**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 polyuria column indicates whether the patient has polyuria or not. The indication is done as ‘“Yes” for having Polyuria and “No” for not having Polyuria.

* Sudden Weight Loss

This column indicates whether the patient has sudden weight loss or not. The values are “Yes” and “No”.

* Weakness

This column indicates whether the patient has weakness or not. The values are “Yes” and “No”.

* Polyphagia

This column indicates whether the patient has Polyphagia or not. The values are “Yes” and “No”.

* Genital Thrush

This column indicates whether the patient has Genital Thrush

or not. The values are “Yes” and “No”.

* Visual Blurring

This column indicates whether the patient has Visual Blurring or not. The values are “Yes” and “No”.

* Itching

This column indicates whether the patient has Itching or not. The values are “Yes” and “No”.

* Irritability

This column indicates whether the patient has Irritability or not. The values are “Yes” and “No”.

* Delayed Healing

This column indicates whether the patient has Delayed Healing

or not. The values are “Yes” and “No”.

* Partial Paresis

This column indicates whether the patient has Partial Paresis or not. The values are “Yes” and “No”.

* Muscle stickiness

This column indicates whether the patient has Muscle stickiness

or not. The values are “Yes” and “No”.

* Alopecia

This column indicates whether the patient has Alopecia or not. The values are “Yes” and “No”.

* Obesity

This column indicates whether the patient has Obesity or not. The values are “Yes” and “No”.

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.

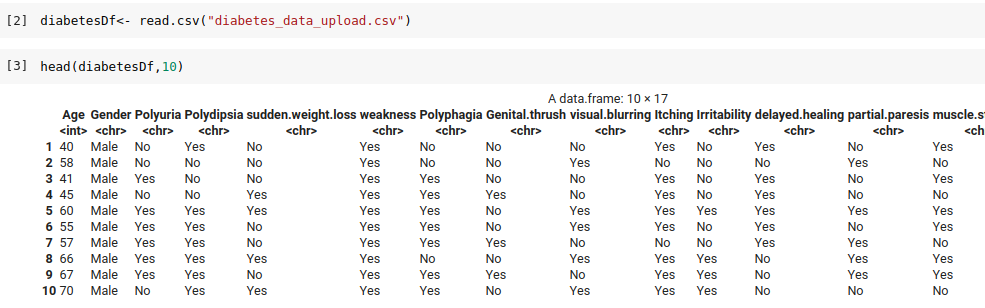


Fig: Screenshot showing first 10 observations in diabetes dataset

**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.

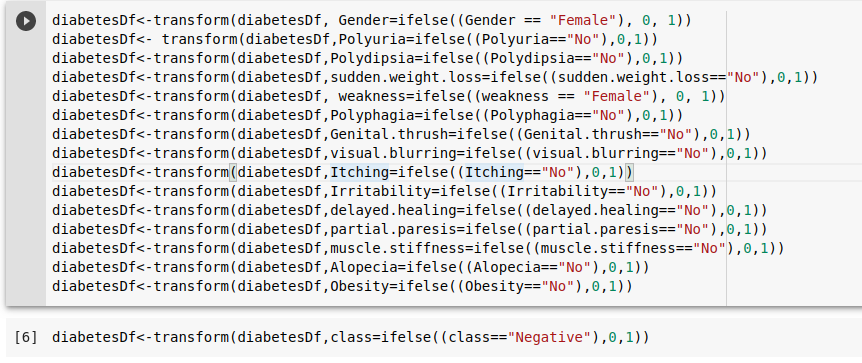


Fig: Code snippet for data preprocessing

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

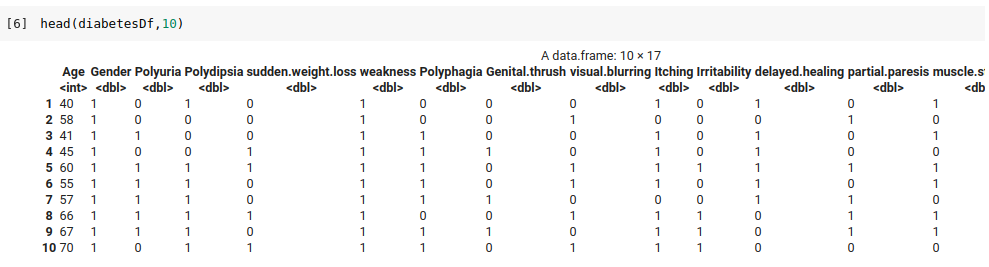


Fig: Screenshot showing first 10 observations of diabetes dataset after conversion to numeric data

**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.

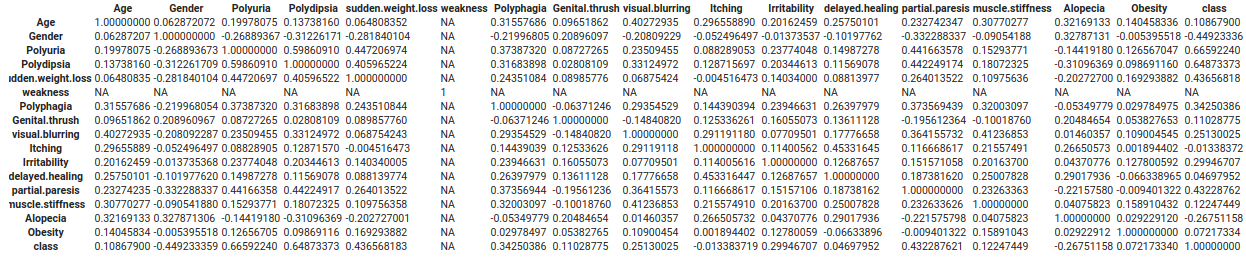


Fig: Correlation plot of the early stage diabetes data

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:

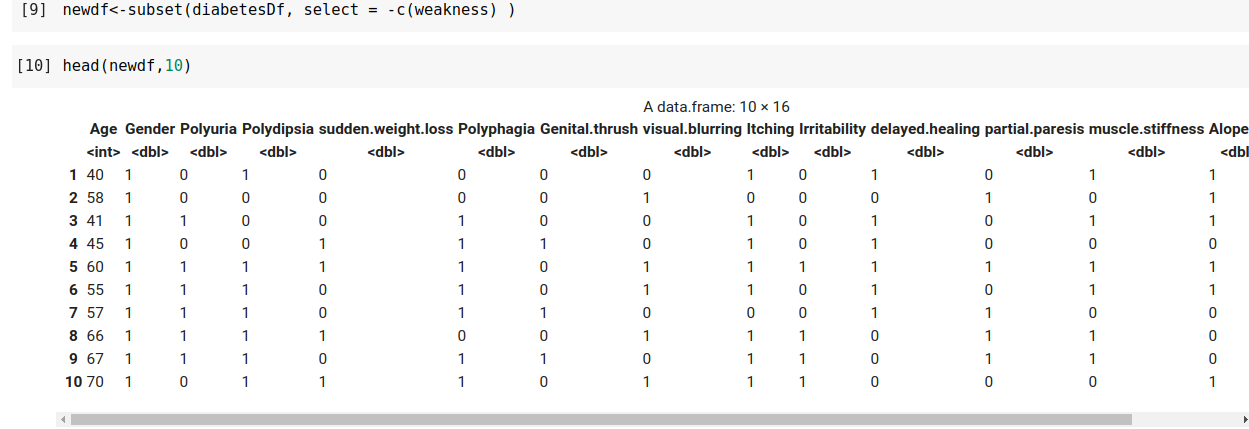


Fig: Code for removing “Weakness” attribute

**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.

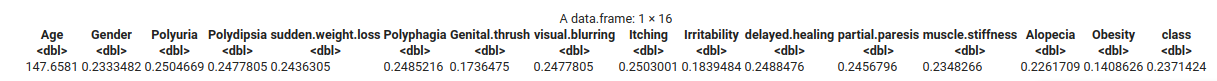


Fig: Variance obtained for each features

**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.

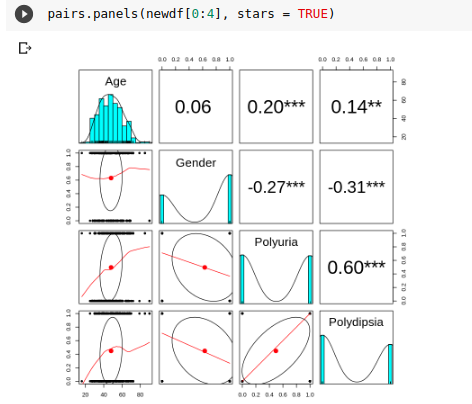


Fig: Scatter plot of the four columns of the dataset.

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.