Probability and Statistics

### Predicting Diabetes in Pima Indians

### In this project, we will design and develop a supervised learning model for predicting whether a patient has diabetes or not. Diabetes is a disease that occurs when blood glucose, also called blood sugar, is too high. Blood glucose is the main source of energy and comes from the food we eat. Insulin is a hormone made by the pancreas that allows our body to use sugar (glucose) from carbohydrates in the food that we eat for energy or to store glucose for future use. Insulin helps keeps our blood sugar level from getting too high (hyperglycemia) or too low (hypoglycemia). When our body does not produce enough insulin to convert the glucose, the excess glucose then stays in our blood and doesn’t reach our cells. Over time, having too much glucose in our blood can cause health problems. Although diabetes has no cure, we can take steps to manage our diabetes and stay healthy. In this project, we will develop a machine learning model to predict whether a patient has diabetes or not based on different characteristics such as Blood Pressure, Skin Thickness, Insulin Level, BMI, Plasma Glucose Level and so on. The data set for this project is taken from the National Institute of Diabetes and Digestive and Kidney Diseases. All patients in this data set are females at least 21 years old of Pima Indian heritage. The Pima Indians (or Akimel Oʼodham, also spelled Akimel Oʼotham, "River People", formerly known as Pima) are a group of Native Americans living in an area consisting of what is now central and southern Arizona. Diabetes is an increasingly prevalent chronic disease characterized by the body’s inability to metabolize glucose. Finding the disease at an early stage helps reduce medical costs and the risk of patients having more complicated health problems. This modeling project will help detect diabetes at early stages and will help us take effective action and to stay healthy.

### References

1. <https://www.kaggle.com/uciml/pima-indians-diabetes-database>
2. <https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes>
3. <https://www.endocrineweb.com/conditions/type-1-diabetes/what-insulin>
4. <https://en.wikipedia.org/wiki/Pima_people>
5. <https://bmcendocrdisord.biomedcentral.com/articles/10.1186/s12902-019-0436-6>

**Data Summary**

# Show the data structure, There are 768 observations of 9 variables  
str(diabetes)

'data.frame': 768 obs. of 9 variables:

$ Pregnancies : int 6 1 8 1 0 5 3 10 2 8 ...

$ Glucose : int 148 85 183 89 137 116 78 115 197 125 ...

$ BloodPressure : int 72 66 64 66 40 74 50 0 70 96 ...

$ SkinThickness : int 35 29 0 23 35 0 32 0 45 0 ...

$ Insulin : int 0 0 0 94 168 0 88 0 543 0 ...

$ BMI : num 33.6 26.6 23.3 28.1 43.1 25.6 31 35.3 30.5 0 ...

$ DiabetesPedigreeFunction: num 0.627 0.351 0.672 0.167 2.288 ...

$ Age : int 50 31 32 21 33 30 26 29 53 54 ...

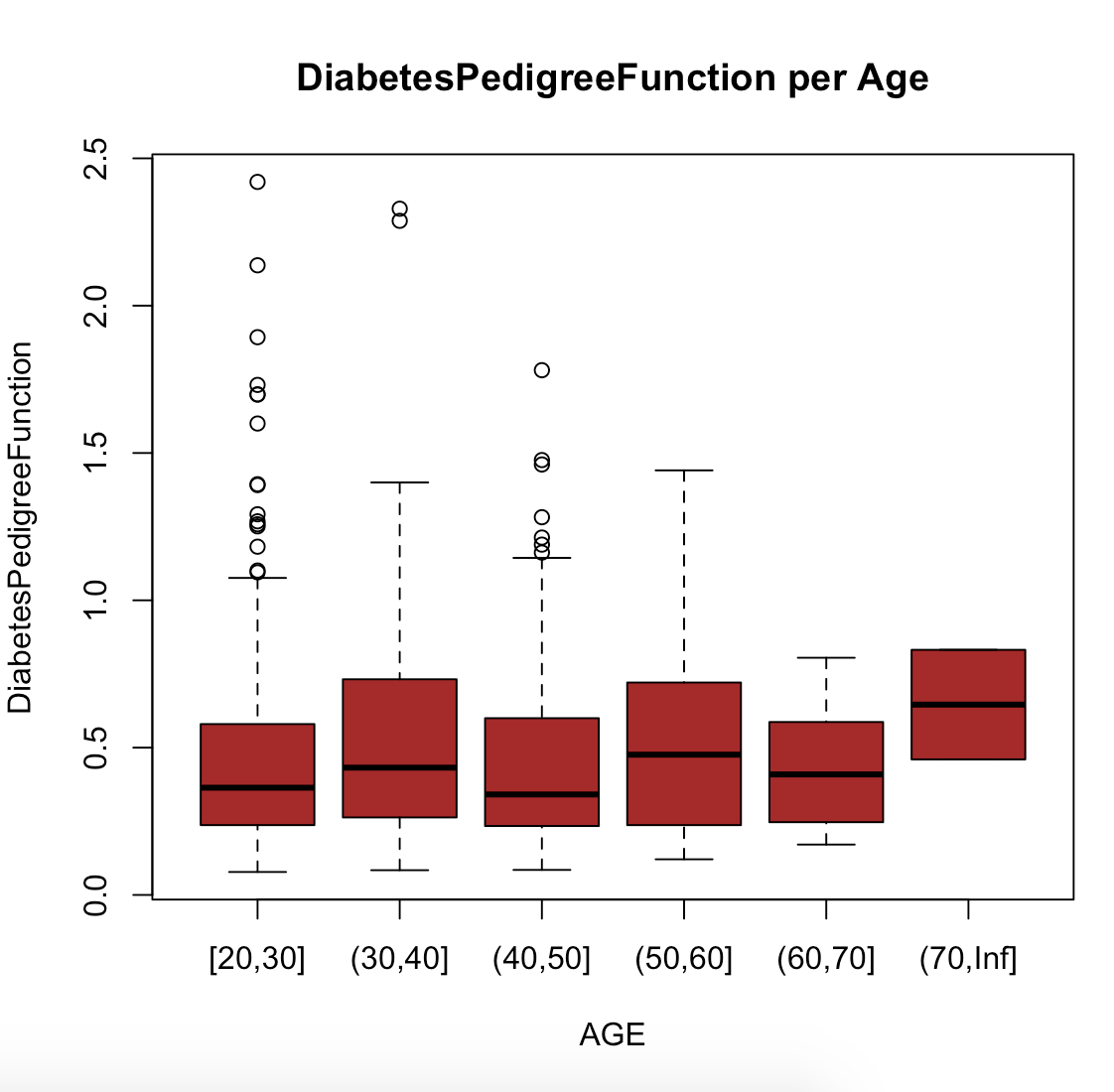
$ Outcome : int 1 0 1 0 1 0 1 0 1 1 …

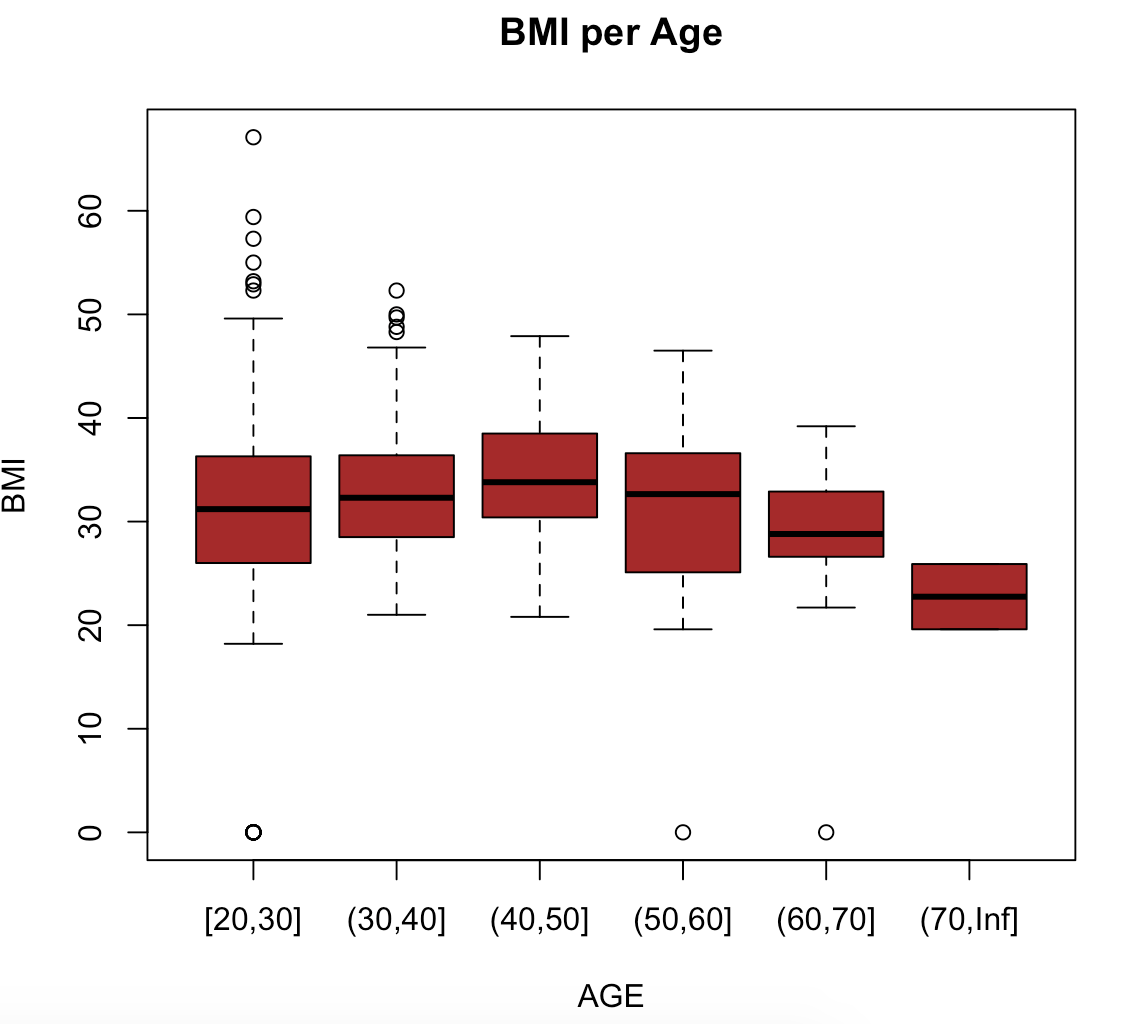
# Show some statistics about our data

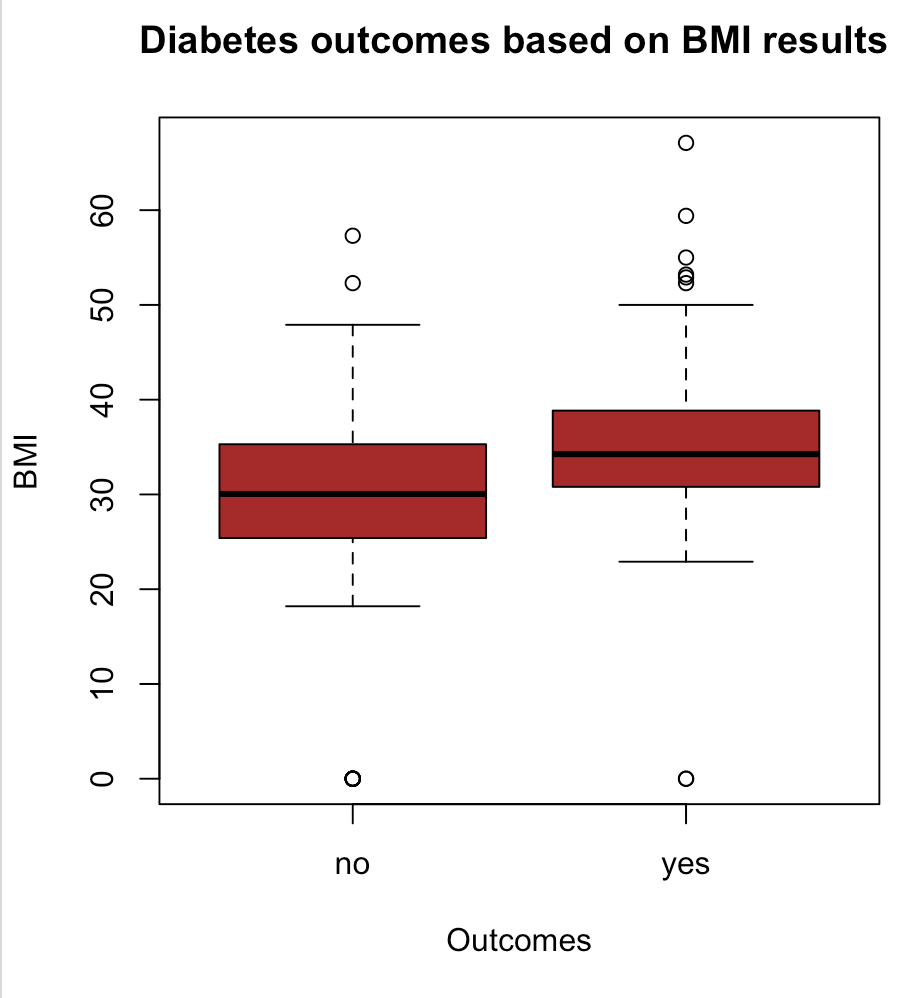
summary(diabetes)

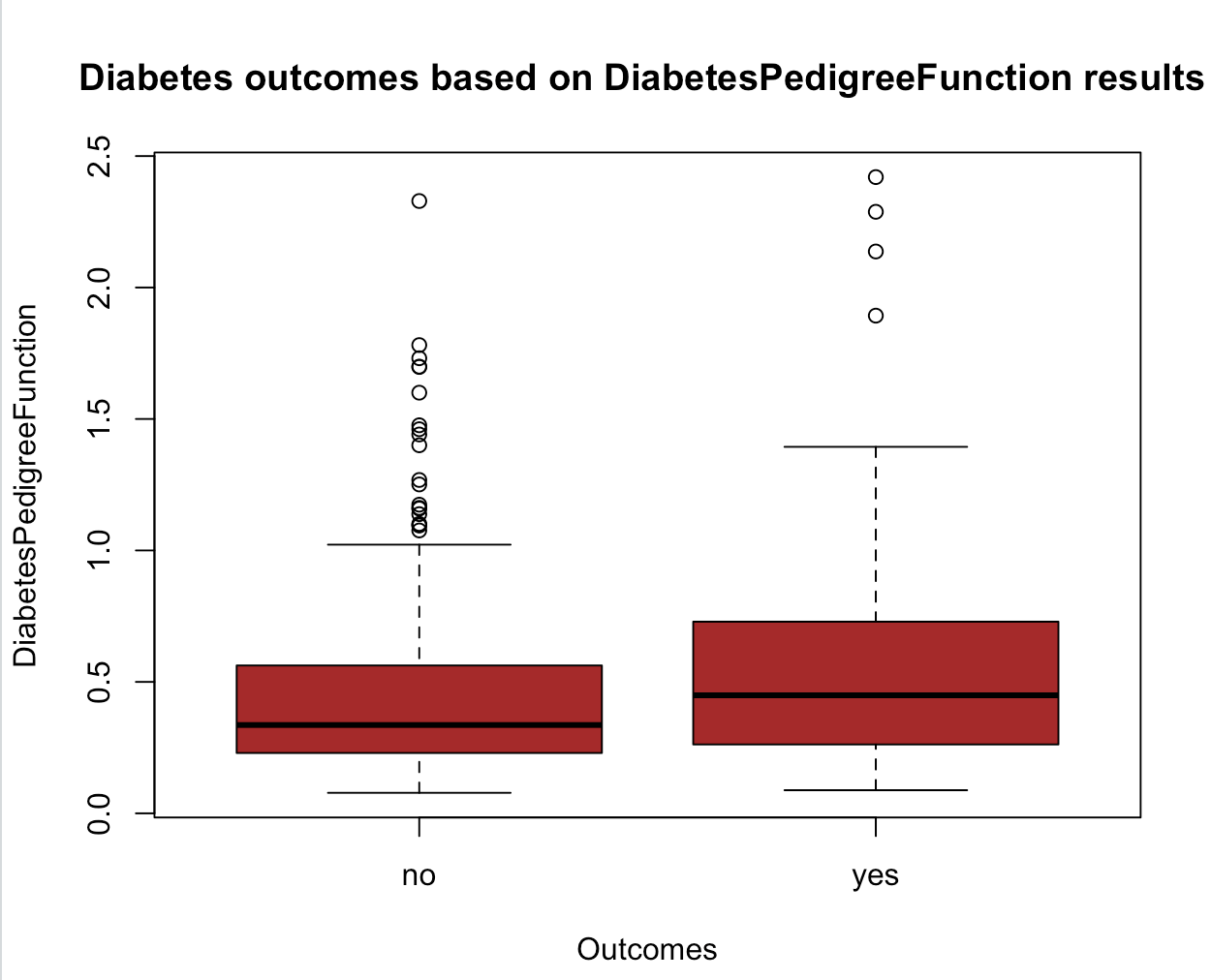
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFunction | Age |
| Min. | 0.000 | 0.0 | 0.00 | 0.00 | 0.0 | 0.00 | 0.0780 | 0.0780 |
| 1st Qu. | 1.000 | 99.0 | 62.00 | 0.00 | 0.0 | 27.30 | 0.2437 | 0.2437 |
| Median | 3.000 | 117.0 | 72.00 | 23.00 | 30.5 | 32.00 | 0.3725 | 0.3725 |
| Mean | 3.845 | 120.9 | 69.11 | 20.54 | 79.8 | 31.99 | 0.4719 | 0.4719 |
| 3rd Qu. | 6.000 | 140.2 | 80.00 | 32.00 | 127.2 | 36.60 | 0.6262 | 0.6262 |
| Max. | 17.000 | 199.0 | 122.00 | 99.00 | 846.0 | 67.10 | 2.4200 | 2.4200 |

**Insights**









**Logistic Model**

# index\_train using 1/3 of the data

index\_train<-sample(768,256)

# Create the TRAINING SET

training\_set <- data[index\_train, ]

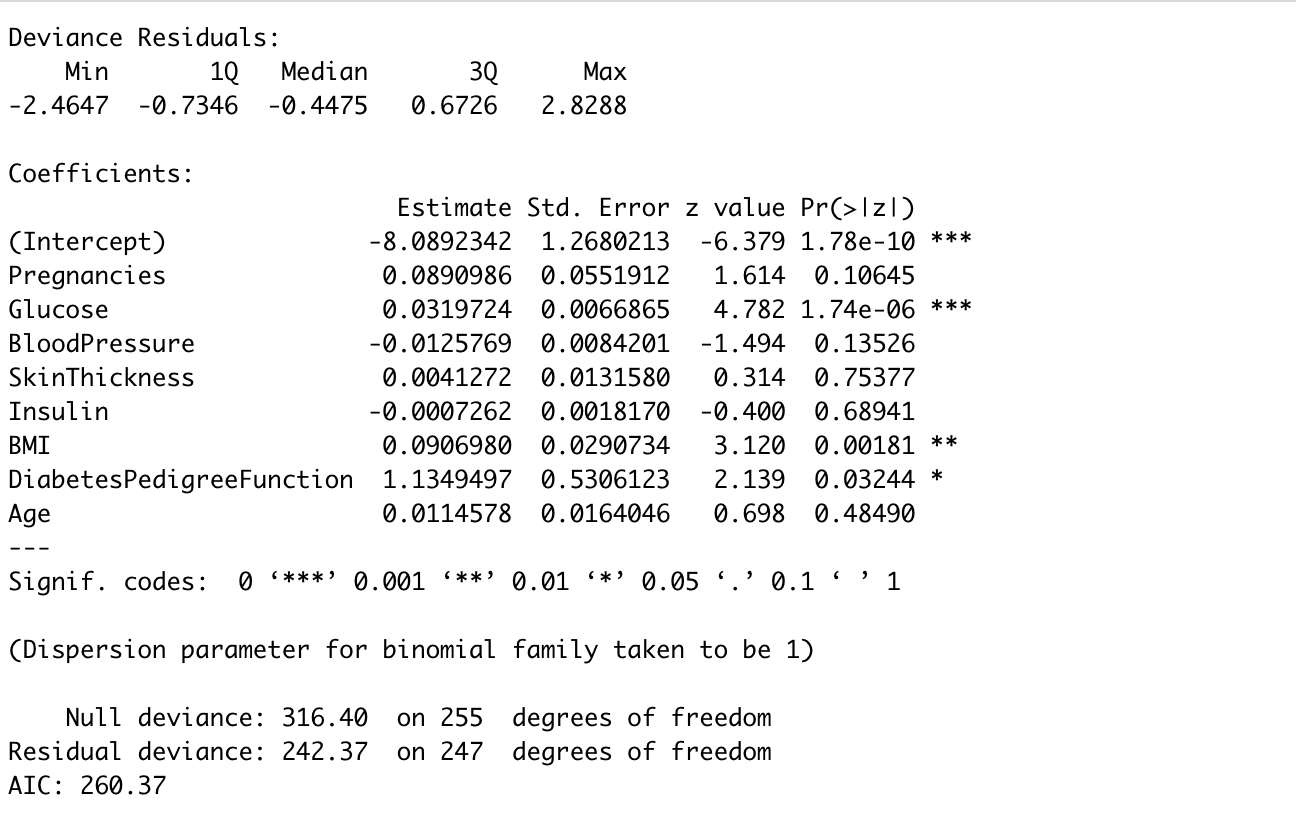
# Create the TEST SET

test\_set <- data[-index\_train, ]

# Fit the Logistic Regression Model using the training\_set

logistic\_model <- glm(Outcome ~ ., family = "binomial", data = training\_set)

# Summary of our logistic model



# Based on the p-value The statistically significant variables: Pregnancies, Glucose, BMI, DiabetesPedigreeFunction

# not statistically significant variables: BloodPressure, SkinThickness, Insulin, and Age are

**Predictions using the Logistic Model**

# predictions using logistic regression model for the test set

predictions\_logistic <- predict(logistic\_model, newdata = test\_set, type = "response")

# wide range means the model is good for predicting diabetes

print(range(predictions\_logistic))  
[1] 0.00473164 0.97076288

#Predict diabetes based on Insulin level

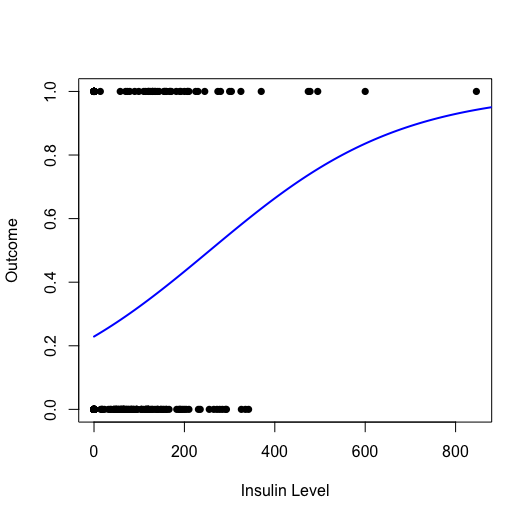
xInsulin <-seq (0, 1000, 10)

Insulin\_logistic\_model <- glm(Outcome ~ Insulin, family = "binomial", data = training\_set)

Insulin\_predictions\_logistic <- predict(Insulin\_logistic\_model, list(Insulin=xInsulin),type="response")

plot(training\_set$Insulin, training\_set$Outcome, pch = 16, xlab = "Insulin Level", ylab = "Outcome")

lines(xInsulin, Insulin\_predictions\_logistic, col = "blue", lwd = 2)

****

**Decision Tree**

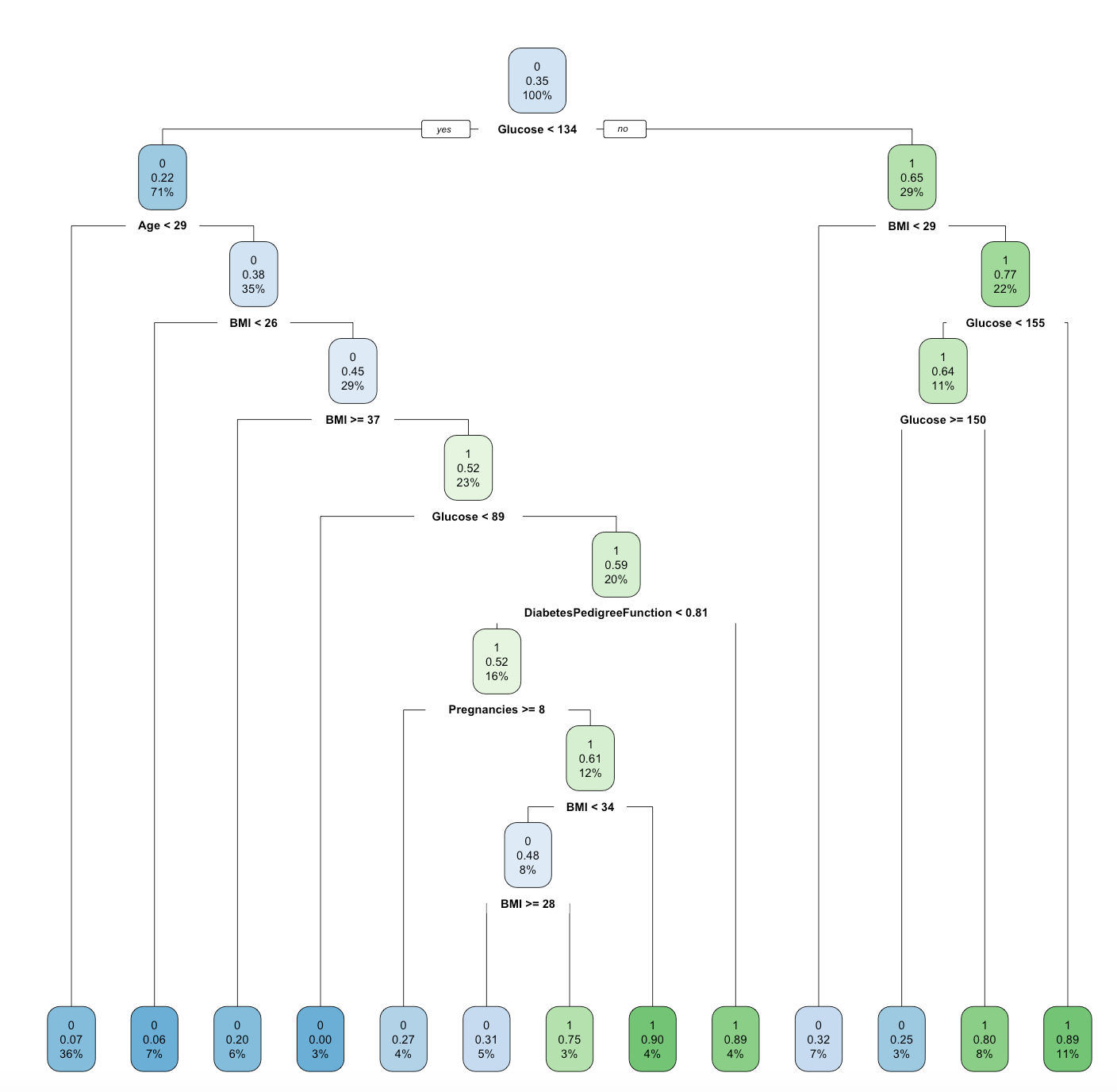
index\_train<-sample(768,256)

training\_set <- data[index\_train, ]

test\_set <- data[-index\_train, ]

model\_dt<-rpart(Outcome~., data=training\_set, method = 'class')

rpart.plot(model\_dt)



**Confusion Matrix**

# Make binary predictions for the original tree using the test set

predictions\_tree <- predict(model\_dt, newdata = test\_set, type = "class")

# Construct confusion matrices using the predictions.

confmatrix\_tree <- table(test\_set$Outcome, predictions\_tree)

|  |  |  |
| --- | --- | --- |
|  | pred | |
|  | 0 | 1 |
| no | 281 | 52 |
| yes | 82 | 97 |

# Calculate for the decision tree accuracy 73.8%

accuracy\_tree <- sum(diag(confmatrix\_tree))/sum(confmatrix\_tree)

[1] 0.7382812