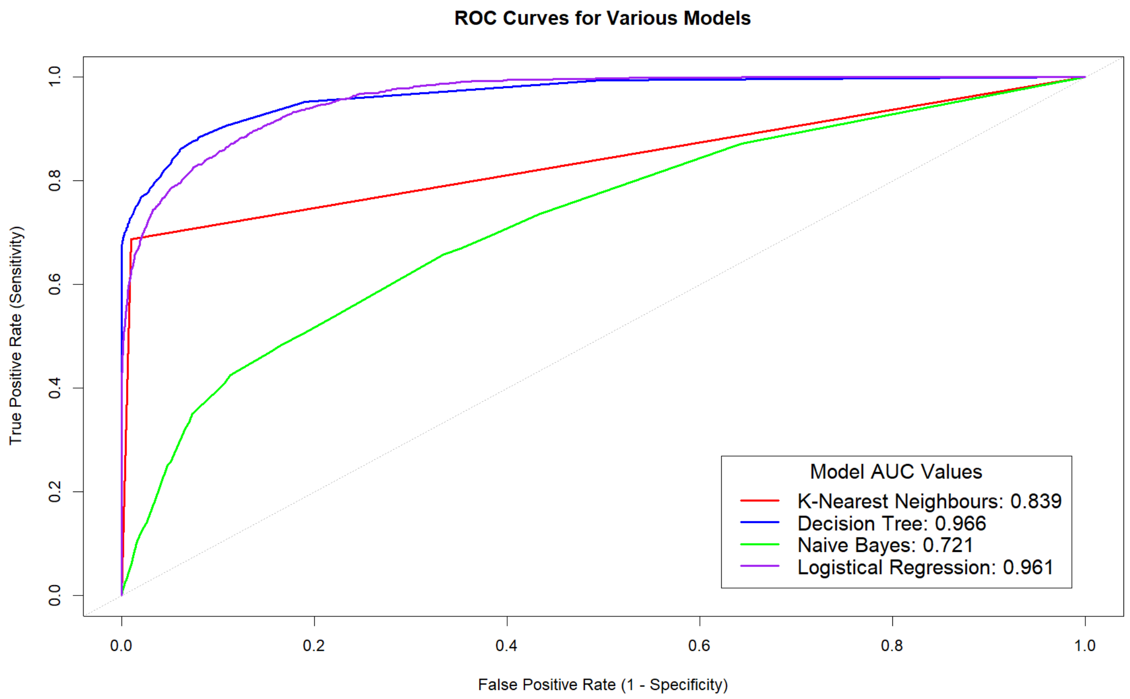

Figure 7: Receiver Operating Characteristic Curves

Model Evaluation and Comparison

Model 1: K-Nearest Neighbors

KNN demonstrated good performance with a moderate Sensitivity (68.7%), a relatively low Type 2 Error (31.3%) and a respectable AUC value (0.839). The ROC curve is positioned relatively near the top-left corner, indicating moderate discriminatory ability. However, KNN's inability to process categorical inputs limits its versatility when it comes to mixed datasets. Additionally, KNN can be computationally intensive, especially when cross-validating or dealing with large datasets and varying K values, leading to longer processing times.

Model 2: Decision Tree

The Decision Tree demonstrated exceptional performance with the highest Sensitivity (97.0%), the lowest Type 2 Error (28.5%), and the top AUC (0.966) among all models. Notably, its ROC curve was nearest to the top-left corner compared to any other model, indicating excellent discriminatory ability. While this showcases the model's effectiveness on the dataset of 100,000 patients, it may also be the result of overfitting, which could compromise the model’s generalizability to unseen data, especially when applied to different patient demographics. Additionally, the visualization of the tree structure may become challenging for complex trees.

Model 3: Naïve Bayes

The Naive Bayes model displayed inferior performance compared to other models, with the lowest Sensitivity (3.41%), the highest Type 2 Error (96.6%), and the worst AUC (0.721). Its ROC curve was also furthest from the top-left corner compared to any other model, indicating poor discriminatory ability. This may be attributed to the model's assumption of feature independence, which does not fully capture the interrelationships observed in related cardiovascular conditions, such as hypertension and heart disease. Additionally, its inability to process quantitative inputs limits its versatility. Overall, the model’s tendency to misclassify actual diabetics makes it a poor choice for screening a population.

Model 4: Logistic Regression

The Logistic Regression model demonstrated competitive performance, with a moderate Sensitivity (62.4%), a decent Type 2 Error (37.6%), and a high AUC (0.961). Despite its slightly lower Sensitivity and higher Type 2 Error compared to other models, its ROC curve and AUC value was similar to the Decision Tree, indicating strong discriminatory ability. However, the model assumes a linear relationship between the features and the log-odds of the outcome, which may not always hold true in complex datasets.

4. Conclusion

Optimal Classifier for Diabetes Screening

Based on the evaluation presented in Section 3, the Decision Tree model emerges as the best choice for diabetes screening. It achieved the highest Sensitivity, the lowest Type 2 Error, and the best ROC curve among all other models. Notably, it also demonstrated the highest Accuracy and Precision, even though these metrics were not the primary focus of the report. Its ability to process both quantitative and categorical data within a reasonable timeframe makes it well-suited for analyzing large and mixed healthcare datasets. Although the risk of overfitting to our dataset may compromise its generalizability to other patient populations, with an appropriately sized and diverse training dataset, the Decision Tree has the potential to be a highly effective tool. Therefore, this report concludes that the Decision Tree is the optimal classification model for diabetes screening.