Business Analytics With R

BUAN(6356.004)

Diabetes Health Indicators Project

Group 9

Aishwarya Balmoori
Krishna Priyanka Challa
Peihua Tsai
Tara Canugovi

1. Executive Summary

Diabetes is a growing global health challenge, with prevalence doubling from 7% in 1990 to 14% in 2022. By 2045, it is projected that 1 in 8 adults will have diabetes. In the United States, over 38 million people—approximately 1 in 10 Americans—are diabetic, and 20% are unaware of their condition. This growing epidemic significantly impacts healthcare systems and economies, with annual costs exceeding $400 billion in the U.S. alone.

Using data from the 2015 Behavioral Risk Factor Surveillance System (BRFSS), which includes over 253,000 responses, this project builds predictive models to identify key health indicators and risk factors for diabetes. By applying techniques such as logistic regression, classification trees, and neural networks, the study aims to enhance early detection and support targeted interventions. Key findings include:

- Critical Predictors: High blood pressure, cholesterol, and BMI strongly predict diabetes risk.
- Model Performance: Logistic regression achieved an AUC of 0.818, classification trees maintained a validation accuracy of 86.52%, and neural networks demonstrated high recall (85%) but require refinement to address false positives.

The insights from this study empower healthcare providers and policymakers to design effective prevention strategies, allocate resources, and improve public health campaigns, contributing to better health outcomes and cost savings.

2. Project Motivation/Background

Diabetes is a chronic disease with far-reaching consequences, including complications like cardiovascular disease, kidney failure, and limb amputation. As the prevalence of diabetes continues to rise, the economic burden of this condition grows, with estimated costs in the United States exceeding $400 billion annually. The disease disproportionately affects populations with lower socioeconomic status and varies by demographic factors such as age, education, and income.

Given these challenges, early detection is essential to mitigate the adverse effects of diabetes through lifestyle modifications and timely medical interventions. Predictive models built from robust datasets like the BRFSS 2015 survey can provide valuable insights into diabetes risk, enabling healthcare providers and policymakers to target prevention strategies effectively.

3. Data Description

The BRFSS dataset includes responses from over 400,000 individuals on health conditions, behaviors, and preventive services.

This project utilizes the 2015 Behavioral Risk Factor Surveillance System (BRFSS) dataset, a comprehensive health-related survey conducted by the CDC that includes 253,680 responses and 21 health and lifestyle indicators. Key variables include BMI, Smoking Status, and Physical Activity, with the target variable **Diabetes_012** categorized into:

- **0**: No diabetes or diabetes only during pregnancy.
- **1**: Diabetes.

The data from Kaggle provides a foundation for building predictive models to classify individuals into these diabetes-related categories, supporting public health efforts.

The dataset used for this analysis contains 218334 instances of class 0 and 35346 instances of class 1, reflecting a natural imbalance in the data. This imbalance highlights the rarity of the minority class, which can present a valuable challenge for the model to identify and predict these instances accurately. Despite the imbalance, the analysis focuses on evaluating the model's performance through metrics such as sensitivity, specificity, and the area under the ROC curve (AUC), ensuring a comprehensive understanding of its ability to handle both classes effectively.

4. BI model, Diagrams, Findings

4.1 Data Cleaning and Preprocessing:

- Removed any records that contained missing values to ensure the integrity and completeness of the dataset.
- Converted the target variable, Diabetes_012, into a factor to facilitate better handling in our modeling process.
- Encoded categorical variables to prepare them for analysis, to be effectively utilized by the model.

From the original set of 21 variables in the dataset, we carefully selected 13 variables that are most predictive and relevant. This selection process aimed to enhance the model's accuracy and interpretability by focusing on features that significantly influence the outcome.

```
diabetes.df <-read.csv("diabetes_012_health_indicators_BRFSS2015.csv")
head(diabetes.df)

#sub-setting data variables -- to consider only useful variables
diabetes.df<-diabetes.df[,-c(4,6,8,10,11,13,14,16,21)]
head(diabetes.df)
```

13 variables that are being used:

```
> head(diabetes.df)
  Diabetes_012 HighBP HighChol BMI Stroke PhysActivity HvyAlcoholConsump GenHlth PhysHlth
1            0      1        1  40      0            0                 0       5       15
2            0      0        0  25      0            1                 0       3        0
3            0      1        1  28      0            0                 0       5       30
4            0      1        0  27      0            1                 0       2        0
5            0      1        1  24      0            1                 0       2        0
6            0      1        1  25      0            1                 0       2        2
  DiffWalk Sex Age Income
1        1   0   9      3
2        0   0   7      1
3        1   0   9      8
4        0   0  11      6
5        0   0  11      4
6        0   1  10      8
```

4.2 Data Summarization:

To gain a comprehensive understanding of the dataset, we thoroughly summarize its various attributes, focusing on both central tendencies and the distribution of values.

Summary Statistics: We utilize the summary() function to generate an overview of each variable. This function reveals essential metrics, including the range, mean, and median values, providing insights into the overall characteristics of the data.

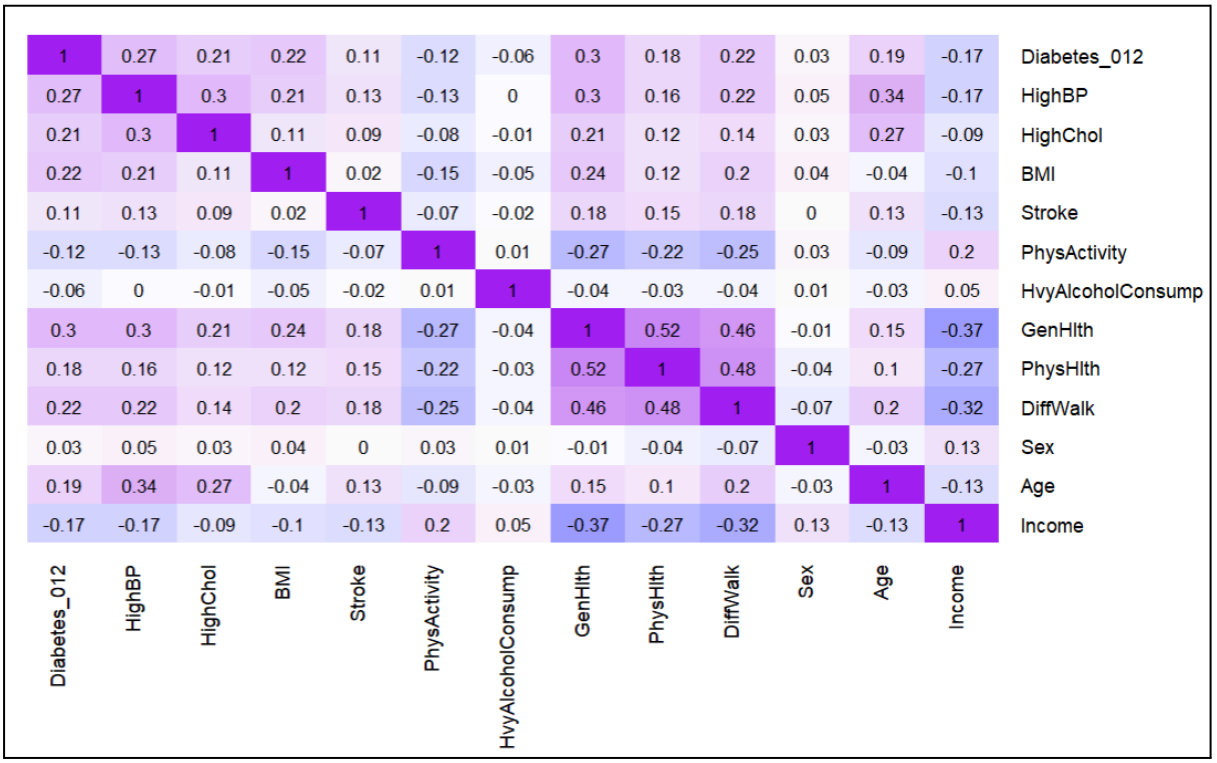
Detailed Metrics: In addition to basic summary statistics, we employ the sapply function to calculate further descriptive measures such as the standard deviation for each variable. This step is crucial as it not only highlights the variability within the data but also allows us to identify any missing values present in each column (if any). By addressing these aspects, we ensure that the dataset is adequately prepared for subsequent analysis while also pinpointing any potential issues related to data quality.

```
                         mean        sd min max median length miss.val
Diabetes_binary    0.13933302 0.3462944   0   1      0 253680        0
HighBP             0.42900110 0.4949345   0   1      0 253680        0
HighChol           0.42412094 0.4942098   0   1      0 253680        0
BMI               28.38236361 6.6086942  12  98     27 253680        0
Stroke             0.04057080 0.1972941   0   1      0 253680        0
PhysActivity       0.75654368 0.4291690   0   1      1 253680        0
HvyAlcoholConsump  0.05619678 0.2303018   0   1      0 253680        0
GenHlth            2.51139231 1.0684774   1   5      2 253680        0
PhysHlth           4.24208057 8.7179513   0  30      0 253680        0
DiffWalk           0.16822375 0.3740656   0   1      0 253680        0
Sex                0.44034216 0.4964292   0   1      0 253680        0
Age                8.03211921 3.0542204   1  13      8 253680        0
Income             6.05387496 2.0711476   1   8      7 253680        0
```

4.3 Correlation Analysis:

To evaluate the relationships among various variables, we construct a correlation matrix. This matrix plays a crucial role in identifying significant associations between different health indicators and diabetes status. By analyzing the correlations within this matrix, we gain

valuable insights into the interactions among these health metrics. This analysis is essential for guiding our selection of relevant predictors for the statistical model.



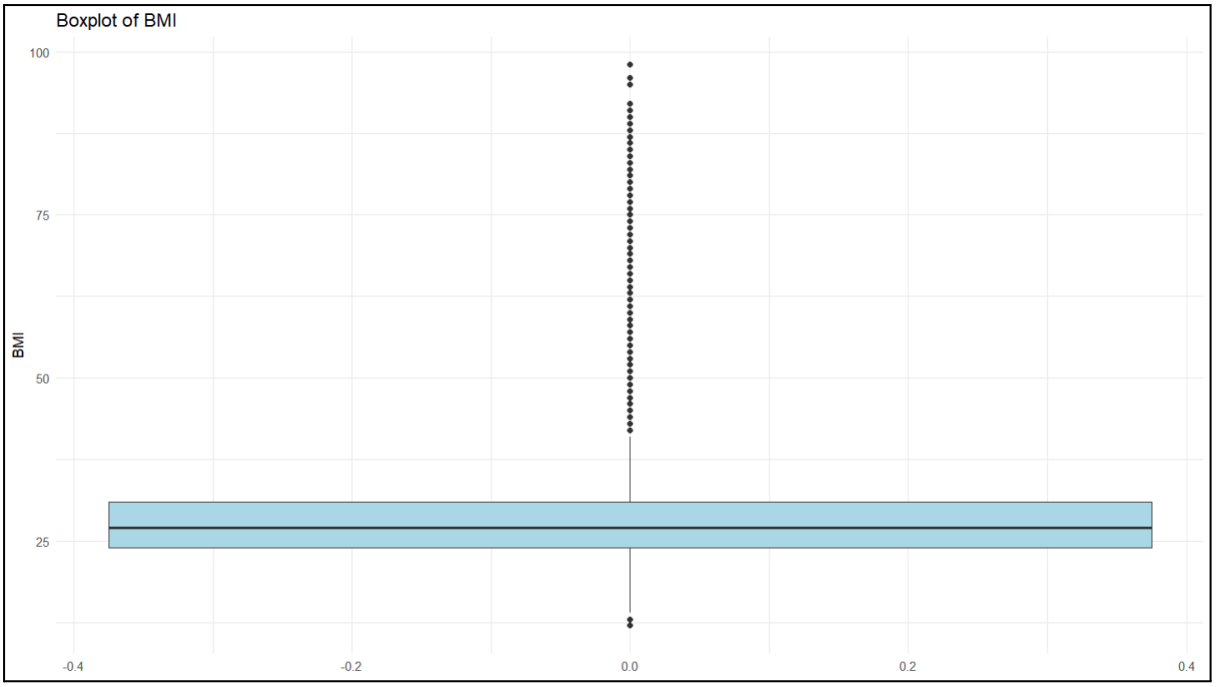
4.4 Data Visualization:

Several visualizations are created to explore and present key relationships in the data:

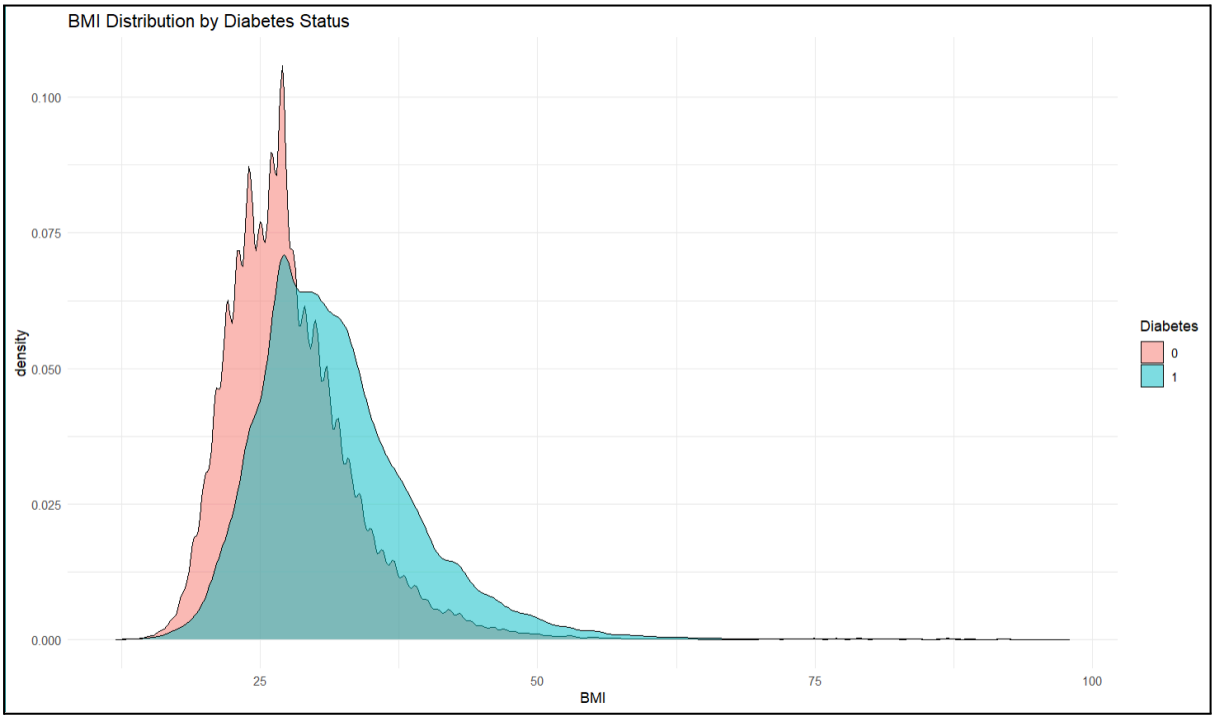
Checking for outliers in BMI with a boxplot:

The majority of the BMI values fall within the range of 25 to 30, which is considered the overweight range. There are a few outliers with BMI values much higher than the rest of the data, indicating the presence of some individuals with very high BMI.

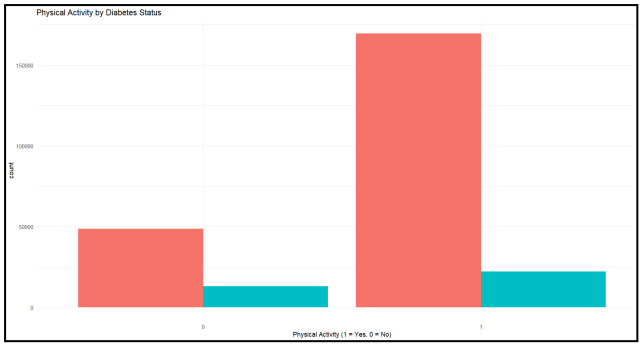
Overall, the boxplot provides a concise and informative visual summary of the distribution of BMI values in the dataset, highlighting the central tendency, spread, and any extreme or unusual values.

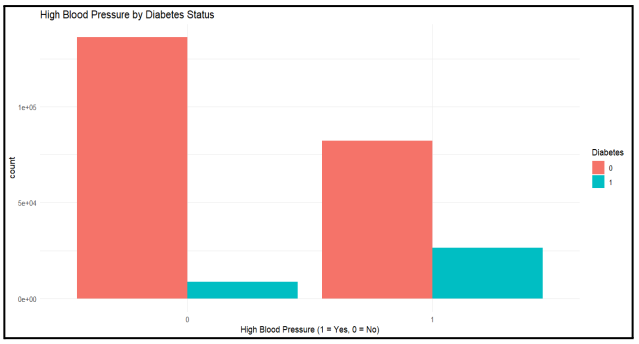


BMI Distribution by Diabetes Status

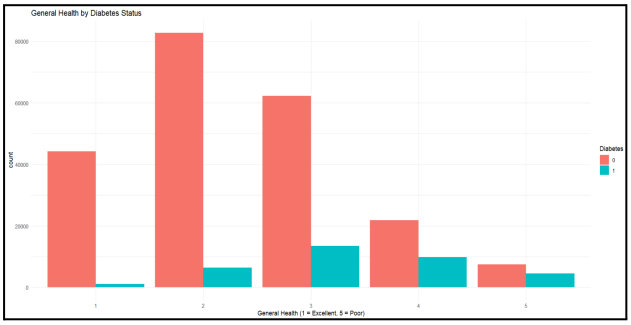


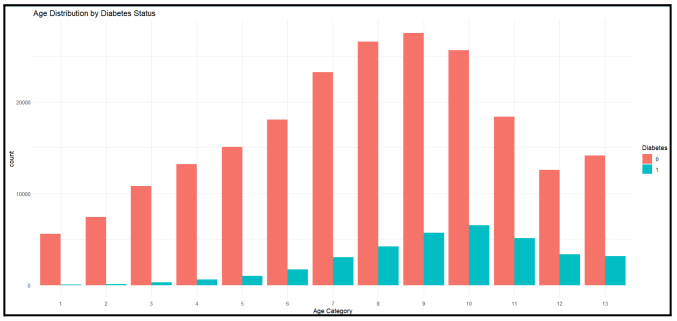
Physical Activity and High Blood Pressure by Diabetes Status





General Health and Age by Diabetes Status





5. Logistic Regression:

The logistic regression model was trained on a 60/40 split of the diabetes dataset, with the training data used to fit the model and the test data used to evaluate its performance. The multinomial logistic regression model was implemented using the nnet::multinom function to analyze the relationship between the outcome variable and predictors.

5.1 Model Coefficients and Interpretations

Intercept: The intercept coefficient of −6.87 represents the baseline log-odds when all predictors are zero.

HighBP and HighChol: The positive coefficients (0.800 and 0.60, respectively) suggest that individuals with high blood pressure or high cholesterol are more likely to fall into certain outcome categories. Both predictors are statistically significant, given their small standard errors.

BMI and Stroke: BMI (0.0610.0610.061) and Stroke (0.1950.1950.195) have positive coefficients, indicating they may slightly increase the likelihood of the outcome.

Age and Income: Age (−0.008) has a negligible negative effect, while income (0.124) exhibits a small positive association with the outcome variable.

5.2 Model Fit

The residual deviance of **97491.51** and the Akaike Information Criterion (AIC) of **97517.51** suggest a reasonably good fit for the model. Lower AIC values indicate better performance, and these values provide a baseline for model comparison.

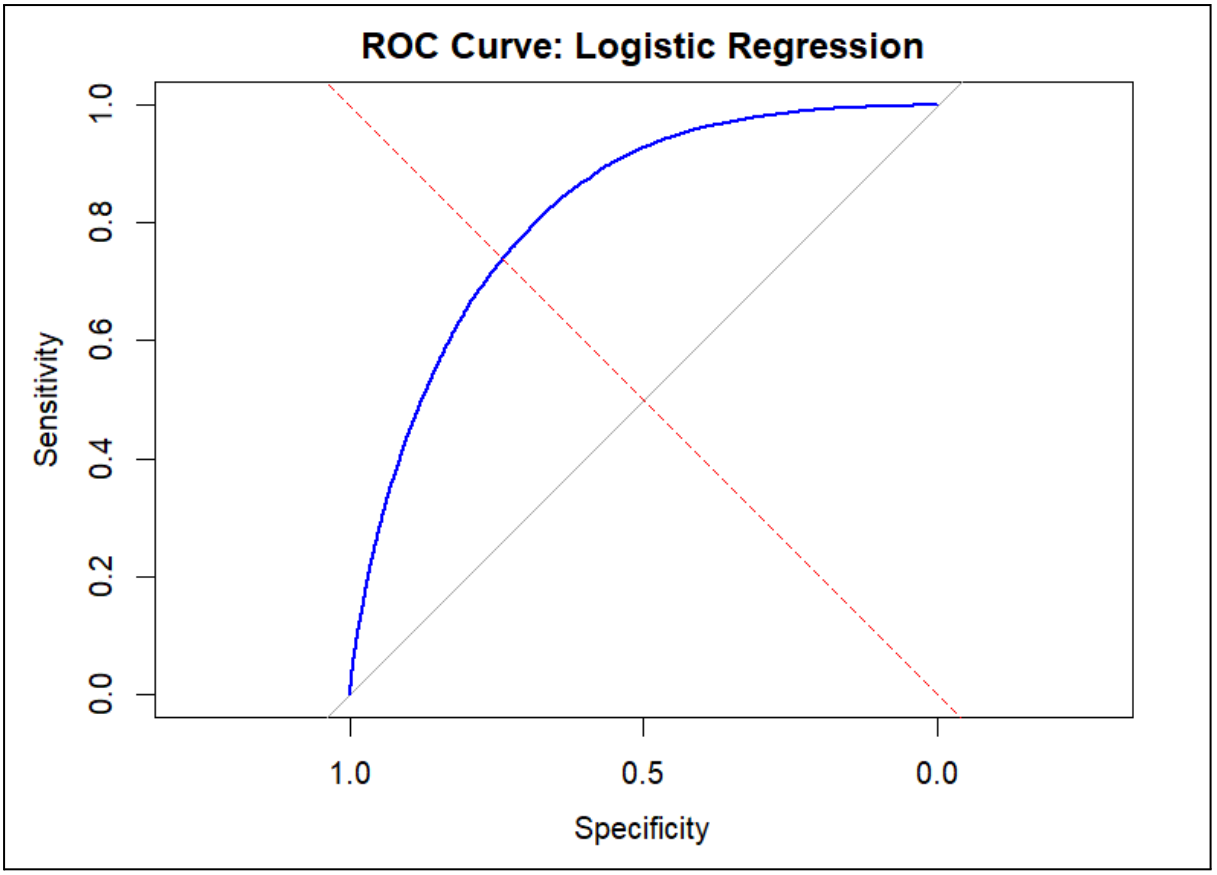
```
> summary(log_model)
Call:
nnet::multinom(formula = .outcome ~ ., data = dat, decay = param$decay)

Coefficients:
                       Values    Std. Err.
(Intercept)       -6.873920410 0.0684533382
HighBP             0.803493194 0.0190243246
HighChol           0.606959929 0.0174491001
BMI                0.061150717 0.0011481988
Stroke             0.194896837 0.0318726919
PhysActivity      -0.063513963 0.0183279484
HvyAlcoholConsump -0.787995129 0.0495340657
GenHlth            0.551922238 0.0102799260
PhysHlth          -0.008048454 0.0009835802
DiffWalk           0.124979799 0.0218073722
Sex                0.274040398 0.0168991529
Age                0.131059281 0.0034143029
Income            -0.054383928 0.0041555235

Residual Deviance: 97491.51
AIC: 97517.51
```

```
> summary(log_pred)
       0                    1
 Min.   :0.01679   Min.   :0.001586
 1st Qu.:0.80093   1st Qu.:0.030763
 Median :0.91988   Median :0.080120
 Mean   :0.86027   Mean   :0.139730
 3rd Qu.:0.96924   3rd Qu.:0.199074
 Max.   :0.99841   Max.   :0.983208
```

The probability distribution shows the model is more confident in predicting class 0 than class 1. The higher mean and quartile values for class 0 indicate a stronger tendency to identify this class.

5.3 ROC Curve Analysis for Logistic Regression Model

The ROC (Receiver Operating Characteristic) curve evaluates the performance of the Logistic Regression model in distinguishing between classes. The x-axis represents 1−Specificity (False Positive Rate), while the y-axis shows Sensitivity (True Positive Rate).



The ROC curve lies above the diagonal (random classifier), indicating that the model performs better than random guessing. The Area Under the Curve (AUC) is **0.818**, suggesting the model has good predictive power. An AUC value closer to 1 indicates a stronger ability to separate the classes, and this score confirms the Logistic Regression model's effectiveness.

5.4 Confusion Matrix

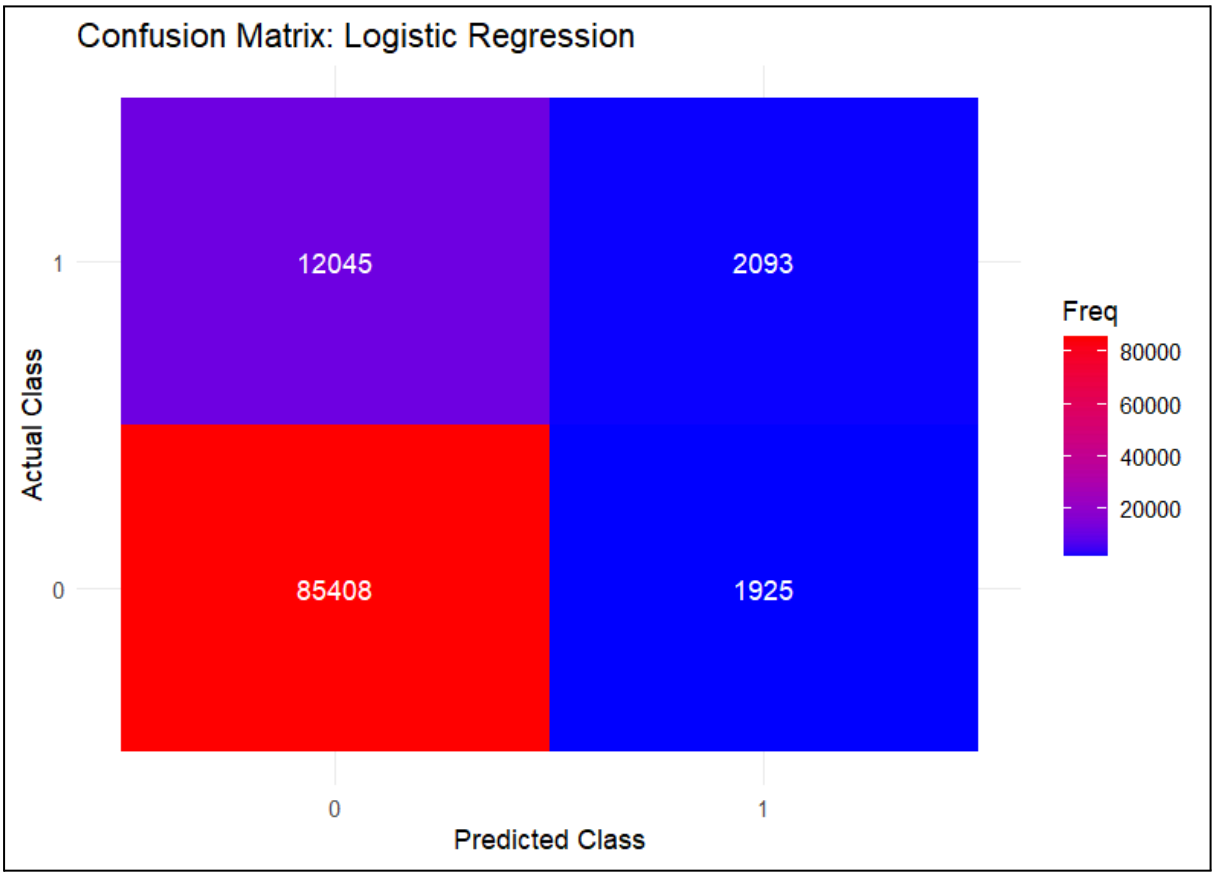
The confusion matrix provides a detailed breakdown of the model's predictions on the test set. The key observations are:

Confusion Matrix (Test Data):

1. **True Positives (TP)**: 2093 cases correctly classified as diabetic.
2. **True Negatives (TN)**: 85408 cases correctly classified as non-diabetic.
3. **False Positives (FP)**: 1925 cases incorrectly classified as diabetic.
4. **False Negatives (FN)**: 12,045 cases incorrectly classified as non-diabetic.

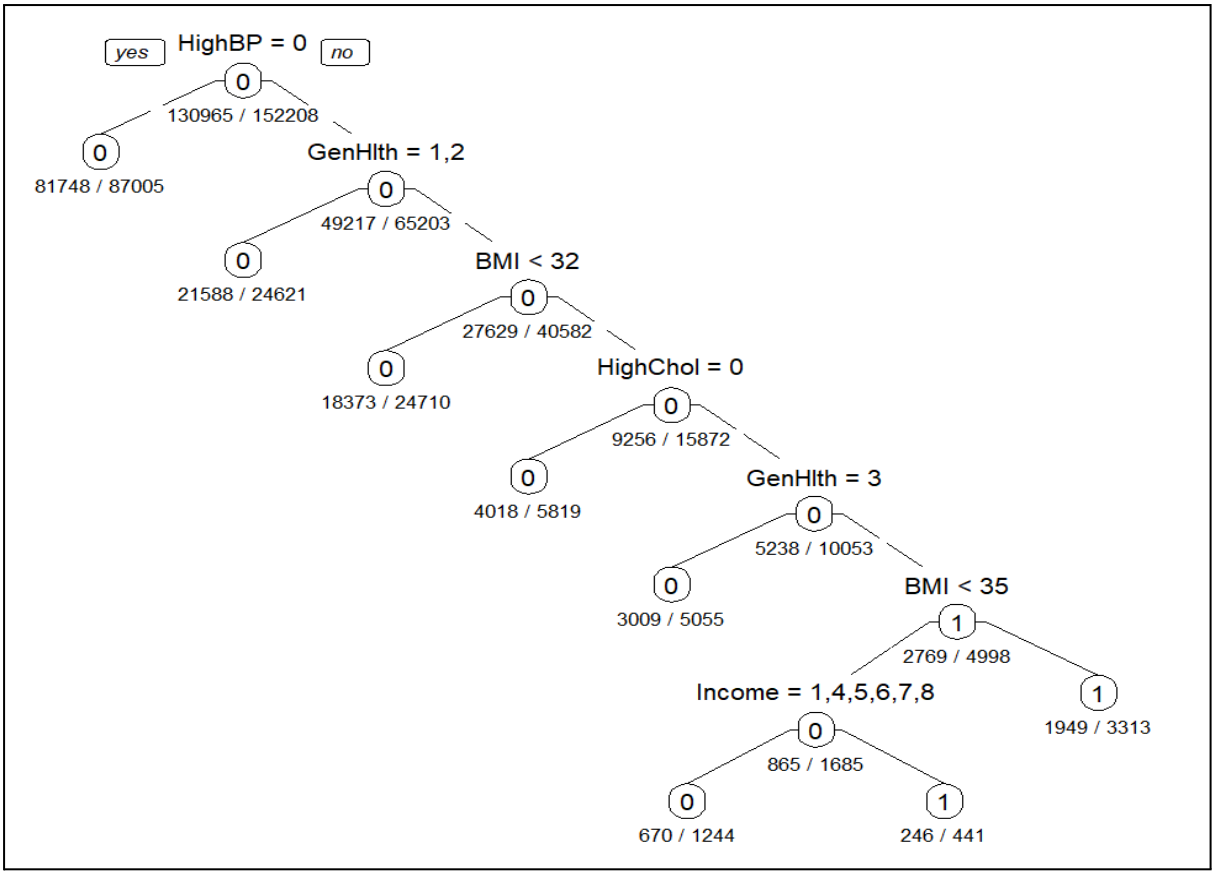
Performance Metrics: Based on the confusion matrix, the following performance metrics were calculated:

- Accuracy: The overall proportion of correct predictions is 86.53%. This indicates that the logistic regression model can correctly classify the diabetes status for about 86.53% of the instances in the test dataset.

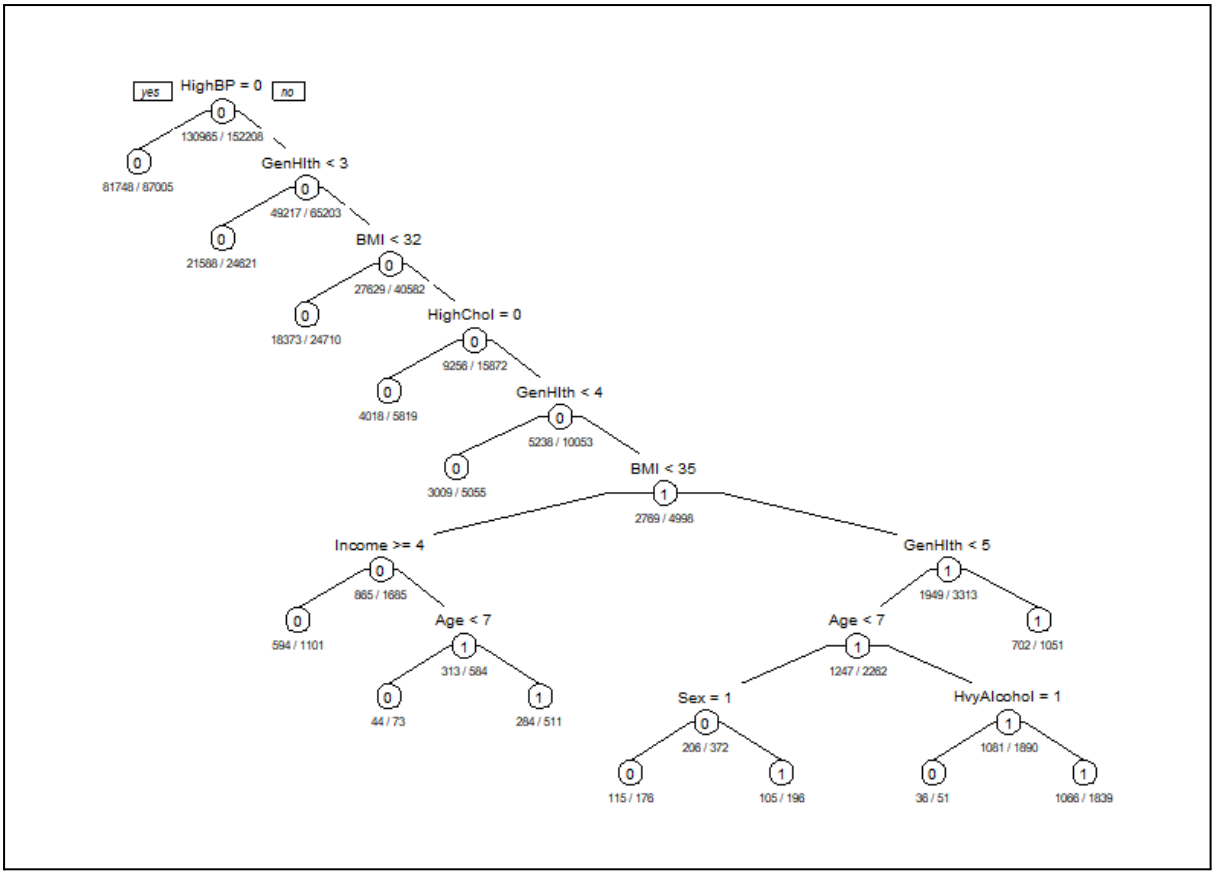


6. Classification tree:

In the part of the Classification tree, we adjust the complexity parameter to view the different trees.
Firstly, we set CP as 0.001. After that, we measure the accuracy of training and validation.
The accuracy of training is 86.46%, and the accuracy of validation is 86.59%.



Secondly, we set the complexity parameter as 0.0005. Then, we measure the accuracy of training and validation. The accuracy of training is 86.51%, and the accuracy of validation is 86.61%.



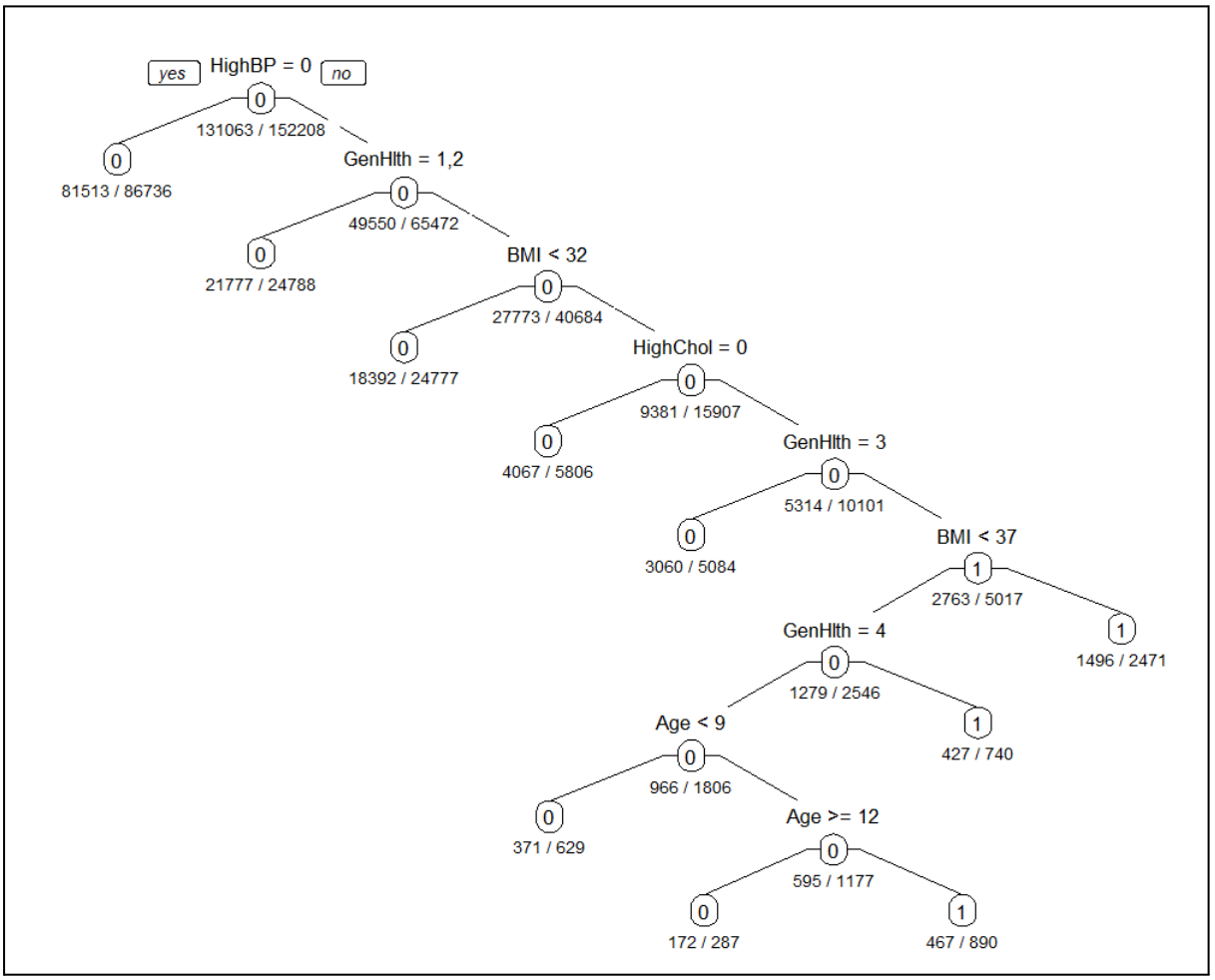
6.1 Dimension Reduction:

To reduce the dimension, we apply Correspondence Analysis (CA) on categorical variables, which are "*GenHlth*", "*PhysHlth*" and "*DiffWalk*". Correspondence Analysis is a multivariate statistical technique used to analyze and visualize the relationships between categorical variables in a contingency table. So we would like to perform Correspondence Analysis in our process.

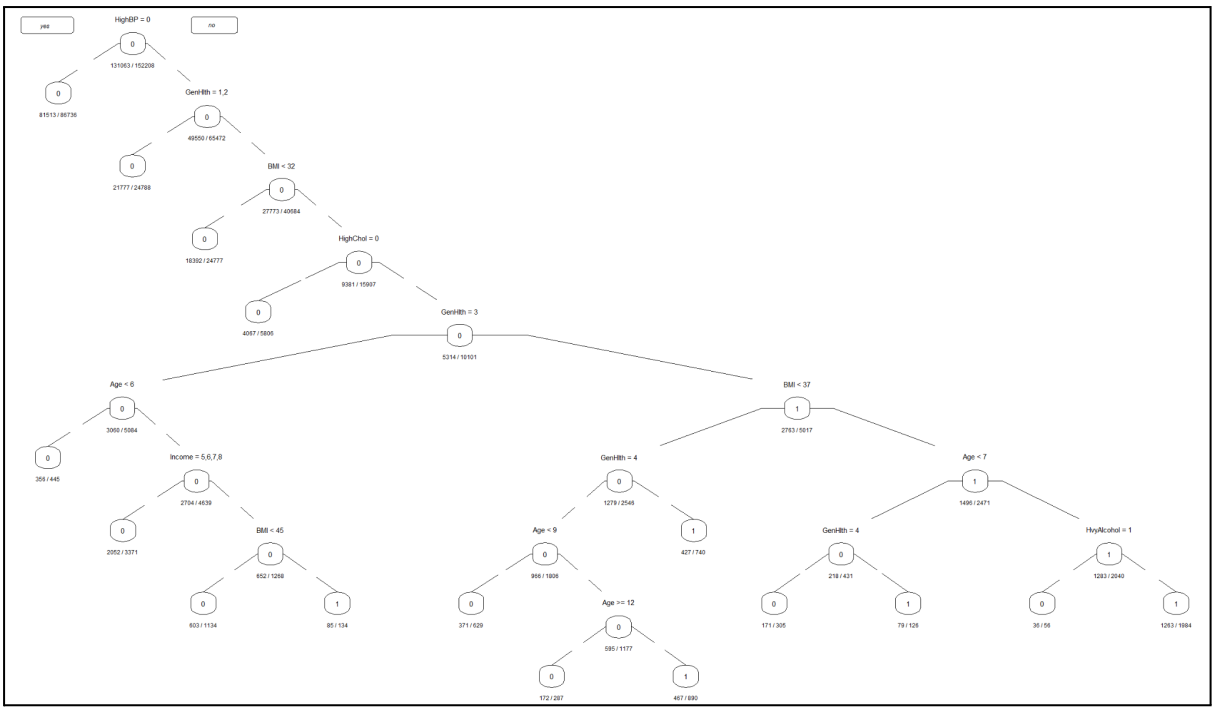
According to the result of CA, we found that "*GenHlth*" contributes the most variance, which is 314.596. Additionally, "*PhysHlth*" contributes 189.771 variance. Hence, we tend to remove "*PhysHlth*" and "*DiffWalk*" to simplify the model.

After reducing the dimension, we tried to repeat the process. Firstly, we set CP as 0.001. After that, we measure the accuracy of training and validation. The accuracy of training is

86.55%, and the accuracy of validation is 86.51%.



Secondly, we set the complexity parameter as 0.0005. Then, we measure the accuracy of training and validation. The accuracy of training is 86.61%, and the accuracy of validation is 86.52%.

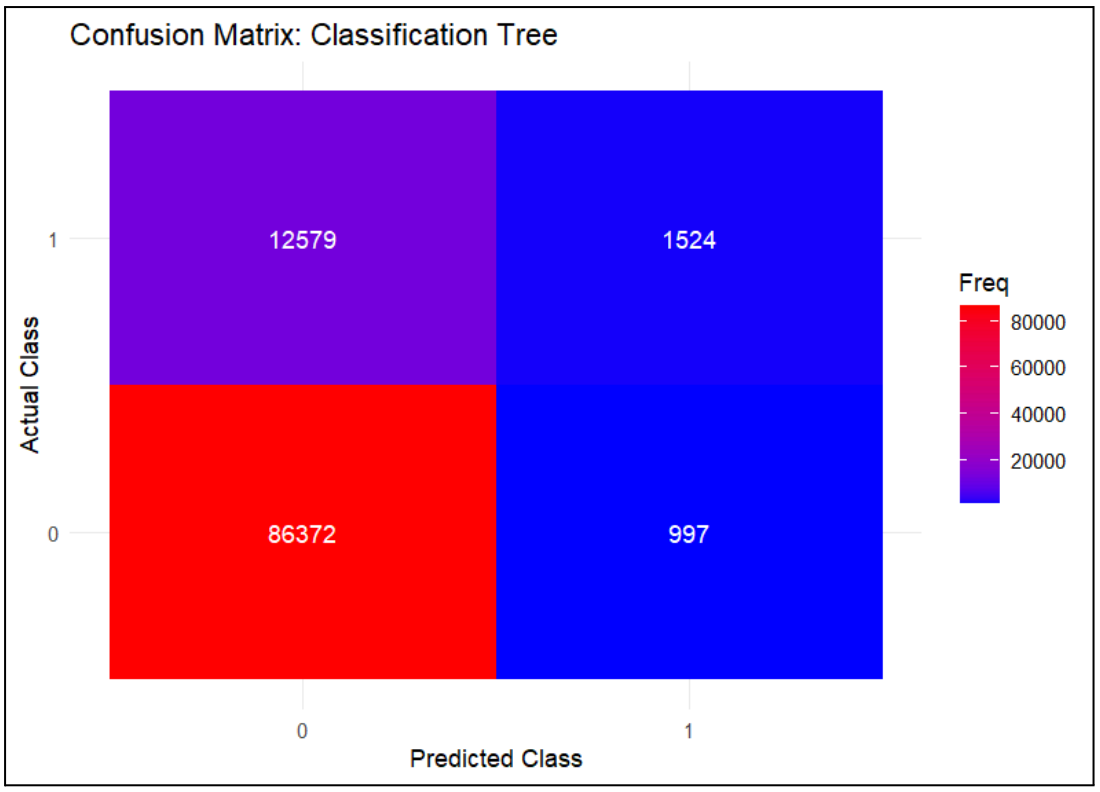


Confusion Matrix Analysis for Classification Tree

The confusion matrix visualizes the performance of a classification tree model in predicting two classes (0 and 1). The rows represent the **actual classes**, while the columns represent the **predicted classes**.

- **True Positives (TP)**: 1,524 instances were correctly predicted as class 1.
- **True Negatives (TN)**: 86,372 instances were correctly predicted as class 0.
- **False Positives (FP)**: 997 instances of class 0 were incorrectly predicted as class 1.
- **False Negatives (FN)**: 12,579 instances of class 1 were incorrectly predicted as class 0.

The accuracy of the model is 86.62%



6.2 The conclusion of the classification tree:

Compared to the two classification trees with different complexity parameters, we observed that the validation accuracy did not significantly increase when the complexity parameter was set to a smaller value. Additionally, after reducing the dimensions, we found that the validation accuracy barely decreased, remaining below 0.1%.

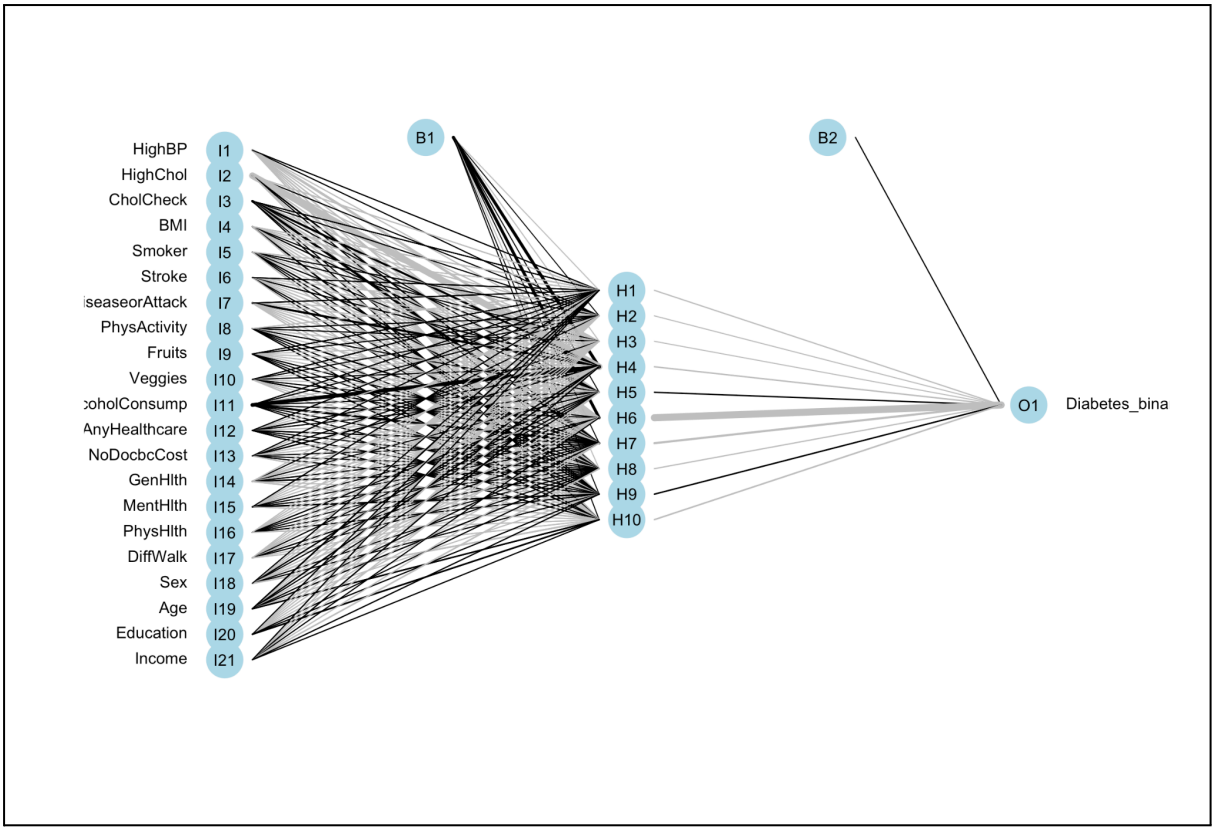
As a result, we decided to use the fourth classification tree, which achieved a validation accuracy of 86.52%. This choice aligns with our goal of minimizing effort by selecting the most straightforward classification tree, which also considers practical real-world considerations.

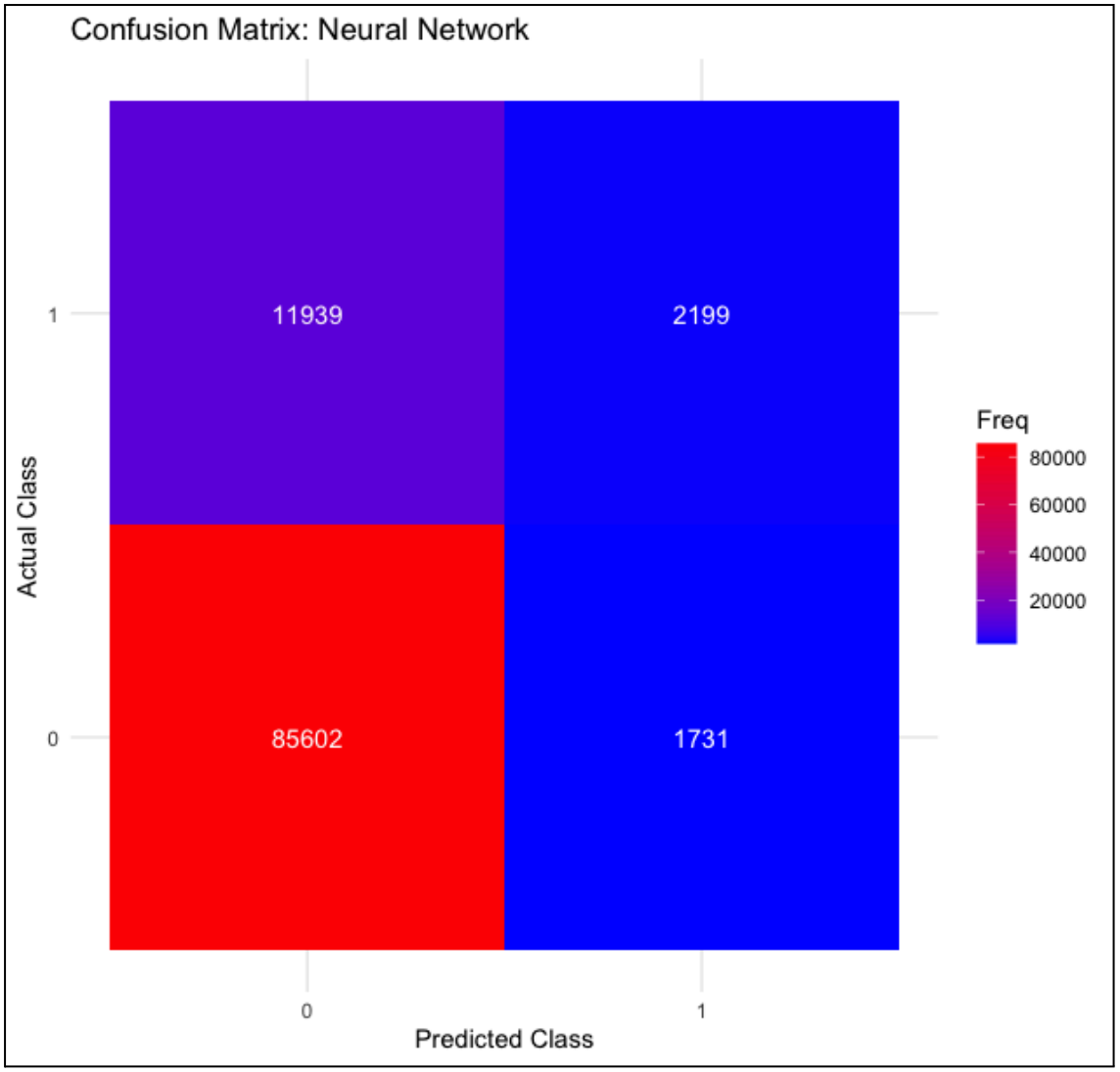
7. Neural network:

A neural network was trained with the following specifications:
- Hidden layer size: 10 neurons
- Regularization parameter (decay): 0.1
- Maximum iterations: 200

The performance of the predictions was evaluated on the test set using a confusion matrix and overall accuracy.

The structure of the neural network was visualized using NeuralNetTools, demonstrating the 10 hidden layer nodes and their connections to the input and output layers.


HighBP
HighChol
CholCheck
BMI
Smoker
Stroke
iseaseorAttack
PhysActivity
Fruits
Veggies
ɔoholConsump
AnyHealthcare
NoDocbcCost
GenHlth
MentHlth
PhysHlth
DiffWalk
Sex
Age
Education
Income
I1
I2
I3
I4
I5
I6
I7
I8
I9
I10
I11
I12
I13
I14
I15
I16
I17
I18
I19
I20
I21
B1
B2
H1
H2
H3
H4
H5
H6
H7
H8
H9
H10
O1
Diabetes_bina



Confusion Matrix (Test Data):

1. **True Positives (TP)**: 2199 cases correctly classified as diabetic.
2. **True Negatives (TN)**: 85,602 cases correctly classified as non-diabetic.
3. **False Positives (FP)**: 1731 cases incorrectly classified as diabetic.
4. **False Negatives (FN)**: 11,939 cases incorrectly classified as non-diabetic.

Performance Metrics:

```
> print(conf_matrix_nn)
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 85602 11939
         1  1731  2199

               Accuracy : 0.8653
                 95% CI : (0.8632, 0.8674)
    No Information Rate : 0.8607
    P-Value [Acc > NIR] : 1.037e-05

                  Kappa : 0.1946

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9802
            Specificity : 0.1555
         Pos Pred Value : 0.8776
         Neg Pred Value : 0.5595
             Prevalence : 0.8607
         Detection Rate : 0.8436
   Detection Prevalence : 0.9613
      Balanced Accuracy : 0.5679

       'Positive' Class : 0
```

Neural Network Accuracy: 86.53 %

Recall (Sensitivity):

Recall = TP/TP+FN = 85,602/(85,602+11,939) = 87.75%

The model has a high overall accuracy (86.53%), indicating a good classification ability on average. The recall is strong, suggesting the model identifies most diabetic cases correctly.

The neural network demonstrated a solid foundation for diabetes risk prediction, achieving high recall and accuracy. However, its high false positive rate and low precision indicate a need for improvement.

8. Managerial Implications

- **Strategic Interventions for Diabetes Management**
    - The insights from this analysis, particularly the identification of key predictors like high blood pressure, cholesterol, and BMI, empower healthcare providers to design targeted prevention programs. By focusing resources on individuals

at higher risk, healthcare systems can reduce the burden of diabetes and its associated costs.
    - The ability to simplify models like the classification tree without sacrificing accuracy ensures that public health teams can use these tools effectively, even in resource-constrained settings.
- **Actionable Insights for Policy and Resource Allocation**
    - High recall rates in models like the neural network indicate a strong potential for early detection, which can guide policy decisions on screening protocols. For example, prioritizing at-risk populations based on age, BMI, and health history.
    - Logistic regression's interpretability provides clear, actionable metrics for healthcare managers to justify budget allocations and improve health outcomes in underserved communities.
- **Enhancing Public Health Campaigns**
    - The findings can inform educational campaigns emphasizing lifestyle changes, such as physical activity and dietary habits, to reduce diabetes risk.
    - The data-driven approach offers quantifiable evidence to advocate for policy changes, such as increased funding for diabetes screening programs.

9. Conclusion

Our analysis of diabetes risk factors, using the 2015 BRFSS dataset with over 253,680 records, highlighted the potential of data-driven models to inform public health strategies.

- **Key Models & Insights**
    - **Classification Tree:** Simplified using Correspondence Analysis, achieving 86.52% validation accuracy, balancing simplicity and effectiveness.
    - **Logistic Regression:** Delivered interpretable insights with strong predictors like high blood pressure and cholesterol, achieving an AUC of 0.818.
    - **Neural Network:** Achieved a high accuracy of 86.53%, but faced challenges such as a recall of 87.75% and a significant false positive rate.
- **Key Takeaways**
    - High blood pressure, cholesterol, and BMI are critical predictors of diabetes.
    - Simpler models, such as logistic regression and classification trees, provide practical and actionable insights, while advanced models like neural networks require further refinement for real-world applications.

This study underscores the power of predictive analytics in addressing critical healthcare challenges, supporting early detection, and targeted interventions for diabetes management.

References

[1] Diabetes Health Indicators Dataset. (n.d.). Kaggle. Retrieved from https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset/code

[2] Diabetic Dataset. (n.d.). Kaggle. Retrieved from https://www.kaggle.com/datasets/shivashankari06/diabetic-datasetDiabetic Dataset

[3] National Diabetes Statistics Report | Diabetes. (2024, May 15). CDC. Retrieved from https://www.cdc.gov/diabetes/php/data-research/index.html

[4] Urgent action is needed as global diabetes cases have increased four-fold over the past decades. (2024, November 13). World Health Organization (WHO). Retrieved from https://www.who.int/news/item/13-11-2024-urgent-action-needed-as-global-diabetes-cases-increase-four-fold-over-past-decades