## STATISTICAL ANALYSIS OPEN ENDED PROJECT REPORT

**ON**

# “DIABETES ANALYSIS USING PIMA INDIAN’S DATASET”

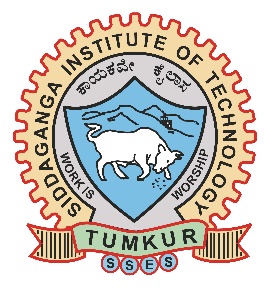
submitted in the partial fulfillment of the requirements for V Semester, Bachelor of Engineering in

**Artificial Intelligence and Data science**

**BY**

**HARIPRASAD N 1SI20AD008**

**MOHITH K G 1SI20AD011**

****

# DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

# SIDDAGANGA INSTITUTE OF TECHNOLOGY

# (An Autonomous institution affiliated to Visvesvaraya Technological University, Belagavi,

# Approved by AICTE, New Delhi, Accredited by NAAC with ‘A Grade, Awarded Diamond

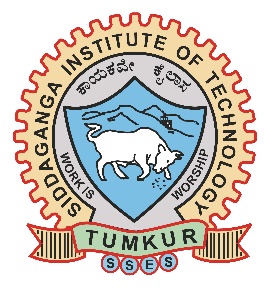
# College Rating by QS I-GAUGE and ISO 9001:2015 certified)

## TUMAKURU-572103

**2022-2023**

**SIDDAGANGA INSTITUTE OF TECHNOLOGY,**TUMAKURU-3

**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

****

## CERTIFICATE

This is to certify that open ended project on **“ DIABETES ANALYSIS USING PIMA INDIAN’S DATASET”** is a Bonafide work carried out by **HARIPRASAD N** (**1SI20AD008), MOHITH K G (1SI20AD011)** of V semester **Bachelor of Engineering in Artificial Intelligence and Data Science** of the **SIDDAGANGA INSTITUTE OF TECHNOLOGY** (An Autonomous Institution, affiliated to VTU, Belagavi & Recognized by AICTE and Accredited by NBA, New Delhi) during the academic year 2022-2023.

**Name of the Teacher Signature with Date**

**Name Signature**

**Dr. Sumalatha Aradhya**

.

**ABSTRACT**

Diabetes is a chronic medical condition that affects millions of people worldwide. The Pima Indian community in Arizona is one of the populations that has a high incidence of diabetes. The aim of this project is to use the Pima Indian Diabetes dataset to analyze the risk factors associated with diabetes in this population. The dataset contains medical records of few patients and includes various parameters

.

The project will employ data and statistical analysis techniques to gain insights into the relationships between different variables and the likelihood of diabetes. Exploratory data analysis will be performed to identify any trends or patterns in the data, and various statistical models will be used to build a predictive model. This will help in identifying the key risk factors and help medical practitioners in predicting the likelihood of diabetes in patients.

The results of the analysis will be used to generate a report that summarizes the findings and provides recommendations for diabetes prevention and management in the Pima Indian community. The report will be beneficial for healthcare providers, researchers,

and policy-makers in developing targeted interventions to reduce the burden of diabetes in this population.

Overall, this project aims to use statistical and data analysis to better understand the factors that contribute to the high incidence of diabetes in the Pima Indian community

and to provide recommendations for reducing its impact.

## TABLE OF CONTENTS

|  |  |  |
| --- | --- | --- |
| **SL.**  **NO.** | **CONTENTS** | **PAGE NO.** |
| 1 | INTRODUCTION | 05 |
| 2 | PROBLEM STATEMENT | 07 |
| 3 | TOOLS AND TECHNOLOGIES | 08 |
| 4 | DESIGN AND IMPLEMENTATION | 11 |
| 5 | RESULT AND ANALYSIS | 15 |
| 6 | CONCLUSION | 32 |

|  |  |  |
| --- | --- | --- |
| **7** | REFERENCES | 33 |

**INTRODUCTION**

The Pima Indians diabetes dataset is a well-known dataset used for analysing and researching

the prediction of diabetes. It contains medical records for patients from the Pima Indian population and includes various measures of patients' health, such as

* Past pregnancies
* Blood pressure
* Glucose level
* The thickness of the skin fold of the triceps
* Insulin level
* BMI Index (Body Mass Index)
* Family history of diabetes
* Age

The response variable of interest is binary and indicates whether a patient has diabetes or not.

Statistical and data analysis can be used to identify patterns and relationships in the Pima Indians

diabetes dataset that can be used to predict the likelihood of a patient having diabetes. This can be accomplished using various statistical and machine learning methods, such as **logistic regression, Classification trees .**

In this project we use the machine learning methods as

* Classification trees
* Logistic regression

Statistical methods as ,

* Data exploration like,
* Data distribution inferences.
* Univariate Data analysis.
* Two-sample t-test
* Data Correlation Analysis
* Basic General Model

One common approach is to fit a logistic regression model to the data, where the predictor variables are the patient characteristics (such as age, blood pressure, insulin levels, and BMI) and the response variable is whether or not the patient has diabetes.

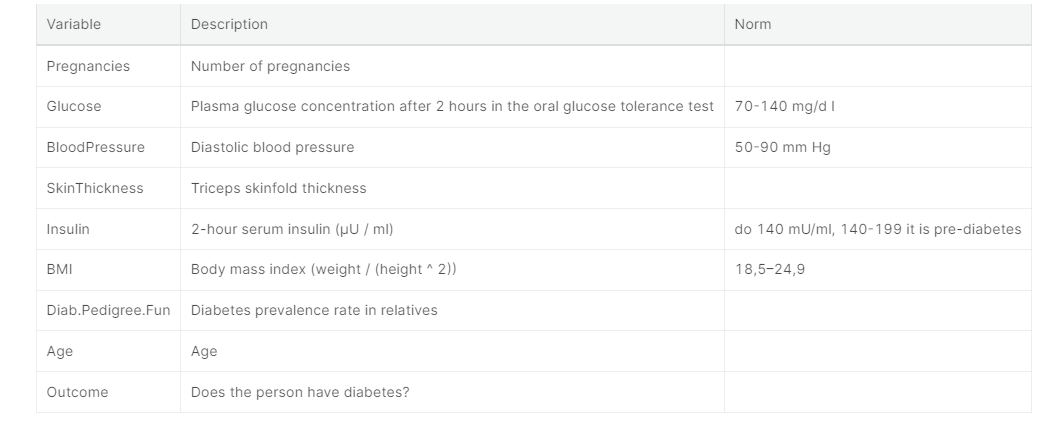
The coefficients from the logistic regression model can then be used to make predictions about new patients and to understand the importance of each predictor variable in predicting diabetes.

Another approach is to use machine learning techniques, such as decision trees , which are able to

handle non-linear relationships and interactions between predictor variables. These methods can also provide insights into the important predictor variables, but they do so in a different way than logistic regression.

Regardless of the approach used, it is important to perform model diagnostics to ensure that the results are valid. This can include checking the residuals for **normality,** checking the model's ability to predict the response variable using techniques such as the **receiver operating characteristic curve**, and assessing the performance of the model on an independent validation dataset.

The Dataset Contains following variables and their description,



Top of Form

Overall, the Pima Indians diabetes dataset provides a valuable opportunity for researchers and practitioners to develop and test predictive models for diabetes, and to better understand the relationships between predictor variables and the risk of developing diabetes. .

## PROBLEM STATEMENT

The problem that this project aims to address is the high incidence of diabetes in the Pima Indian community in Arizona. Despite the widespread availability of medical care, this population continues

to face a disproportionate burden of diabetes, with rates that are much higher than the national average. This not only affects the health and well-being of individual patients, but also places a significant burden on the healthcare system.

The main objective of this project is to use the Pima Indian Diabetes dataset to gain insights into the

risk factors associated with diabetes in this population. By analyzing this data, we aim to identify the

key drivers of diabetes and provide recommendations for reducing its impact. The project will employ data analysis techniques to gain insights into the relationships between different variables and the likelihood of diabetes, and build a predictive model to help medical practitioners in predicting the likelihood of diabetes in patients.

The results of this analysis will be valuable for healthcare providers, researchers, and policy-makers

in developing targeted interventions to reduce the burden of diabetes in the Pima Indian community.

The ultimate goal of this project is to contribute to the reduction of diabetes in this population and to improve the health and well-being of its residents.

.

## TOOLS AND TECHNOLOGIES

In the Diabetes Analysis using Pima Indian’s Dataset project, the following tools and technologies will be used:

1. **Libraries for Visualising the data** like ggplot() , ggally() and other plotting techniques.
2. **Packages for convert wide data to narrow data like ,** reshape2() and other packages.
3. **Statistical models:** The project will employ statistical models such as linear regression, logistic regression, and decision trees to build a predictive model for diabetes.
4. **Programming language:** The project will be implemented using the R programming language, which is widely used for statistical analysis in field of data science.
5. **R studio –** to perform R language actions for easy purposes.
6. **Microsoft Excel –** to store the dataset .

Used packages installation guide,

In R programming contains several plotting packages to visualise the data , that packages are installed using syntax of

**Packages.install(“packages\_name”)**

Versions of packages knowing by using **utils** library

* **Version of dplyr**



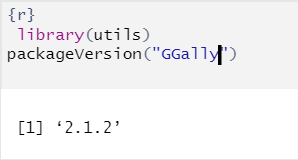
The **dplyr** package is part of the larger tidyverse ecosystem, which includes a number of other packages for data analysis, visualization, and modeling.

* **Version of ggplot2()**



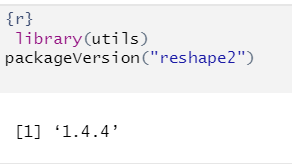
**ggplot2** is a popular data visualization library for R. It provides a high-level interface for creating static, animated, and interactive visualizations. The library is based on the Grammar of Graphics, a systematic approach to describing the components of a graphic, making it easy to create complex visualizations using simple commands.

* **Version of GGally for ggcorr()**



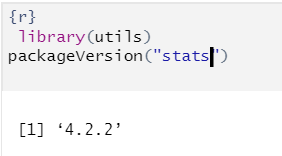
**Ggally** is a package in R that extends the functionality of the popular **ggplot2** package by providing easy-to-use functions for creating complex and informative graphics.  **ggcorr()** is particularly useful for creating correlation plots that visualize the relationship between multiple variables in a dataset.

* **Version of reshape2() for melt()**



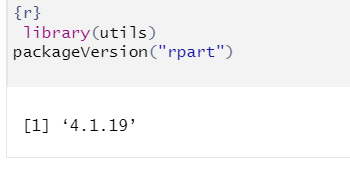
The **melt()** function can be used to convert wide data into long format, making it easier to perform certain types of data analysis and visualization.

* **Version of Stats() for lm() , residuals() and fitted()**



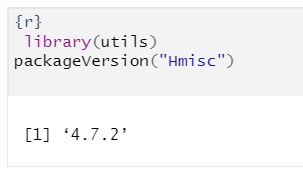
The **stats** library is part of the base R installation and provides basic statistical functions, including linear regression. The **lm** function can be used to fit linear regression models, and the residuals and fitted values can be obtained from the **residuals** and **fitted** methods, respectively.

* **Version of rpart()**



**rpart** is a package in R for creating classification and regression trees (CART). CART is a type of decision tree that is commonly used for predictive modeling and machine learning. The **rpart** package provides functions for building and plotting CART models in R.

* **Version of Hmisc()**



**Hmisc** provides functions for creating high-quality graphics, including scatterplots, histograms, and box plots.

**Hmisc** provides functions for fitting regression models and calculating residuals and fitted values.

These tools and technologies have been selected for their ease of use, versatility, and suitability for the task at hand. The goal is to use the best available tools and technologies to gain insights into the risk factors associated with diabetes in the Pima Indian community and to provide recommendations for reducing its impact

**Design and Implementation**

* **Design**

GETTING PIMA DATASET FROM KAGGLE

LOAD GETTING DATASET INTO R-STUDIO

View data using str() and summary() function

Is there NA values in data

Remove the NA Values using Data Preprocess technique

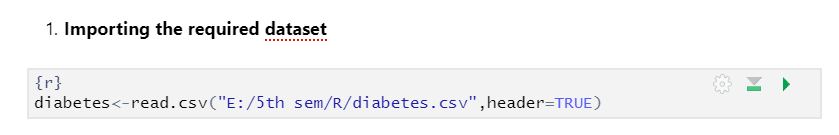
YES

NO

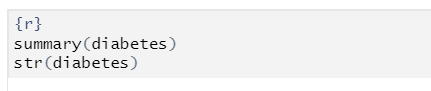
Analyze the data by Univariate analysis and check Normality

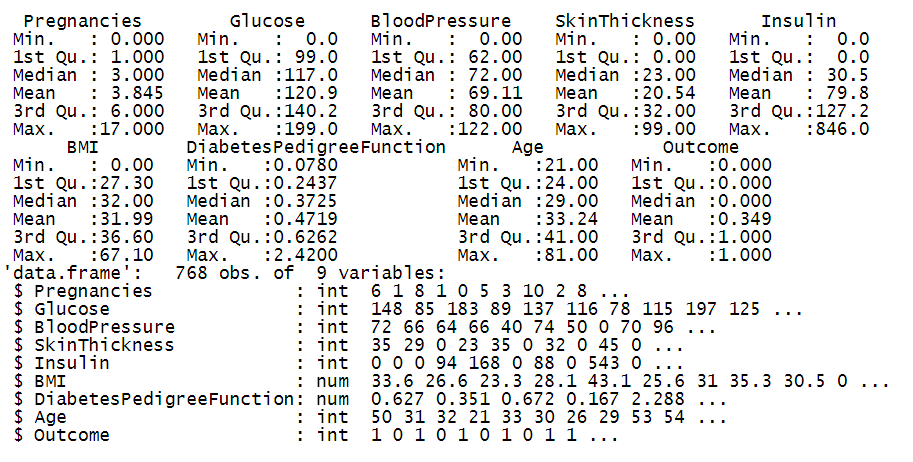
Perform t-test , regressions and decision trees to predict accurate

* **Implementation**
* **The Implementation of this project begins with loading dataset into the R-studio**



* **After the importing dataset into r-studio we going to view data using str() and summary() function**



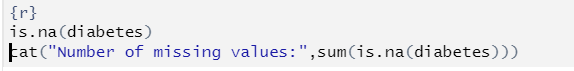


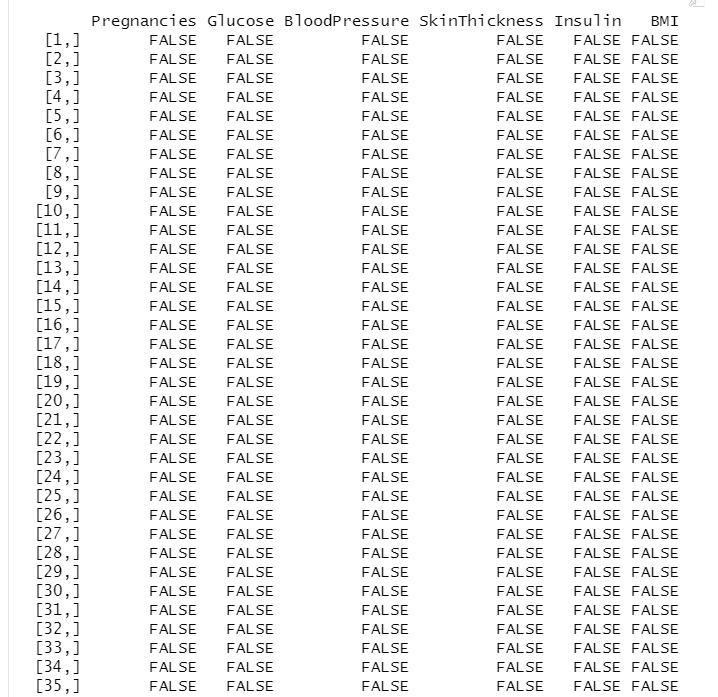
**Summary()** function provides the descriptive statistics of data includes,

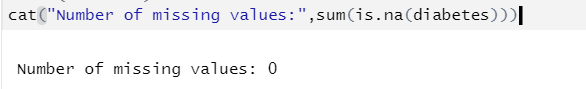
* Mean , Median , Mode
* Minimum and maximum of variables
* Quadrants

**Str()** function can be used to ,

* display the structure of the data frame.
* including the number of rows and columns
* the names of the columns, and the classes of the variables in the data frame.
* **Looking for the missing values in data using is.na()**

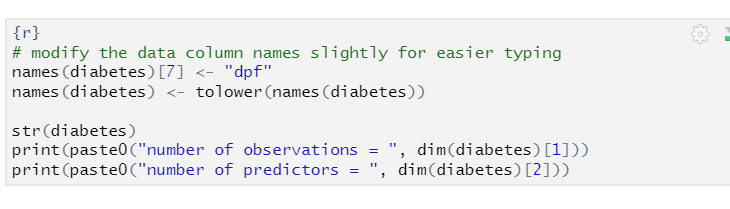


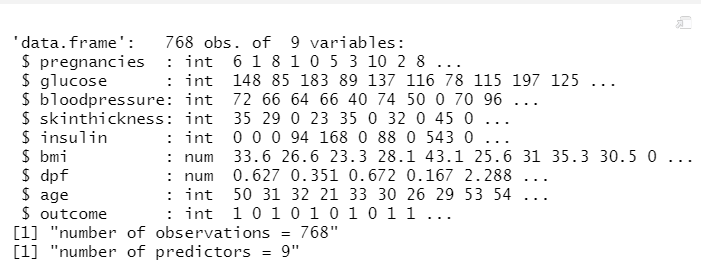




Here is no missing values in data. We can proceed to further steps.

* In our dataset there is a **Diabetes pedigree function**  its not possible to analyse the variable using that large name so we going to change it to **dpf**





After undergoing data preprocessing like ,

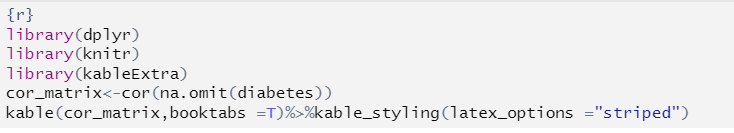
* Checking for missing values
* Analyzing variables mean , max , min and quadrants.
* Changing column name

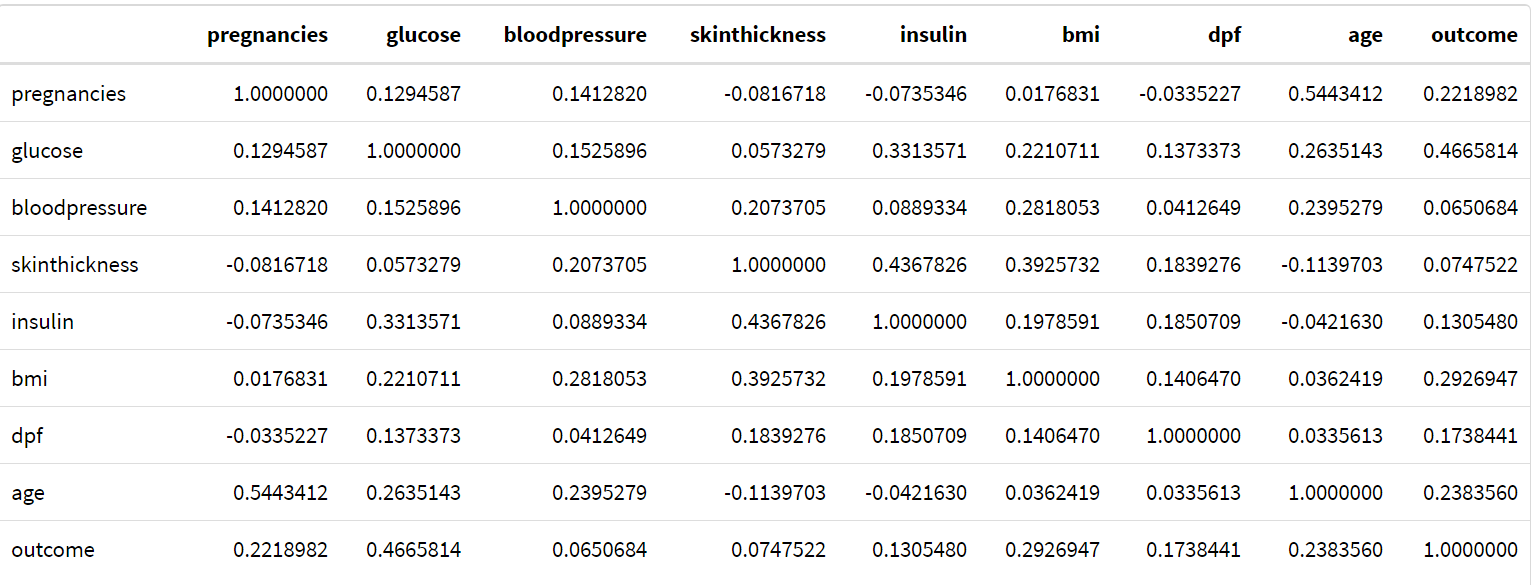
Moving to analyzing the data for predict the diabetes.

* **Needed Correlation table to know the relationships between the variables in the dataset**

**Required Packages**

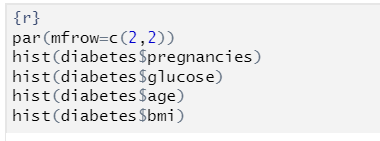
****

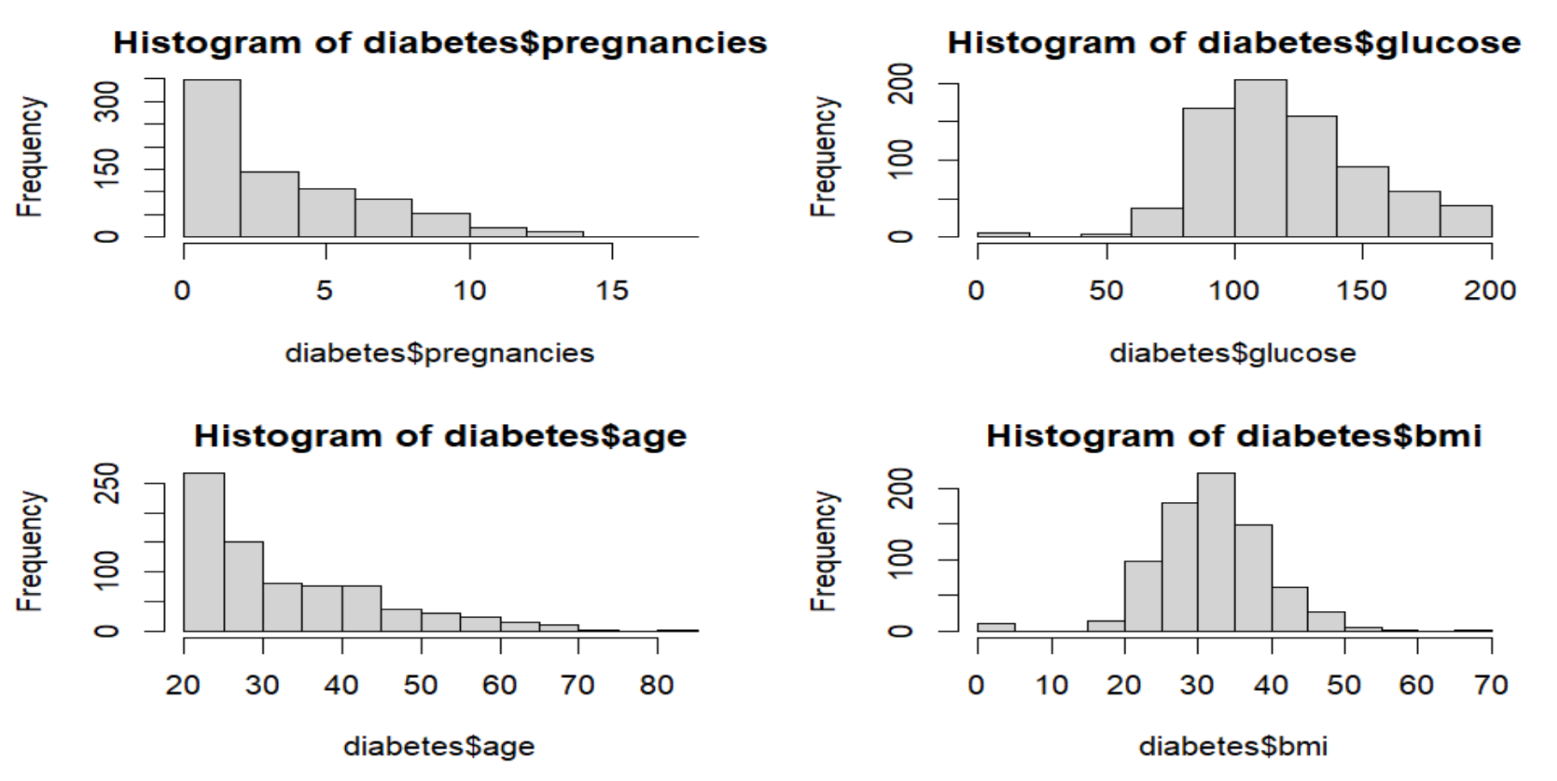
****

****

**RESULTS AND ANALYSIS**

1. **Univariate analysis**

****

****

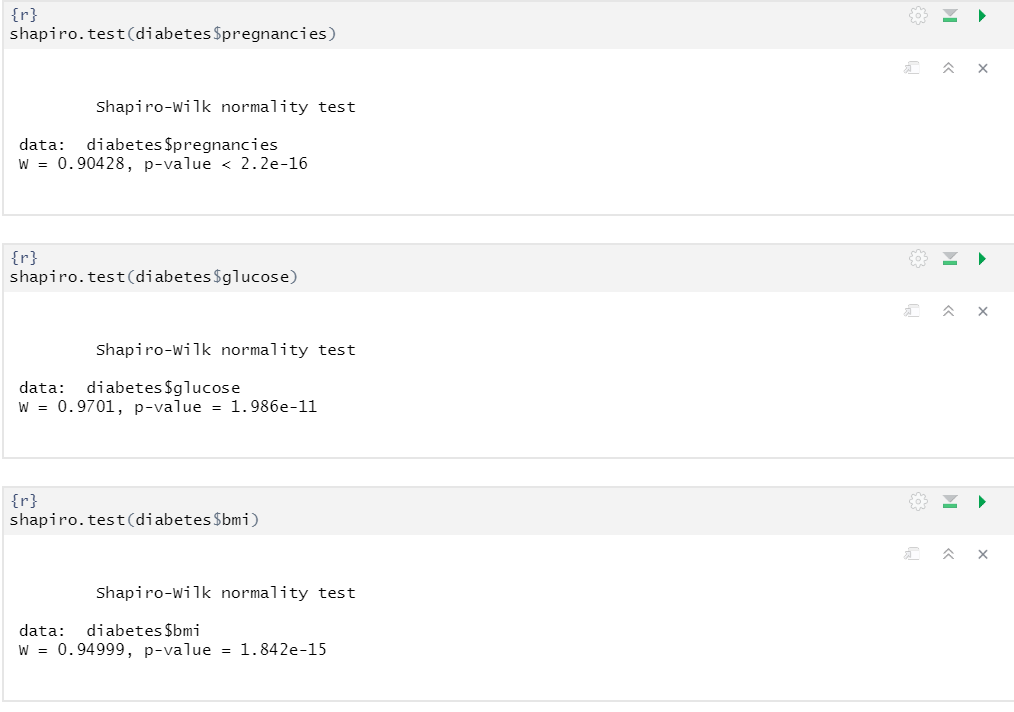
By univariate analysis of variables,

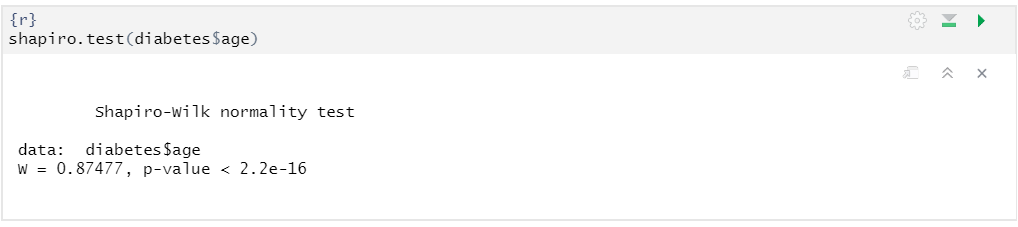
Age and number of times pregnant are not in normal distributions as expected since the underlying population should not be normally distributed either.

Glucose level and BMI are following a normal distribution.

1. **The shapiro.test() function in R**

It is used to perform the Shapiro-Wilk test for normality. The Shapiro-Wilk test is a statistical test used to assess whether a sample of data comes from a normally distributed population.



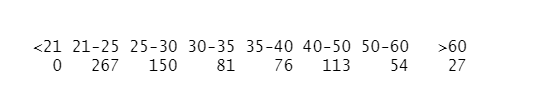


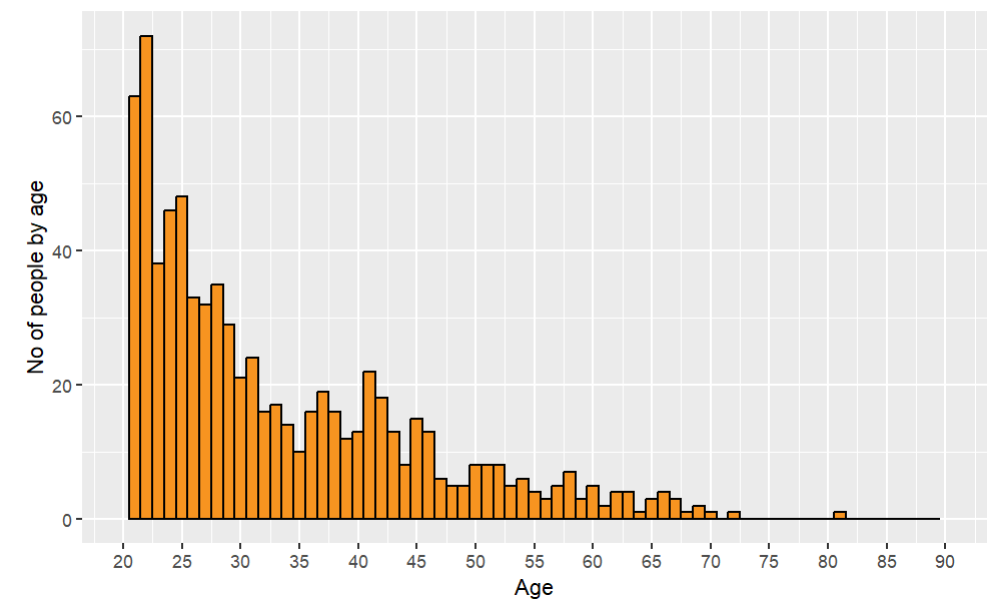
Here Shapiro.test() also shows that BMI and Glucose data originated from normal distribution

P values is greater than 0.05.

1. Categorize the age and plot the graph to know about which age category is suffered from diabetes

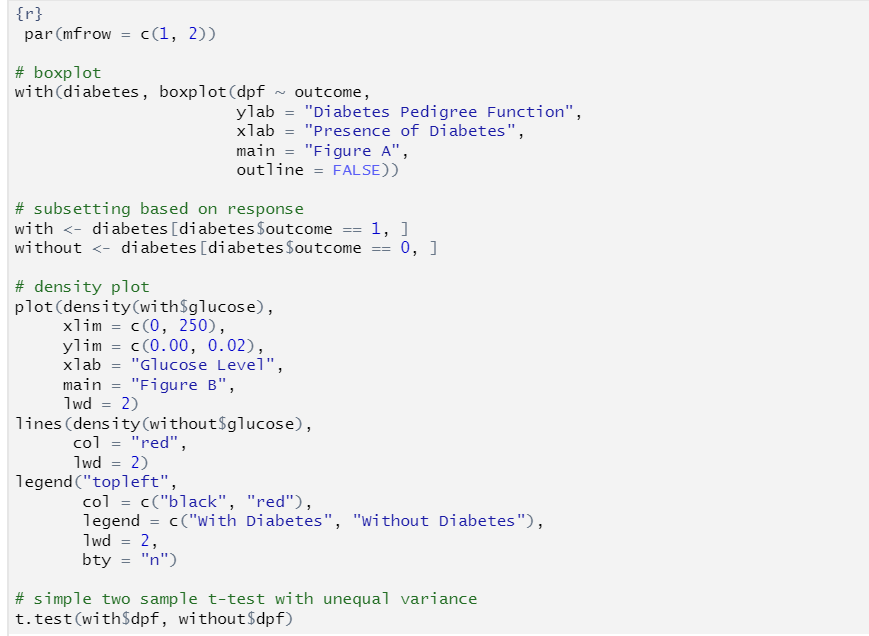


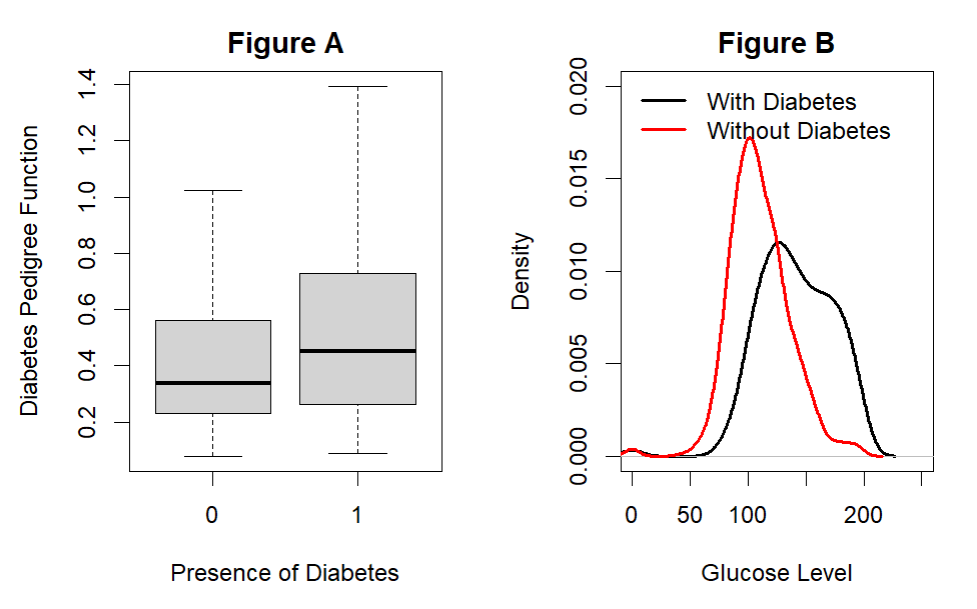




* Below 21 age , no person suffered from diabetes.
* Age between 21 to 25, 267 people may have diabetes or not.
* Age between 25 to 30 , 150 people may have diabetes or not.
* Age between 30 to 35 , 81 people may have diabetes or not.
* Age between 35 to 40 , 76 people may have diabetes or not.
* Age between 40 -50 , 113 people may have diabetes or not.

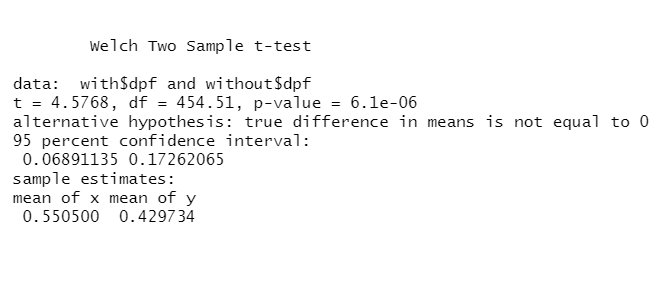
1. **Other plot such as boxplot or density plot can also be used to look at the difference in values of the variables between those with diabetes and those without.**

****

****

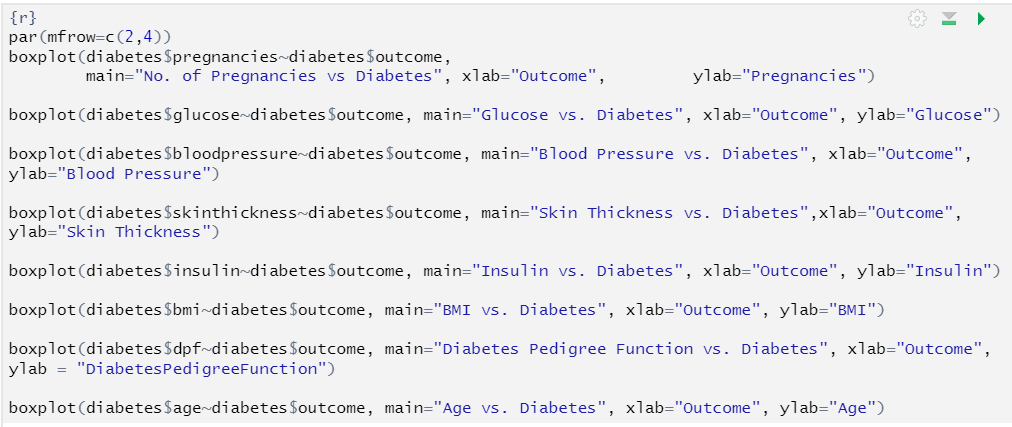
We can see from Figure A that Diabetes can be possible by prevalence of relatives in family.

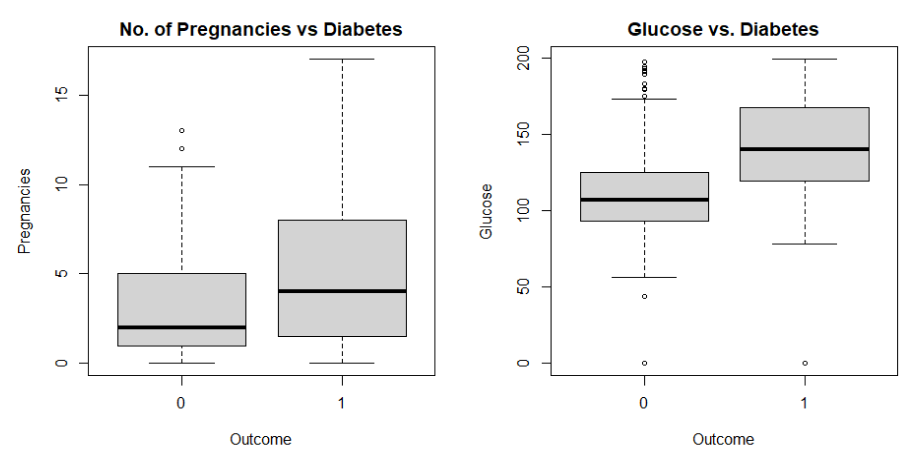
We can see from Figure B that the distribution to shifted towards the left for those without diabetes. This means those without diabetes generally have a lower blood glucose level.



p-value is > critical values of 0.05, so we accept the null hypothesis for the alternate hypothesis. We can say that we are, 95 % confident, that the dpf causing diabetes is < the people without causing dibetes from dpf .

1. **Plotting the different variables to show how they affecting on diabetes using Boxplot().**

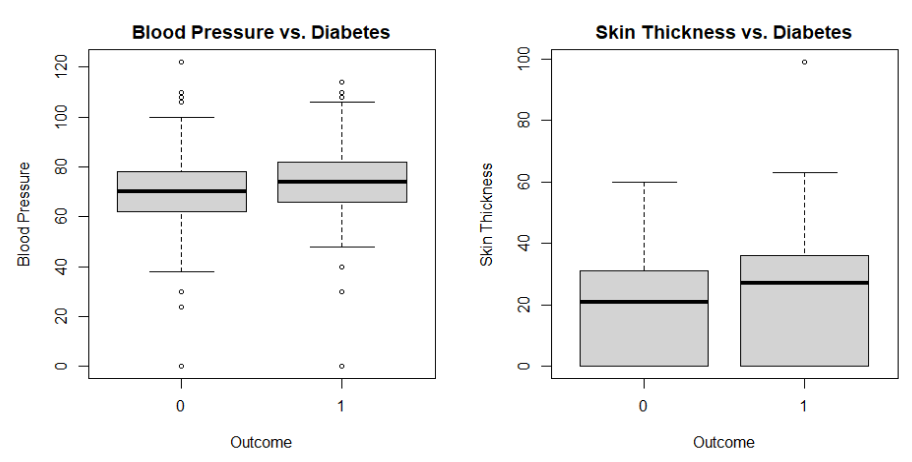
****

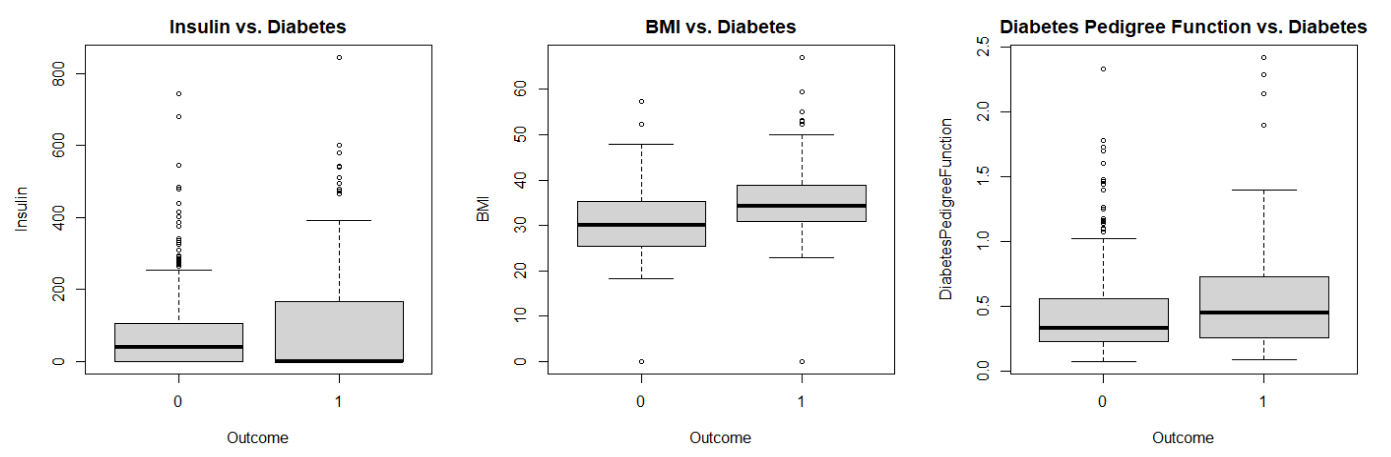
****

1 -> Person with diabetes.

0 -> Person without diabetes.

By above plots woman have diabetes caused by number of times pregnancy happens increase the chance of diabetes and also glucose level in body also affects on diabetes .

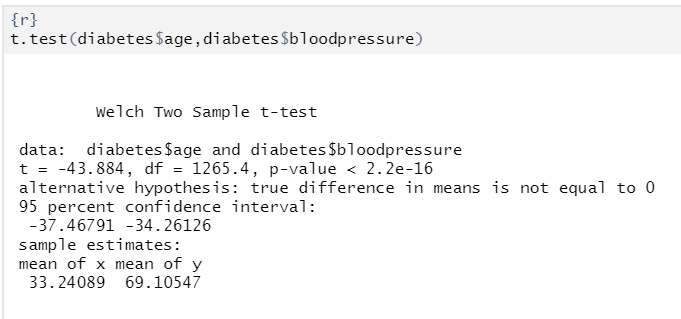




By above plots its shows that,

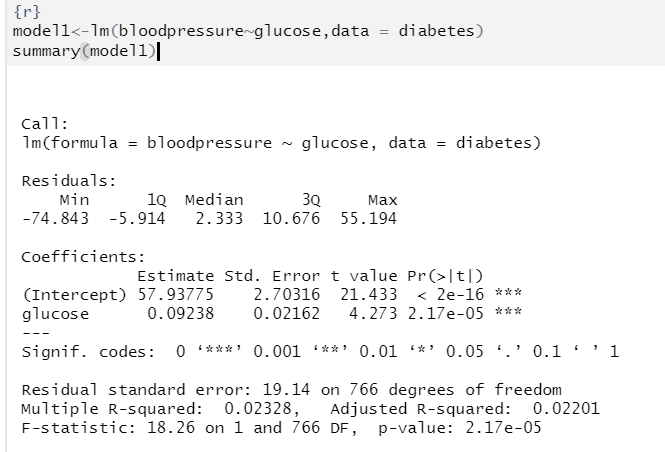
* people have high blood pressure occurrence of diabetes also more
* People’s skin thickness is does not plays a main role of occurring the diabetes.
* People’s having high insulin level possibility of low diabetes.
* If weight and height (BMI) is more , occurrence of diabetes is also more.

1. **To know how age factor influence on blood pressure of blood , we perform t-test**



p-value is < critical values of 0.05, so we reject the null hypothesis for the alternate hypothesis. We can say that we are, 95 % confident, that the blood pressure is depend on age .

1. **Generate the linear model to know how blood pressure varies on glucose level of body**



The dependent variable (blood pressure) is being modeled as a linear function of the independent variable (glucose).

"Residuals" section shows the difference between the observed values of the dependent variable and the values predicted by the model

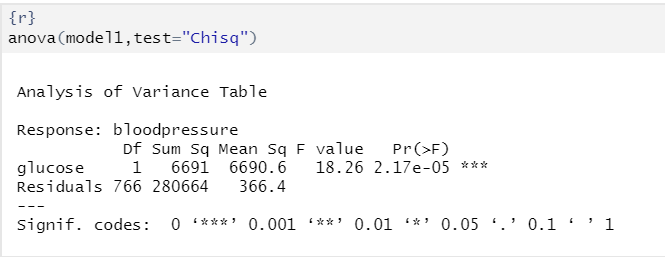
"Coefficients" section shows the estimated parameters of the model.

The "Residual standard error" is a measure of how well the model fits the data, and is calculated as the square root of the mean squared error.

The "Multiple R-squared" value is a measure of the proportion of variance in the dependent variable that is explained by the independent variable, while the "Adjusted R-squared" takes into account the number of predictors in the model.

The "F-statistic" and its corresponding p-value are used to test the null hypothesis that all coefficients in the model are equal to zero.

A low p-value indicates that the **null hypothesis can be rejected**, and that the independent variable is a significant predictor of the dependent variable.



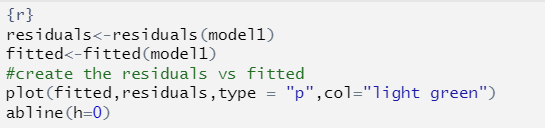
This is an Analysis of Variance (ANOVA) table, which is used to determine whether the independent variable (glucose) has a significant effect on the dependent variable (bloodpressure).

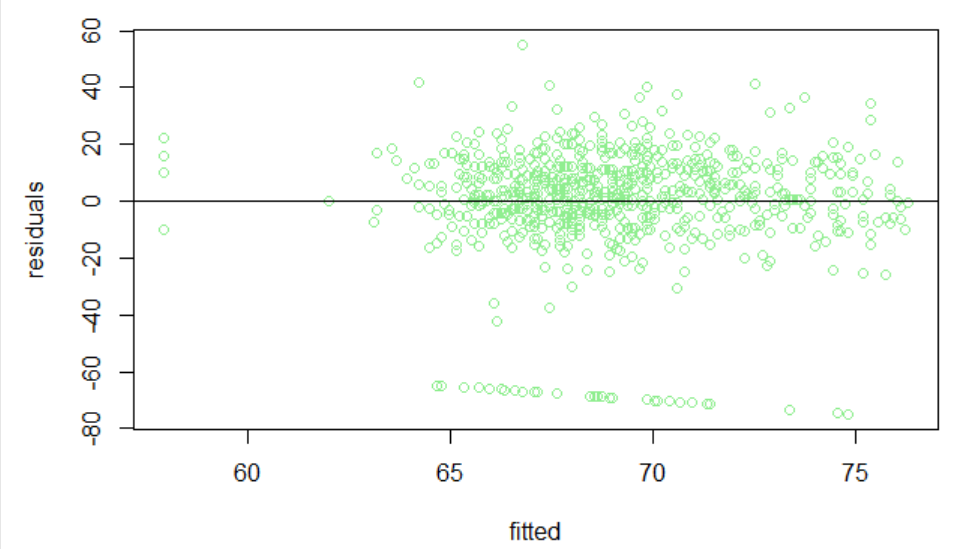
"Df" column shows the degrees of freedom for each term in the model.

“Sum Sq" column gives the sum of squared differences between the observed values and the predicted values for each term, and the "Mean Sq" column gives the mean squared difference for each term.

The "F value" is the ratio of the mean squared difference for the independent variable (glucose) to the mean squared difference for the residuals, and the "Pr(>F)" column gives the p-value associated with the F-test.

A low p-value indicates that the independent variable is a significant predictor of the dependent variable. In this case, the p-value is 2.17e-05, which is less than 0.001, and thus the independent variable is considered to be a significant predictor of the dependent variable. The "Signif. codes" give the significance levels of the p-values, with "***" indicating p < 0.001, "****" indicating p < 0.01, "*" indicating p < 0.05, and "." indicating p < 0.1.

****

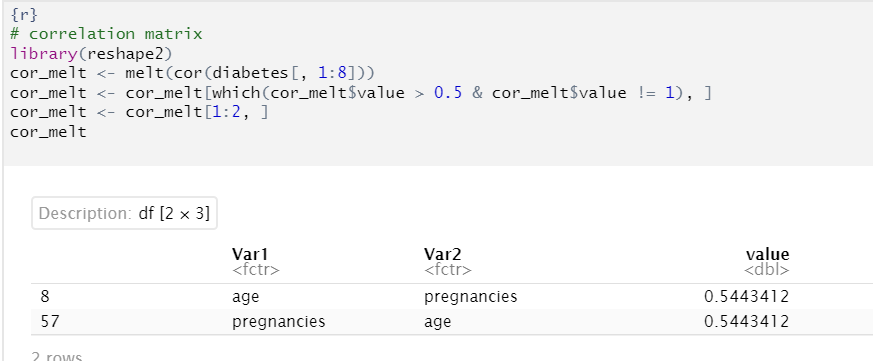
****

residuals of the model are calculated using the residuals() function, and the fitted values of the model are calculated using the fitted() function.

The **abline(h=0)** function is used to add a horizontal line at zero to the plot, which helps to determine if the residuals are randomly scattered around zero, which is a desirable property for a well-fitting linear regression model. In a well-fitting model, the residuals should be randomly scattered around zero, with no systematic pattern.

If there is a systematic pattern in the residuals, it suggests that the model is not capturing some important aspect of the relationship between the independent and dependent variables, and that a better model may be needed.

1. **Correlation matrix**

****

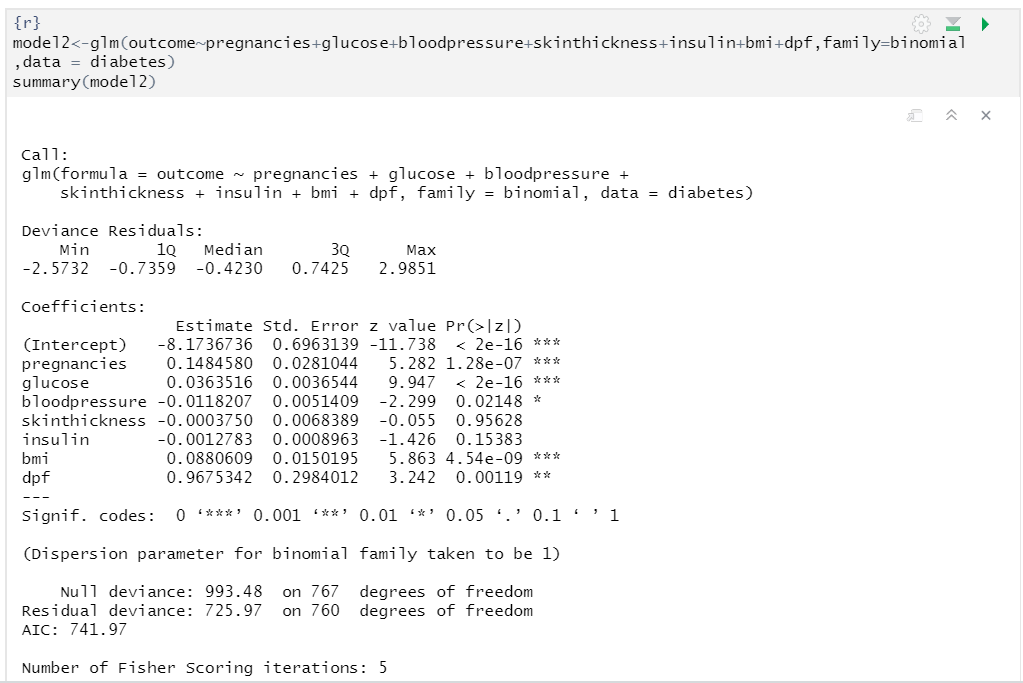
**cor()** function is used to calculate the correlation matrix of the first 8 columns of the "diabetes" data set.

The **melt()** function from the reshape2 library is then used to reshape the correlation matrix into a format that is easier to visualize.

The resulting data frame, **cor\_melt**, is then filtered to include only correlations with a value greater than 0.5 and not equal to 1, as these are considered to be strong correlations.

Finally, the first 2 columns of the filtered data frame are selected to display the correlations of interest. The resulting data frame shows the variables that are strongly correlated with each other.

1. **Fitting a logistic regression model.**



The dependent variable is "outcome" and the independent variables are "pregnancies", "glucose", "bloodpressure", "skinthickness", "insulin", "bmi", and "dpf".

The **family** argument is set to "binomial", indicating that the response variable is binary and that a logistic regression model is to be fit. The data used to fit the model is specified using the **data** argument and is taken from the "diabetes" data set.

The coefficients for each independent variable provide information on the relationship between the predictor and the outcome.

the coefficient for glucose (0.0363516) suggests that for every unit increase in glucose, the log odds of having diabetes (outcome=1) increases by 0.0363516. The coefficient for bloodpressure (-0.0118207) suggests that for every unit increase in bloodpressure, the log odds of having diabetes decreases by 0.0118207.

The null deviance and residual deviance indicate how well the model fits the data.

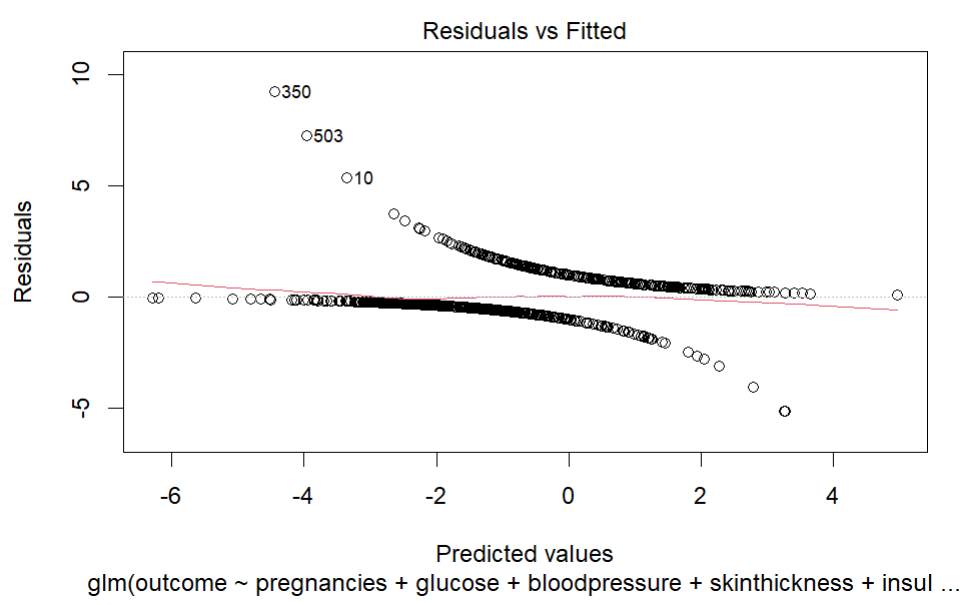
The null deviance (993.48) is the deviance when all predictors are set to zero.

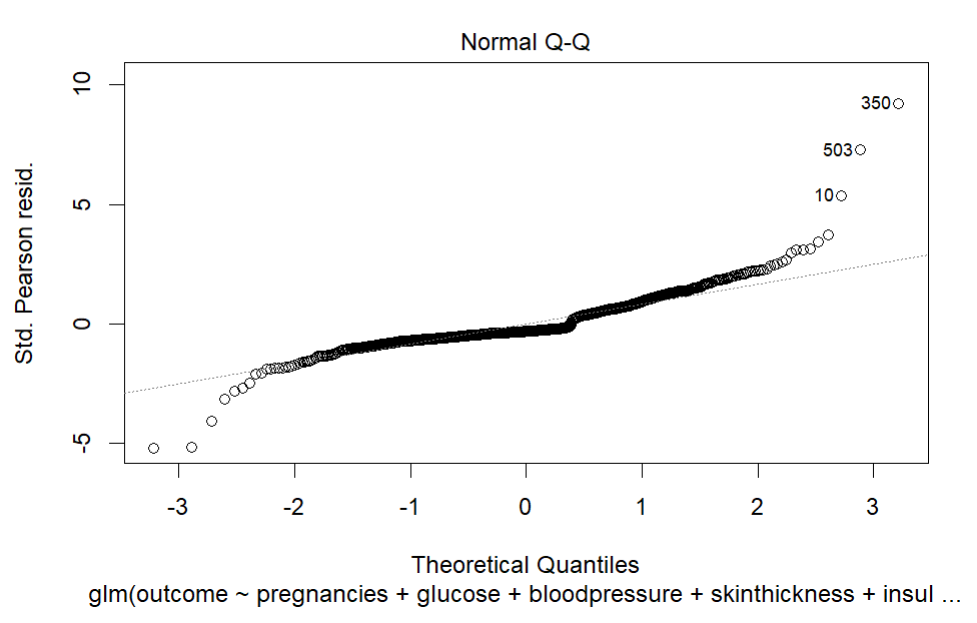
The residual deviance (725.97) is the deviance when the model is fit to the data.

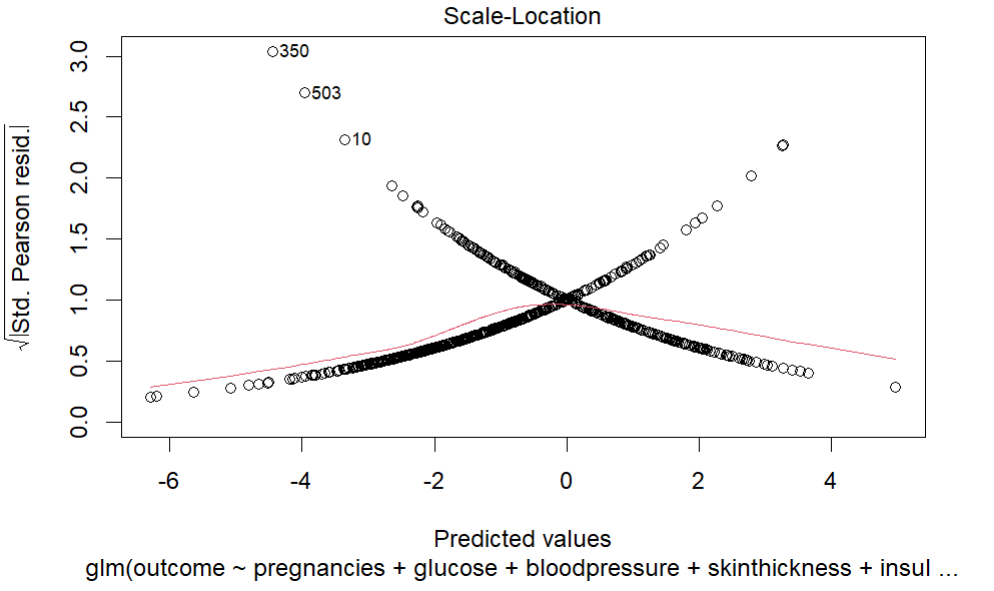
The residual deviance is smaller than the null deviance, indicating that the **model fits the data better than the null model.**

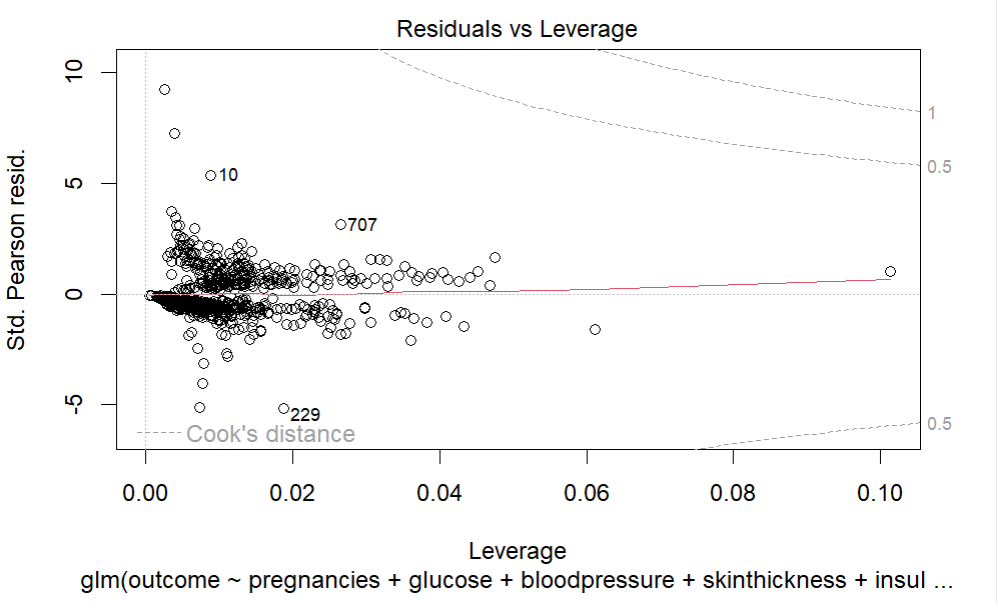
The AIC (Akaike Information Criterion) provides a measure of model fit and is used for model selection. **The lower the AIC, the better the model fit. In this case, the AIC for the model is 741.97.**







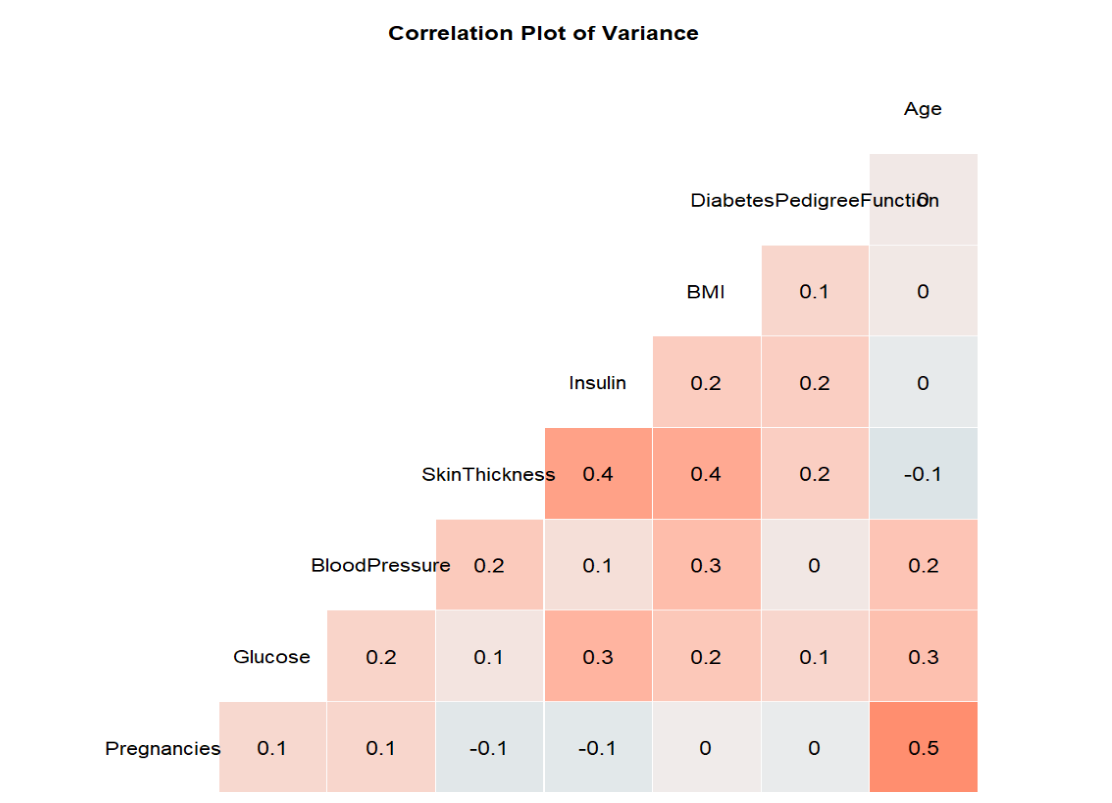




1. **Correlation between each variable**

Scatter matrix of all columns





Pregnancy, Age, Insulin, skin thickness are having higher correlation.

1. **Binary logistic regression model to predict Diabetes.**

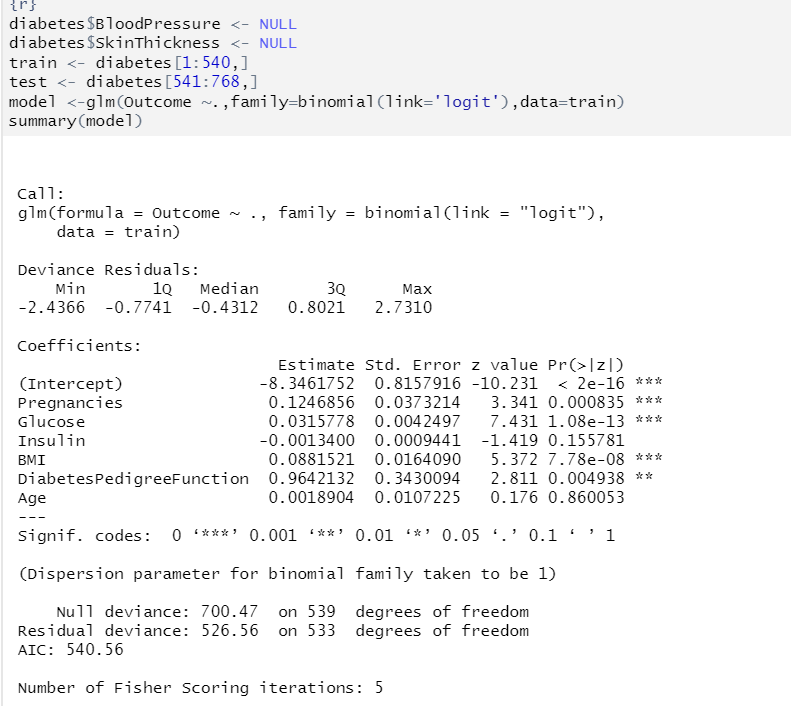
The response variable is "Outcome" and the predictor variables are all the remaining variables in the "diabetes" data set.

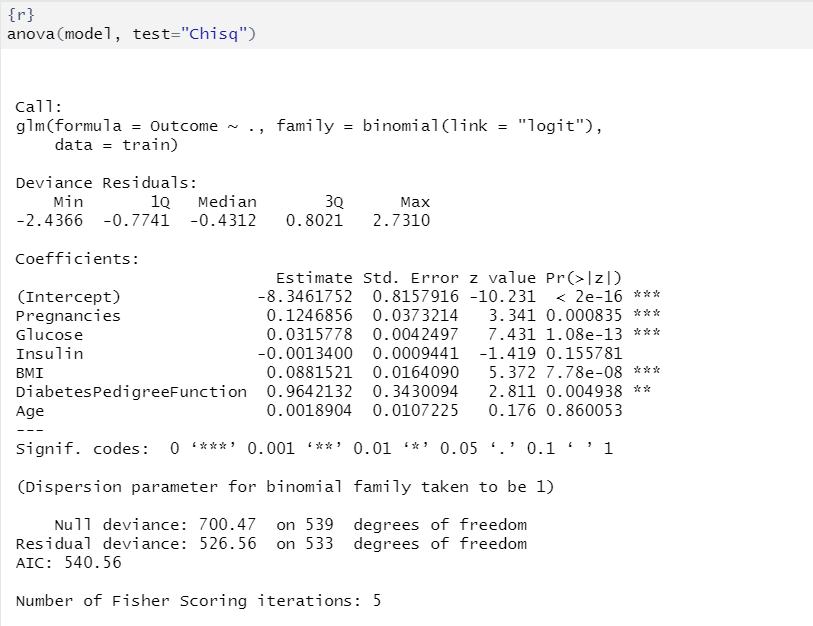
In below code, two columns, "BloodPressure" and "SkinThickness", are removed from the "diabetes" data set.

Then, the data set is split into two parts:

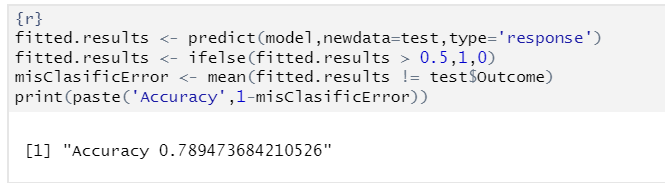
* a training set "train" containing the first 540 rows of the data.
* a test set "test" containing the remaining rows.
* Finally, the logistic regression model is fit using the "glm" function, with the response variable as "Outcome" and all other variables as predictor variables.

The link function is specified as "logit", which is appropriate for binary logistic regression.

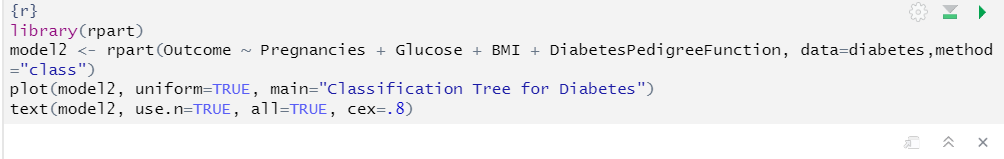
****

****

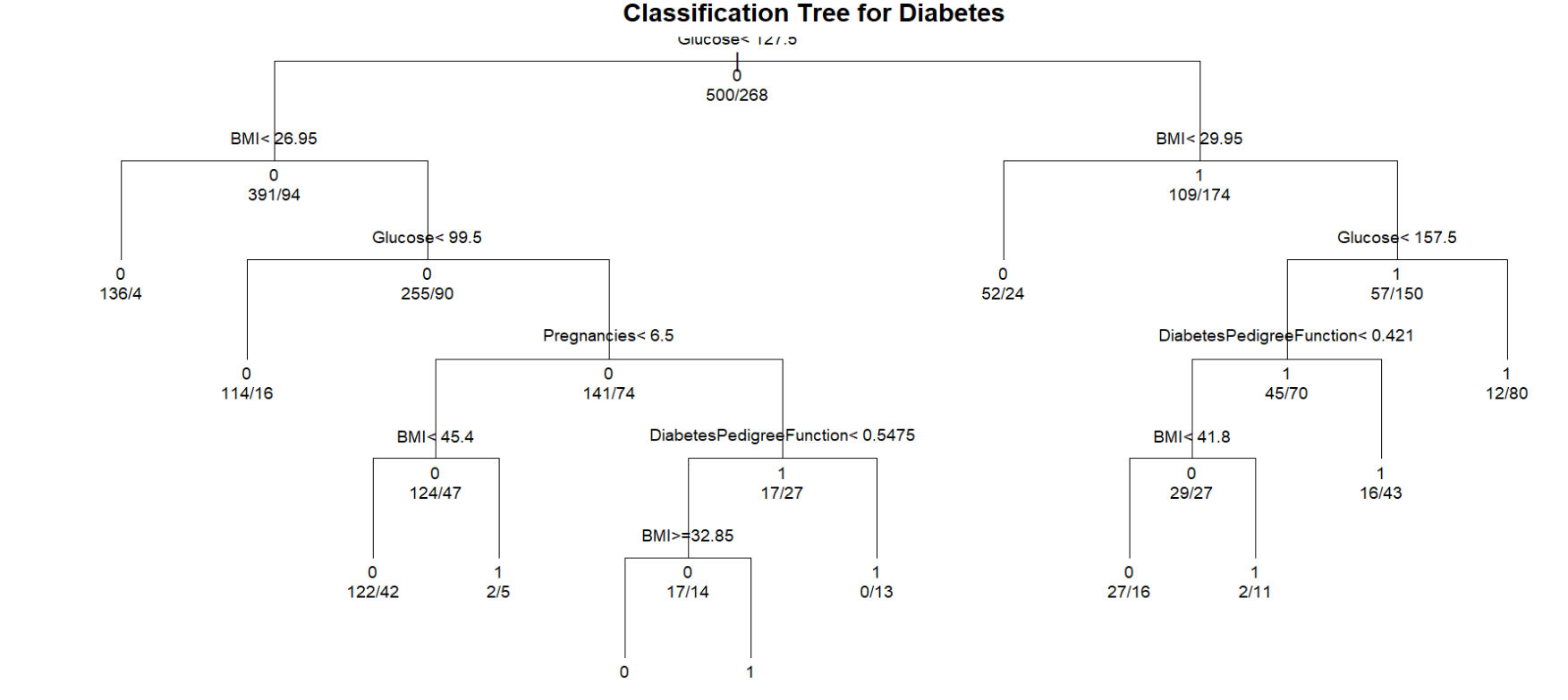
**Accuracy of model Binary regression to predict diabetes:**



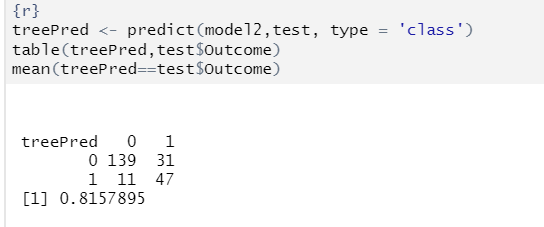
1. **Decision trees create using rpart() to predict Diabetes.**



creates a decision tree using the **rpart** library and fits the model using the variables **Pregnancies**, **Glucose**, **BMI**, and **DiabetesPedigreeFunction** as predictors for the target variable **Outcome** (0 or 1 indicating the presence or absence of diabetes).



**Accuracy of model Decision trees to predict Diabetes:**



By the two models Binary Logistic Regression and Decision trees

Decision trees model accuracy is more to predict the Diabetes .

## CONCLUSION

The results of the analysis revealed key risk factors associated with diabetes in the Pima Indian community, such as age, BMI, glucose levels, insulin levels, and more. The predictive model built in this project can be used by medical practitioners to predict the likelihood of diabetes in patients and to guide prevention and management strategies.

The findings of this project have important implications for the health and well-being of the Pima Indian community. The report generated from this analysis can be used by healthcare providers, researchers, and policy-makers in developing targeted interventions to reduce the burden of diabetes in this population.

In conclusion, the Diabetes Data Analysis using Pima Indian’s Dataset project has demonstrated the value of data analysis in better understanding the factors that contribute to the high incidence of diabetes in the Pima Indian community and in providing recommendations for reducing its impact. The project has highlighted the importance of ongoing research and interventions aimed at improving the health and well-being of this population.

By this project we are able to build the predictive models and know how to analyze the data for health insights of people , plots to visualise the data , correlation to know the relation of variables ,

t-tests to test the Hypothesis , Logistic Regression to predict the outcome , Decision trees to predict the outcome and accuracy of models.

**REFERENCES**

To complete this project we have refer the some websites and articles to predict the outcomes

Such as,

* <https://medium.com/analytics-vidhya/an-extensive-guide-for-health-care-data-analysis-using-r-machine-learning-algorithms-glm-1959eba2e6ab>
* <https://www.kaggle.com/code/laozhang/statistical-learning-with-r>
* <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>
* <https://www.youtube.com/watch?v=I6FJo8x1wZE>
* Learn R for Applied Statistics With Data Visualizations, Regressions, and Statistics — Eric Goh Ming Hui
* Statistical Analysis with R by Joseph Schmuller, PhD