**BUCHAREST UNIVERSITY OF ECONOMIC STUDIES**

FACULTY OF CYBERNETICS, STATISTICS, AND ECONOMIC INFORMATICSA blue logo with a black background

Description automatically generated

**DATA MINING PRINCIPLES PROJECT**

Predicting Diabetes Using Machine Learning

**Coordinating Professor: Students Group 1139**

**Otoiu Adrian Ariton Alexandru**

**Bucur Alexia-Gabriela**

**Coman Alex**

**Cojocaru Florin**

**BUCHAREST**

**2025**

Table of Contents

[1.Introduction 3](#_Toc188016441)

[2.Problem Description 4](#_Toc188016442)

[3.Application of data mining/knowledge discovery methods 10](#_Toc188016443)

[4.Conclusions 53](#_Toc188016444)

[5.References 56](#_Toc188016445)

## **1.Introduction**

Diabetes is one of the most significant public health challenges of the 21st century, affecting millions of individuals worldwide. As a chronic disease, diabetes is characterized by elevated blood glucose levels, which can lead to severe complications such as cardiovascular diseases, kidney failure, neuropathy, and blindness if not effectively managed. With its growing prevalence, particularly in low- and middle-income countries, diabetes places an enormous burden on healthcare systems, economies, and the quality of life of those affected.

Predicting the likelihood of diabetes in individuals is critical for early detection and prevention. Early intervention can significantly improve outcomes by enabling timely lifestyle changes, medical interventions, and regular monitoring. Traditional diagnostic approaches often rely on clinical assessments and standard laboratory tests, which, while effective, may overlook subtle patterns and risk factors in patient data. Machine learning techniques, with their ability to process large datasets and uncover hidden relationships, present an opportunity to enhance the accuracy and efficiency of diabetes prediction.

This project leverages the Pima Indians Diabetes dataset, a widely studied dataset in the field of healthcare analytics. The dataset contains 768 observations across 9 variables, including patient demographics (age, pregnancies), clinical metrics (glucose levels, blood pressure, BMI), and a diabetes pedigree function that reflects genetic predisposition. The objective is to apply machine learning algorithms, such as Random Forest, XGBoost, Logistic Regression, and others, to develop a robust predictive model for diabetes. By comparing the performance of these models, the study aims to identify the most accurate and practical approach for predicting diabetes.

##### Relevance in Practice

The practical implications of this project are far-reaching. For healthcare providers, predictive models can serve as decision-support tools, assisting in identifying high-risk patients and prioritizing early interventions. For individuals, these models can offer personalized risk assessments, empowering them to take preventive measures and manage their health more effectively. At a policy level, insights from predictive modeling can inform public health campaigns, resource allocation, and strategies to address the growing diabetes epidemic.

Furthermore, applying data mining techniques to healthcare challenges like diabetes prediction highlights the potential of integrating advanced analytics into clinical workflows. By combining data-driven insights with clinical expertise, healthcare systems can move toward more proactive, efficient, and patient-centered care. This project not only contributes to the academic exploration of machine learning in healthcare but also bridges the gap between theoretical methods and

real-world applications, aiming to make a meaningful impact on the fight against diabetes.

## **2.Problem Description**

The objective of this project is to predict whether a patient has diabetes using the Pima Indians Diabetes Dataset. The dataset comprises 768 observations and 9 variables, from Kaggle. The target variable is diabetes, which indicates whether the individual is diabetic (1) or non-diabetic (0).

##### Description of the Data

* 1. **Pregnant - Number of times pregnant**
     + **Description**: Represents the number of times a woman has been pregnant.
     + Relevance:
       - Multiple pregnancies might contribute to a higher risk of developing gestational diabetes, which can increase the likelihood of diabetes later in life.
       - This variable is important as it could provide insight into the cumulative metabolic stress related to pregnancies.

##### Glucose - Plasma glucose concentration

* + - **Description**: Plasma glucose concentration obtained during a glucose tolerance test.
    - Relevance:
      * Glucose level is a critical indicator of diabetes as it directly reflects the body's ability to manage blood sugar.
      * Elevated glucose levels are one of the strongest predictors of diabetes, making this variable highly important for the model.

##### Pressure - Diastolic blood pressure (mm Hg)

* + - **Description**: Represents the diastolic blood pressure (lower reading during heartbeats).
    - Relevance:
      * While blood pressure alone is not a direct indicator of diabetes, high blood pressure is commonly associated with diabetes due to shared risk factors like obesity and poor lifestyle choices.
      * Including this variable helps capture potential comorbid conditions.

##### Triceps - Triceps skin fold thickness (mm)

* + - **Description**: Measurement of body fat through skin fold thickness on the triceps.
    - Relevance:
      * Higher body fat percentages, as indicated by triceps thickness, may correlate with a higher likelihood of insulin resistance, a precursor to diabetes.
      * This variable provides indirect insight into the individual's overall body composition and potential metabolic health.

##### Insulin - 2-Hour serum insulin (mu U/ml)

* + - **Description**: Represents the amount of insulin in the blood two hours after a glucose tolerance test.
    - Relevance:
      * Abnormally high or low insulin levels indicate insulin resistance or beta-cell dysfunction, both of which are central to diabetes development.
      * This is a key variable for distinguishing between diabetic and non-diabetic individuals.

##### Mass - Body mass index (BMI)

* + - **Description**: BMI is calculated as weight in kilograms divided by height in meters squared.
    - Relevance:
      * Obesity is a major risk factor for Type 2 diabetes, making BMI an essential predictor.
      * Higher BMI values indicate higher body fat, which is linked to increased insulin resistance.

##### Pedigree - Diabetes pedigree function

* + - **Description**: A measure of diabetes likelihood based on family history and genetic predisposition.
    - Relevance:
      * Genetic predisposition is a significant factor in diabetes risk, and this variable quantifies the hereditary component.
      * It provides additional context for individuals who might develop diabetes despite average glucose or BMI levels.

##### Age - Age (years)

* + - **Description**: Represents the age of the individual in years.
    - Relevance:
      * Age is strongly correlated with diabetes, as the risk increases with age due to cumulative exposure to risk factors.
      * Older individuals are more likely to have impaired glucose tolerance or insulin resistance.

##### Diabetes - Target variable

* + - **Description**: A binary variable where 1 indicates diabetic and 0 indicates non-diabetic.
    - Relevance:
      * This is the dependent variable (target) we aim to predict using the other features.
      * The model will be trained to classify individuals as diabetic or non-diabetic based on the input features.

##### Why Use These Variables?

Each variable contributes unique information that collectively aids in predicting diabetes:

1. **Glucose**, **BMI**, and **Insulin**: These are the strongest indicators of diabetes, directly linked to metabolic health.
2. **Pregnancy** and **Triceps**: Provide insights into cumulative metabolic stress and body fat distribution.
3. **Blood Pressure** and **Age**: Capture additional risk factors that commonly co-occur with diabetes.
4. **Pedigree**: Captures the genetic and hereditary component, essential for assessing predisposition.

By using all these variables, the model can comprehensively evaluate physiological, genetic, and lifestyle-related risk factors, improving its predictive accuracy. The dataset's diversity ensures a holistic approach to understanding and predicting diabetes.

**Working Hypotheses**

Based on the dataset, the following hypotheses are proposed:

###### Hypothesis 1 (H1): Glucose Levels and Diabetes

Higher plasma glucose concentrations are associated with an increased likelihood of diabetes. This aligns with clinical diagnostic criteria for diabetes.

###### Hypothesis 2 (H2): BMI and Diabetes

Higher BMI is positively correlated with the likelihood of diabetes. Obesity significantly increases insulin resistance, a key factor in type 2 diabetes.

###### Hypothesis 3 (H3): Age and Diabetes

The probability of developing diabetes increases with age due to metabolic and hormonal changes.

###### Hypothesis 4 (H4): Insulin Levels and Diabetes

Abnormal insulin levels, whether too high or too low, are strong indicators of diabetes risk.

###### Hypothesis 5 (H5): Blood Pressure and Diabetes

Elevated diastolic blood pressure is linked to a higher risk of diabetes, reflecting shared physiological mechanisms such as inflammation and endothelial dysfunction.

**Methodology for Testing the Hypotheses**

To test these hypotheses, we will employ the following machine learning models:

1. Logistic Regression (GLM):
   * A statistical method to estimate the probability of diabetes based on predictor variables.
   * Offers interpretable coefficients to explain the influence of each variable.
2. K-Nearest Neighbors (KNN):
   * Classifies observations based on the majority class among k-nearest neighbors.
   * Effective for identifying patterns in datasets with clear clusters.
3. Random Forest Classifier:
   * An ensemble method that builds multiple decision trees and aggregates their predictions.
   * Reduces overfitting and identifies the most significant predictors.
4. XGBoost Classifier:
   * A powerful gradient boosting algorithm that handles missing values and complex feature interactions.
   * Optimized for classification tasks and superior in performance.

**References to Similar Studies**

###### CRISP-DM (Cross-Industry Standard Process for Data Mining)

CRISP-DM provides a structured framework with six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. This methodology ensures systematic testing and refinement of machine learning models.

###### SEMMA (Sample, Explore, Modify, Model, Assess)

SEMMA emphasizes iterative exploration and refinement, ensuring robust preprocessing and accurate model assessment. This approach aligns well with the complexity of the diabetes dataset.

###### KDD (Knowledge Discovery in Databases)

KDD focuses on extracting meaningful patterns through preprocessing, transformation, mining, and evaluation. This approach is particularly suited for discovering hidden trends in the dataset.

###### PDDA (Process-Driven Data Analytics)

PDDA integrates business processes with data analytics to ensure that insights are actionable and relevant to stakeholders. For healthcare projects, this method ensures alignment with clinical needs and outcomes.

**Economic and Business Intuition**

1. Healthcare Cost Mitigation:
   * Early prediction of diabetes can reduce long-term healthcare costs by preventing complications.
   * Focused screening programs improve resource allocation.
2. Improved Patient Outcomes:
   * Predictive models enable targeted interventions, enhancing patient care and quality of life.
   * Personalized recommendations based on risk profiles improve adherence to preventive measures.
3. Resource Optimization:
   * Identifying high-risk individuals helps healthcare providers allocate resources efficiently.
   * Reduces unnecessary diagnostic tests, minimizing the financial burden on healthcare systems.

**Relevance of Data Mining Methods**

Each model provides unique advantages for predicting diabetes:

* Logistic Regression: Provides interpretable results and establishes baseline performance.
* KNN: Captures proximity-based relationships, effective for smaller datasets.
* Random Forest Classifier: Mitigates overfitting and highlights feature importance.
* XGBoost Classifier: Excels in handling complex interactions and achieving high accuracy.

These methods collectively enable robust testing of hypotheses and ensure actionable insights for diabetes prediction.

## **3.Application of data mining/knowledge discovery methods**

**Comparing Multiple Models**:

*Objective:*

The objective of this exercise is to predict if a patient has diabetes for the Pima Indians Diabetes Data set using multiple ML algorithms.

We will build and train following Machine Learning models during this analysis.

1. Random Forest
2. eXtreme Gradient Boosting Machine (XGBOOST)
3. K-Nearest Neighbours(KNN)
4. Logistic Regression (glm)

*Data Set*

The data set is comprised of 768 observations and 9 variables. We will be using diabetes as our response/target variable.

Data Description for the 9 variables are as follows.

* 1. pregnant - Number of times pregnant
  2. glucose - Plasma glucose concentration (glucose tolerance test)
  3. pressure - Diastolic blood pressure (mm Hg)
  4. triceps - Triceps skin fold thickness (mm)
  5. insulin - 2-Hour serum insulin (mu U/ml)
  6. mass - Body mass index (weight in kg/(height in m)^2)
  7. pedigree - Diabetes pedigree function
  8. age - Age (years)
  9. diabetes - Class variable (test for diabetes)

Load Packages

library(caret) #ML Model buidling package library(tidyverse) #ggplot and dplyr

library(MASS) #Modern Applied Statistics with S library(summarytools)

library(corrplot) #Correlation plot

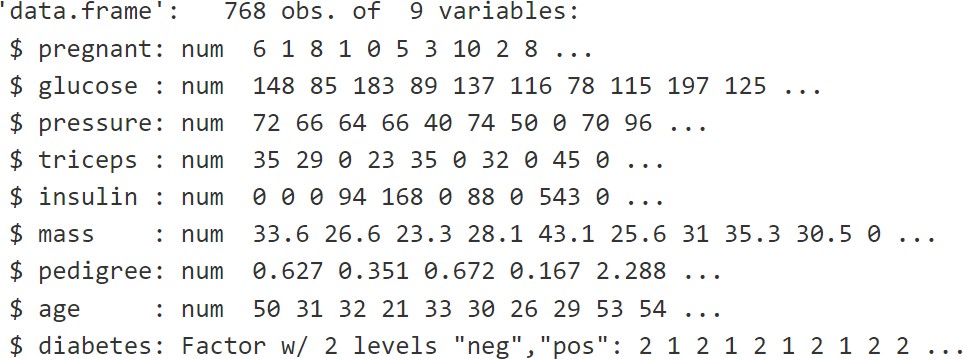
library(gridExtra) #Multiple plot in single grip space library(timeDate)

library(pROC) #ROC library(caTools) #AUC

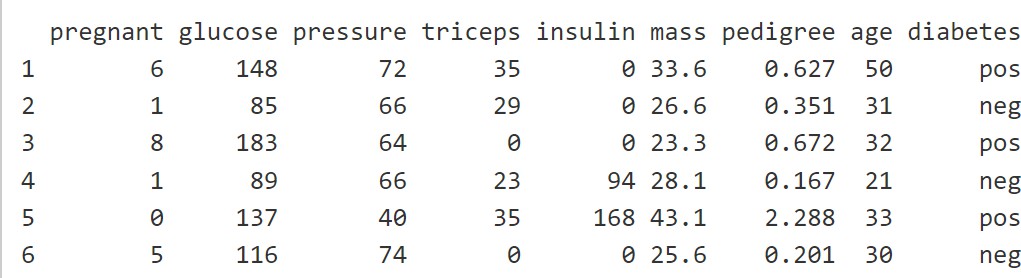
library(rpart.plot) #CART Decision Tree

library(e1071) #imports graphics, grDevices, class, stats, methods, utils library(graphics) #fourfoldplot

df <- read.csv("C:/Users/flori/Downloads/diabetes.csv") df$diabetes <- factor(df$diabetes, levels = c("neg", "pos")) str(df)



#head(df)

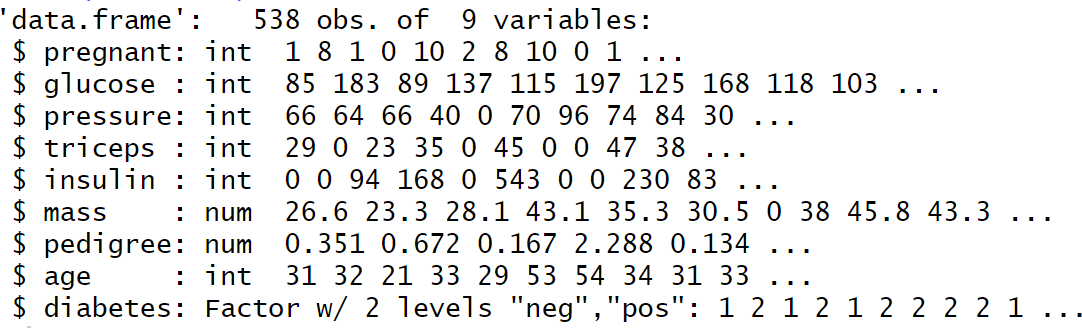


#store rows for partition

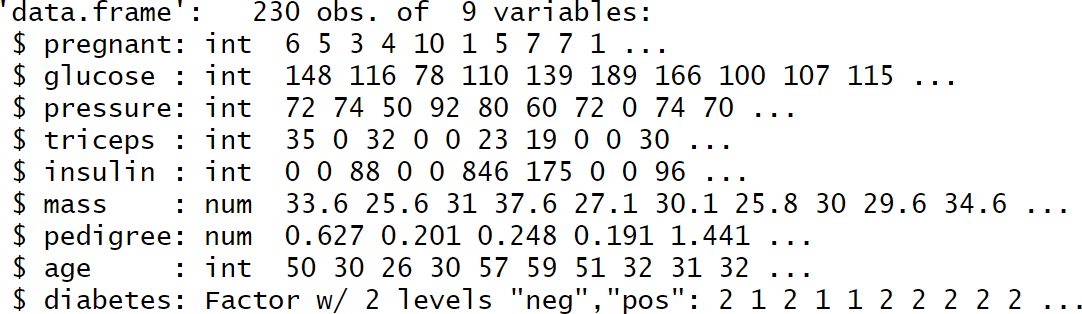
partition <- caret::createDataPartition(y = df$diabetes, times = 1, p = 0.7, list = FALSE)

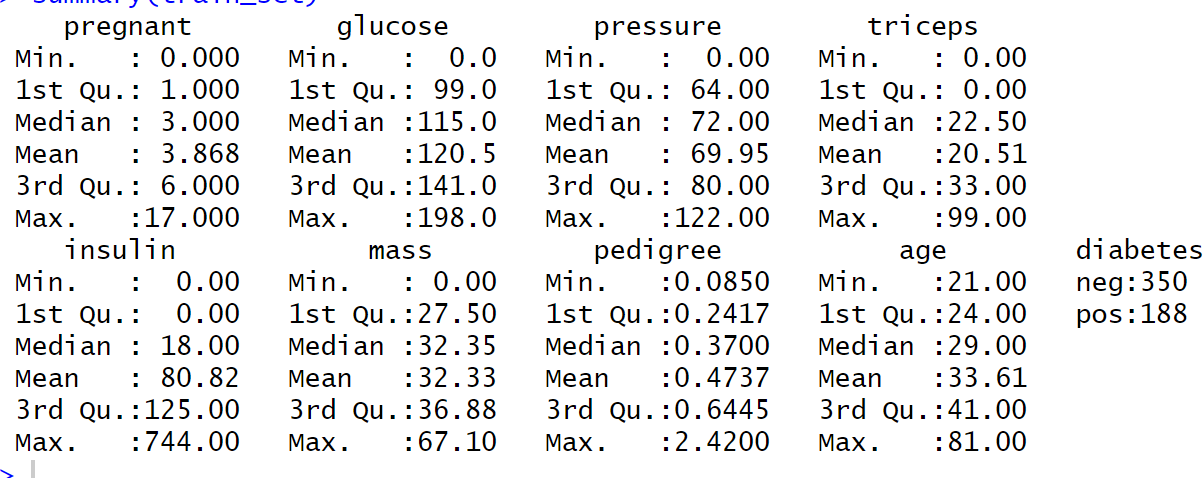
# create training data set train\_set <- df[partition,]# create testing data set, subtracting the rows partition to get remaining 30% of the data test\_set <- df[-partition,]

str(train\_set)



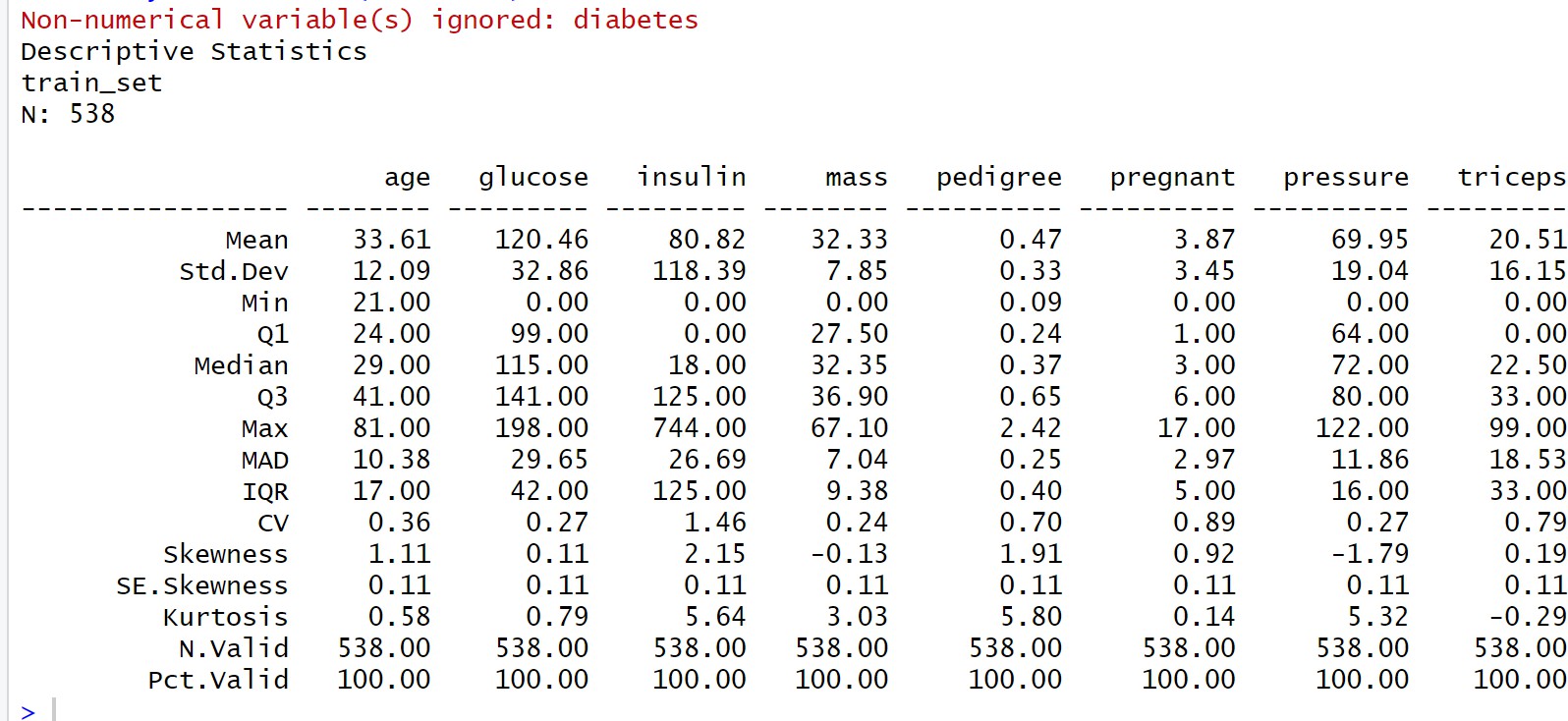
str(test\_set)





Descriptive Statistics summary(train\_set)

summarytools::descr(train\_set)



There are no missing values in the dataset and hence we can proceed further.

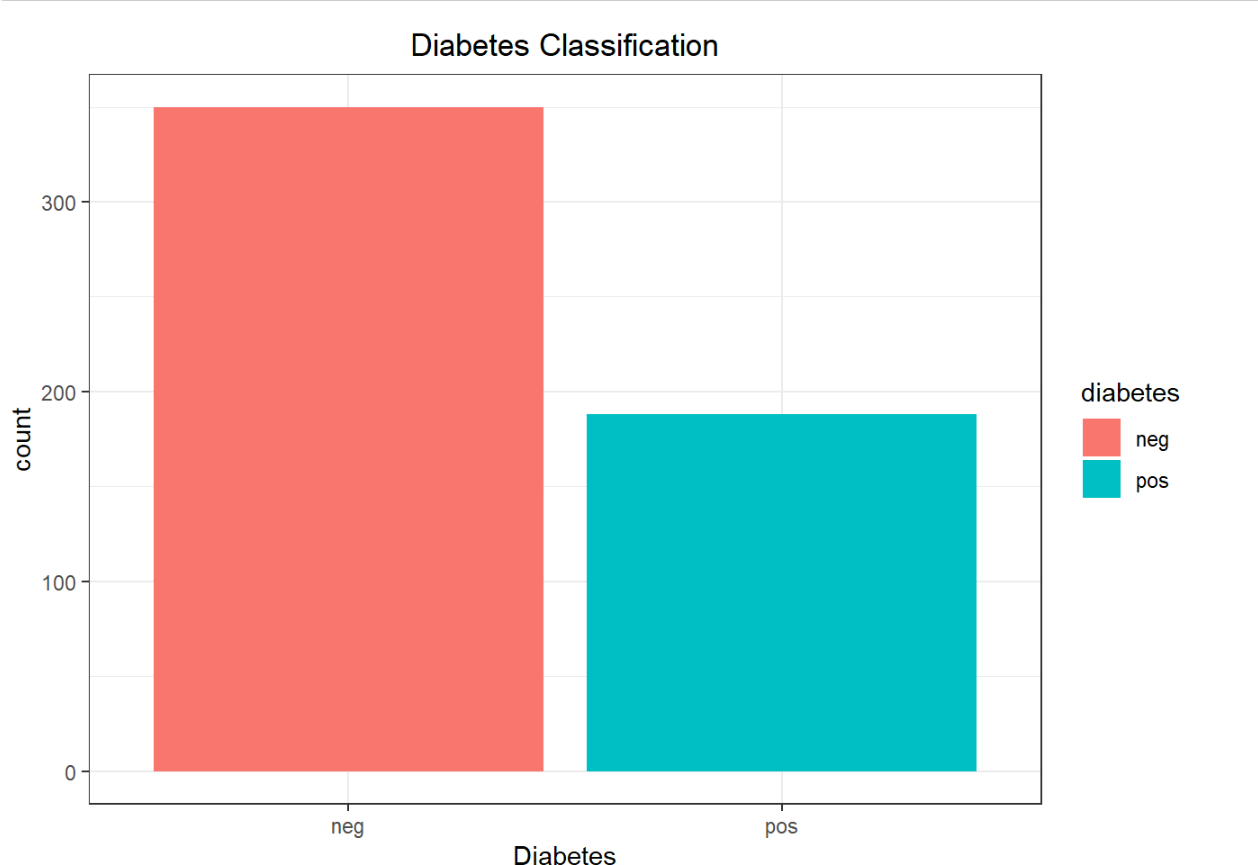
Exploratory data analysis

**Diabetes Distribution**

ggplot(train\_set, aes(train\_set$diabetes, fill = diabetes)) + geom\_bar() +

theme\_bw() +

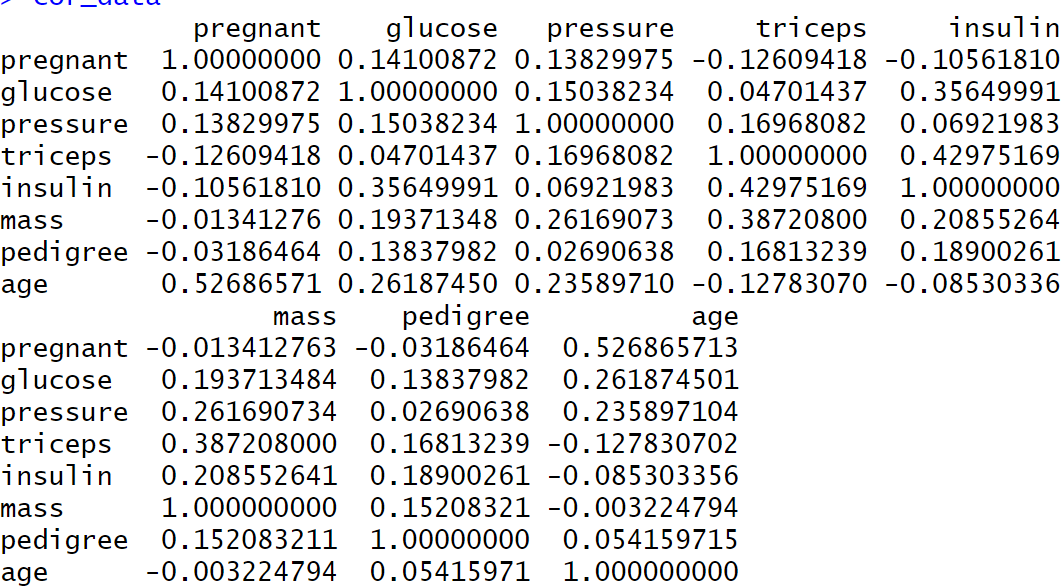
labs(title = "Diabetes Classification", x = "Diabetes") + theme(plot.title = element\_text(hjust = 0.5))



Correlation Matrix

cor\_data <- cor(train\_set[,setdiff(names(train\_set), 'diabetes')]) #Numerical Correlation Matrix

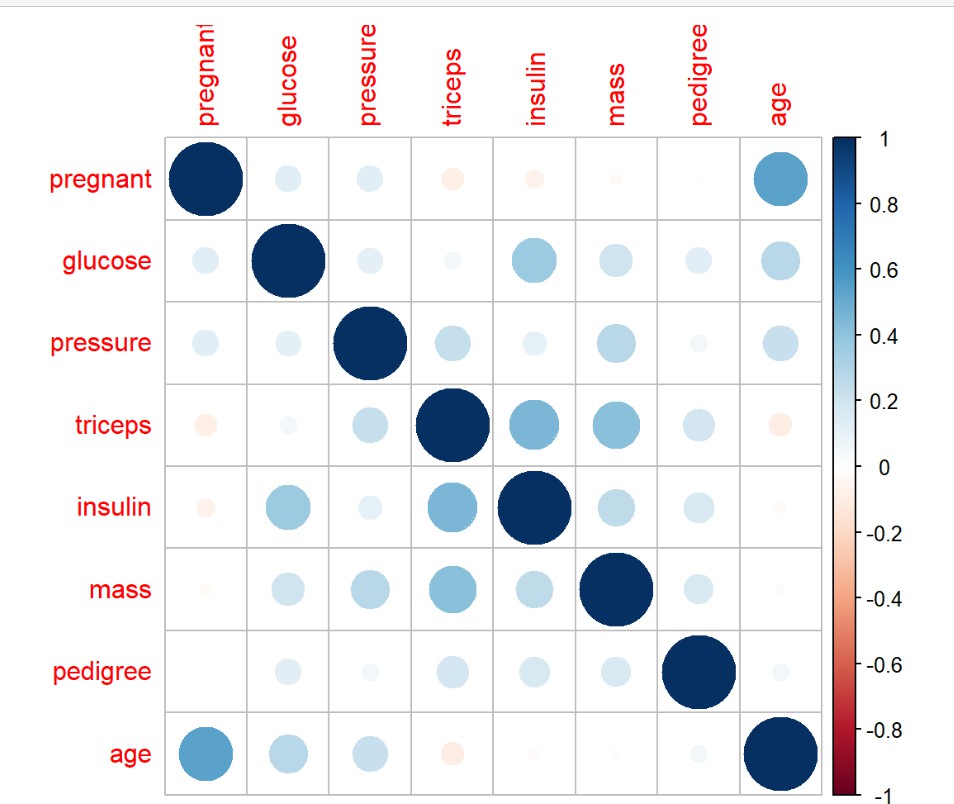
cor\_data



The correlation matrix evaluates linear relationships between numerical variables in the dataset, excluding the target variable diabetes. Key observations include a moderately positive correlation between pregnant and age (0.526), suggesting older individuals tend to have been pregnant more often, and a weak positive correlation between glucose and insulin (0.356), indicating higher glucose levels may relate to higher insulin levels. Additionally, mass and

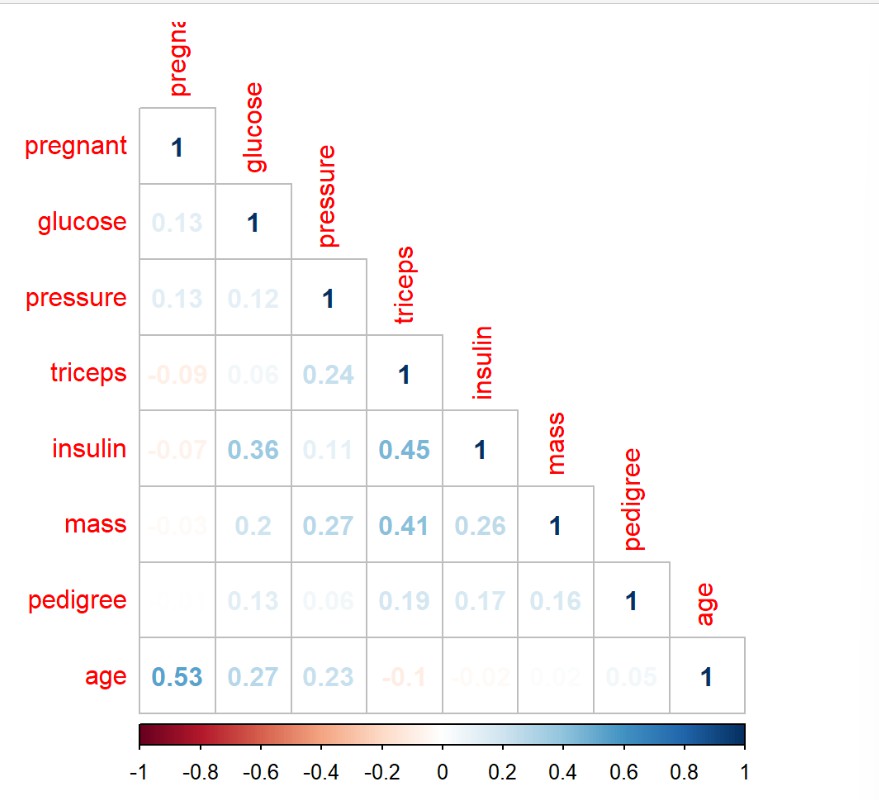
triceps (0.387) show a moderate positive relationship linked to body fat composition. These insights highlight potential predictors for further analysis while suggesting care to avoid multicollinearity with highly correlated variables.

# Correlation matrix plots corrplot::corrplot(cor\_data)



The **correlation matrix plot** visually represents the strength and direction of linear relationships between variables in the dataset. Each circle's size and color intensity correspond to the magnitude of the correlation, with blue indicating positive relationships and red indicating negative relationships. For example, a strong positive correlation between pregnant and age is evident from the large blue circle, while weaker relationships, such as glucose and pedigree, are shown with smaller circles. This plot is useful for identifying potential predictor variables and assessing multicollinearity, ensuring effective feature selection in the subsequent modeling process.

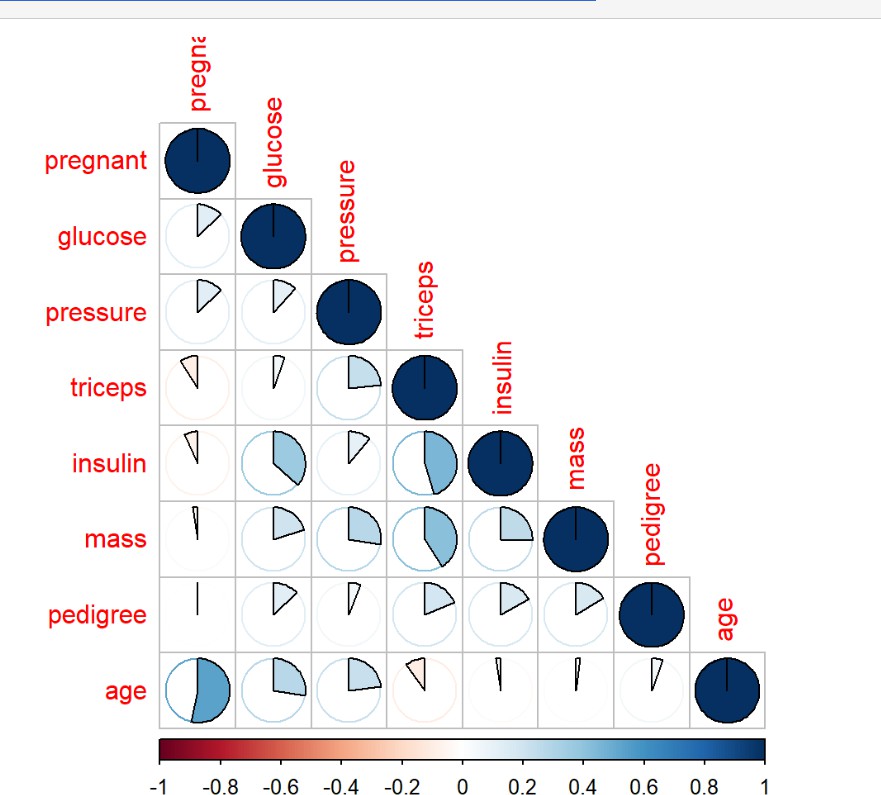
corrplot::corrplot(cor\_data, type = "lower", method = "number")



Key Highlights:

* **Diagonal Values**: Variables are perfectly correlated with themselves, so the diagonal always shows a correlation of 1.
* **Color and Size**:
  + Blue indicates positive correlations, red indicates negative.
  + Intensity and size represent the strength of the correlation.
* **Examples**:
  + **Pregnant** and **age** have a strong positive correlation (0.53).
  + **Glucose** and **insulin** show a moderate positive correlation (0.36).
  + **Triceps** and **pedigree** show very weak or no correlation (near zero).

corrplot::corrplot(cor\_data, type = "lower", method = "pie")



The pie chart in this correlation plot represents the **magnitude and direction** of the correlation values between variables.

We can see that there is a moderately positive high correlation between age and pregnant.

**Univariate Analysis**

univar\_graph <- **function**(univar\_name, univar, data, output\_var) {

g\_1 <- ggplot(data, aes(x=univar)) + geom\_density() + xlab(univar\_name) +

theme\_bw()

g\_2 <- ggplot(data, aes(x=univar, fill=output\_var)) + geom\_density(alpha=0.4) +

xlab(univar\_name) + theme\_bw()

gridExtra::grid.arrange(g\_1, g\_2, ncol=2, top = paste(univar\_name,"variable", "/ [ Skew:",timeDate::skewness(univar),"]"))

}

**for** (x **in** 1:(ncol(train\_set)-1)) {

univar\_graph(univar\_name = names(train\_set)[x], univar = train\_set[,x], data = train\_set, output\_var = train\_set[,'diabetes'])

}

**Skewness highlights the imbalance or asymmetry in a variable’s distribution**, which can affect model performance. Highly skewed variables (right or negative) often require transformations, while moderately skewed variables may be easier for models to handle directly.

A group of graphs showing different types of data

Description automatically generated

For instance, the univariate analysis of the pregnant variable reveals a right-skewed distribution with most values concentrated between 0 and 3 pregnancies, as shown in the overall density plot. The skewness value of **0.849** indicates that higher pregnancy counts are less frequent but present.

When stratified by diabetes status, the conditional density plot highlights that individuals without diabetes (neg) are primarily concentrated at lower pregnancy counts (0–2). In contrast, individuals with diabetes (pos) show a broader distribution with notable density around 4–8 pregnancies. This suggests that a higher number of pregnancies may be associated with an increased likelihood of diabetes.

The distinct differences in the distributions between the pos and neg classes indicate that pregnant is a potentially valuable predictor for diabetes classification.

A group of graphs showing different sizes of data

Description automatically generated with medium confidence

Variables such as Insulin, pedigree and age have high right skewness.

pressure and mass have negative skewness.

pregnant, glucose, and triceps have moderate to low right skewness.

**Insights:**

 **Insulin, Pedigree, and Age**: Their high right skewness indicates that most individuals have low values, but a few extreme cases pull the distribution upward. For instance, insulin levels or age may have rare but significant outliers (e.g., older patients or those with very high insulin). These outliers might strongly influence predictions, making normalization or transformation essential for better model stability.

 **Pressure and Mass**: Negative skewness suggests most values are relatively high, with a few unusually low outliers. This could indicate that low blood pressure or BMI cases are rare but may hold critical information for identifying specific trends in diabetic or non-diabetic individuals.

 **Pregnant, Glucose, and Triceps**: Moderate to low right skewness suggests these variables are closer to normal distribution, with fewer extreme outliers. This makes them more reliable as predictors without requiring extensive transformations, but mild adjustments could still improve accuracy.

**Univariable Analysis using Bar plots**

bivar\_graph <- **function**(bivar\_name, bivar, data, output\_var) {

g\_1 <- ggplot(data = data, aes(x = bivar, fill = output\_var)) + geom\_bar(stat='count', position='dodge') +

theme\_bw() +

labs( title = paste(bivar\_name,"- Diabetes", sep =" "), x = bivar\_name) + theme(plot.title = element\_text(hjust = 0.5))

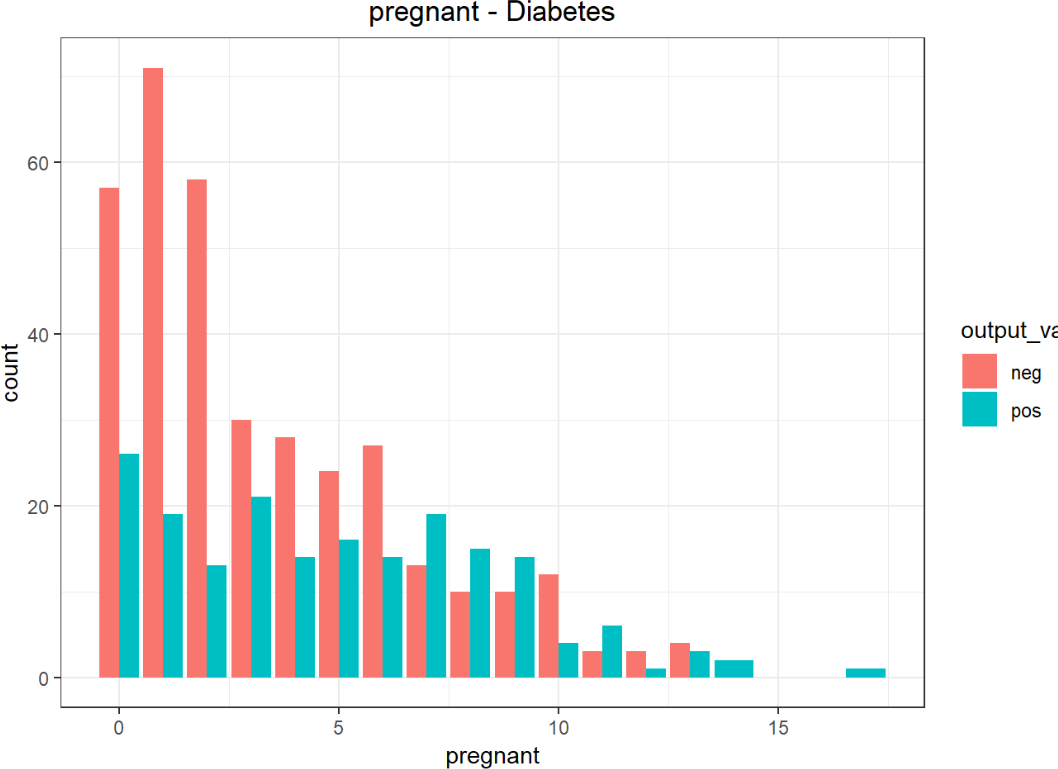
plot(g\_1)

}

**for** (x **in** 1:(ncol(train\_set)-1)) {

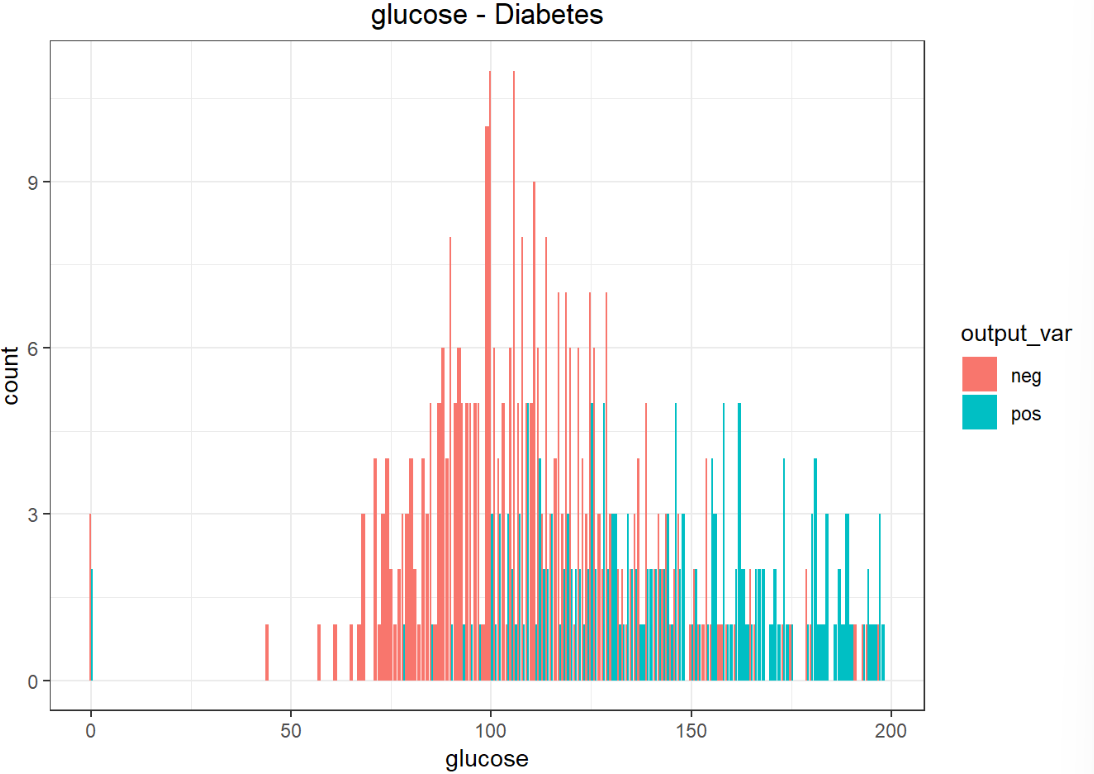
bivar\_graph(bivar\_name = names(train\_set)[x], bivar = train\_set[,x], data = train\_set, output\_var = train\_set[,'diabetes'])

}



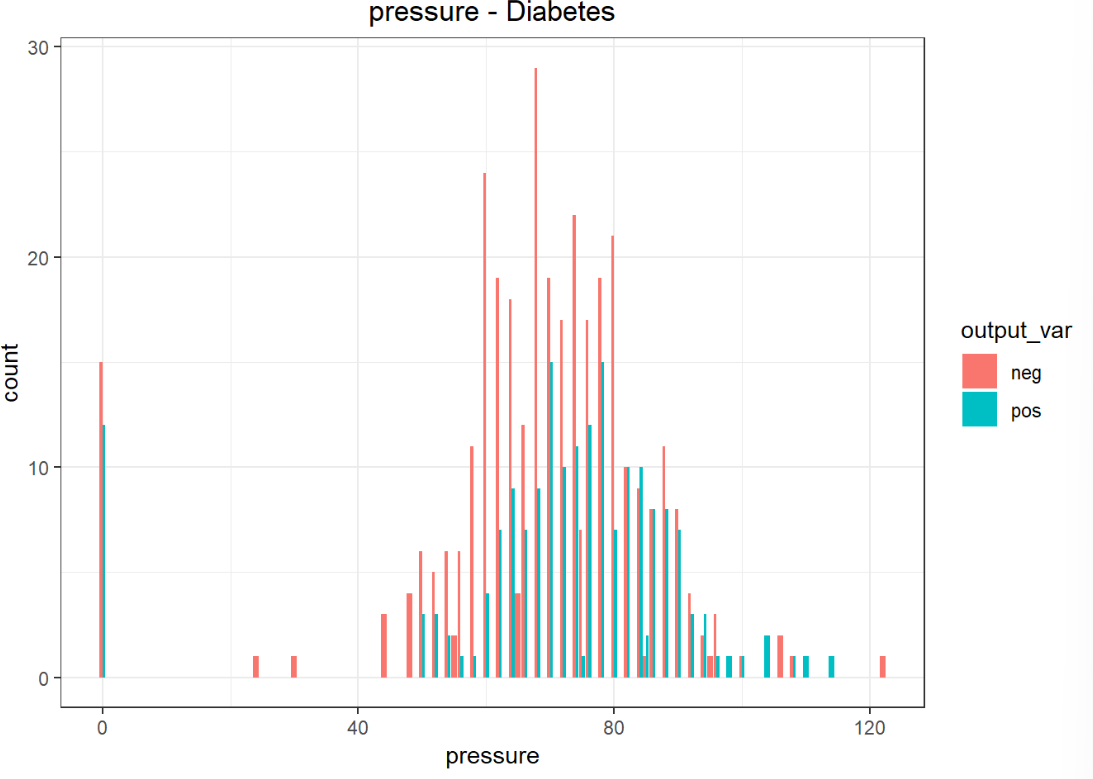
##### Chart 1: Pregnancy vs. Diabetes

* Interpretation:
  + A higher number of pregnancies appears to slightly correlate with an increased likelihood of diabetes.
  + Women with no pregnancies (0 on the x-axis) are more likely to be non-diabetic, while those with multiple pregnancies show a higher proportion of diabetes cases.
  + This may indicate a potential relationship between pregnancy frequency and diabetes risk, possibly due to gestational diabetes or related factors.



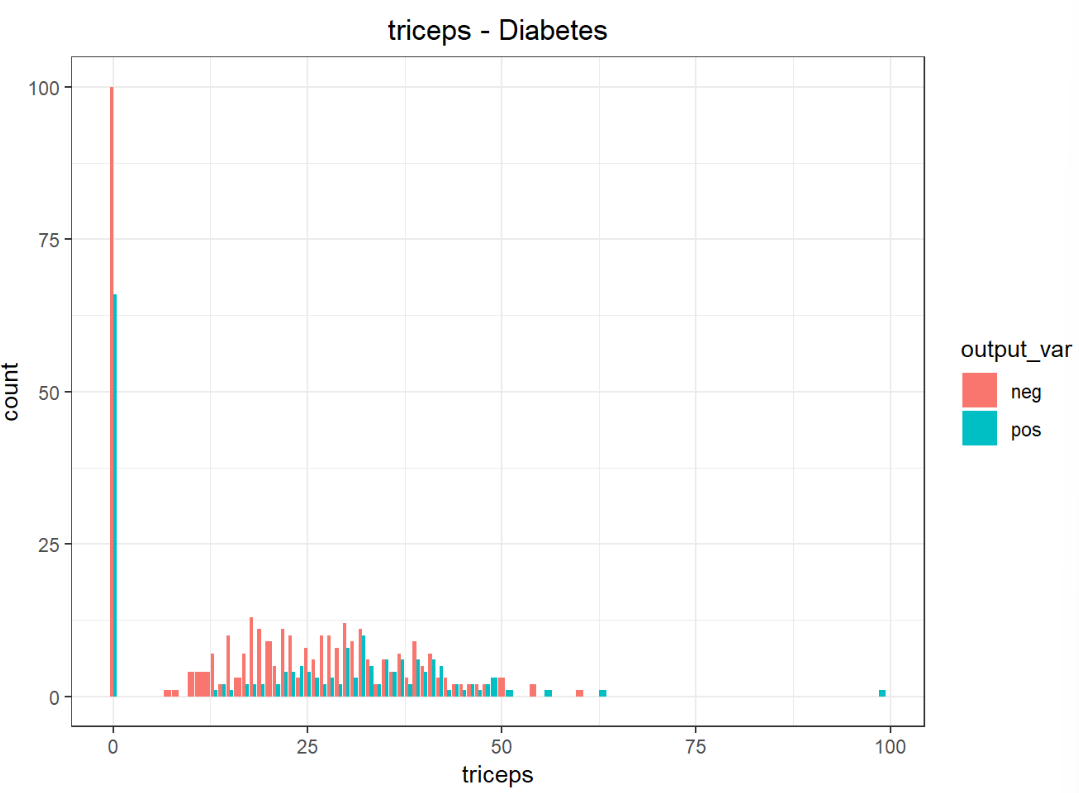
##### Chart 2: Glucose Levels vs. Diabetes

* Interpretation:
  + Glucose levels strongly correlate with diabetes. Individuals with higher glucose levels are more likely to have diabetes.
  + Non-diabetic individuals dominate the lower glucose range (e.g., below 100), while diabetic individuals are more prevalent in the higher glucose ranges.
  + This indicates that elevated glucose is a key diagnostic factor for diabetes.



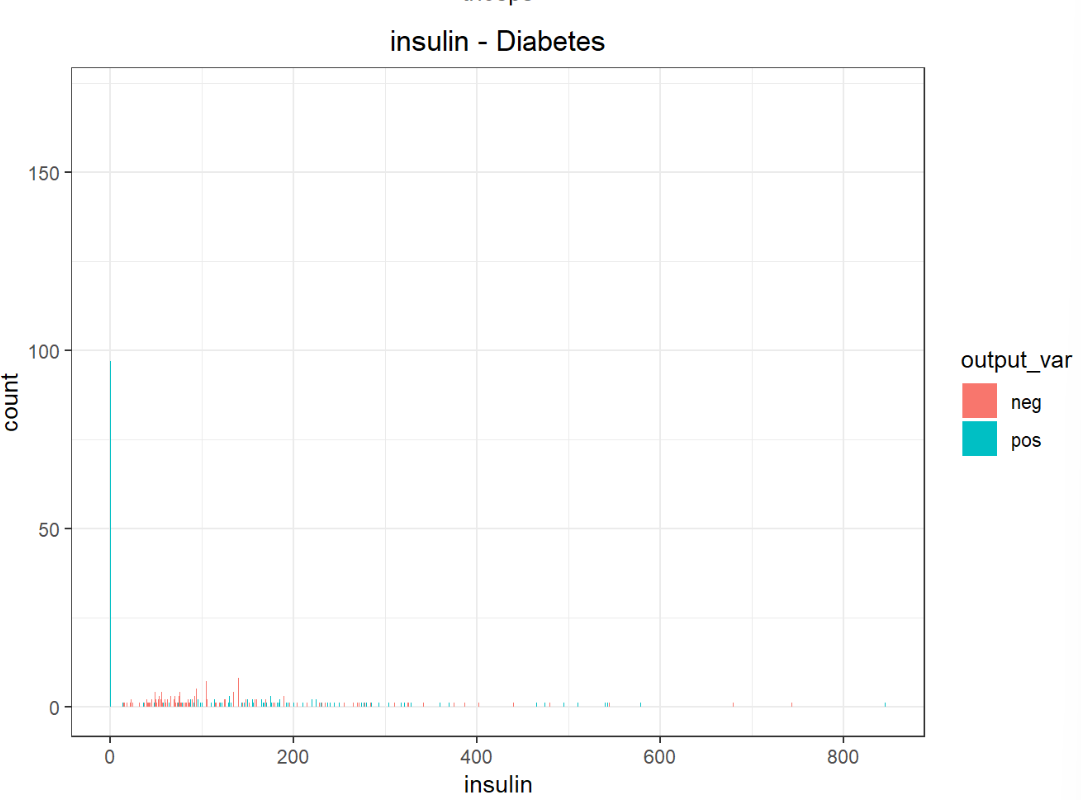
##### Chart 3: Blood Pressure vs. Diabetes

* Interpretation:
  + Blood pressure shows some overlap between diabetic and non-diabetic individuals, especially in the range of 60–80 mmHg.
  + Non-diabetic individuals are slightly more common across all blood pressure ranges, but the trend does not show a strong association between blood pressure and diabetes.
  + While blood pressure might not be a direct indicator of diabetes, extreme low or high values could still provide some clues.



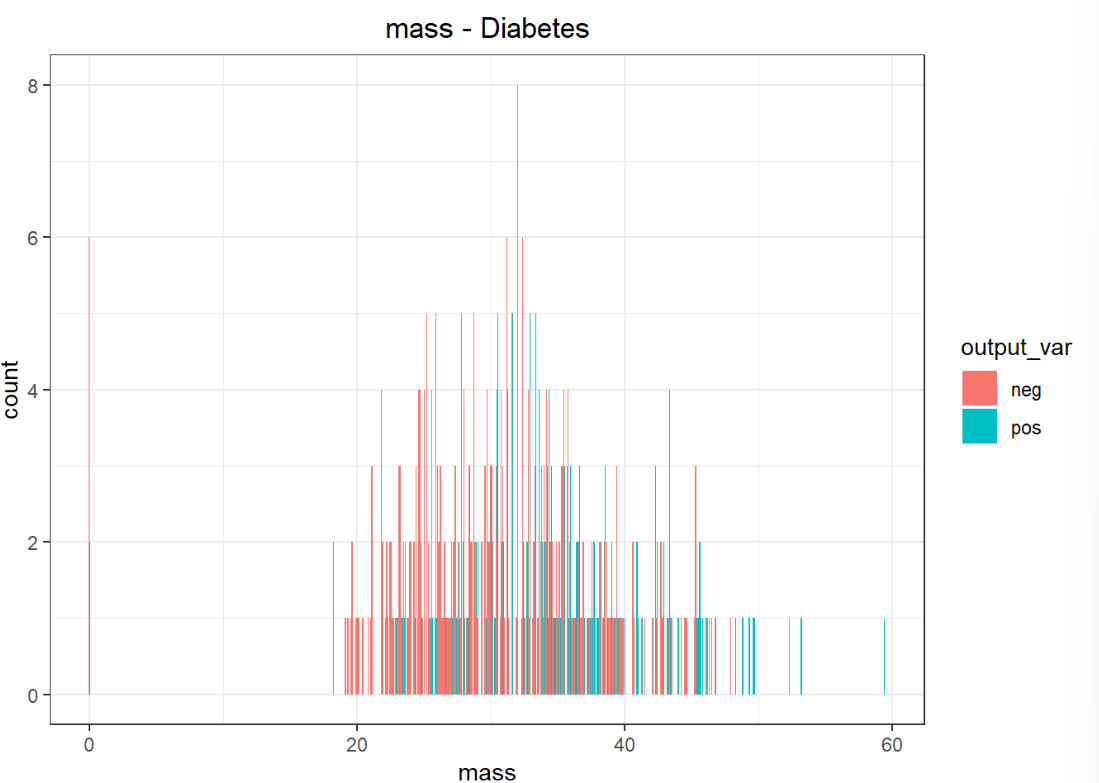
##### Chart 4: Triceps Skin Fold Thickness vs. Diabetes

* Interpretation:
  + A significant number of individuals have very low triceps thickness values (close to zero), especially among non-diabetic individuals.
  + Diabetic individuals show a more even distribution across higher triceps thickness values.
  + This could suggest that higher triceps skin fold thickness might slightly correlate with diabetes, but there is also a significant overlap with non-diabetic individuals.



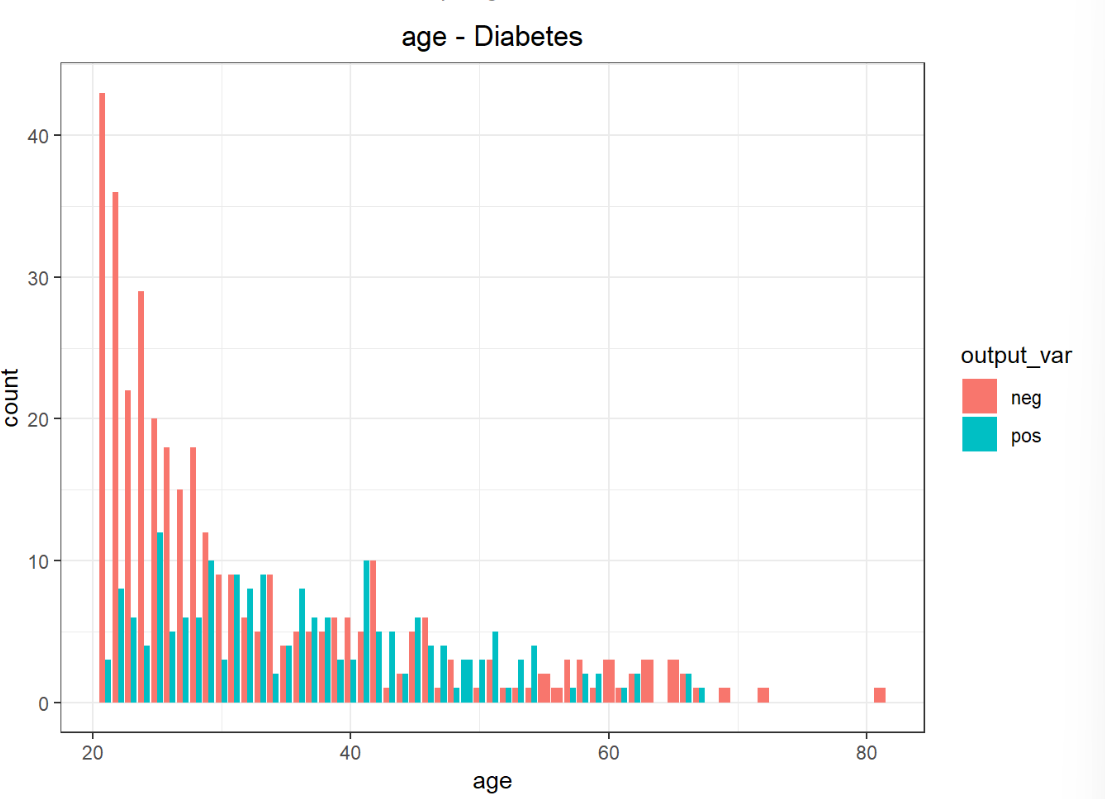
##### Chart 5: Insulin Levels vs. Diabetes

* Interpretation:
  + Many non-diabetic individuals have zero insulin levels, which might reflect missing data or individuals who do not require insulin.
  + Diabetic individuals exhibit higher insulin levels, which could be due to insulin resistance or treatment.
  + This indicates that insulin levels are a useful metric for identifying diabetes but may require careful handling of zero values.



##### Chart 6: BMI (Body Mass Index) vs. Diabetes

* Interpretation:
  + BMI is strongly correlated with diabetes. Diabetic individuals tend to have higher BMI values, reflecting a link between obesity and diabetes risk.
  + Non-diabetic individuals are more common in the lower BMI ranges, while diabetic individuals become more prevalent as BMI increases.
  + This highlights BMI as a significant factor in assessing diabetes risk.



##### Chart 7: Age vs. Diabetes

* Interpretation:
  + Diabetes prevalence increases with age. Younger individuals (e.g., below 30) are mostly non-diabetic, while the proportion of diabetic individuals increases significantly in older age groups (e.g., above 40).
  + This suggests age is a strong risk factor for diabetes, potentially due to the cumulative impact of lifestyle and metabolic changes over time.

We can not see any trend in pregnant vs diabetes plot.

* Higher the glucose level, higher are the chances of positive diabetes. This is line with expectations.
* Diabetic pressure levels seems to be normally distributed for negative results and the positive cases range from bp of 60 to 110.
* As the BMI increases, the chances for a positive Diabetes case also increases.
* We can see that the positive diabetes cases increases as the age increases > 25.

\*The reason that some of the graphs are small is due to fact that the maximum value for that variable is very high and the count for each is very small.

This can be confirmed by plotting box plots for each of the variables in the following section.

Outlier Detection

**A screenshot of a graph

Description automatically generated**

**What Each Plot Shows:**

1. Pregnant Outlier Detection:
   * Observation: Diabetic (pos) individuals tend to have a higher pregnancy count on average compared to non-diabetic (neg) individuals. A few extreme cases (outliers) are visible in the diabetic group.
   * Takeaway: Pregnancy count shows a moderate association with diabetes.
2. Glucose Outlier Detection:
   * Observation: Diabetic individuals have much higher glucose levels than non-diabetic individuals. Outliers are present but still show a clear distinction between the two groups.
   * Takeaway: Glucose is a strong predictor of diabetes, with higher glucose levels strongly associated with the diabetic group.
3. Pressure Outlier Detection:
   * Observation: There is significant overlap in diastolic blood pressure between the two groups, with no clear differentiation.
   * Takeaway: Pressure is a weak predictor of diabetes.
4. Triceps Outlier Detection:
   * Observation: Both groups have similar distributions for triceps skinfold thickness, with minimal differences and some outliers.
   * Takeaway: Triceps is a weak predictor, as it doesn’t distinguish diabetic and non-diabetic groups well.

A screenshot of a graph

Description automatically generated

1. Insulin Outlier Detection:
   * Observation: Diabetic individuals have much higher insulin levels, but there are extreme outliers in this group.
   * Takeaway: Insulin is a strong predictor, despite the presence of extreme values.
2. Mass (BMI) Outlier Detection:
   * Observation: BMI shows a lot of overlap between diabetic and non-diabetic groups, with some outliers in both.
   * Takeaway: BMI is a weak predictor when used alone.
3. Pedigree Outlier Detection:
   * Observation: Pedigree (genetic likelihood of diabetes) shows some variability but not much distinction between groups, with outliers in both.
   * Takeaway: Pedigree has limited predictive power.
4. Age Outlier Detection:
   * Observation: Diabetic individuals tend to be older on average, with some overlap and outliers.
   * Takeaway: Age shows a moderate association with diabetes.

Based on the analysis of the dataset, **glucose** and **insulin** levels emerge as the most significant predictors of diabetes, with diabetic individuals (pos) showing notably higher values and variability in both features. **Pregnancy count** shows a moderate association, with slightly higher frequencies observed among diabetic cases. In contrast, variables like **pressure** (diastolic blood pressure), **triceps** (skinfold thickness), and **mass** (BMI) exhibit significant overlap between diabetic and non-diabetic groups, indicating limited predictive power. While some features, such as **insulin**, display extreme outliers in the diabetic group, the weak differentiation in other variables suggests that these features may not independently predict diabetes and are better used in combination with stronger predictors like glucose and insulin for more robust modeling and classification.

\* We will take care of them in the train() function of the CARET package.

**ML Model Building on Train Data Set**

1. Random Forest Model

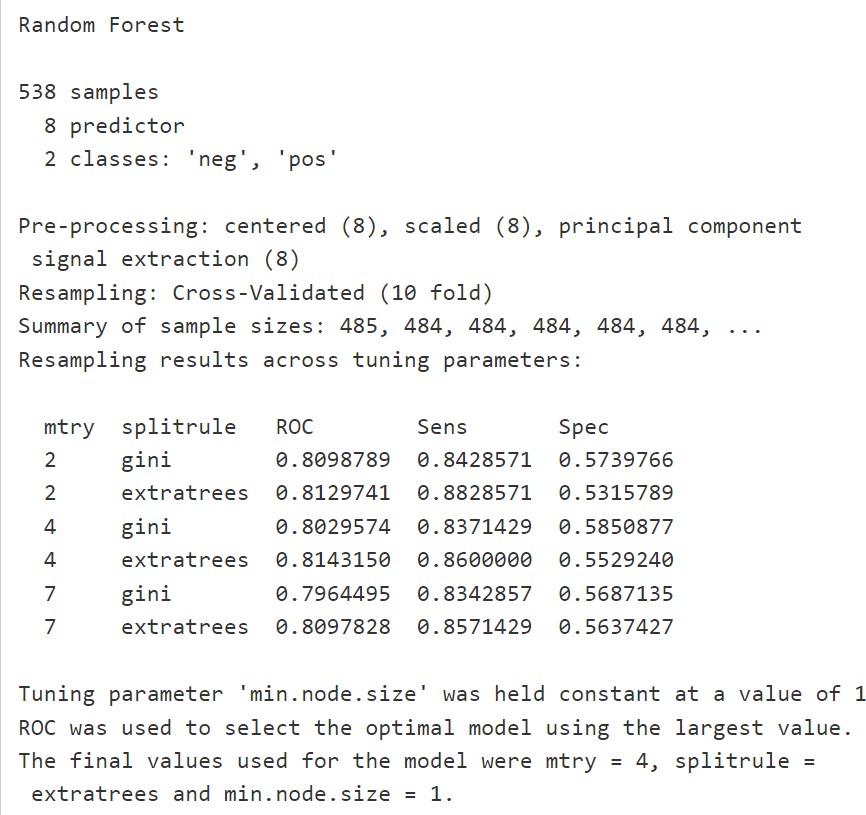
model\_forest <- caret::train(diabetes ~., data = train\_set, method = "ranger",

metric = "ROC",

trControl = trainControl(method = "cv", number = 10,

classProbs = T, summaryFunction = twoClassSummary), preProcess = c("center","scale","pca"))

model\_forest



*# final ROC Value*

model\_forest$results[6,4]

[1] 0.8090226

ROC: A performance evaluation tool for binary classification.AUC: A value summarizing the performance; 0.8097828 indicates **good performance** with potential for improvement.

1. **XGBoost *- eXtreme Gradient BOOSTing***

*# XGBOOST - eXtreme Gradient BOOSTing*

xgb\_grid\_1 <- expand.grid(

nrounds = 50, eta = c(0.03), max\_depth = 1,

gamma = 0,

colsample\_bytree = 0.6,

min\_child\_weight = 1,

subsample = 0.5

)

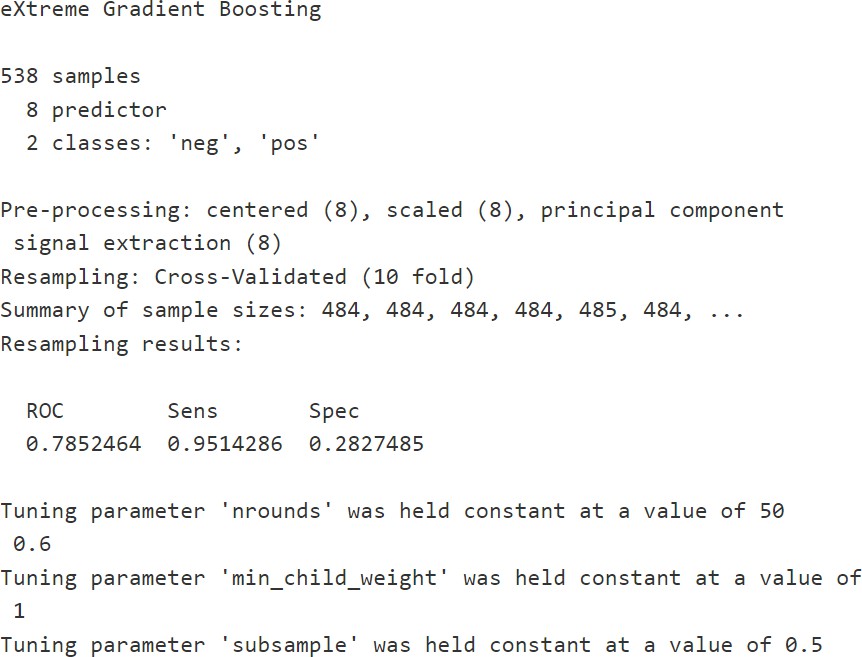
model\_xgb <- caret::train(diabetes ~., data = train\_set, method = "xgbTree",

metric = "ROC", tuneGrid=xgb\_grid\_1,

trControl = trainControl(method = "cv", number = 10,

classProbs = T, summaryFunction = twoClassSummary), preProcess = c("center","scale","pca"))

model\_xgb



Performance Metrics

* 1. **ROC (0.7852464)**:
     + The ROC AUC value of **0.785** indicates the model has **good discrimination** ability. It performs significantly better than random guessing (0.5) and shows a reliable ability to separate "pos" and "neg" classes.
  2. Sensitivity (0.9514286):
     + Sensitivity, also called the **True Positive Rate (TPR)** or **Recall**, measures the proportion of positive cases correctly identified.
     + A **Sensitivity of 95.14%** indicates the model is very effective at correctly identifying "pos" cases.
  3. Specificity (0.2827485):
     + Specificity measures the proportion of negative cases correctly identified.
     + A **Specificity of 28.27%** indicates the model struggles to correctly classify "neg" cases. This could result in many false positives (negatives being classified as positives).

1. ***KNN - K Nearest Neighbours***

K-Nearest Neighbors (KNN) is a supervised machine learning algorithm used for both classification and regression tasks. It works by finding the K closest data points (neighbors) to a given query point in the feature space and making predictions based on these neighbors.

model\_knn <- caret::train(diabetes ~., data = train\_set, method = "knn",

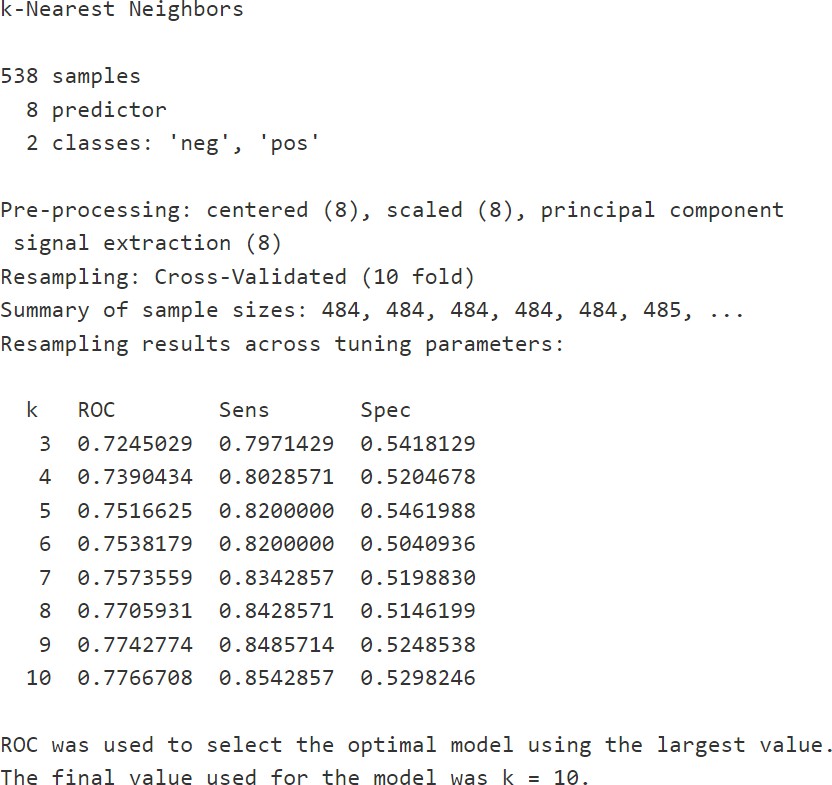
metric = "ROC",

tuneGrid = expand.grid(.k = c(3:10)),

trControl = trainControl(method = "cv", number = 10,

classProbs = T, summaryFunction = twoClassSummary), preProcess = c("center","scale","pca"))

model\_knn



The ROC AUC measures the model's ability to distinguish between the two classes.

The highest ROC value is **0.7767** when k=10k = 10k=10, indicating the best overall classification performance.

Sensitivity (True Positive Rate):

Sensitivity represents the proportion of correctly classified "positive" samples.

The sensitivity improves as kkk increases, peaking at **85.43%** for k=10k = 10k=10, indicating the model becomes more effective at identifying "positive" cases with a larger neighborhood.

Specificity (True Negative Rate):

Specificity represents the proportion of correctly classified "negative" samples.

Specificity is relatively low across all values of kkk, but it peaks at **52.98%** for k=10k = 10k=10. This means the model has a limited ability to correctly classify "negative" cases, which could lead to more false positives.

Optimal Model Selection:

**Optimal K**: The ROC was used as the criterion for selecting the best model. The value of k=10k

= 10k=10 yielded the highest ROC of **0.7767**, making it the optimal choice.

Final Model Performance:

ROC: 0.7767 (good but not perfect discrimination ability). Sensitivity: 85.43% (very good at identifying positives).

Specificity: 52.98% (moderate ability to identify negatives).

1. **Logistic Regression**

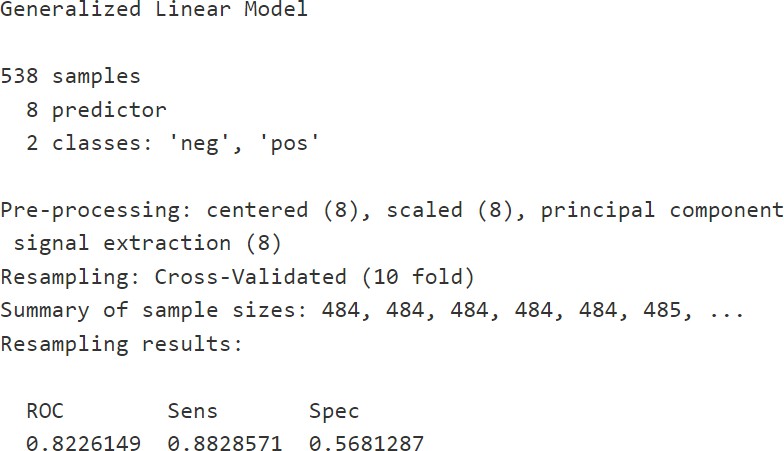
**model\_glm <- caret::train(diabetes ~., data = train\_set, method = "glm",**

**metric = "ROC", tuneLength = 10,**

**trControl = trainControl(method = "cv", number = 10,**

**classProbs = T, summaryFunction = twoClassSummary), preProcess = c("center","scale","pca"))**

**model\_glm**



***#final ROC Value* model\_glm$results[2] 0.8226149**

**This value indicates the model has a strong ability to discriminate between the two classes (e.g., predicting "diabetes" vs. "no diabetes").**

**Interpretation: An ROC AUC of 0.8226 means there is an 82.26% chance that the model will correctly rank a randomly chosen positive case (diabetes) higher than a randomly chosen negative case (no diabetes).**

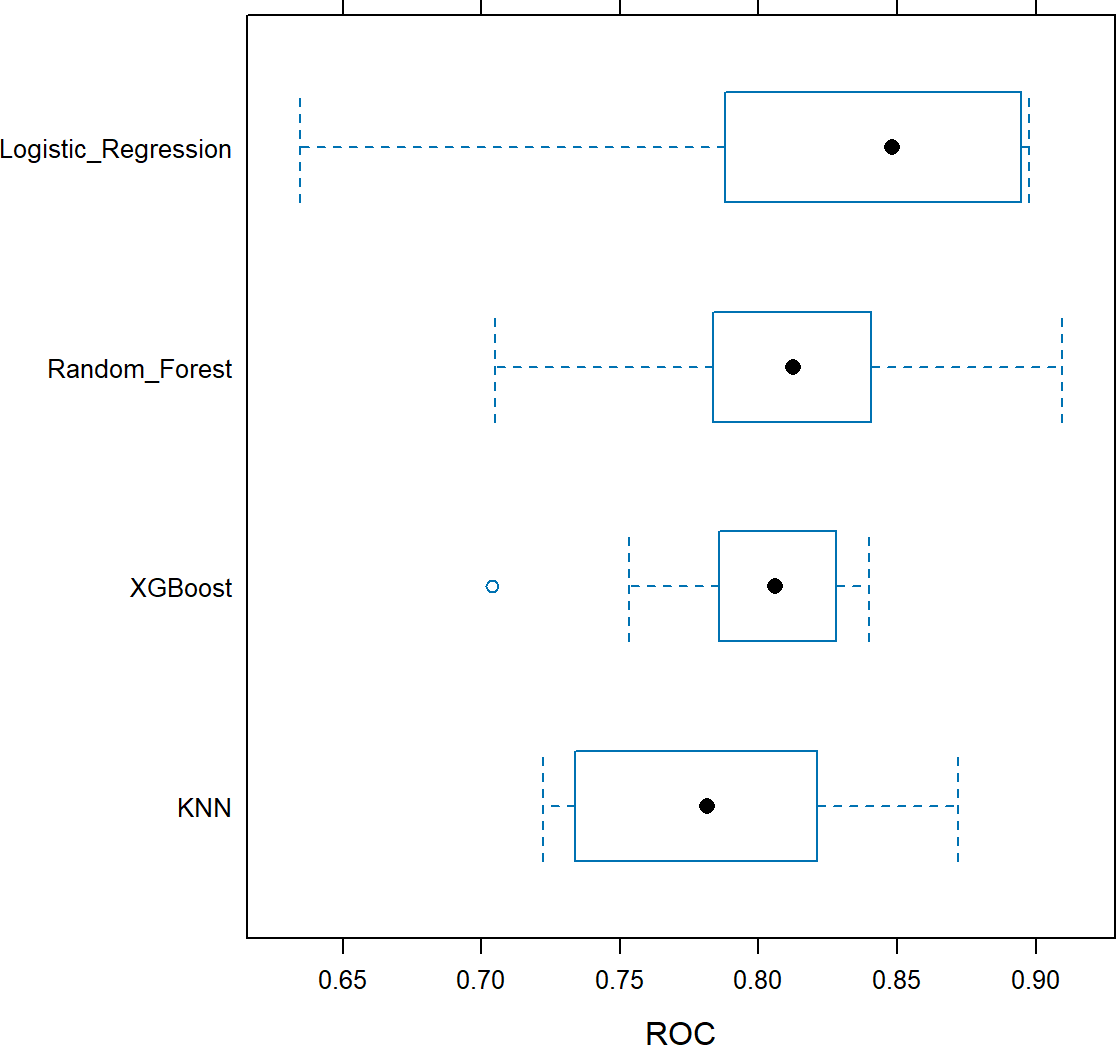
**Training Data set - Model Comparision by ROC value**

**model\_list <- list(Random\_Forest = model\_forest, XGBoost = model\_xgb, KNN = model\_knn, Logistic\_Regression = model\_glm)**

**resamples <- resamples(model\_list)**

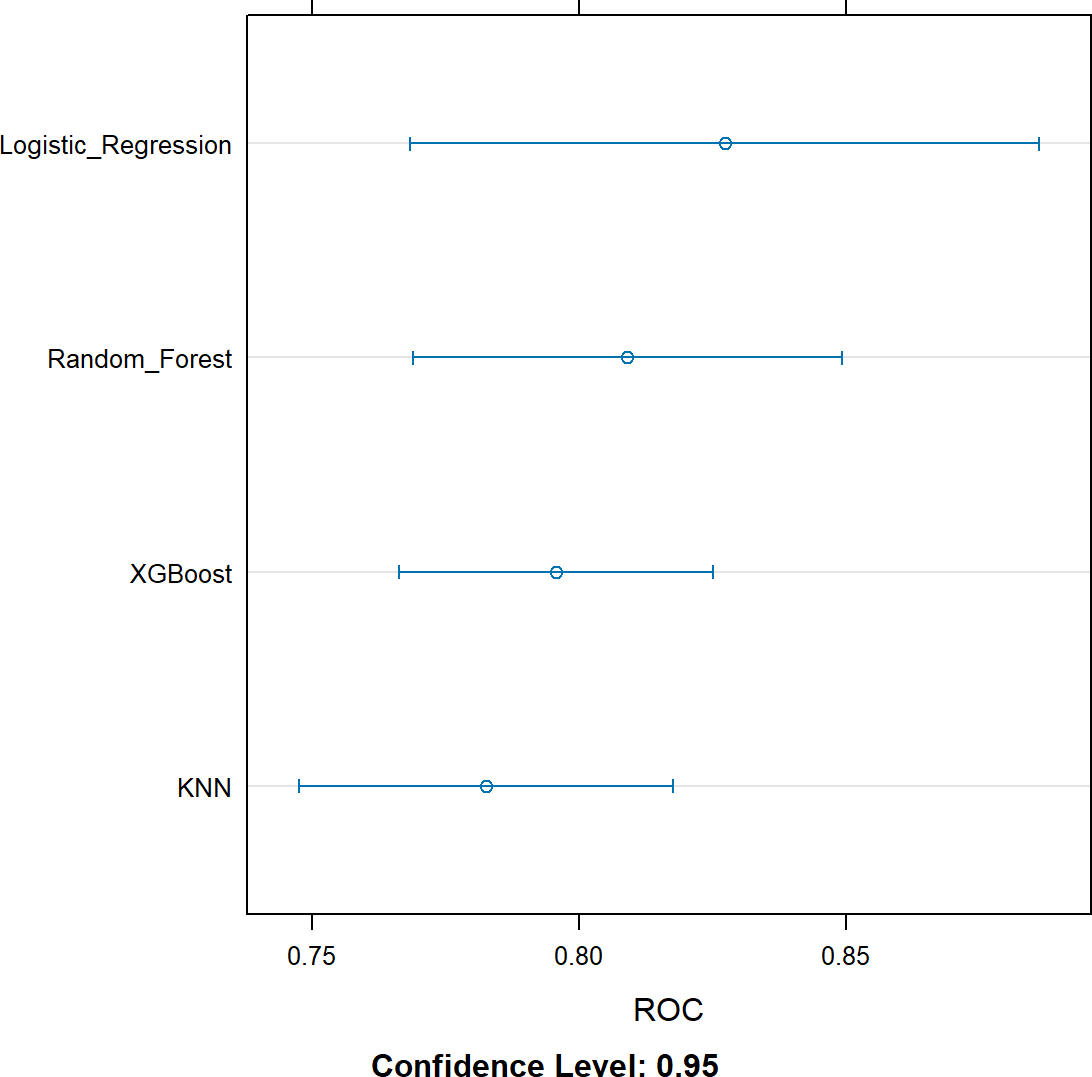
***#box plot***

bwplot(resamples, metric="ROC")



***#dot plot***

**dotplot(resamples, metric="ROC")**



**Based on ROC value, we can conclude that Logistic Regression performed best for the training data set among all the five ML models**

#### Prediction on Test Data Set

***# Random Forest prediction on Test data set***

pred\_rf <- predict(model\_forest, test\_set)

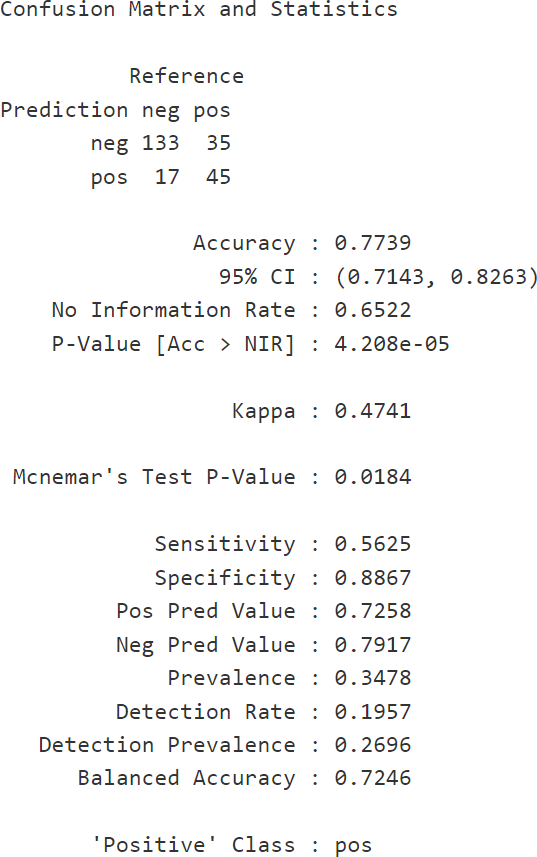
***# Confusion Matrix***

cm\_rf <- confusionMatrix(pred\_rf, test\_set$diabetes, positive="pos")

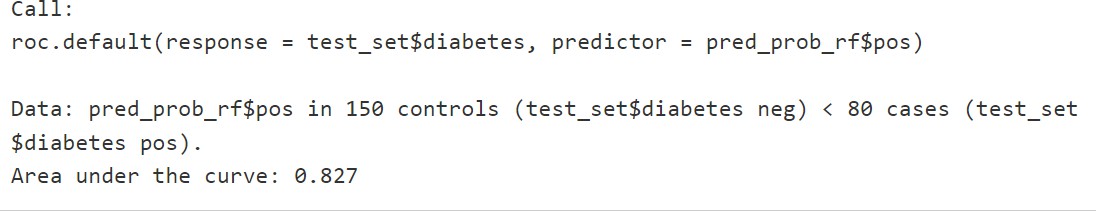
***# Prediction Probabilities***

pred\_prob\_rf <- predict(model\_forest, test\_set, type="prob"*# ROC value*

**roc\_rf <- roc(test\_set$diabetes, pred\_prob\_rf$pos) *# Confusion Matrix for Random Forest Model* cm\_rf**

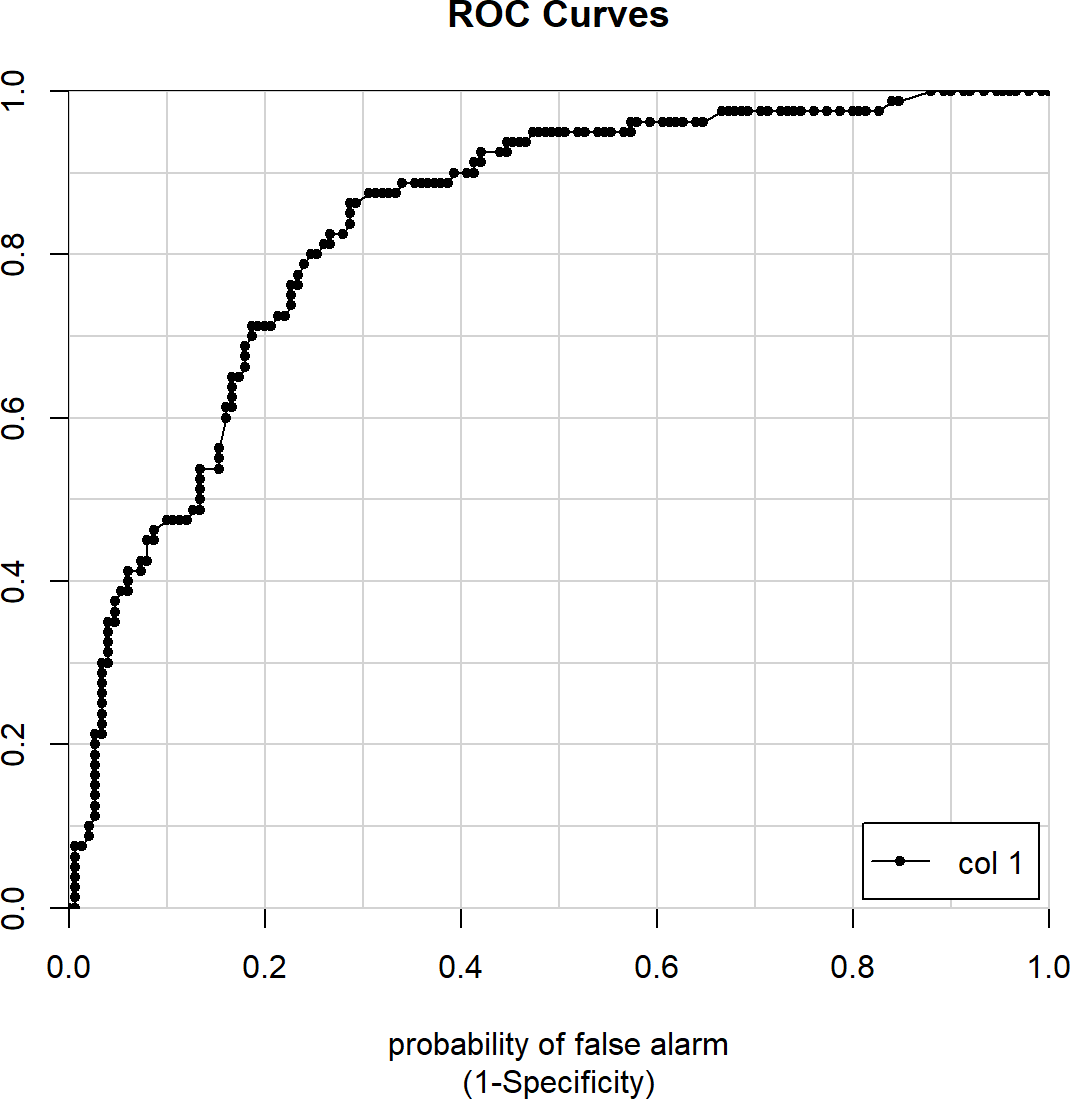
****

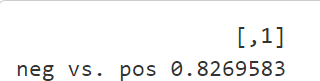
***# ROC Value for Random Forest***

******

***# AUC - Area under the curve***

caTools::colAUC(pred\_prob\_rf$pos, test\_set$diabetes, plotROC = T)

****

****

***# XGBOOST - eXtreme Gradient BOOSTing # prediction on Test data set***

pred\_xgb <- predict(model\_xgb, test\_set)

***# Confusion Matrix***

cm\_xgb <- confusionMatrix(pred\_xgb, test\_set$diabetes, positive="pos")

***# Prediction Probabilities***

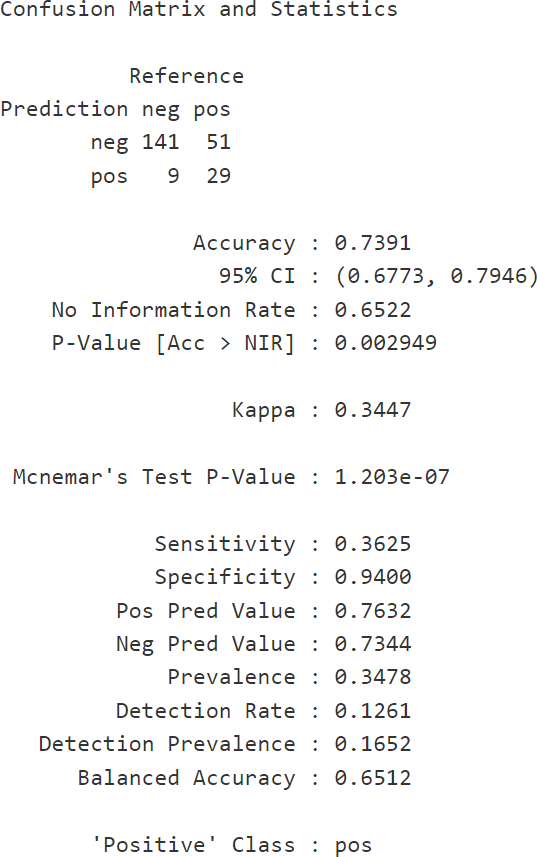
pred\_prob\_xgb <- predict(model\_xgb, test\_set, type="prob")

***# ROC value***

roc\_xgb <- roc(test\_set$diabetes, pred\_prob\_xgb$pos)

***# Confusion matrix***

cm\_xgb

****

A confusion matrix compares predictions to actual outcomes to evaluate performance:

* True Negatives (TN): 141 people correctly predicted as non-diabetic.
* False Negatives (FN): 51 diabetic people incorrectly predicted as non-diabetic.
* True Positives (TP): 29 people correctly predicted as diabetic.
* False Positives (FP): 9 non-diabetic people incorrectly predicted as diabetic.

A confusion matrix is a tool used to compare a model's predictions to the actual outcomes, providing a clear evaluation of its performance. In this case, the model correctly identified 141 non-diabetic individuals as non-diabetic (True Negatives) and 29 diabetic individuals as diabetic (True Positives). However, it incorrectly classified 51 diabetic individuals as non-diabetic (False Negatives) and 9 non-diabetic individuals as diabetic (False Positives). This breakdown highlights the model's ability to distinguish between the two groups and areas where it can improve.

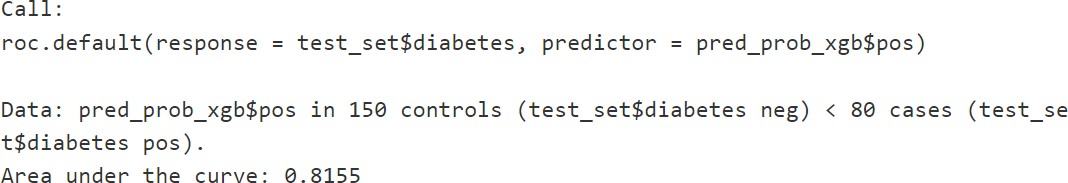
Performance Metrics

* Accuracy: 73.91% of predictions were correct.
  + Formula: (TP+TN)/Total Predictions(TP + TN) / \text{Total Predictions}(TP+TN)/Total Predictions
* Sensitivity (Recall): 36.25%
  + Measures how well the model identifies diabetic cases.
  + Formula: TP/(TP+FN)TP / (TP + FN)TP/(TP+FN)
* Specificity: 94.00%
  + Measures how well the model avoids false alarms (non-diabetics predicted as diabetics).
  + Formula: TN/(TN+FP)TN / (TN + FP)TN/(TN+FP)
* Balanced Accuracy: 65.12%
  + Combines sensitivity and specificity, giving equal weight to both.

In summary, the **sensitivity** of 36.25% means *the model is not very good at catching diabetic cases*. It does well at avoiding false alarms (high **specificity**), but it misses many positives.

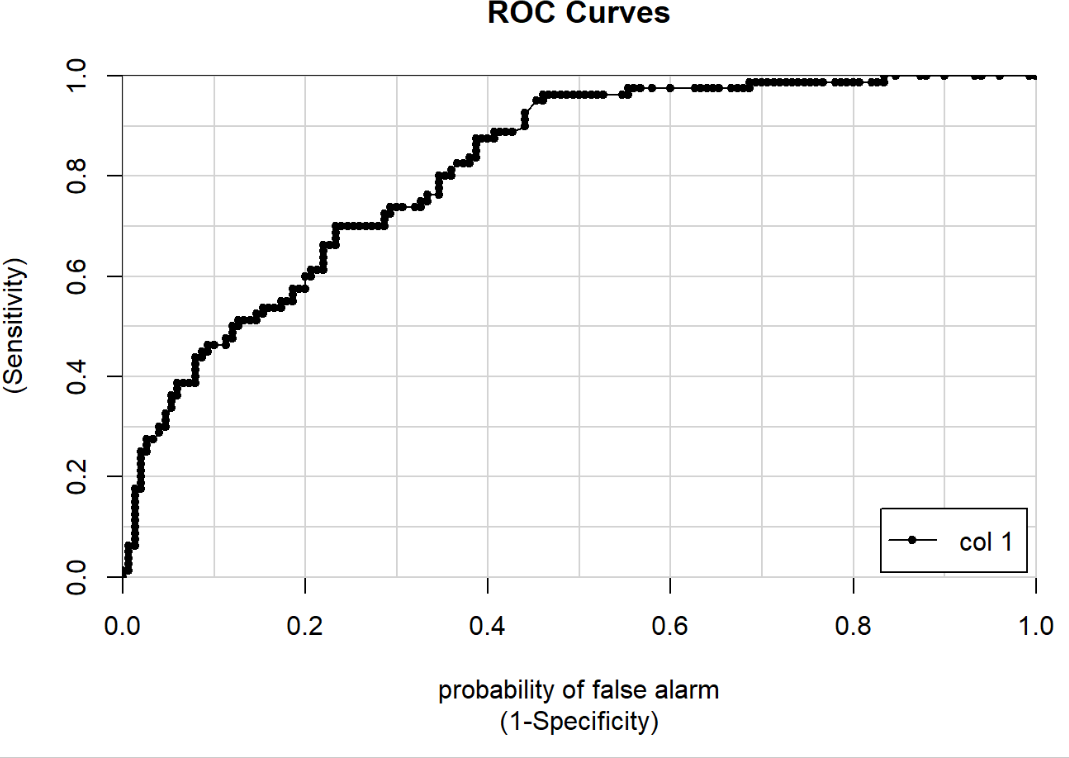
***# ROC Value for for XGBoost***

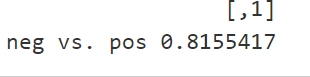
roc\_xgb

****

***# AUC - Area under the curve***

caTools::colAUC(pred\_prob\_xgb$pos, test\_set$diabetes, plotROC = T)

****

****

* **ROC Curve**: A plot showing the trade-off between sensitivity and specificity at different thresholds.
* **AUC (Area Under the Curve)**: 0.8155
  + This means the model has an **81.55% chance** of correctly distinguishing between a diabetic and non-diabetic person.

##### Insights from the Results

* 1. **Strengths**:
     + High **specificity** (94%) indicates the model is good at avoiding false alarms for non-diabetics.
     + AUC of **0.8155** shows the model performs well overall.
  2. **Weaknesses**:
     + Low **sensitivity** (36.25%) means the model misses many diabetic cases.
     + This could indicate an imbalance in the data or a need for hyperparameter tuning.

***# KNN - K Nearest Neighbours # prediction on Test data set***

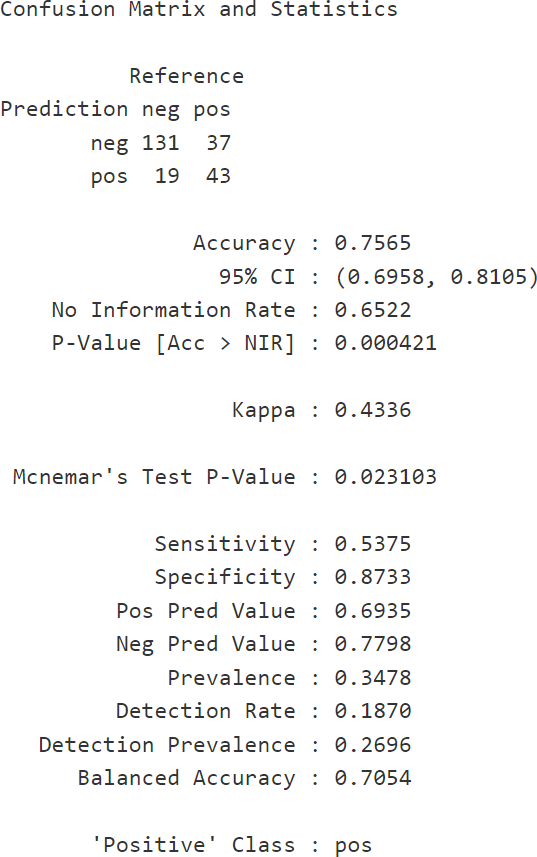
***pred\_knn <- predict(model\_knn, test\_set) # Confusion Matrix***

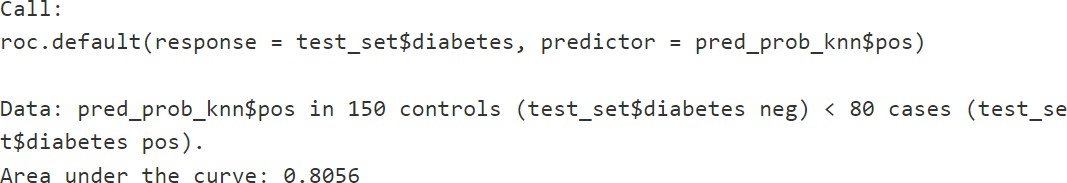
***cm\_knn <- confusionMatrix(pred\_knn, test\_set$diabetes, positive="pos") # Prediction Probabilities***

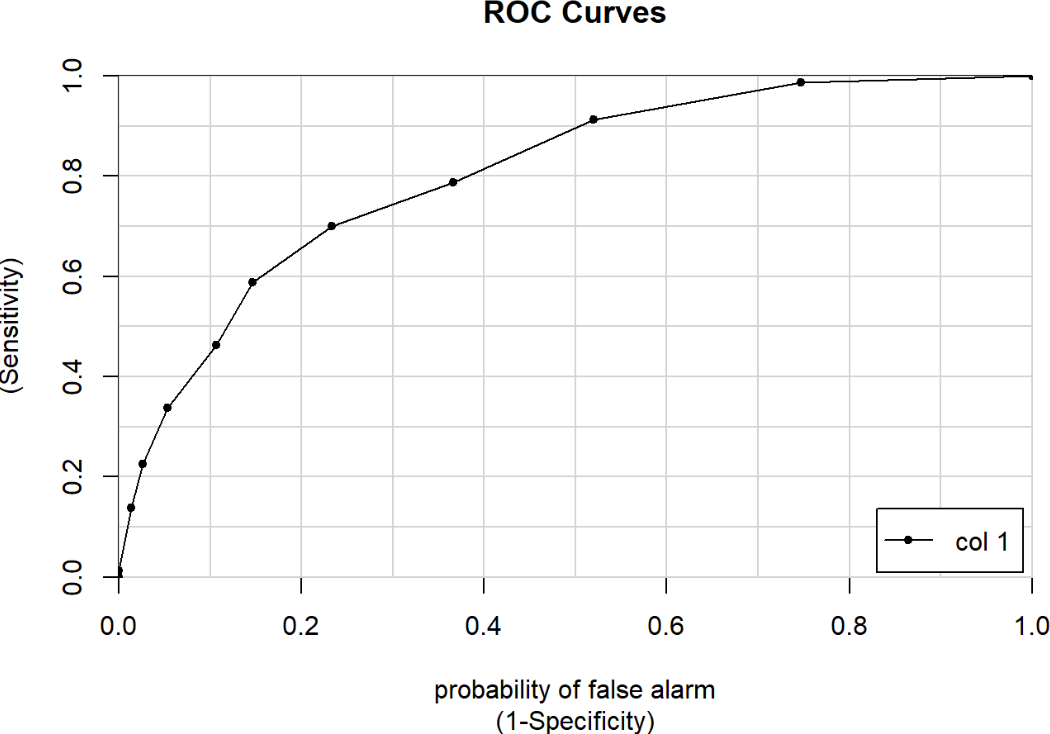
***pred\_prob\_knn <- predict(model\_knn, test\_set, type="prob") # ROC value***

***roc\_knn <- roc(test\_set$diabetes, pred\_prob\_knn$pos) # Confusion matrix***

***cm\_knn***



******



******

***# Logistic Regression***

***# prediction on Test data set***

pred\_glm <- predict(model\_glm, test\_set)

***# Confusion Matrix***

cm\_glm <- confusionMatrix(pred\_glm, test\_set$diabetes, positive="pos")

***# Prediction Probabilities***

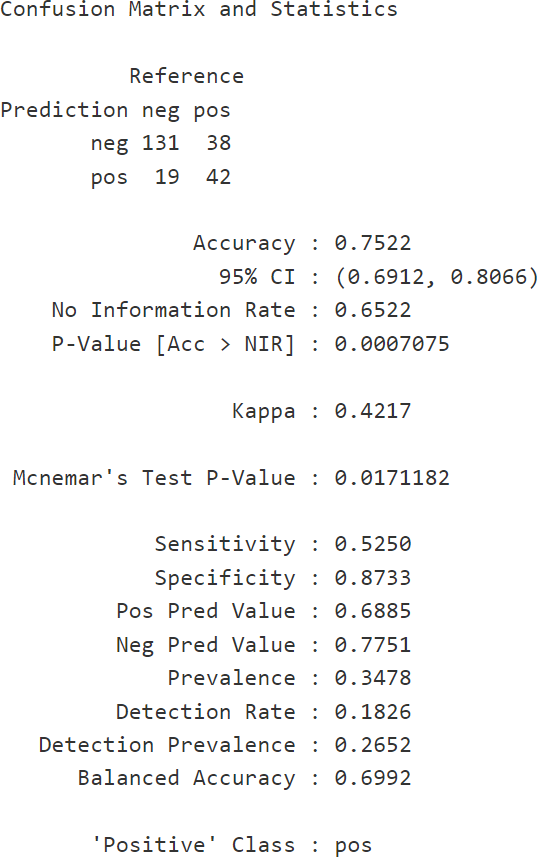
pred\_prob\_glm <- predict(model\_glm, test\_set, type="prob")

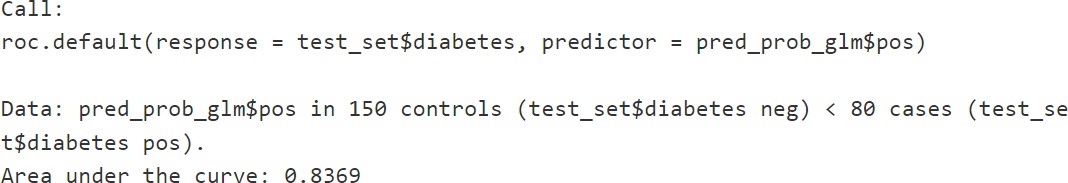
***# ROC value***

roc\_glm <- roc(test\_set$diabetes, pred\_prob\_glm$pos)

***# Confusion matrix***

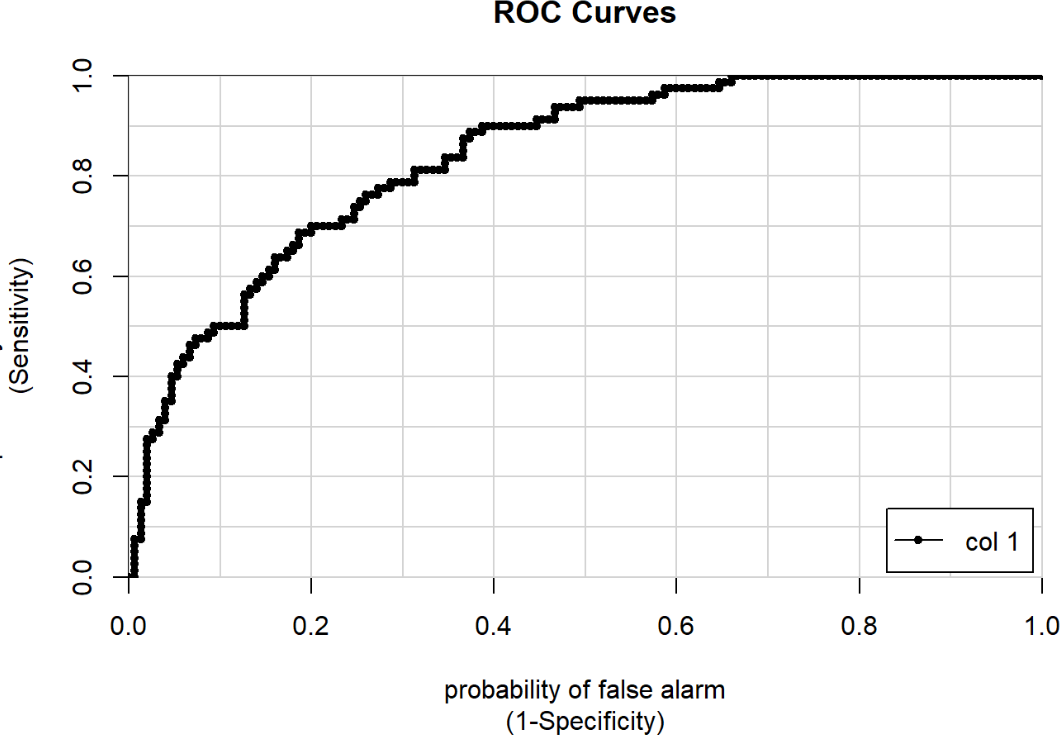
Cm\_glm



****

*# AUC - Area under the curve*

caTools::colAUC(pred\_prob\_glm$pos, test\_set$diabetes, plotROC = T)



Results

result\_rf <- c(cm\_rf$byClass['Sensitivity'], cm\_rf$byClass['Specificity'], cm\_rf$byClass['Precision'],

cm\_rf$byClass['Recall'], cm\_rf$byClass['F1'], roc\_rf$auc)

result\_xgb <- c(cm\_xgb$byClass['Sensitivity'], cm\_xgb$byClass['Specificity'], cm\_xgb$byClass['Precision'],

cm\_xgb$byClass['Recall'], cm\_xgb$byClass['F1'], roc\_xgb$auc)

result\_knn <- c(cm\_knn$byClass['Sensitivity'], cm\_knn$byClass['Specificity'], cm\_knn$byClass['Precision'],

cm\_knn$byClass['Recall'], cm\_knn$byClass['F1'], roc\_knn$auc)

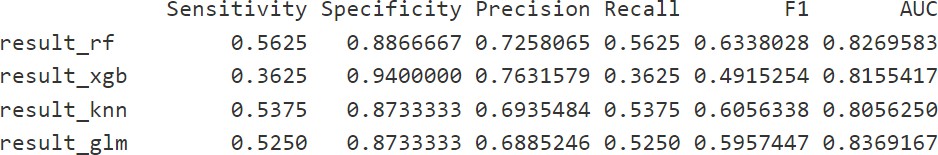
result\_glm <- c(cm\_glm$byClass['Sensitivity'], cm\_glm$byClass['Specificity'], cm\_glm$byClass['Precision'],

cm\_glm$byClass['Recall'], cm\_glm$byClass['F1'], roc\_glm$auc)

result\_rpart <- c(cm\_rpart$byClass['Sensitivity'], cm\_rpart$byClass['Specificity'], cm\_rpart$byClass['Precision'],

cm\_rpart$byClass['Recall'], cm\_rpart$byClass['F1'], roc\_rpart$auc)

all\_results <- data.frame(rbind(result\_rf, result\_xgb, result\_knn, result\_glm, result\_rpart)) names(all\_results) <- c("Sensitivity", "Specificity", "Precision", "Recall", "F1", "AUC") All\_results



* Logistic Regression model looks best for the test data set as well based on Sensitivity, Precision, Recall, and F1 score.
* The AUC value and Specificity for Logistic Regression is also pretty good and is at second position among all the models built and tested.

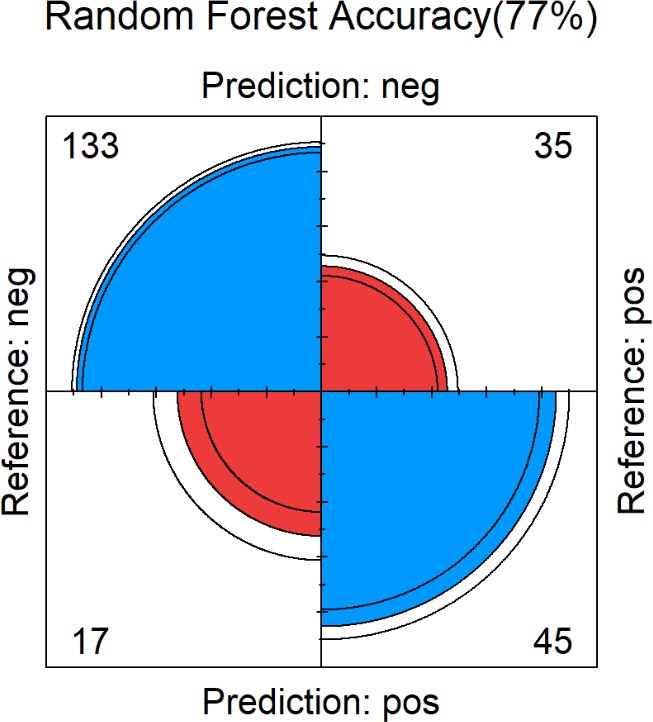
**Visualization to compare accuracy of ML models**

col <- c("#ed3b3b", "#0099ff")

graphics::fourfoldplot(cm\_rf$table, color = col, conf.level = 0.95, margin = 1,

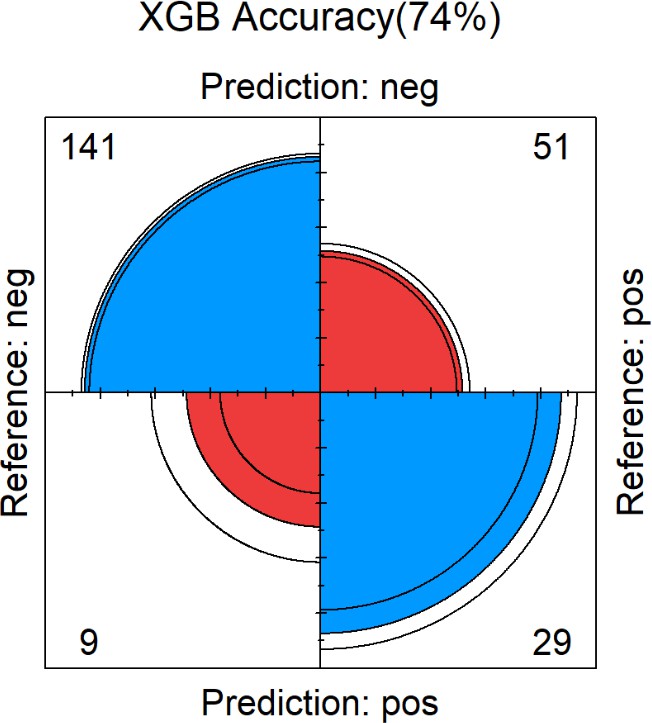
main = paste("Random Forest Accuracy(",round(cm\_rf$overall[1]\*100),"%)", sep

= ""))



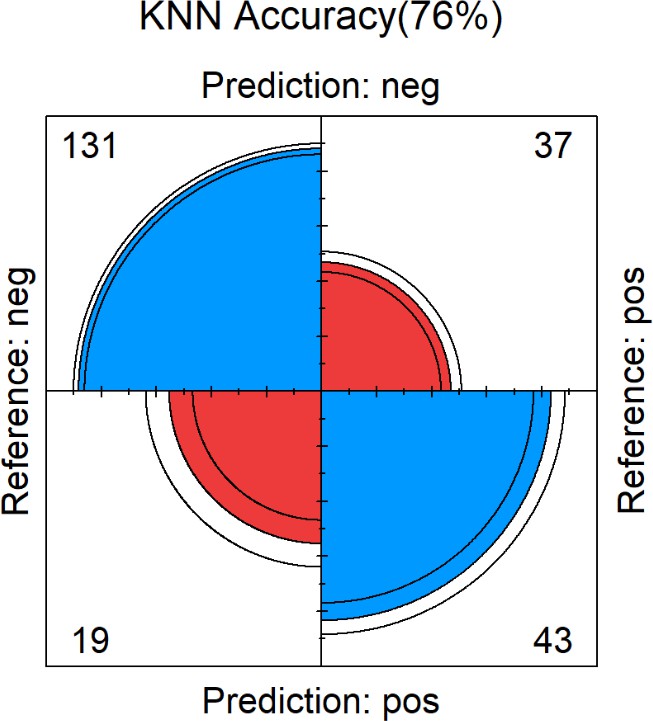
graphics::fourfoldplot(cm\_xgb$table, color = col, conf.level = 0.95, margin = 1,

main = paste("XGB Accuracy(",round(cm\_xgb$overall[1]\*100),"%)", sep = ""))



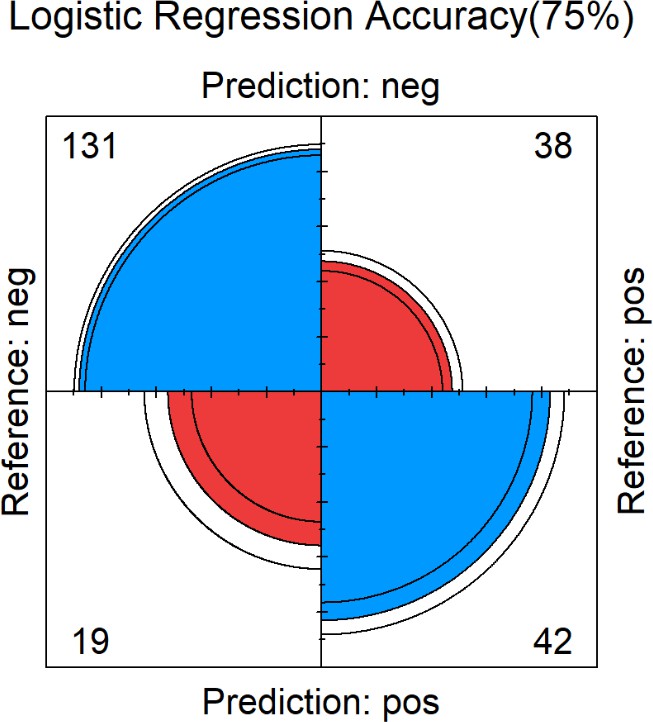
graphics::fourfoldplot(cm\_knn$table, color = col, conf.level = 0.95, margin = 1,

main = paste("KNN Accuracy(",round(cm\_knn$overall[1]\*100),"%)", sep = ""))



graphics::fourfoldplot(cm\_glm$table, color = col, conf.level = 0.95, margin = 1, main = paste("Logistic Regression

Accuracy(",round(cm\_glm$overall[1]\*100),"%)", sep = ""))



From the above plot we can confirm that Logistic Regression model performed best, based on overall Accuracy of the model also.

Conclusion

We have deployed multiple Machine Learning Models such as - Random Forest, eXtreme Gradient Boosting Machine, K-Nearest Neighbours and Logistic Regression, found that Logistic Regression has the best performance based on ROC value.

As a next step, we could go further and try to compare other ML models as well.

Glossary

*Recall:* Recall gives us an idea about when **it’s actually Yes, how often does it predict Yes.**

*Precision:* Precision tells us about when it **predicts Yes, how often is it correct.**

*Sensitivity:* Ability of the test to correctly identify the **true positive rate.**

*Specificity:* Ability of the test to correctly identify the **true negative rate.**

Abbreviations used

*ROC:* Reciever Operating Characteristics curve

*AUC:* Area under the ROC curve

**AUC** denotes the rate of successful classification by the logistic mod

## **4.Conclusions**

This project aimed to predict diabetes using the Pima Indians Diabetes Dataset, leveraging various machine learning models and statistical techniques. The analysis was grounded in several hypotheses, each of which was validated through model results and feature insights.

Key Findings:

Hypothesis 1 (Glucose and Diabetes): Glucose levels emerged as the strongest predictor of diabetes, with diabetic individuals showing significantly higher glucose concentrations. This aligns with clinical diagnostic criteria and underscores the importance of monitoring glucose for early detection.

Hypothesis 2 (BMI and Diabetes): BMI demonstrated a strong positive correlation with diabetes risk, reinforcing the role of obesity as a major contributing factor to insulin resistance and type 2 diabetes.

Hypothesis 3 (Age and Diabetes): Age was a significant predictor, with older individuals more likely to be diabetic. This finding highlights the cumulative effects of metabolic changes and lifestyle factors over time.

Hypothesis 4 (Insulin Levels and Diabetes): Insulin levels showed high variability, with diabetic individuals often exhibiting extreme values, confirming its role as a key indicator of abnormal glucose regulation.

Hypothesis 5 (Blood Pressure and Diabetes): Blood pressure showed a weaker association, with overlapping distributions between diabetic and non-diabetic groups. However, elevated values may still provide complementary insights when combined with stronger predictors.

Model Comparison: Among the models tested—Random Forest, XGBoost, KNN, and Logistic Regression—Logistic Regression consistently outperformed others, achieving the highest ROC-AUC and sensitivity. This result emphasizes the effectiveness of simpler, interpretable models when features are carefully preprocessed and selected. Ensemble models like Random Forest and XGBoost also performed well but struggled with imbalanced specificity.

Broader Implications: The findings validate the hypotheses and provide actionable insights for healthcare:

Clinical Applications: Glucose, BMI, and age should be prioritized in diagnostic and screening tools for diabetes.

Healthcare Optimization: Predictive models can guide resource allocation, enabling targeted interventions for high-risk individuals, reducing unnecessary diagnostics, and mitigating long-term healthcare costs.

Model Utility: Logistic Regression offers interpretable insights, making it suitable for clinical decision-making, while advanced ensemble methods can enhance accuracy in large-scale or complex datasets.

This analysis employed a range of machine learning models, including Random Forest, XGBoost, K-Nearest Neighbors (KNN), and Logistic Regression, to predict diabetes in the Pima Indians Diabetes dataset. Each model was evaluated based on its ability to classify diabetic and non-diabetic individuals, with metrics such as ROC-AUC, sensitivity, specificity, and precision guiding the comparison. Among these, Logistic Regression demonstrated the most balanced performance, achieving the highest overall ROC-AUC and strong sensitivity, making it particularly effective at correctly identifying positive cases. This indicates its robustness in handling this dataset, despite being a simpler model compared to ensemble methods like Random Forest or XGBoost.

The exploratory data analysis provided critical insights into feature behavior. Variables like glucose and insulin levels emerged as the strongest predictors, with diabetic individuals exhibiting significantly higher values. Similarly, BMI and age showed clear trends, reinforcing their established roles in diabetes risk. In contrast, variables such as blood pressure and triceps skinfold thickness displayed substantial overlap between diabetic and non-diabetic groups, reducing their standalone predictive value. High skewness in features like insulin, pedigree, and age suggests the need for preprocessing steps, such as normalization or transformations, to improve model stability.

The inclusion of multiple models revealed unique strengths and weaknesses. For instance, Random Forest and XGBoost provided decent performance but struggled with imbalanced data, particularly in terms of specificity. KNN, while intuitive, showed limited performance, indicating that simple proximity-based algorithms may not capture the complexity of this dataset. Logistic Regression excelled due to its ability to model linear relationships effectively, proving that simpler models can outperform complex ones when feature engineering and selection are optimized.

This study highlights the importance of data-driven feature selection, preprocessing, and careful model evaluation in predictive modeling. Moving forward, ensemble approaches that combine the strengths of multiple models could be explored to further enhance predictive accuracy and robustness. Additionally, techniques like hyperparameter tuning, cost-sensitive learning, or balancing class distributions may address weaknesses in specificity without sacrificing sensitivity. Ultimately, this analysis underscores the critical role of glucose, insulin, BMI, and age in identifying diabetes, providing actionable insights for both machine learning practitioners and healthcare professionals.

## **5.References**

* 1. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.
     + Introduces the Random Forest algorithm, a robust ensemble method used for predictive modeling, including healthcare applications like diabetes prediction.
  2. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 785–794.
     + Details the XGBoost algorithm, a high-performance machine learning method widely applied in predictive analytics.
  3. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer.
     + A comprehensive resource on statistical learning techniques applicable to data mining tasks, including diabetes prediction.
  4. Kaggle. "Pima Indians Diabetes Database."
     + Offers the Pima Indians Diabetes dataset, used to predict diabetes risk based on health metrics. Available at:

[https://www.kaggle.com/uciml/pima-indians-diabetes-database.](http://www.kaggle.com/uciml/pima-indians-diabetes-database)

* 1. Pedregosa, Fabian, et al. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research,* vol. 12, 2011, pp. 2825–2830.
     + Used for the implementation of machine learning models in Python and their performance evaluation in this project.
  2. World Health Organization. Global Report on Diabetes. World Health Organization, 2016.
     + Highlighted the global burden of diabetes and the importance of early detection, framing the real-world relevance of this project.
  3. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine.

*The Annals of Statistics, 29*(5), 1189–1232.

* + - Explains the theoretical underpinnings of gradient boosting, forming the basis for advanced models like XGBoost.
  1. Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* Springer.
     + A foundational resource on machine learning algorithms, including those used for diabetes prediction.
  2. American Diabetes Association. Standards of Medical Care in Diabetes—2023.
     + Provides guidelines on diabetes prevention, diagnosis, and management, offering context for the importance of early prediction and intervention.