Predicting Diabetes Onset: A Predictive Modeling Approach Using the Pima Indians Dataset.

Predictive Modeling Algorithms

Youngstown State University

Spring 2025

Rengitha Dowlings & Felix Kina

Table of Contents

1. ABSTRACT

Diabetes is a chronic metabolic disorder with significant health and economic impacts. Early prediction and diagnosis are critical for effective disease management and prevention of complications. In this project, we used the publicly available Pima Indians Diabetes dataset, which includes clinical and personal attributes for 768 female patients of Pima Indian heritage, to predict the onset of Type 2 diabetes. Three supervised classification models, Logistic Regression, Random Forest, and Classification Tree were developed to evaluate their predictive power and interpretability. Each model was trained on 70% of the data and tested on the remaining 30%. Model performance was assessed using test error rates and confusion matrices. Logistic Regression achieved an accuracy of 78.8% and an error rate of 21.2%, highlighting glucose, BMI, and family history as key predictors. Random Forest demonstrated robustness in capturing complex interactions with an accuracy of approximately 78.35% and an error rate of 21.65%. The pruned Classification Tree outperformed the others with the highest accuracy (81.4%) and lowest error rate (18.6%), while also offering clear decision rules. The results confirm the effectiveness of predictive modeling in identifying diabetes risk and reinforce the clinical importance of glucose levels, BMI, and genetic predisposition as significant indicators of disease onset.

2. INTRODUCTION

Diabetes mellitus, commonly known as diabetes, is a metabolic disorder characterized by elevated blood glucose (sugar) levels. It occurs when the body cannot produce enough insulin, a hormone critical for regulating blood sugar or cannot effectively use the insulin it produces. Diabetes is a chronic and potentially life-threatening disease that affects millions of people worldwide. Over time, uncontrolled diabetes can lead to serious complications affecting the heart, kidneys, eyes, nerves, and other organs. Early diagnosis plays a crucial role in effective treatment and management, helping to reduce long-term complications. In this project, we explore the use of predictive modeling to identify the likelihood of Type 2 diabetes onset based on medical and personal attributes.

The dataset used is the well-known Pima Indians Diabetes dataset, which contains diagnostic measurements for female patients of Pima Indian heritage aged 21 and older. The goal is to apply predictive modeling techniques to address the problem of predicting diabetes onset among women of Pima Indian heritage. Through data preprocessing, feature selection, and model evaluation, this project aims to uncover insights that not only demonstrate technical understanding of predictive analytics but also underscore the real-world importance of data-driven healthcare solutions.

3. DATA COLLECTION AND PREPROCESSING

3.1 Data Description

The dataset used in this project is the Pima Indians Diabetes dataset, originally collected by the National Institute of Diabetes and Digestive and Kidney Diseases. It was obtained from an online source. It contains medical diagnostic measurements for 768 female patients of Pima Indian heritage, all aged 21 years and older. The dataset contains 9 variables including Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, DiabetesPedigreeFunction, Age and Outcome. The response variable for the dataset is Outcome**,** which is binary with 0 if non-diabetic, 1 if diabetic and the remaining 8 as the predictor or independent variables.

3.2 Data Processing and Data Training

We performed a thorough exploration to understand the data distribution and identify any anomalies. The data contained no missing data, however features like Glucose, Blood Pressure, Skin Thickness, Insulin, and BMI had zero values which are medically impossible. These values were replaced with median values to ensure data quality. We chose three supervised learning algorithms: Logistic Regression, Classification Tree, and Random Forest. Logistic Regression was selected for its simplicity, interpretability, and strong theoretical foundation in modeling binary outcomes. The Classification Tree was included for its ability to provide clear, rule-based decision paths that are easy to visualize and explain to non-technical audiences. Finally, the Random Forest algorithm was selected for its robustness and high predictive power. All three models were trained on 70% of the data, with 30% reserved for testing. Model performance is assessed using the mean test error rate and confusion matrix.

4. EXPLORATORY DATA ANALYSIS

Before applying predictive models to the data, an initial exploration of the dataset reveals key characteristics of the Pima Indian female population studied.

Variable	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Pregnancies	0.00	1.000	3.000	3.85	6.00	17.000
Glucose	44.00	99.75	117.00	121.66	140.25	199.00
Blood Pressure	24.00	64.00	72.00	72.39	80.00	122.00
Skin Thickness	7.00	23.00	23.00	27.33	32.00	99.00
Insulin	14.00	30.50	31.25	94.65	127.25	846.00
BMI	18.20	27.50	32.00	32.45	36.60	67.10
DiabetesPedigreeFunction	0.08	0.24	0.37	0.47	0.63	2.42
Age	21.00	24.00	29.00	33.24	41.00	81.00
Outcome	0.000	0.00	0.000	0.349	1.000	1.000

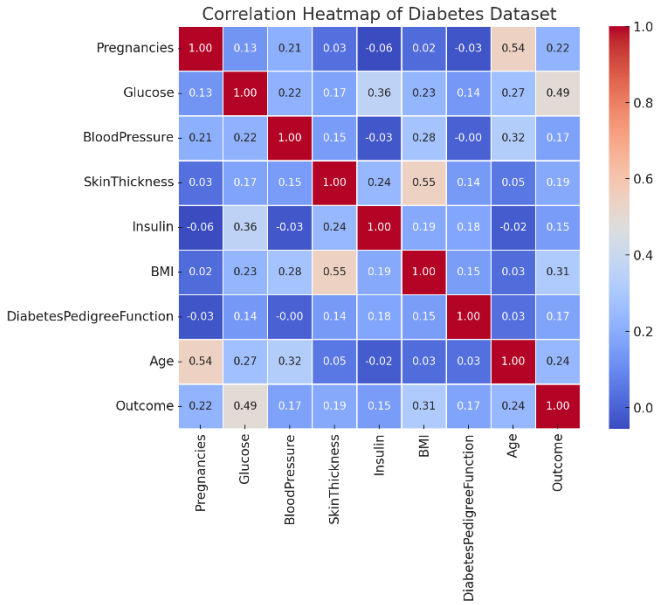
Table 1: Table of summary statistics.

Based on the summary statistics, the average age is approximately 33 years, and the BMI mean of 32.45 suggests a generally overweight cohort. Participants had an average of 3.8 pregnancies, with values ranging from 0 to 17. The mean glucose level is 121.66 mg/dL, with some individuals showing significantly high levels (up to 199), indicating possible risk of hyperglycemia. Insulin and Skin Thickness variables displayed high variability, with insulin levels reaching a maximum of 846 µU/mL.

The Diabetes Pedigree Function, which estimates genetic predisposition to diabetes, has a mean of 0.47, with values extending up to 2.42, showing considerable variation across the sample. Finally, the Outcome variable indicates that approximately 35% of the patients were diagnosed with diabetes and 65% not having diabetes.

4.1 Correlation Heatmap of the Dataset.

To assess the relationships among predictor variables before fitting the models, we examined the correlation matrix of the dataset. The outcome variable was changed to numeric to create the heatmap. From the map, most variables showed weak to moderate correlations, indicating minimal risk of multicollinearity.



The strongest correlation observed was between Skin Thickness and BMI ($r \approx 0.65$), suggesting a moderate association, but not high enough to warrant exclusion of either variable.

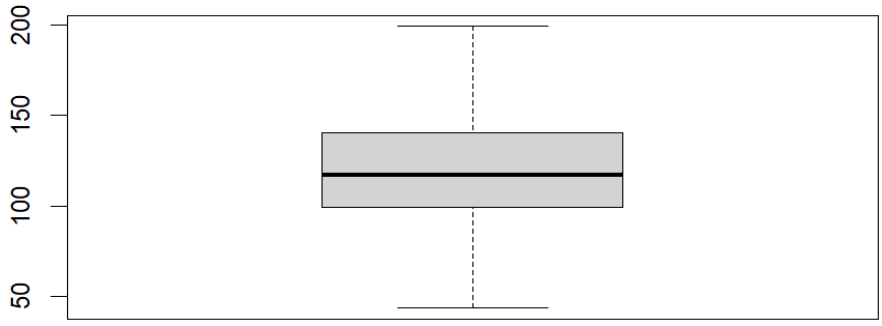
Glucose, Age, and Diabetes Pedigree Function showed relatively low correlations with other predictors, supporting their inclusion in the model.

Graph 1: Correlation heatmap of the independent variables and outcome variable.

4.2 Distribution of Independent Variables

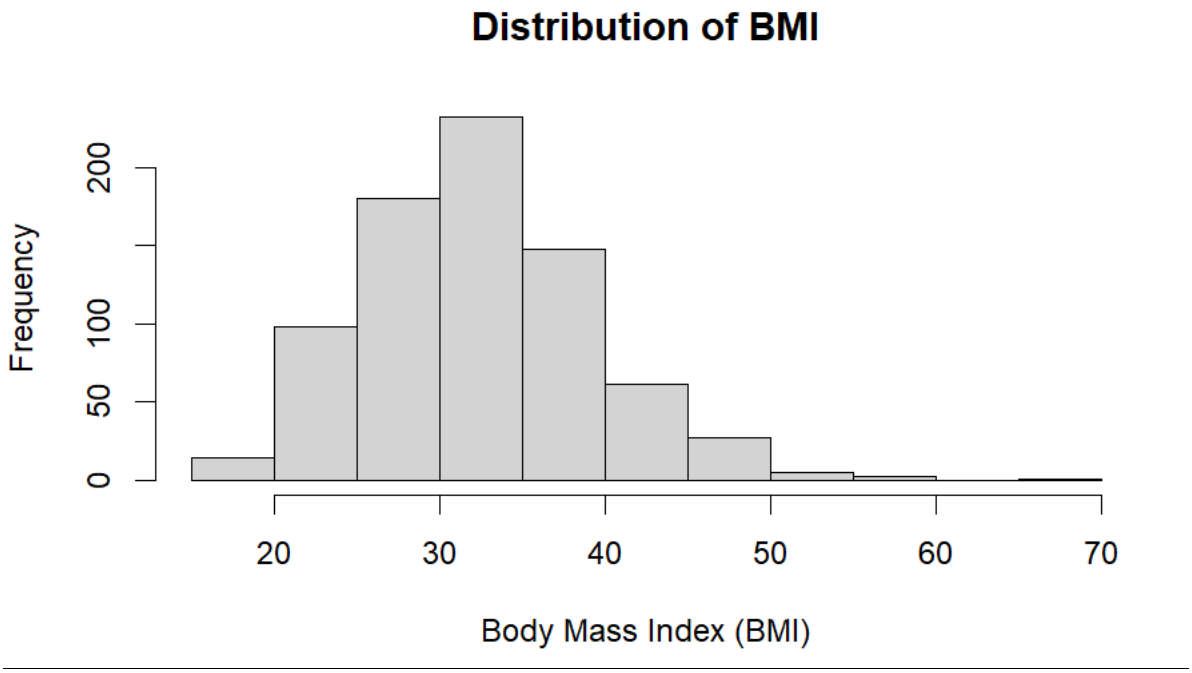
Understanding the distribution of independent variables is essential for identifying patterns, potential outliers, and preparing data for modeling. We examined key predictors such as Glucose, Pregnancies, and BMI, all of which are clinically relevant to diabetes risk.

Graph 2: Histogram showing the distribution of Glucose



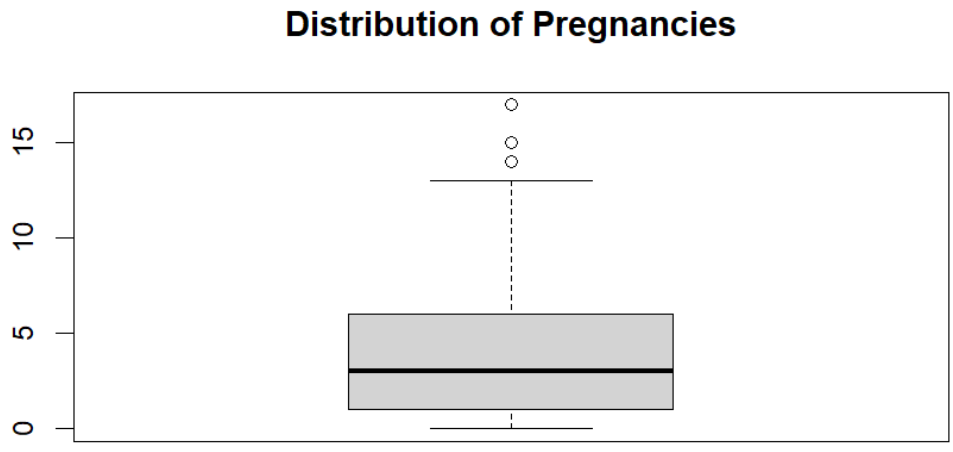
The boxplot illustrates the distribution of glucose levels among participants in the dataset. The median glucose value lies around 120 mg/dL, with most values falling between approximately 90 and 150 mg/dL. The interquartile range (IQR) suggests a moderate spread in the central 50% of the data. The distribution appears slightly skewed to the right, which may suggest the presence of higher glucose values in a subset of the population, consistent with individuals at risk of or already experiencing hyperglycemia.

Graph 3: Distribution of Body Mass Index (BMI)



The distribution of BMI is approximately right skewed, with most values falling between 25 and 40, peaking around 30, which is classified as overweight according to standard health guidelines. The overall shape suggests that most participants fall within the overweight to obese range, reinforcing BMI as a relevant predictor of diabetes risk in this dataset.

Graph 4: Boxplot Showing the Distribution of Pregnancies



Based on the boxplot, the median is around 3, with most values falling between 1 and 7. The distribution is right-skewed, and some outliers are present above 12, indicating that a few individuals reported a significantly higher number of pregnancies, up to 17.

5. Data Modelling

5.1 Logistic Regression

To ensure the appropriateness of logistic regression for modeling the relationship between clinical predictors and the likelihood of diabetes, we conducted a thorough assessment of the model assumptions. The response variable is binary, with values coded as 0 (no diabetes) and 1 (diabetes). This satisfies the fundamental requirement for binary logistic regression. Each observation in the dataset corresponds to a unique patient record hence there is no indication of repeated measurements or clustering within the dataset, and thus the assumption of independence is met. A correlation matrix was generated to assess potential multicollinearity among the predictor variables. The analysis revealed moderate correlations between a few variables (e.g., Skin Thickness and BMI, $r \approx 0.65$), but no correlation exceeded the commonly accepted threshold. This indicates that multicollinearity is not a significant concern in this dataset. Two predictors, Diabetes Pedigree Function and Age, showed significant interaction effects, indicating a violation of the linearity assumption. To address this, both variables were log-transformed prior to model fitting to improve the linearity of their relationship with the logit.

Fig 1: Summary of full logistic training model

```
Call:
glm(formula = Outcome ~ ., family = "binomial", data = train)

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -9.416901   1.545409  -6.093 1.10e-09 ***
Pregnancies               0.123599   0.038013   3.252  0.00115 **
Glucose                   0.035708   0.004389   8.135 4.12e-16 ***
BloodPressure            -0.013008   0.010390  -1.252  0.21062
SkinThickness            -0.009907   0.013891  -0.713  0.47574
Insulin                  -0.002040   0.001149  -1.776  0.07582 .
BMI                       0.105339   0.020528   5.131 2.88e-07 ***
DiabetesPedigreeFunction  0.477107   0.180616   2.642  0.00825 **
Age                       0.643917   0.423261   1.521  0.12818
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 704.80  on 536  degrees of freedom
Residual deviance: 520.52  on 528  degrees of freedom
AIC: 538.52

Number of Fisher Scoring iterations: 5
```

A logistic regression model was fitted using the training data to assess the relationship between the predictors and the likelihood of diabetes onset. The model included all eight predictors. Among these, Glucose, BMI, Pregnancies, and Diabetes Pedigree Function were statistically significant predictors of diabetes risk. Insulin showed borderline significance, while Blood Pressure**,** Skin Thickness, and Age were not statistically significant.

Based on the model, a one-unit increase in glucose level is associated with an estimated 0.0357 increase in the log-odds of diabetes, holding all other variables constant. Also, for each one-unit increase in BMI, the odds of being diagnosed with diabetes increase by approximately 11.1%, holding all other variables constant. The model achieved a residual deviance of 520.52 on 528 degrees of freedom and an AIC of 538.52, indicating a reasonably good fit to the training data.

Model Selection

A backward elimination method was employed to identify the most parsimonious and statistically robust logistic regression model for predicting diabetes onset. This iterative process began with the full model containing all predictors and systematically removed variables that did not show statistical significance. Three predictors, Blood Pressure, Skin Thickness, and Age were excluded from the model due to their lack of statistical significance and minimal impact on overall model performance.

Fig 2: Summary of final model after elimination.

```
Call:
glm(formula = Outcome ~ Pregnancies + Glucose + Insulin + BMI +
    DiabetesPedigreeFunction, family = "binomial", data = train)

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -8.066591   0.843798  -9.560  < 2e-16 ***
Pregnancies               0.144430   0.031559   4.577 4.73e-06 ***
Glucose                   0.036240   0.004222   8.584  < 2e-16 ***
Insulin                  -0.002230   0.001114  -2.002  0.04531 *
BMI                       0.091745   0.017489   5.246 1.55e-07 ***
DiabetesPedigreeFunction  0.482794   0.179758   2.686  0.00724 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 704.80  on 536  degrees of freedom
Residual deviance: 524.08  on 531  degrees of freedom
AIC: 536.08

Number of Fisher Scoring iterations: 4
```

The final model retained five key predictors: Pregnancies, Glucose, Insulin, BMI, and DiabetesPedigreeFunction. All retained variables were statistically significant at the 5% level, indicating strong associations with the likelihood of diabetes diagnosis. This refined model achieved a residual deviance of 524.08 on 531 degrees of freedom and an AIC of 536.08, indicating an improved and more efficient fit compared to the full model.

Final Logistic Model Performance

To evaluate the performance of the final logistic regression model, predictions were made on the test dataset, and a confusion matrix was generated to assess classification outcomes. The confusion matrix is as follows:

	Predicted: 0	Predicted: 1
Actual: 0	145	14
Actual: 1	35	37

This indicates that 145 non-diabetic and 37 diabetic individuals were correctly classified. However, 14 non-diabetic and 35 diabetic individuals were falsely predicted as diabetic (false positives). The overall test error rate was calculated as 21.2%, indicating the final model misclassified approximately 21.2% of the test cases, which corresponds to an accuracy of about 78.8%.

5.2 Random Forest

To further improve predictive performance and capture relationships among variables, a Random Forest classifier was trained using the training dataset. The model was built with 500 decision trees, and at each split, 3 predictors were randomly selected for consideration. Variable importance was enabled to assess the contribution of each predictor to model performance. The OOB estimate of error rate for the model was estimated at 26.44%, reflecting the internal cross-validation error provided by the random forest algorithm with a test error rate of 21.65%. The confusion matrix from the predictions is as follows:

Fig 3: Summary of training model for Random Forest.

```
Call:
 randomForest(formula = Outcome ~ ., data = train, ntree = 500,      mtry = 3, importance = T)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 26.44%
Confusion matrix:
    0   1 class.error
0 278  63   0.1847507
1  79 117   0.4030612
> |
```

This indicates that the model correctly identified 278 non-diabetic and 117 diabetic individuals. It also misclassified 63 non-diabetic cases as diabetic (false positives), and 79 diabetic cases as non-diabetic (false negatives). The model was more accurate in classifying non-diabetic individuals (class error = 18.45%) than diabetic individuals (class error = 40.31%).

5.3 Classification Tree

A classification tree model was developed using the training dataset to predict the likelihood of diabetes onset. The tree was constructed using the *rpart* method with default settings for classification tasks. The unpruned tree had an error rate of 19.91% and a confusion matrix which indicates that indicates that the model correctly identified 141 non-diabetic and 44 diabetic individuals. It also misclassified 28 non-diabetic cases as diabetic (false positives), and 18 diabetic cases as non-diabetic (false negatives). The tree was then pruned with a cp value of 0.015306. The test error rate of the pruned was slightly lower than that of the unpruned tree with a rate of 18.61% and for the confusion matrix, the model correctly predicted 136 non-diabetic and 52 diabetic individuals correctly and misclassified 23 non-diabetic and 20 diabetic individuals.

6.0 CONCLUSION

In this study, we developed and evaluated three supervised learning models—logistic regression, random forest, and classification tree—to predict the likelihood of diabetes onset using clinical and demographic features. Each model offered unique strengths: logistic regression provided interpretability and statistical insight into predictor significance; random forest demonstrated robustness and the ability to model complex interactions; and the classification tree offered intuitive, rule-based decision paths. Among the three, the pruned classification tree emerged as the best-performing model, achieving the lowest error rate (18.6%) and the highest classification accuracy (81.4%). It balanced predictive performance with clarity, making it especially suitable for real-world application in clinical settings where decisions must be transparent and explainable. Logistic regression achieved an accuracy of 78.8% with an error rate of 21.2%, and provided valuable interpretability and statistical inference, identifying glucose, BMI, pregnancies, and family history as significant predictors. Random forest showed strong potential for capturing complex patterns and variable interactions, with an accuracy of approximately 78.65% and a test error rate of 21.65%, although it struggled slightly with class imbalance. Overall, the models consistently highlighted glucose levels, BMI, and family history as key predictors of diabetes, reinforcing their relevance in early detection and screening efforts.

While the models provided useful insights and strong predictive performance, there are several limitations to consider. First, the dataset used is specific to a single ethnic group, which may limit the generalizability of the findings to broader populations. Second, several important predictors of diabetes—such as diet, physical activity, genetic markers, and HbA1c levels—were not included in the dataset, potentially affecting the completeness of the models. Lastly, class imbalance (with more non-diabetic than diabetic cases) may have affected sensitivity, particularly in the random forest model, which showed a higher false negative rate. Future work could include using a larger, more diverse dataset, incorporating additional clinical features, and applying advanced techniques to improve sensitivity and model robustness.

7.0 References

Şahin, Ş. (2022). *Pima Indians diabetes classification project*. Medium.

Pradaschnor, N. (n.d.). *Pima Indians diabetes dataset* [Dataset]. GitHub. Retrieved from https://github.com/npradaschnor/Pima-Indians-Diabetes-Dataset/blob/master/diabetes.csv

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. *Proceedings of the Annual Symposium on Computer Application in Medical Care*, American Medical Informatics Association, 261–265.

8.0 APPENDIX

List of Variables in the Dataset

- **Pregnancies** – Number of times pregnant
- **Glucose** – Plasma glucose concentration (2 hours in an oral glucose tolerance test)
- **BloodPressure** – Diastolic blood pressure (mm Hg)
- **SkinThickness** – Triceps skin fold thickness (mm)
- **Insulin** – 2-Hour serum insulin (mu U/ml)
- **BMI** – Body mass index (weight in kg/(height in m)^2)
- **DiabetesPedigreeFunction** – A function that scores likelihood of diabetes based on family history
- **Age** – Age in years
- **Outcome** – Class variable (0 if non-diabetic, 1 if diabetic)

Graph 4: Pruned Classification Tree

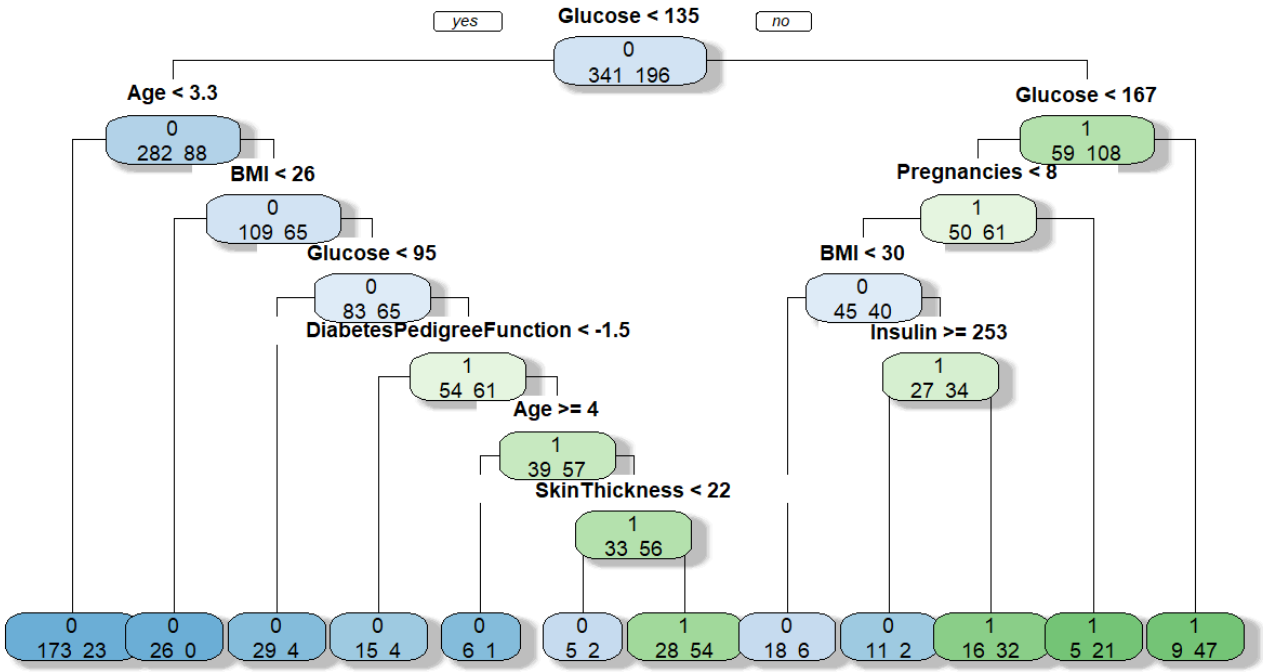


Fig 4: Complexity Parameter values for constructing the Pruned Classification Tree

```
Classification tree:
rpart(formula = Outcome ~ ., data = train, method = "class")

Variables actually used in tree construction:
[1] Age                      BMI                      DiabetesPedigreeFunction
[4] Glucose                  Insulin                  Pregnancies
[7] SkinThickness

Root node error: 196/537 = 0.36499

n= 537

        CP nsplit rel error  xerror     xstd
1 0.250000      0   1.00000 1.00000 0.056920
2 0.026786      1   0.75000 0.89796 0.055497
3 0.022959      5   0.64286 0.85204 0.054729
4 0.015306     10   0.52551 0.82653 0.054266
5 0.013605     11   0.51020 0.82143 0.054171
6 0.010204     14   0.46939 0.78571 0.053471
7 0.010000     15   0.45918 0.81122 0.053976
```