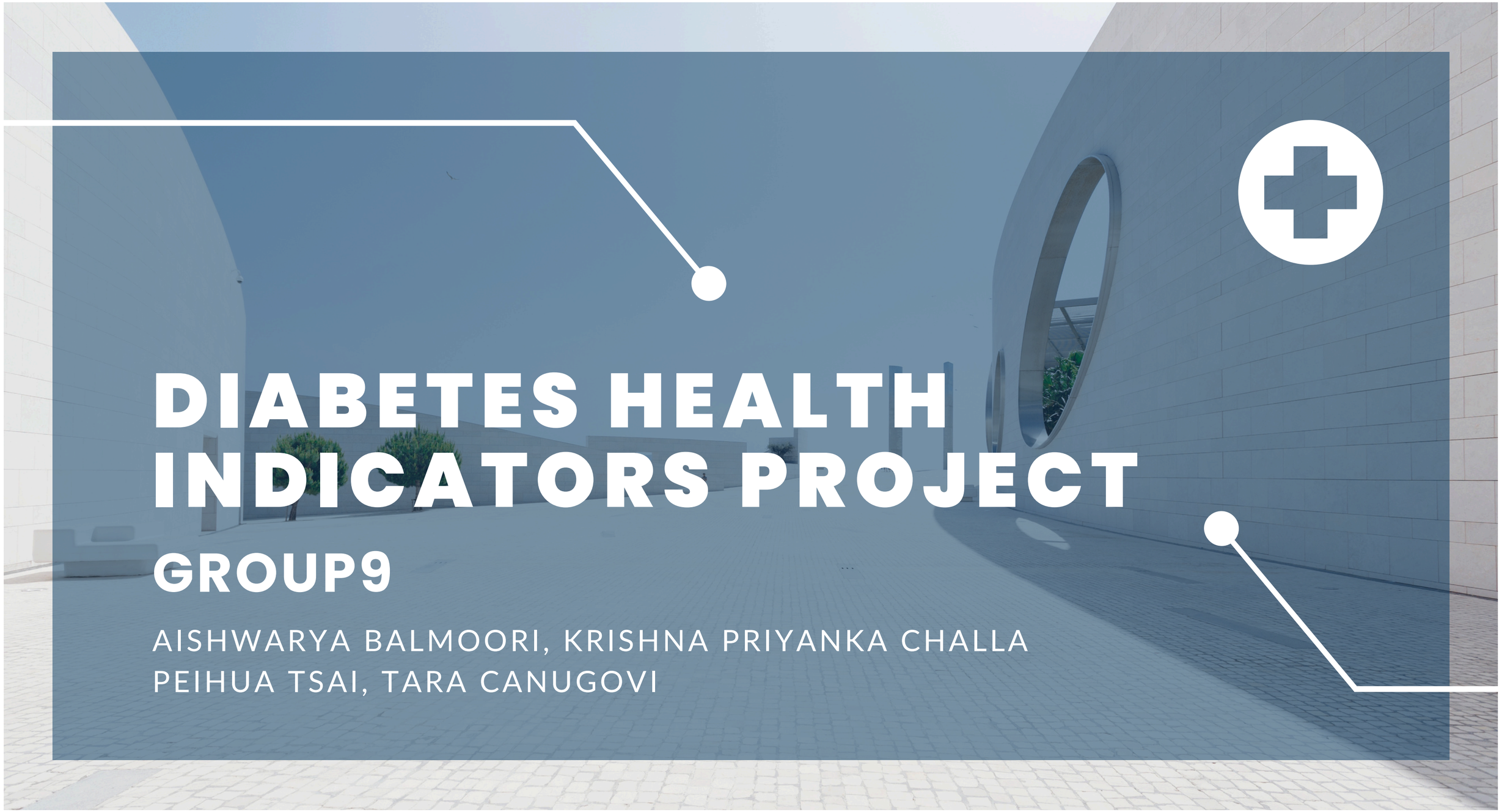


DIABETES HEALTH INDICATORS PROJECT

GROUP9

AISHWARYA BALMOORI, KRISHNA PRIYANKA CHALLA
PEIHUA TSAI, TARA CANUGOVI

CONTENT INDEX

- DATA RESOURCE
- DATA INTRODUCTION

- CLASSIFICATION TREE
- NEURAL NETWORK
- MULTINOMIAL LOGISTIC REGRESSION MODEL

OBJECTIVE

To analyze the 2015 BRFSS dataset to predict diabetes prevalence by:

1. **Identifying Risk Factors:** Key health indicators like BMI, blood pressure, and cholesterol.
2. **Developing Models:** Build and evaluate machine learning models to classify diabetes risk.
3. **Providing Insights:** Support healthcare interventions, policy decisions, and public health campaigns to reduce the burden of diabetes.

DATA RESOURCE & INTRODUCTION

Data comes from

- Kaggle
- Diabetes Indicators
- 253,680 records
- 21 variables

Variables Selection

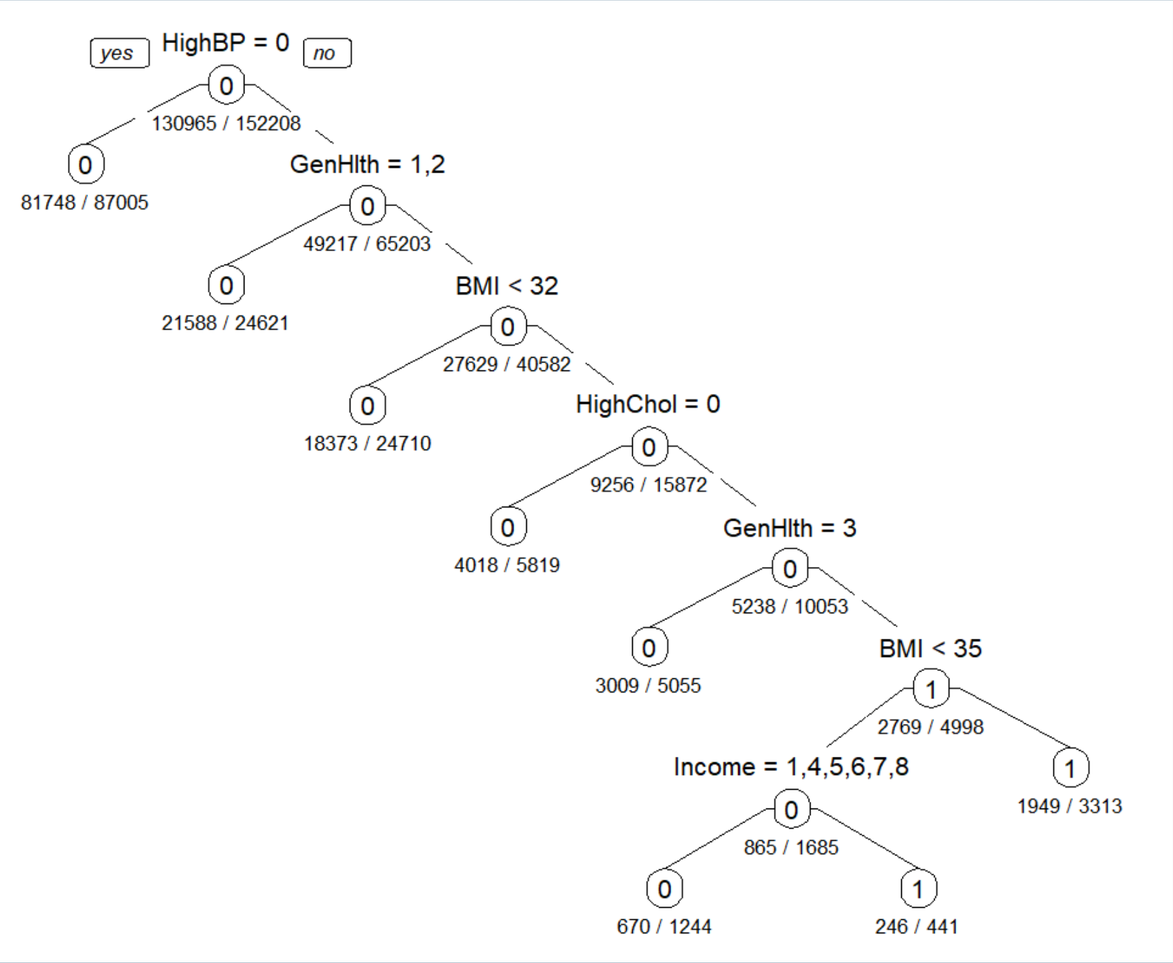
- 12 independent variables
- 1 dependent variable
- “Diabetes_binary" is target variable

Data Type

NUMERIC VARIABLES

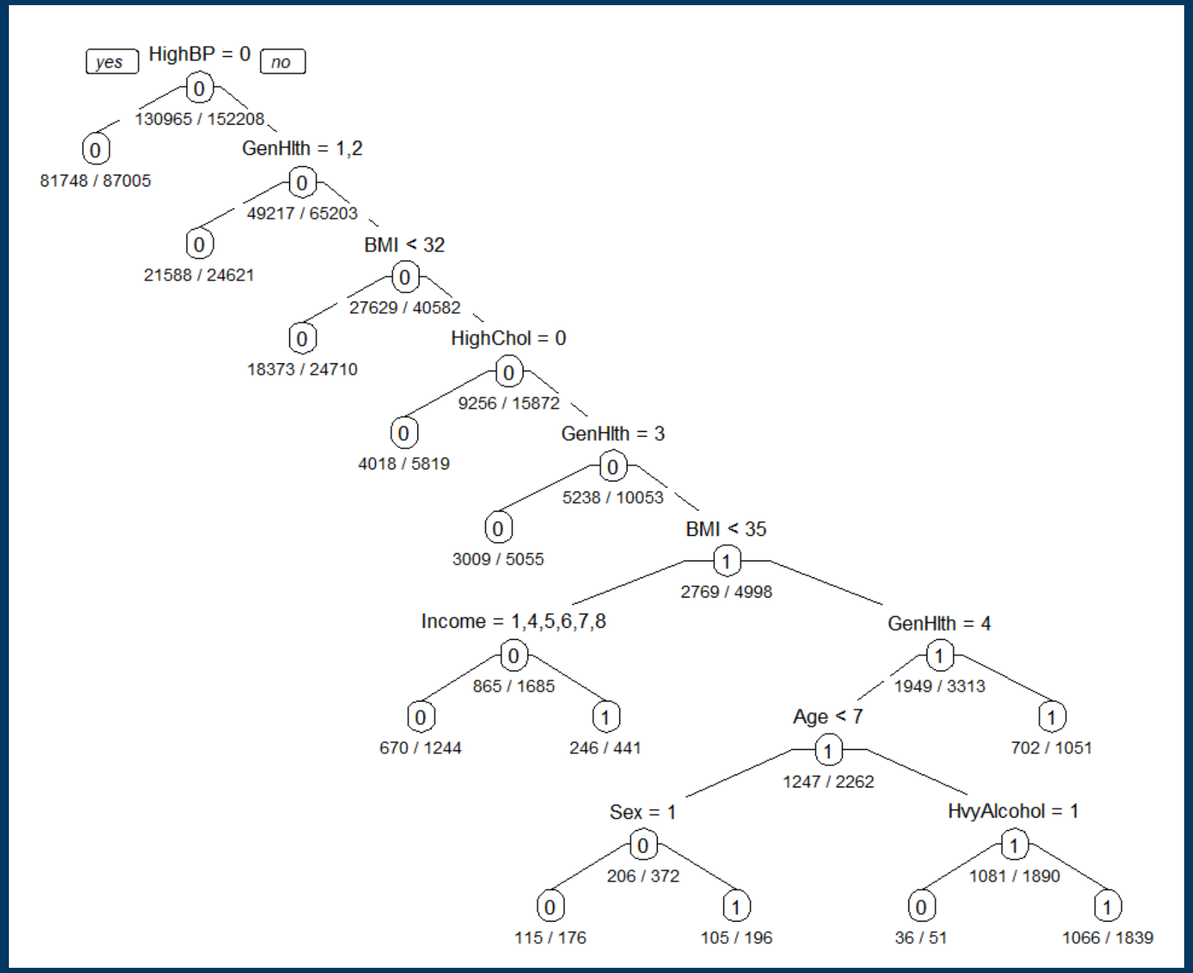
1. BMI
2. PhysHlth
3. Age

CLASSIFICATION TREE



- When CP=0.001
- Training Accuracy: 86.46%
- Validating Accuracy: 86.59%

CLASSIFICATION TREE



- When CP=0.0005
- Training Accuracy: 86.51%
- Validating Accuracy: 86.61%

DIMENSION REDUCTION

CORRELATION ANALYSIS

1

We found that there are three variables, which have a positive relationship.
"GenHlth", "PhysHlth" and "DiffWalk".

CORRESPONDENCE ANALYSIS (CA)

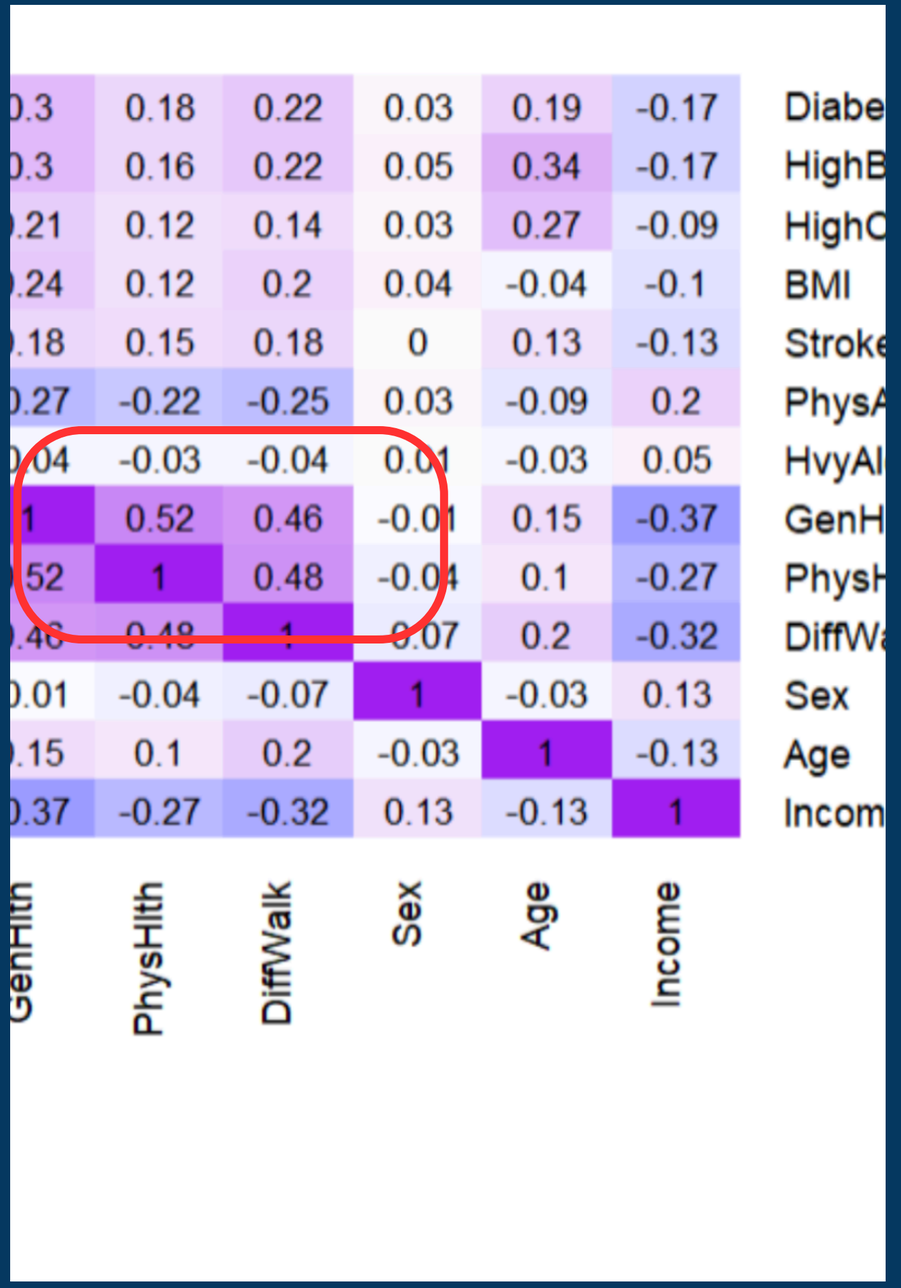
2

The result shows that "GenHlth" contributes the most variance(314.596).

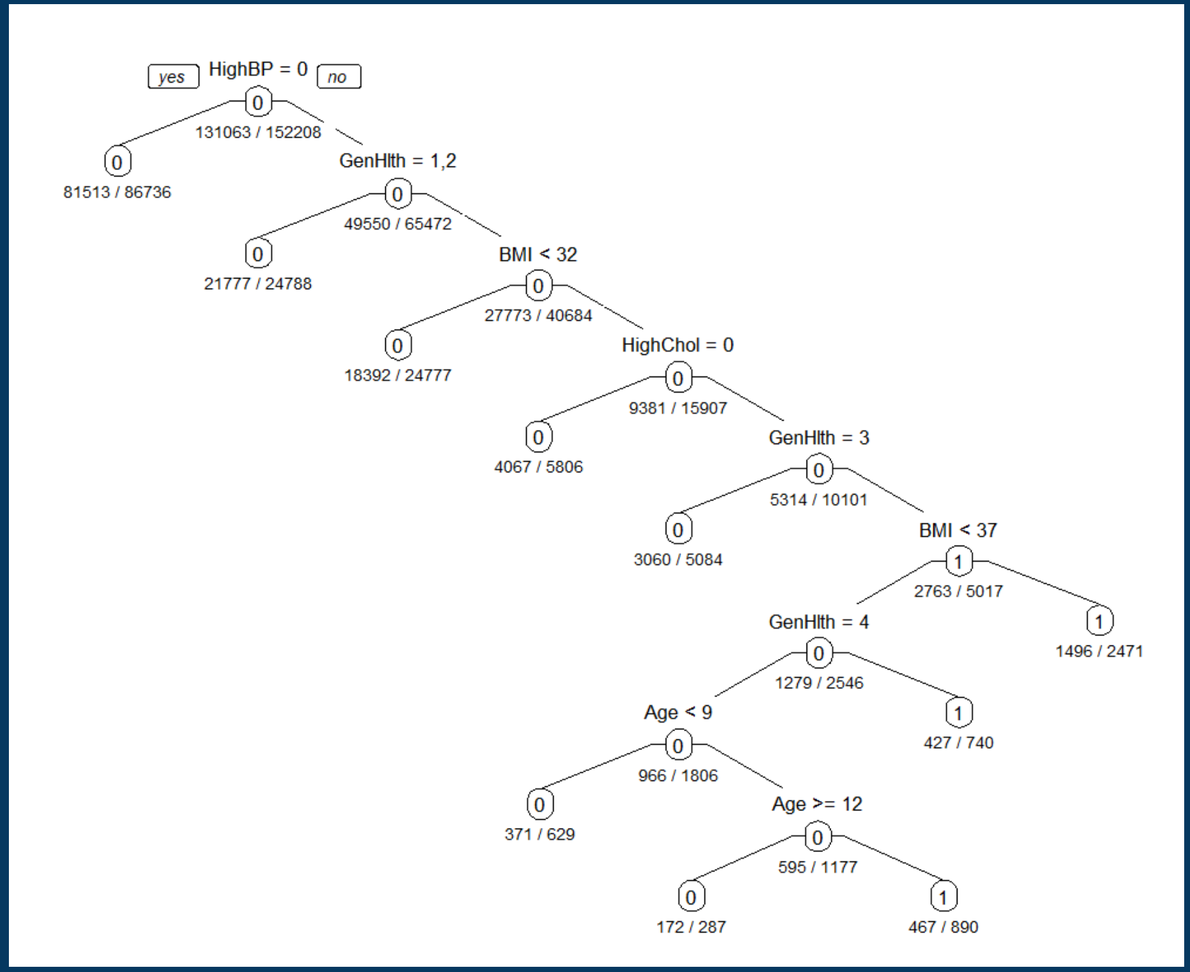
REMOVE UNNECESSARY VARIALBES

3

Therefore, we prefer to remove "PhysHlth" and "DiffWalk".

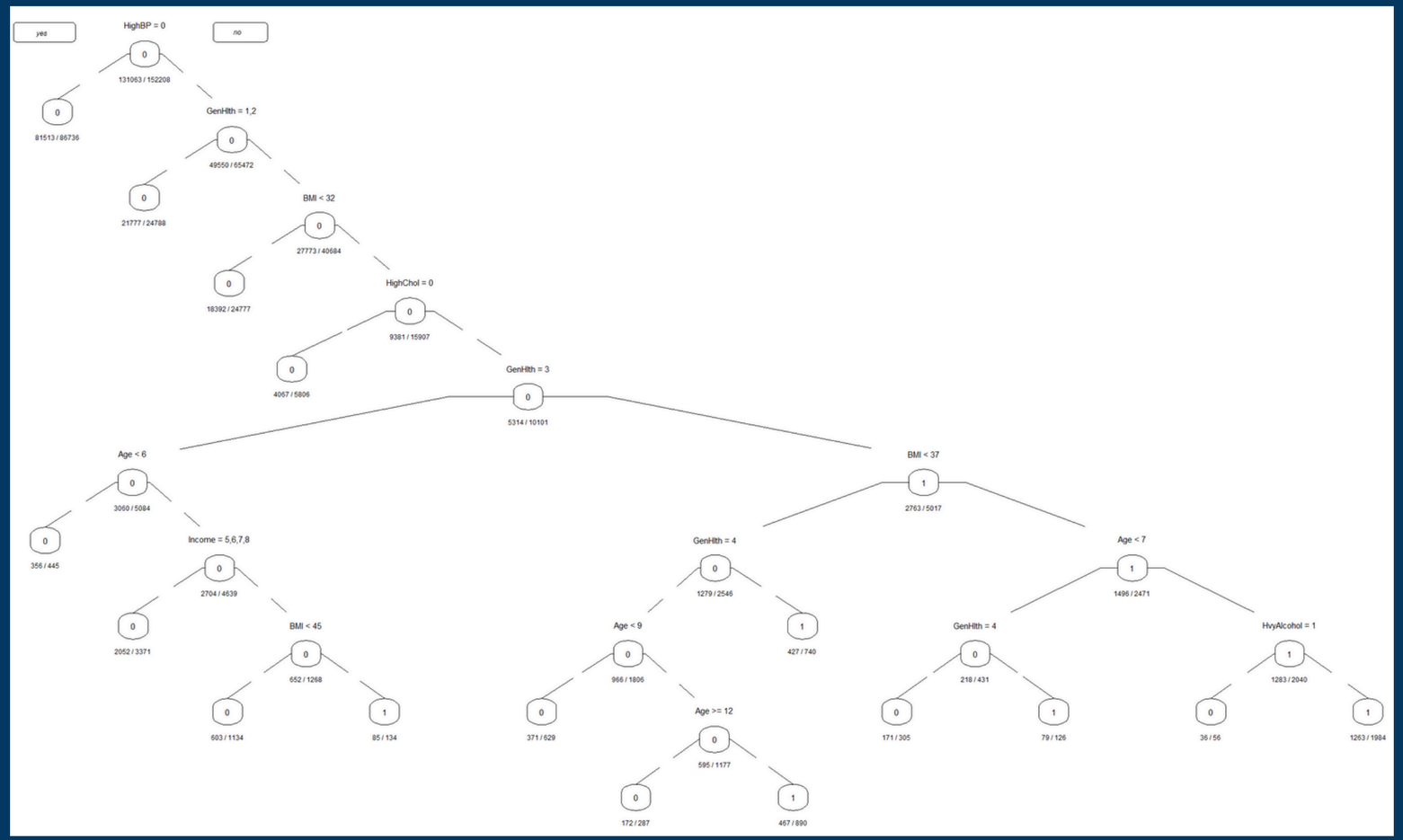


CLASSIFICATION TREE



- When CP=0.001
- Training Accuracy: 86.55%
- Validating Accuracy: 86.51%

CLASSIFICATION TREE



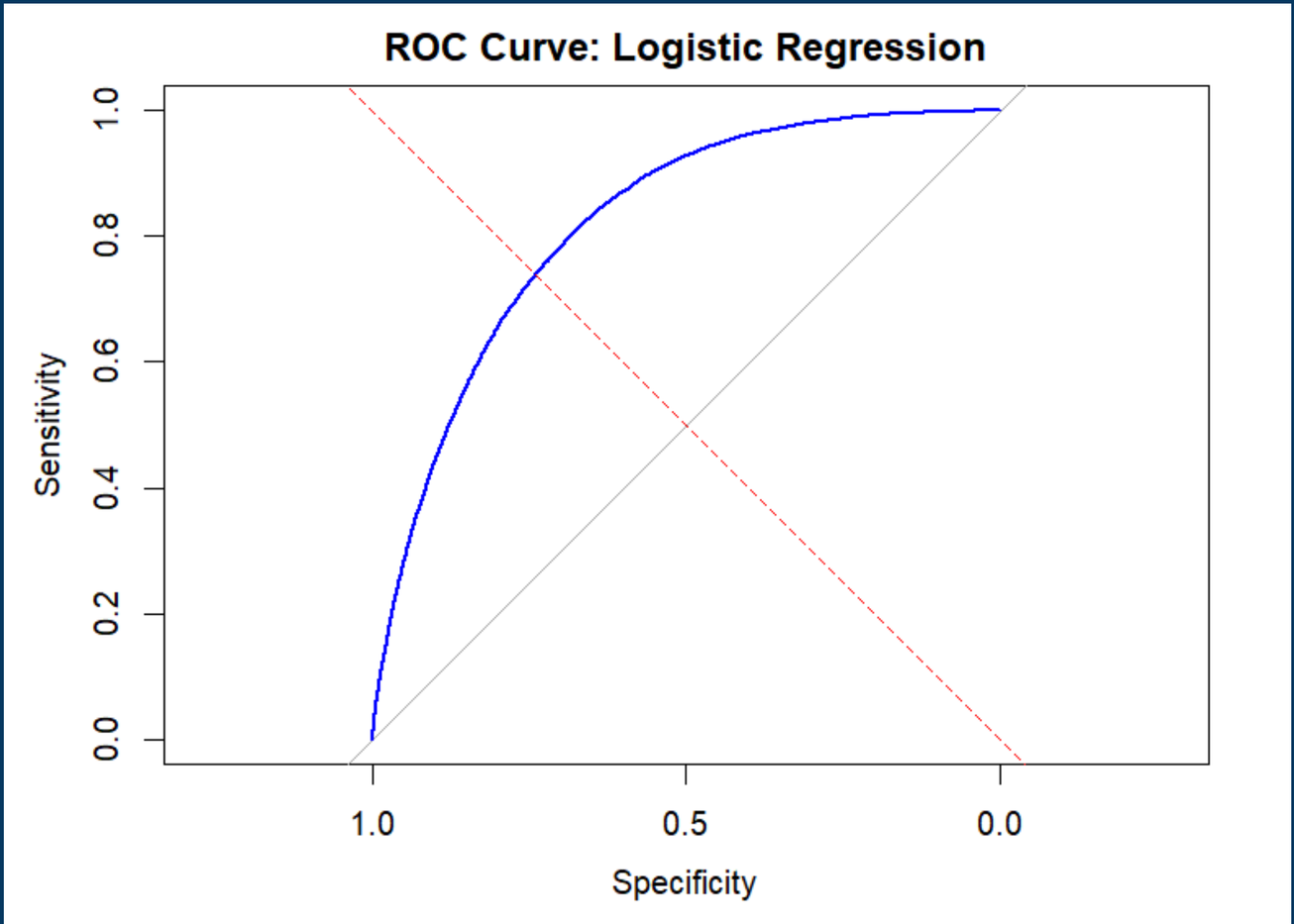
- When CP=0.0005
- Training Accuracy: 86.61%
- Validating Accuracy: 86.52%

LOGISTIC REGRESSION

```
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
```

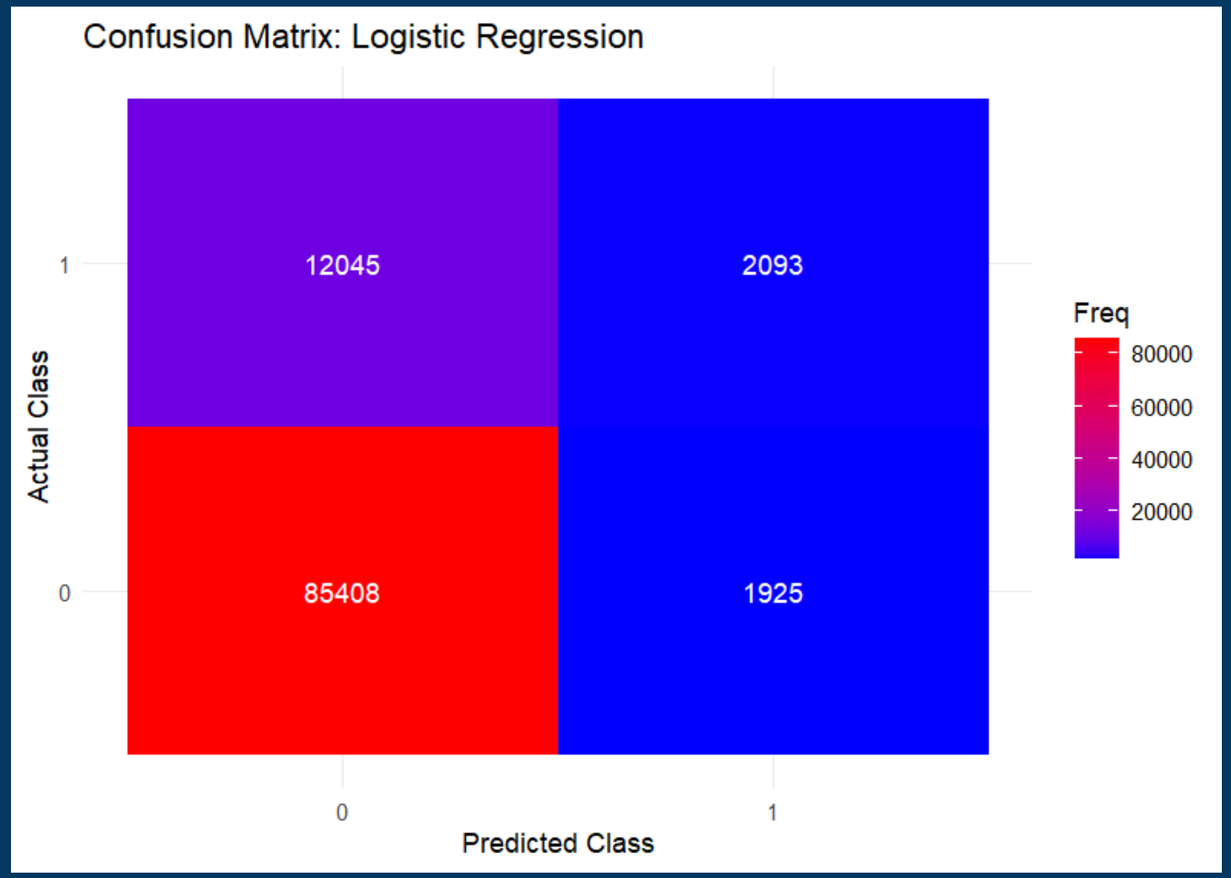
```
> summary(log_pred)
       0                    1
 Min.   :0.01679   Min.   :0.001586
 1st Qu.:0.80093   1st Qu.:0.030763
 Median :0.91988   Median :0.080120
 Mean   :0.86027   Mean   :0.139730
 3rd Qu.:0.96924   3rd Qu.:0.199074
 Max.   :0.99841   Max.   :0.983208
```

LOGISTIC REGRESSION



- The Area Under the Curve (AUC) is 0.818
- Good predictive power

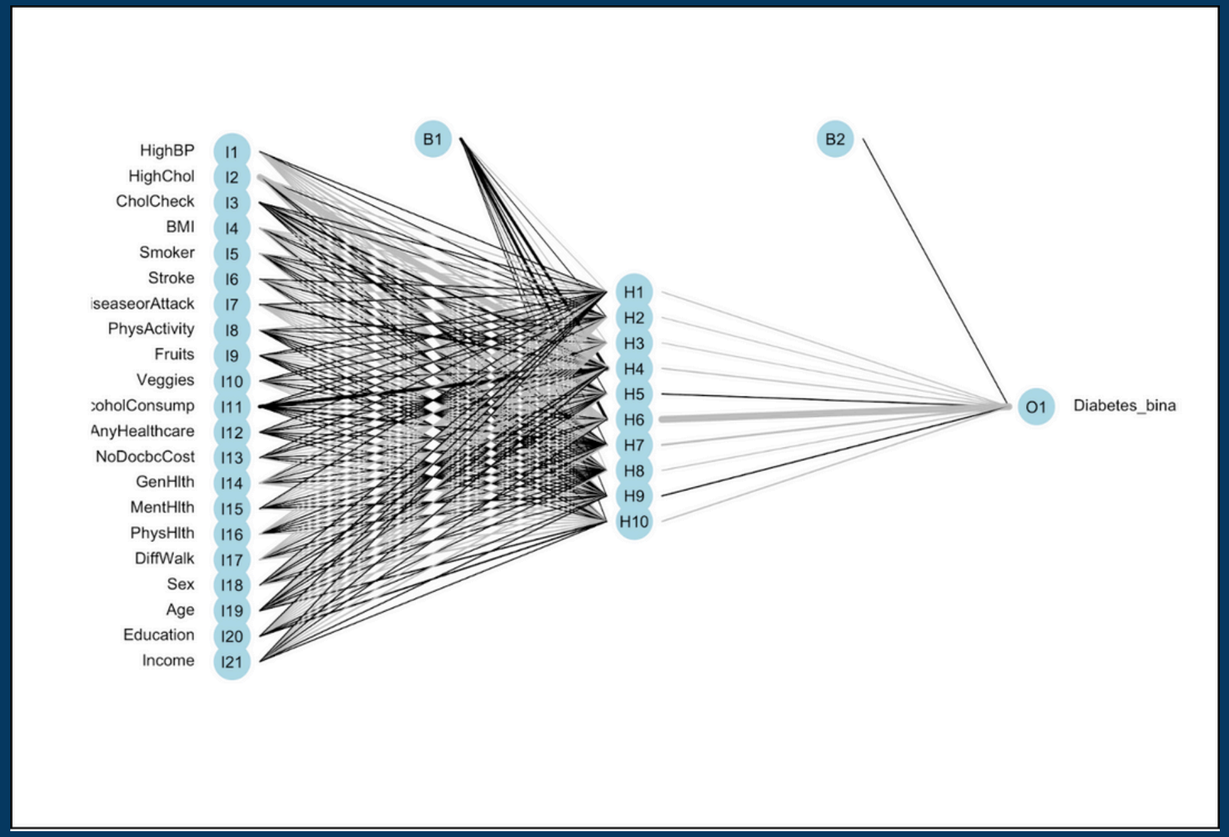
LOGISTIC REGRESSION



Performance Metrics:

Logistic Regression Accuracy: 86.53 %

NEURAL NETWORK

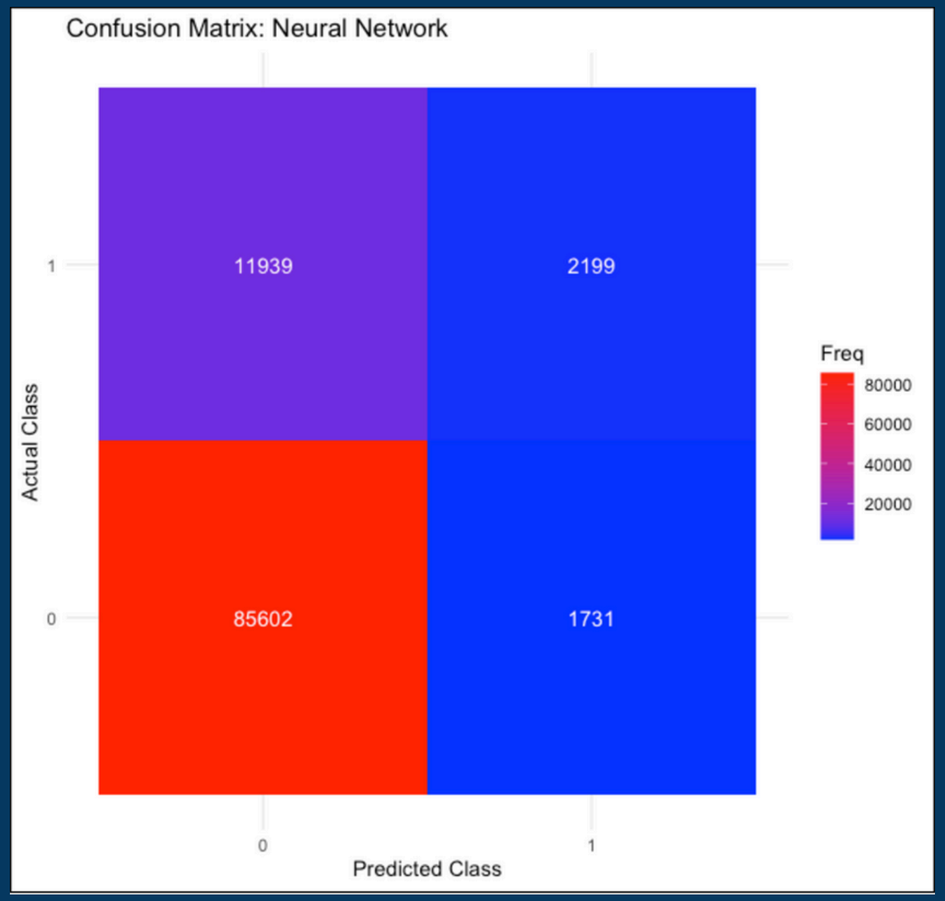


- Hidden layer size: 10 neurons
- Regularization parameter (decay): 0.1
- Maximum iterations: 200

NEURAL NETWORK

```
> 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



Performance Metrics:

- Neural Network Accuracy: 86.53 %
- Recall (Sensitivity): 87.75%

COMPARATIVE MODEL PERFORMANCE

Model	Accuracy	Key Features	Challenges
Logistic Regression	86.53%	High interpretability, AUC = 0.818.	Moderate performance for complex data.
Classification Tree	86.51%	Simplified, CP = 0.001.	Limited flexibility for advanced scenarios.
Neural Network	86.53%	High recall, robust learning.	High false positive rate.

MANAGERIAL IMPLICATIONS

- Strategic Interventions for Diabetes Management
- Actionable Insights for Policy and Resource Allocation
- Enhancing Public Health Campaigns

CONCLUSION

Key Finding:

- High blood pressure, cholesterol, and BMI identified as critical predictors of diabetes.

Model Performance:

- Classification Tree: Balanced simplicity and effectiveness with 86.52% validation accuracy.
- Logistic Regression: Strong AUC of 0.818 with interpretable and actionable insights.
- Neural Network: High accuracy (86.53%) and recall (85.0%), but challenges with false positives.

CONCLUSION

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.

THANKS FOR LISTENING