Diabetes Prediction Project

Introduction:

The goal of this project is to predict the diabetes for PIIMA Women Indians by using attributes like Insulin, Blood Pressure, Pregnancy, Skin Thickness, Age etc., Target variable(Output) is considered as “0” and “1” as yes and no respectively.

One dataset in csv format is considered for this project to train and test the models. The dataset contains 9 variables which 768 observations. The dataset is segregated into 70% and 30% for training and testing.

The tool used is R along with its libraires and packages such as dplyr,e1071,c50,ggplot2 and Performance Analytics for data wrangling, data cleaning, data visualization, building predictive model and finding the correlation between the attributes.

Data Cleaning:

The data was checked for missing and duplicate values. The dataset was cleaned with no missing and unidentified values.

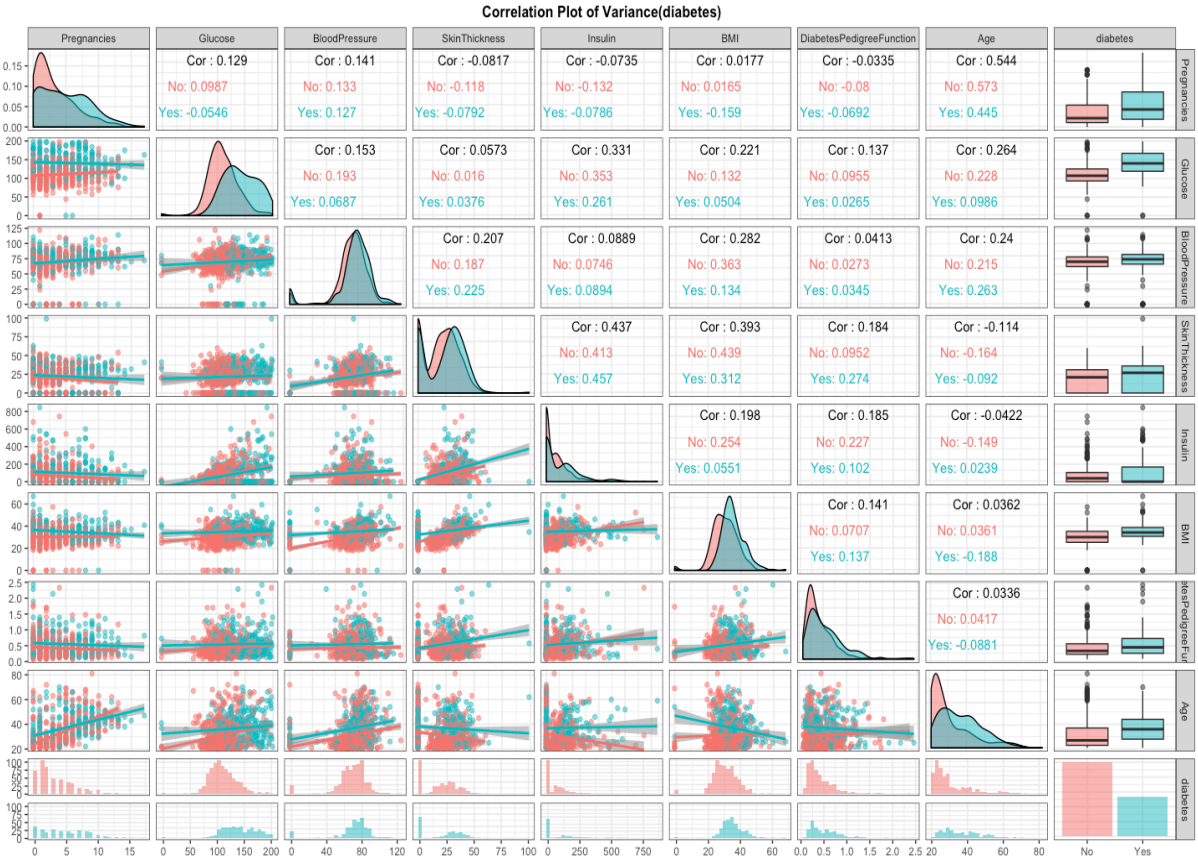
Exploratory data analysis(EDA):

Numerical and categorical variables were identified and summarised to get an overview of the dataset. In this case, 8 numerical variables and one categorical variables are considered. Numerical variables are Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function and Age.

Output attribute is the categorical variable. Here, Output is renamed as “diabetes” for better understanding.

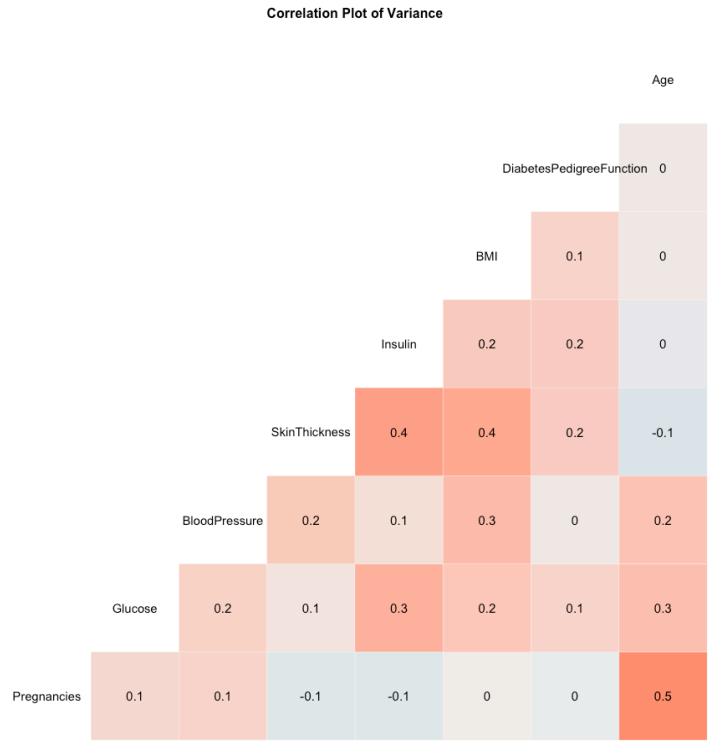
Relationship between variables

The below figure depicts explains the correlation value of each and every attributes in the dataset when compared to the target variable. This shows that every attributes has its own dependency with the target variable. Age, glucose and pregnancies are different according to the target variable outcome.



Correlation of Variance's Diabetes

The below figure of correlation plot shows the correlation variance. This shows that Insulin, BMI and Skin Thickness is 0.4. Age and Pregnancies has 0.5 which conveys that both variables are 50 % correlated to each other.



Model Selection and Evaluation:

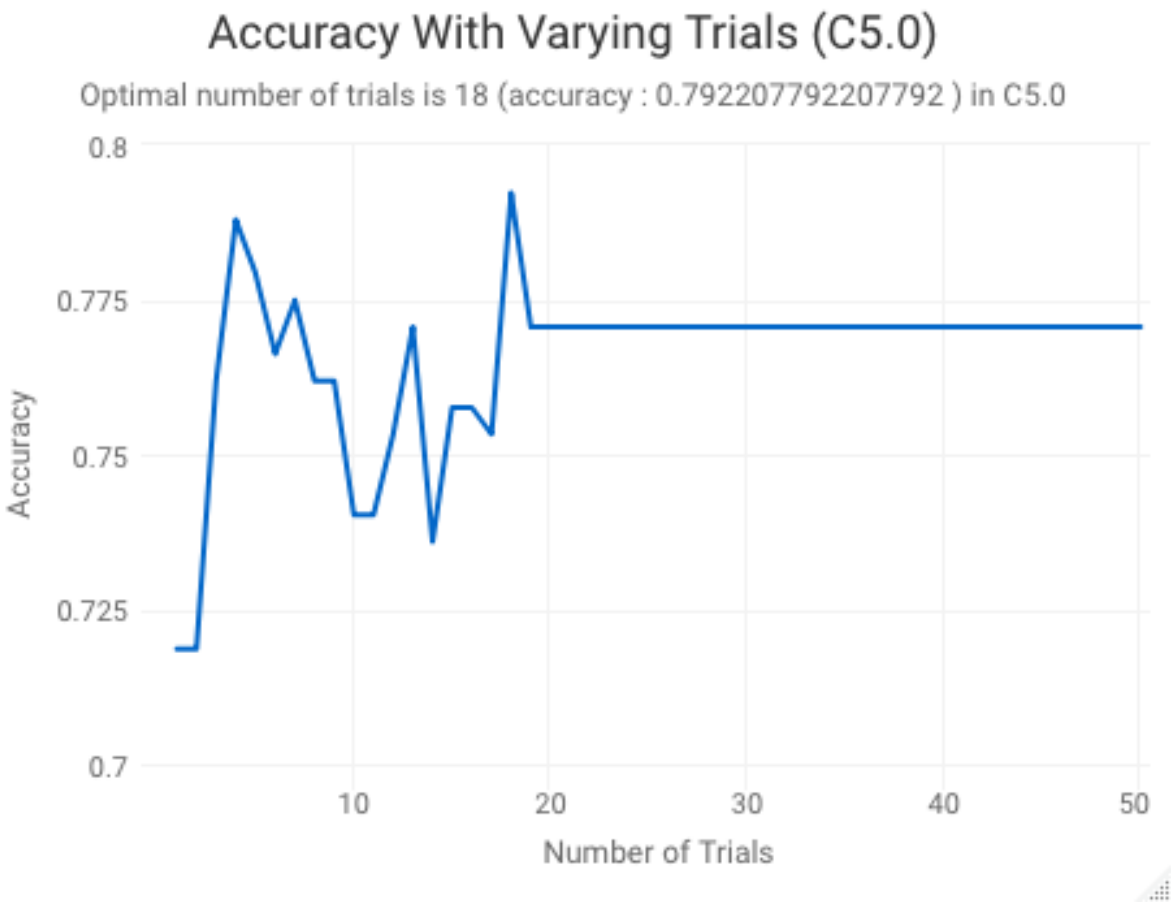
The two different types of classification algorithms selected are:

1. Support Vector Machine(SVM)
2. C5.0 (Decision Tree)

Accuracy and the confusion matrix is considered as the evaluation metric. The model with maximum accuracy and the best confusion matrix is selected as the best model for diabetes prediction.

Best Model:

The accuracy for C5.0 79% which is higher compared to SVM accuracy of 78%. Though, the accuracy is near to each other's prediction the confusion matrix of C5.0 is better. Hence C5.0 is considered as the best model for the diabetes prediction.



The above image shows the highest iteration which hits the best accuracy is 18. Hence the 18th iteration is fixed as the best optimization parameter for this model. The model was trained on the 7:3 of training and testing dataset.

Conclusion:

The C5.0 model is the best model for this prediction. The parameter can be tuned and optimised by using various trials of the decision tree algorithms. Though SVM is considered as the best for classification algorithm, in this case c5.0 with various parameter tuning is set to be best model for diabetes prediction with Age and pregnancies as the important attributes for the detection of diabetes.