**Description**

This dataset is the data for the prediction of early stage diabetes. This dataset was prepared from a direct questionnaire from the patients of Sylhet Disease Hospital which is in Bangladesh. This dataset is prepared for a classification task for determining whether a patient is suffering from early stage diabetes or not.

The dataset consists of the following columns representing features of the dataset.

* Age

The values of the age column ranges from 20 to 60.

* Sex

The sex column values is indicated as Male and Female

* Polyuria

The state of sufferring from excessive urination is known as polyuria. The polyuria column indicates whether the patient has polyuria or not.

* Sudden Weight Loss

Loosing of body muscles and fats in very short period of time is sudden weight loss. This column indicates whether the patient has sudden weight loss or not.

* Weakness

The feeling of fatigue, lethargicness iin an individual is weakness.This column indicates whether the patient has weakness or not.

* Polyphagia

Polyphagia is the state of suffering from extreme hunger situations. This column indicates whether the patient has Polyphagia or not.

* Genital Thrush

Itching on genttal bosy parts due to fungal infections is known as genital thrush. This column indicates whether the patient has Genital Thrush

or not.

* Visual Blurring

The state of vision being blurred or unclear can be referred as visual blurring.This column indicates whether the patient has Visual Blurring or not.

* Itching

Itching can be referred as skin irritation folllowed by infections ldue to lesions, rashes in skin. This column indicates whether the patient has Itching or not.

* Irritability

Unpleasant disturbed feeling is called iritability. This column indicates whether the patient has Irritability or not.

* Delayed Healing

The condition in which healing process gets delayed can be refered as delayed healing.This column indicates whether the patient has Delayed Healing

or not.

* Partial Paresis

The condition in which mild weakening of body muscle or mild paralysis is called partial paresis. This column indicates whether the patient has Partial Paresis or not.

* Muscle stickiness

The condition in which muscle tissues gets adhesive and stick to each other is called muscle stickiness. This column indicates whether the patient has Muscle stickiness

or not.

* Alopecia

Loss of hair can be refered as alopecia. This column indicates whether the patient has Alopecia or not.

* Obesity

Gaining of excessive body weight is called obesity.This column indicates whether the patient has Obesity or not.

According to these features the class variable is differentiated into two classes:

1. Positive

This class indicates that the patient has risk of early stage diabetes.

1. Negative

This class indicates that the patent has no risk of early stage diabetes.

**Data Preparation**

A csv format data of early stage diabetes dataset was downloaded from the UCI machine learning repository website using read.csv method in R.

Following data were obtained from the dataset.

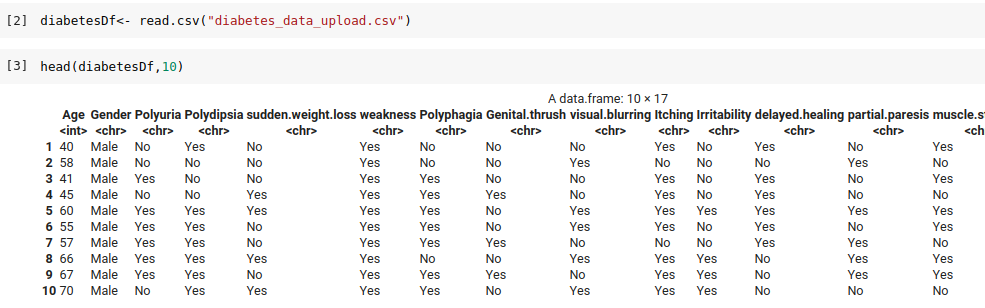


Fig: Screenshot showing first 10 observations in diabetes dataset

**Data Preprocessing**

The features of the dataset such as Gender, Polyuria and others have character datatype in their data observations. As a first step, we convert such character type data into numeric form for further processing and easy data wrangling.

For example, the values in the Gender column like “Male” are replaced with 1 and “Female” is replaced with 0.

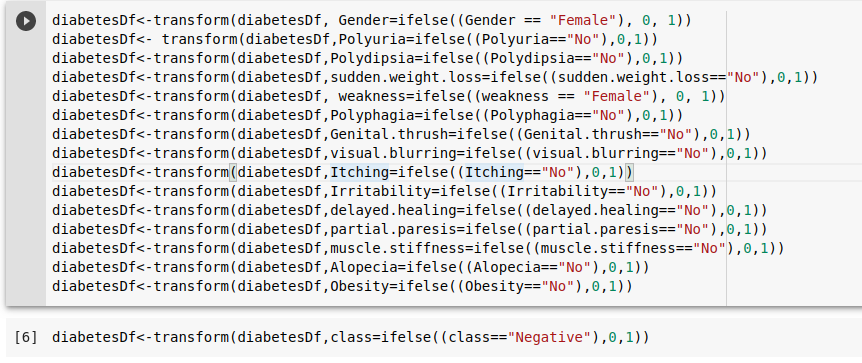


Fig: Code snippet for data preprocessing

After performing such conversion of values in the feature columns of the dataset, the first 10 instances of the dataset are shown below:

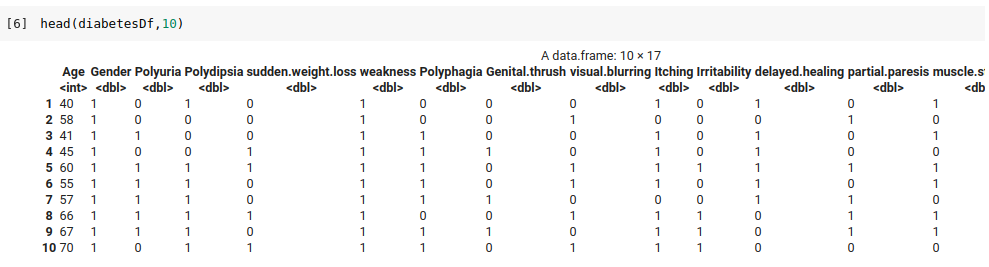


Fig: Screenshot showing first 10 observations of diabetes dataset after conversion to numeric data

**Data Analysis**

**Correlation plot**

Correlation plot is generated for the dataset using the “corrr” package in R. Pearson coefficient of correlation is developed. The correlation plot was obtained using the following block of code.



Following correlation plot was obtained.

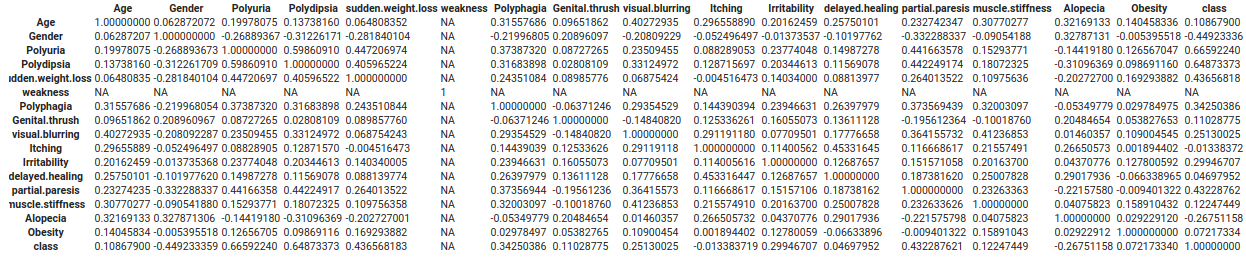


Fig: Correlation plot of the early stage diabetes data

The correlation coefficient close to -1 indicates strong negative linear relationship whereas close to +1 indicates strong positive linear relationship. From the correlation plot obtained for the data, we can evaluate that the attributes like “Polyuria” and “Polydipsia” have correlation values very close to +1 with the class variable thus indicating that these attributes have strong positive relationship and they affect the prediction of the class variable highly. Similarly, attributes such as “Gender” and “Alopecia” have correlation values very close to -1 indicating that these attributes have a strong negative relationship with the class column.

Similarly, from the correlation plot, we can observe that the correlation coefficient between the features are less , that means each feature is different from each other and has an effect on the output. Thus, we should keep all features. Similarly, we should also remove the “Weakness” attribute. We remove the “Weakness” attribute and observe the first 10 observation using following code:

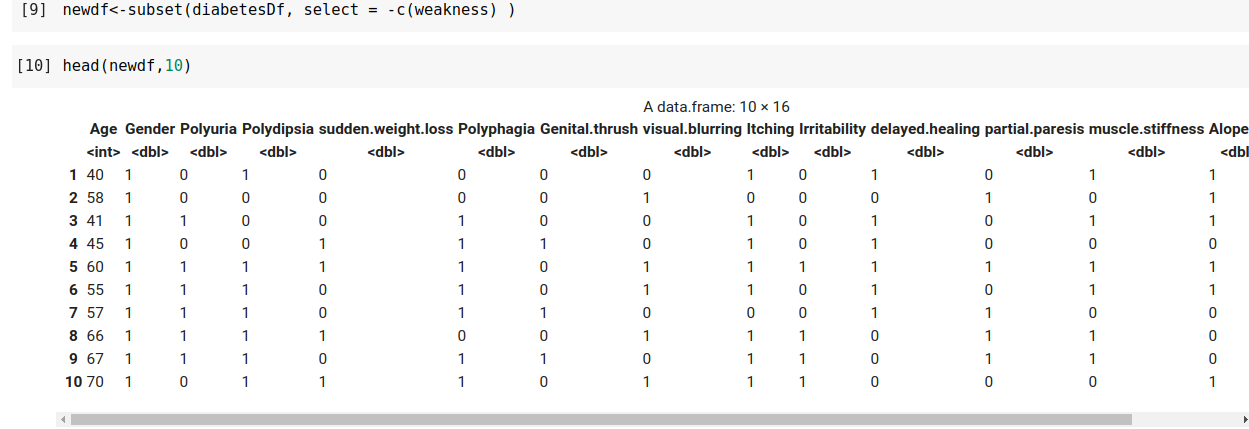


Fig: Code for removing “Weakness” attribute

**Correlation visualization using corrplot package**

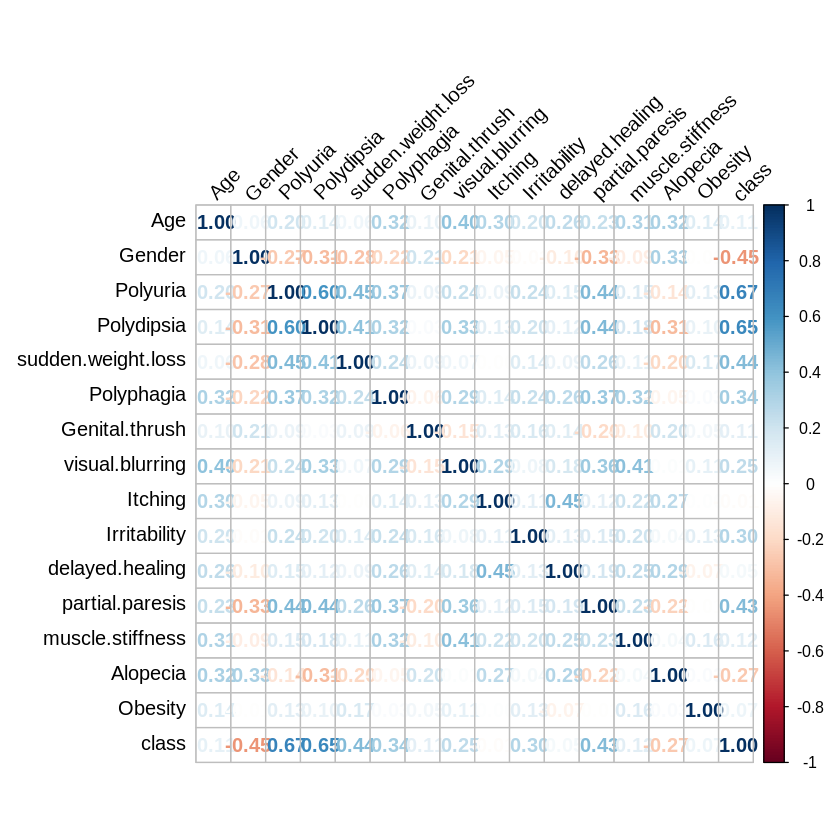
The correlation plot of above data was visualized using the corrplot package in R. The conversion of every “Yes”, “No” values in every columns to categorical numeric values like “0” and “1” was necessary for the correlation plot. Following code was used to install the corrplot package and use it in the subset of the whole preprocessed dataframe we obtained after removing the “Weakness” attribute.

install.packages("corrplot")

library(corrplot)

corrplot(cor(newdf), method="number",shade.col=NA, tl.col="black", tl.srt=45)

Following plot was obtained.

Fig: Correlation visualization using corrplot package

**Variance**

Variance of each and every features or columns of the dataset was calculated and evaluated using the “dplyr” package in R.



Following results were obtained.

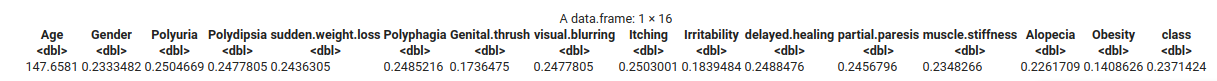


Fig: Variance obtained for each features

**Data Visualization**

**Scatter Plot**

For the visualization of data, scatter plots of each and every features of the dataset was plotted using the “psych” package in R. The pair\_plot() function of the package was used for plotting the scatter plot. The number of features being high, visually effective scatter plot was not obtained. Thus, scatter plots of only 4 columns were plotted and the relation between data was analyzed. After plotting the scatter plot for four columns, the following result was obtained.

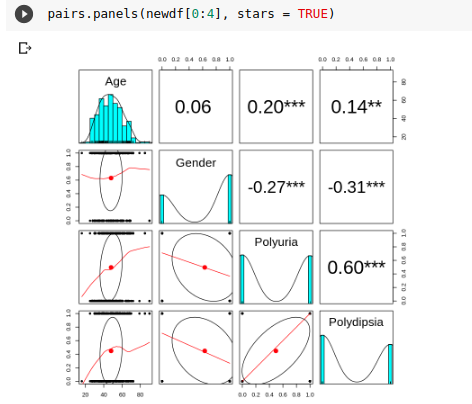


Fig: Scatter plot of the four columns of the dataset.

Scatter plot is done by using the pairs.panels() feature of the “psych” package. This feature is used to produce a matrix scatter plot with bivariate scatter plots below the diagonal, histograms on the diagonal, and the correlation of Pearson above the diagonal. From the above scatter plot, we can evaluate that the “Age” attribute follows normal distribution whereas other attributes are mainly categorical in the dataset.

**Histogram of the age feature in the dataset**

Histogram of the age feature in the dataset was developed to observe the data distribution in this feature. Following code was used to develop the histogram.

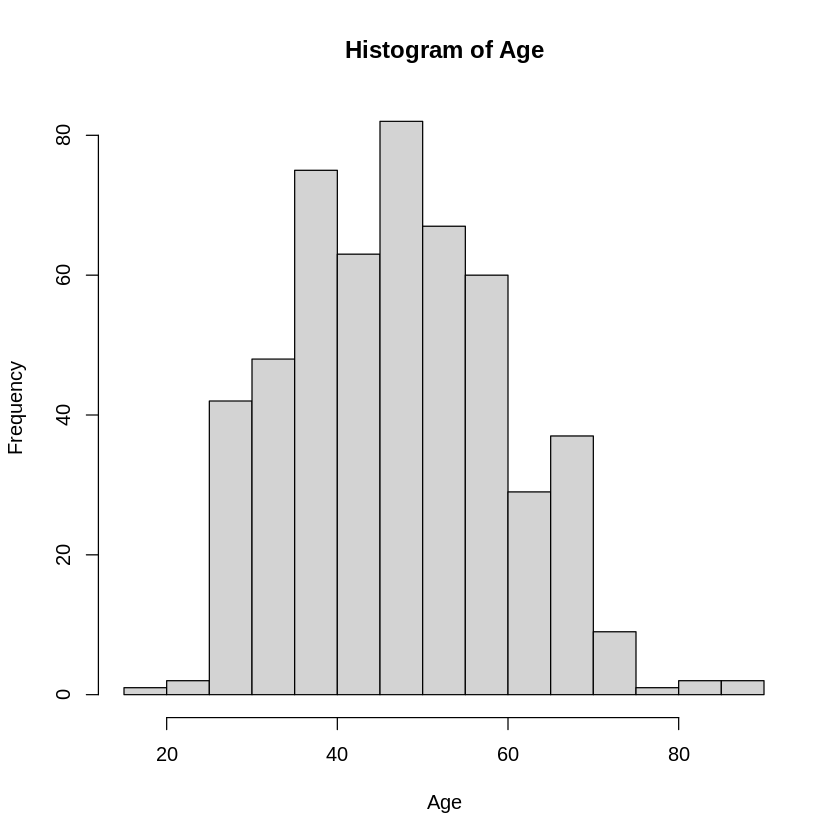
###Reading dataframe

diabetesDf<- read.csv("diabetes\_data\_upload.csv")

attach(diabetesDf)

names(diabetesDf)

hist(Age)

Fig: Histogram of the Age feature in the dataset

From above visualization, we can observe that most of the data observations in the dataset represent the individuals in the age above 25 and below 70. The age group with the highest frequency of data observations is 45-50.

**Mosaic plot**

Mosaic plots can be used to visualize the distribution of categorical data with respect labels or classes. Mosaic plot of different features in the datast was developed and visualized. Following code was used to develop a mosaic plot of the gender feature with the output class.

gendertable <- table(Gender,class)

gendertable

class

Gender Negative Positive

Female 19 173

Male 181 147

mosaicplot(gendertable, main = "Gender Mosaic Plot",

sub = "Early stage diabetics condition based on gender",

xlab = "Gender",

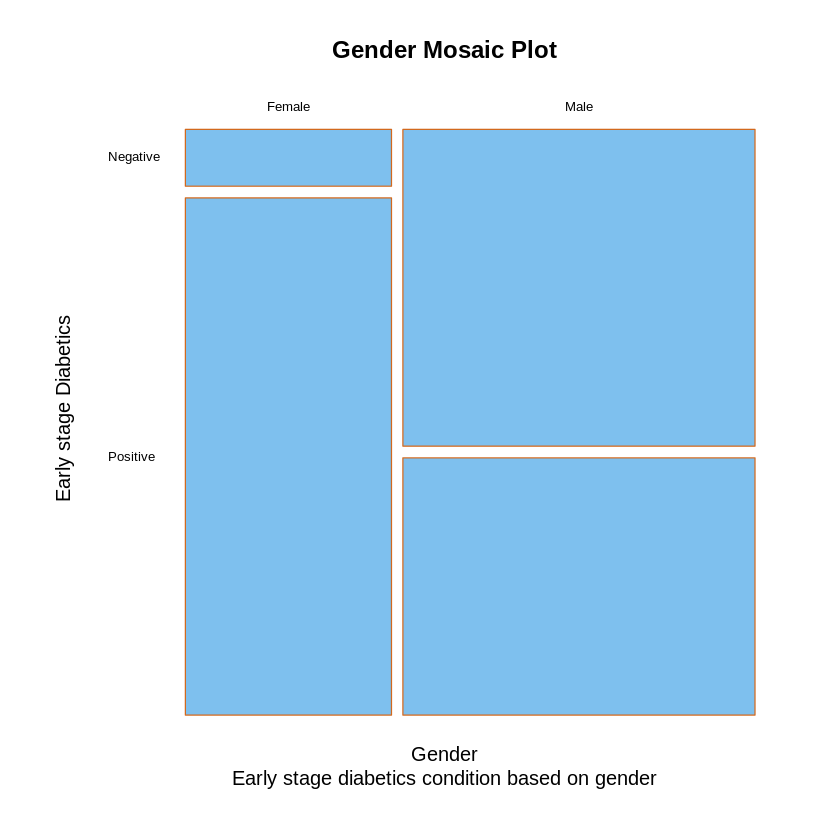
ylab = "Early stage Diabetics",

las = 1,

color = "skyblue2",

border = "chocolate")

Following mosaic plot was obtained:

Fig: Gender and class mosaic plot

Similarly, mosaic plot between polyuria and class of the dataset was visualized to gain insight on whether early stage diabeties patient had Polyuria frequently or not.

Following code was used:

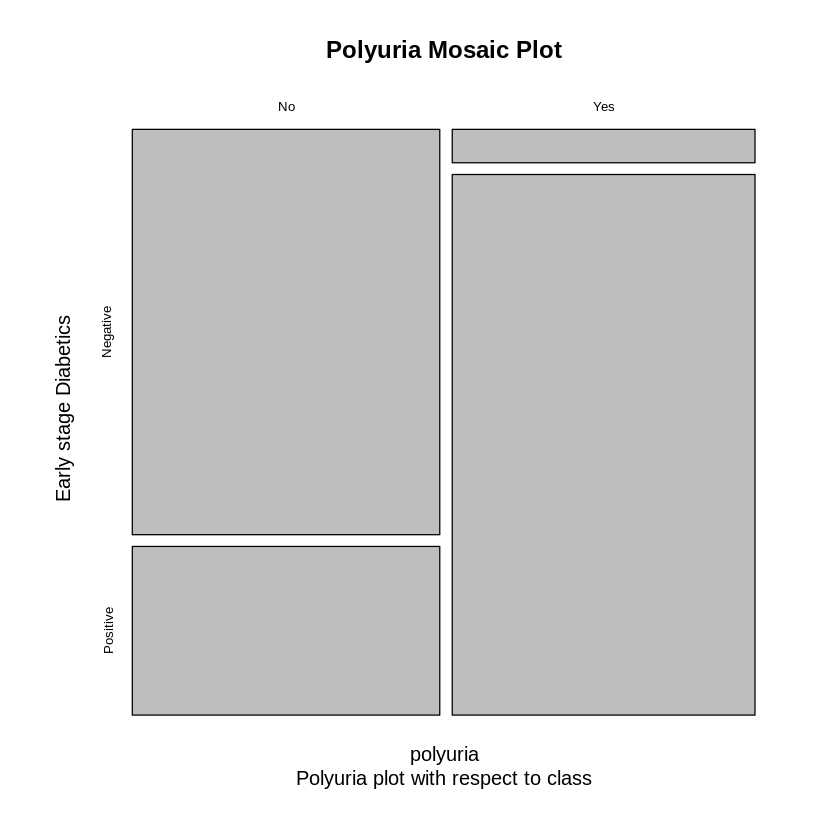
polyuriatable <- table(Polyuria,class)

mosaicplot(polyuriatable, main = "Polyuria Mosaic Plot",

sub = "Polyuria plot with respect to class",

xlab = "polyuria",

ylab = "Early stage Diabetics")

Fig: Mosaic plot between polyiria and class