EM-623 Data Science & Knowledge Discovery

Analysing Pima Indians Diabetes

Student – Smit Mehta

Semester – Spring 2018

Instructor – Dr. Carlo Lippizi

OVERVIEW

CRISP-DM METHODOLOGY :-

1) Project Goals
2) Business Understanding
3) Data Understanding
4) Data Preparation
5) Data Modelling
6) Evaluation Phase
7) Deployment Phase
8) Conclusion

Data Set Source - https://www.kaggle.com/uciml/pima-indians-diabetes-database

Project Goals

Our Goal is to predict if a patient has diabetes using the dataset given. Data set consist of 768 data of females at least 21 years old of PIMA Indian Heritage. Dataset contains 9 Columns which has Pregnancy, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age and Outcome.

Business Understanding

We need to perform various analysis on the data such as Correlation matrix and Decision Tree to predict if a patient has diabetes or not. We will analyse different attributes such as Pregnancy, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age and we will use Outcome to predict the symptoms of Diabetic Patient. This outcome will be used as a result to prevent diabetes by analysing symptoms.

Tools Used - Rattle

Data Understanding

After Exploring data, we found all the values as integer and there are no missing values in data as shown in Figure 1.

```
Data frame:crs$dataset[, c(crs$input, crs$risk, crs$target)]    768 observations and 9 variables    Maximum # NAs:0


                          Storage
Pregnancies               integer
Glucose                   integer
BloodPressure             integer
SkinThickness             integer
Insulin                   integer
BMI                        double
DiabetesPedigreeFunction  double
Age                       integer
Outcome                   integer

For the simple distribution tables below the 1st and 3rd Qu.
refer to the first and third quartiles, indicating that 25%
of the observations have values of that variable which are
less than or greater than (respectively) the value listed.
```

FIGURE-1

Further, Describing Min, Median and Max values of all the variable shown in Figure-2.

```
  Pregnancies        Glucose      BloodPressure    SkinThickness
 Min.    : 0.000   Min.    :  0.0   Min.    :  0.00   Min.    : 0.00
 1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00
 Median : 3.000   Median :117.0   Median : 72.00   Median :23.00
 Mean    : 3.845   Mean    :120.9   Mean    : 69.11   Mean    :20.54
 3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00
 Max.    :17.000   Max.    :199.0   Max.    :122.00   Max.    :99.00
    Insulin           BMI         DiabetesPedigreeFunction      Age
 Min.    :  0.0   Min.    : 0.00   Min.    :0.0780           Min.    :21.00
 1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437           1st Qu.:24.00
 Median : 30.5   Median :32.00   Median :0.3725           Median :29.00
 Mean    : 79.8   Mean    :31.99   Mean    :0.4719           Mean    :33.24
 3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262           3rd Qu.:41.00
 Max.    :846.0   Max.    :67.10   Max.    :2.4200           Max.    :81.00
    Outcome
 Min.    :0.000
 1st Qu.:0.000
 Median :0.000
 Mean    :0.349
 3rd Qu.:1.000
 Max.    :1.000

Rattle timestamp: 2018-05-08 18:20:47 Smit
========================================================================
```

FIGURE-2

Here,

Pregnancies – Number of pregnancies of females in dataset. (Average- 3.845)
Glucose – Glucose level of patient(Average- 120.9)
Blood Pressure - Average Blood Pressure is 69.11
Skin Thickness – Average Skin Thickness of patient is 20.54
Insulin – Average Insulin is Found to be 79.8
BMI – Body Mass Index (Average – 31.99)

Outcome – Number of patient with diabetes is 0.349.

In Outcome, 1 means patient are Diabetic.

Now, We would have closer look at the histogram of all the data for further understanding.

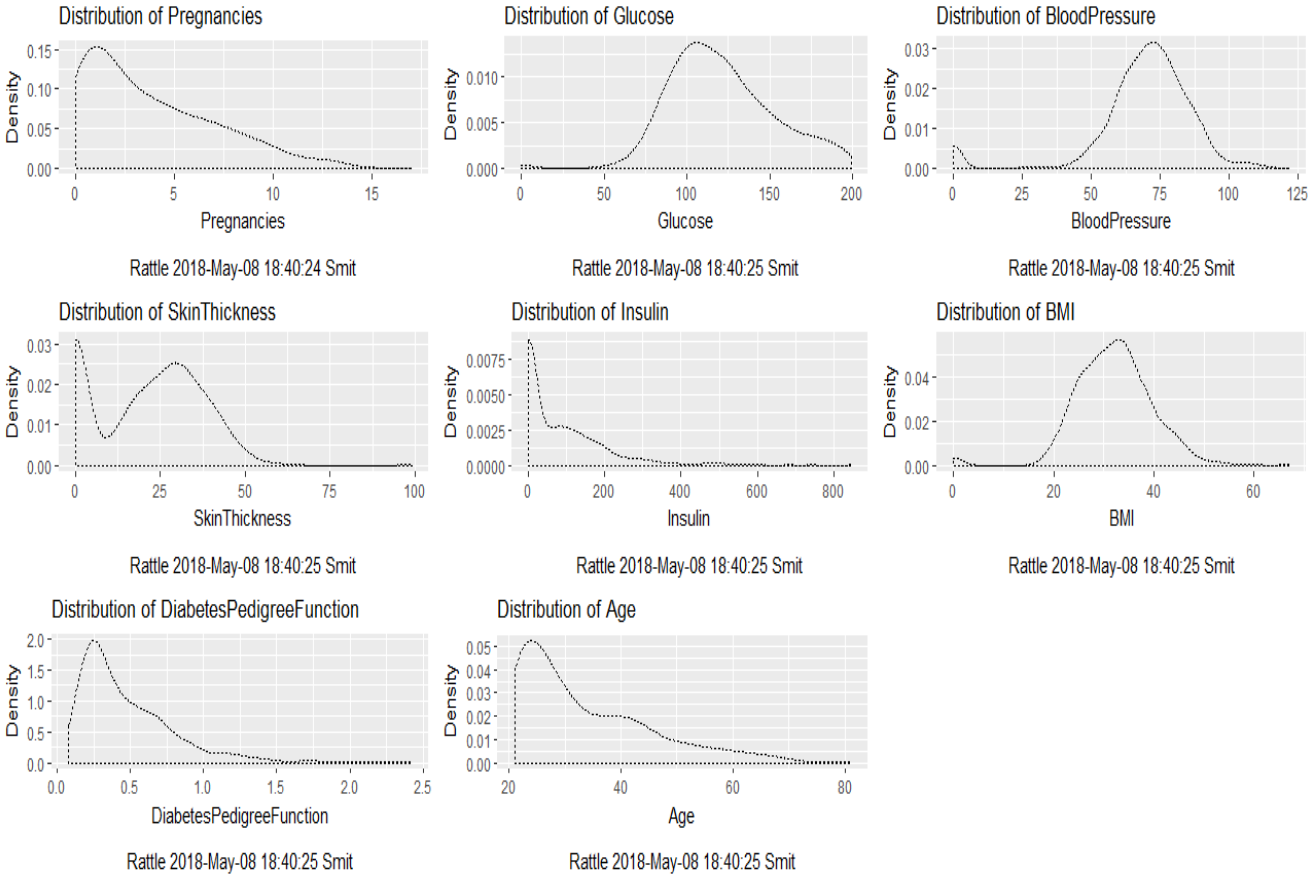


FIGURE-3

Performing Correlation Matrix to watch the correlation between different variable.

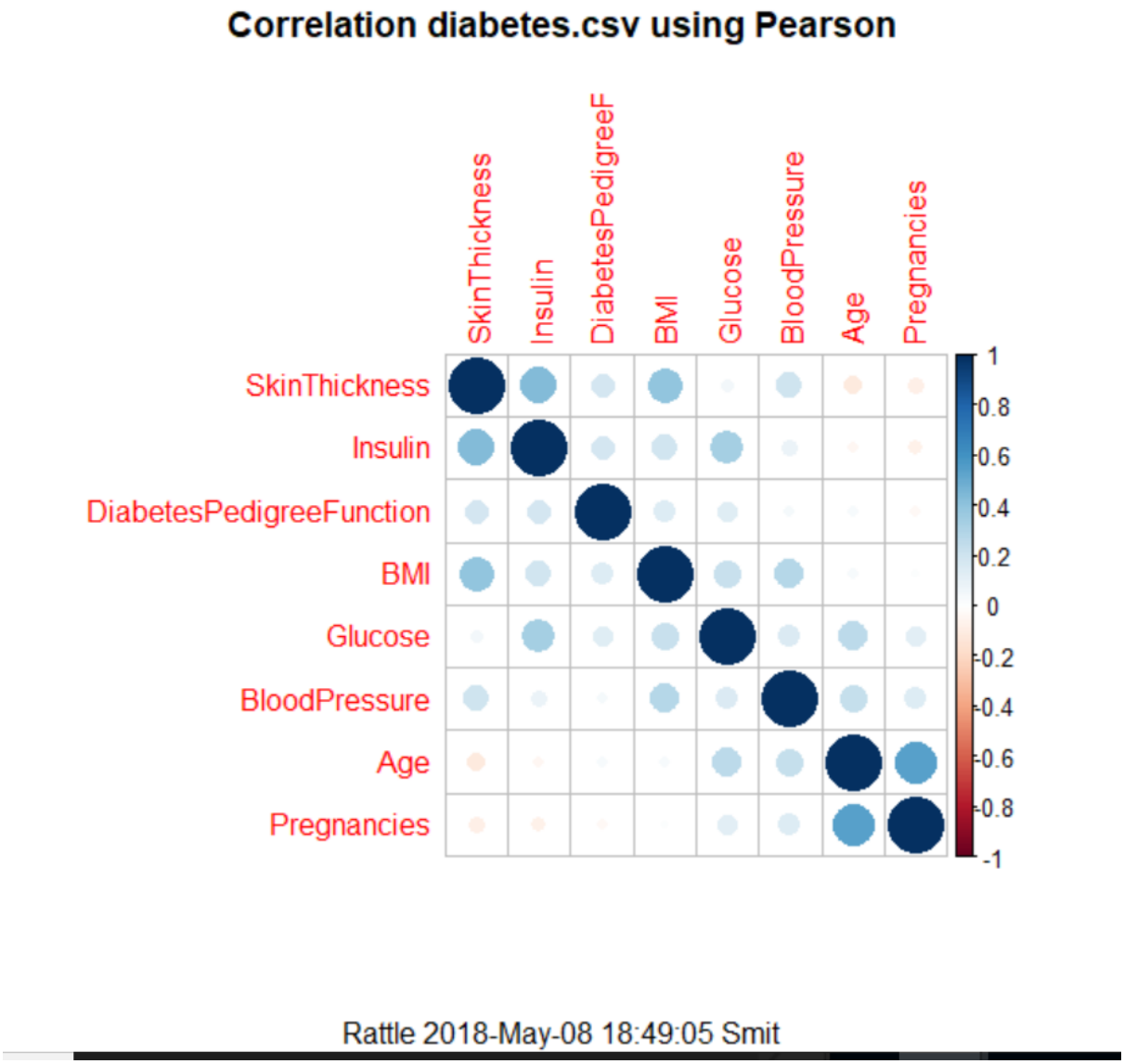


FIGURE-4

FIGURE-4 shows the correlation between different variables. It shows that Age and Pregnancies are correlated. Skin thickness has no correlation with Age and Pregnancies. But, Skin Thickness is correlated with Insulin. BMI(Body Mass Index) is also correlated with Skin Thickness. Glucose is also related with Skin Thickness. This means Higher Skin thickness can have higher Insulin.

Let’s get number of people having diabetes correlated with all other variables.

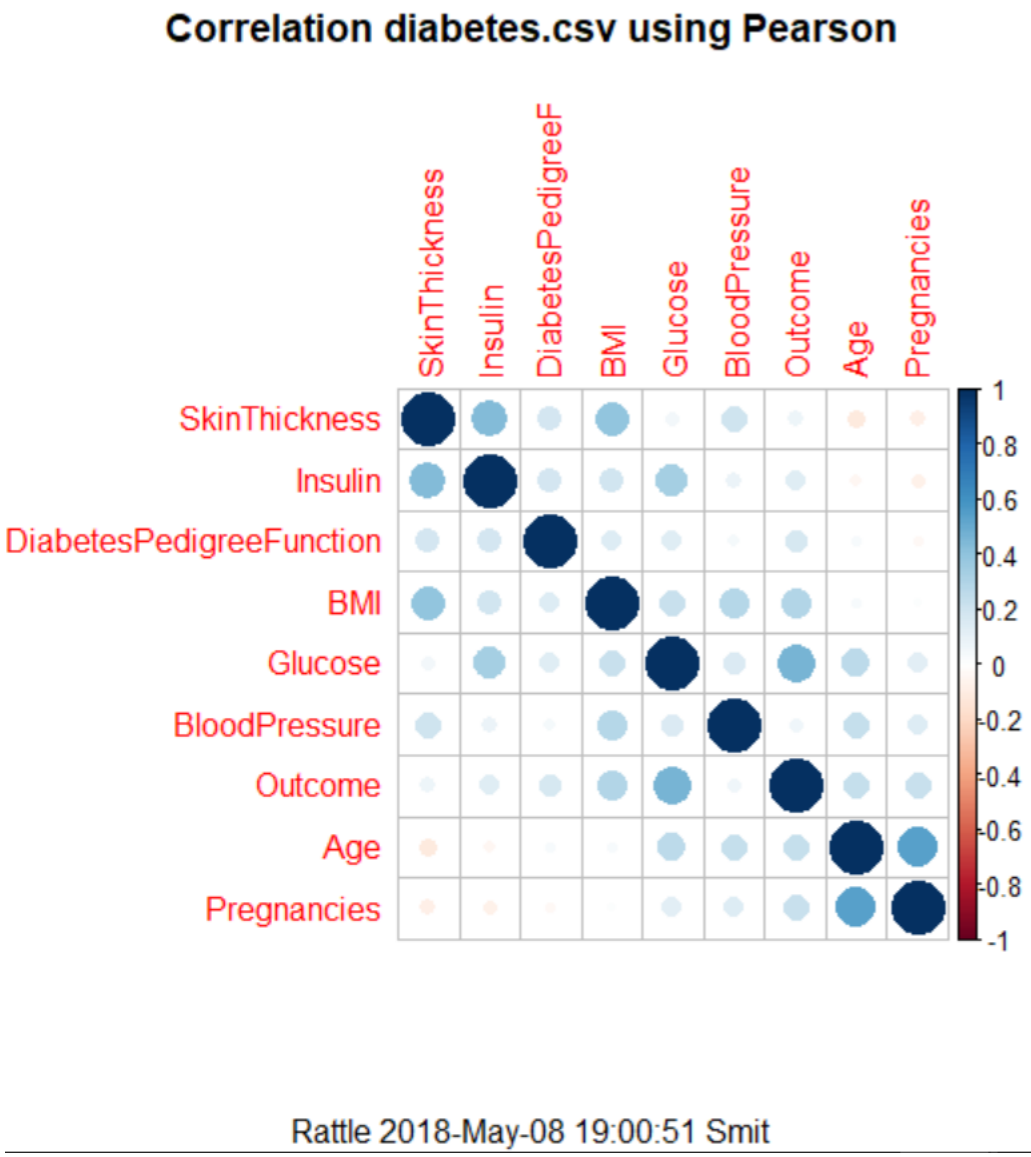


FIGURE-5

Applying Correlation Matrix with the Outcome variable, it is seen that Outcome is 46% correlated with the amount of glucose in patient as shown in Figure-6.

```
                          SkinThickness     Insulin DiabetesPedigreeFunction
SkinThickness                1.00000000  0.43678257                0.18392757
Insulin                      0.43678257  1.00000000                0.18507093
DiabetesPedigreeFunction     0.18392757  0.18507093                1.00000000
BMI                          0.39257320  0.19785906                0.14064695
Glucose                      0.05732789  0.33135711                0.13733730
BloodPressure                0.20737054  0.08893338                0.04126495
Outcome                      0.07475223  0.13054795                0.17384407
Age                         -0.11397026 -0.04216295                0.03356131
Pregnancies                 -0.08167177 -0.07353461               -0.03352267
                                  BMI    Glucose BloodPressure    Outcome
SkinThickness             0.39257320 0.05732789    0.20737054 0.07475223
Insulin                   0.19785906 0.33135711    0.08893338 0.13054795
DiabetesPedigreeFunction  0.14064695 0.13733730    0.04126495 0.17384407
BMI                       1.00000000 0.22107107    0.28180529 0.29269466
Glucose                   0.22107107 1.00000000    0.15258959 0.46658140
BloodPressure             0.28180529 0.15258959    1.00000000 0.06506836
Outcome                   0.29269466 0.46658140    0.06506836 1.00000000
Age                       0.03624187 0.26351432    0.23952795 0.23835598
Pregnancies               0.01768309 0.12945867    0.14128198 0.22189815
                                   Age Pregnancies
SkinThickness             -0.11397026 -0.08167177
Insulin                   -0.04216295 -0.07353461
DiabetesPedigreeFunction   0.03356131 -0.03352267
BMI                        0.03624187  0.01768309
Glucose                    0.26351432  0.12945867
BloodPressure              0.23952795  0.14128198
Outcome                    0.23835598  0.22189815
Age                        1.00000000  0.54434123
Pregnancies                0.54434123  1.00000000

Rattle timestamp: 2018-05-08 19:00:51 Smit
=====================================================
```

FIGURE-6

Transforming Outcome column in to indicator Variable by first transforming it to category and then to Indicator variable which gives us much more Understanding about the data.

We Start performing correlation matrix with the Indicator variable where TIN_TFC_OUTCOME_.0.0 indicates number of non-diabetic patient while TIN_TFC_OUTCOME_.0.1 indicate diabetic patient.

They both would contradict each other.

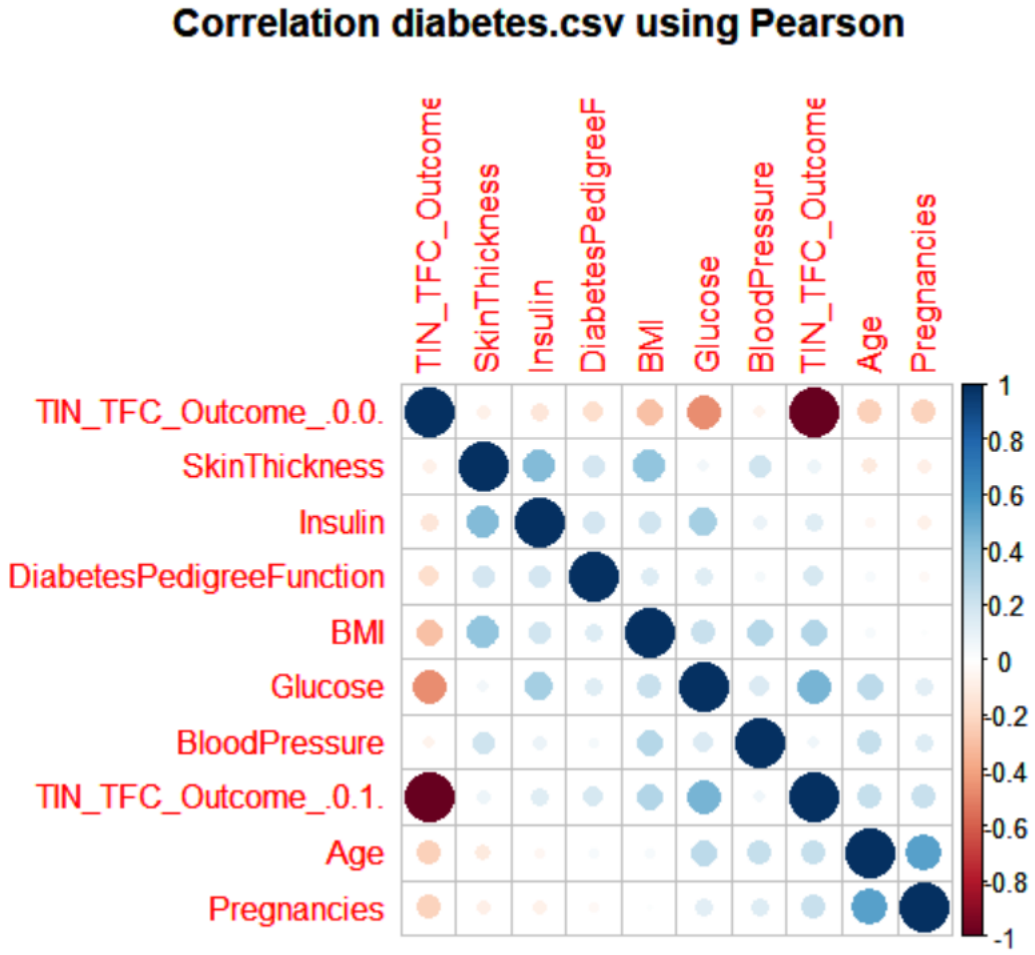


FIGURE--7

Figure-7 shows the correlation of People with diabetes and without diabetes with all other variables.

They are mostly contradict to each other.

Data Understanding

We start with eliminating observations which is not required for data analysis. But after analysing data we got to know that data is clean and it does not have any missing value.

After analysing correlation, it was found that Diabetes patient are highly correlated with glucose level.

Glucose level plays a major role and it is co-related to Insulin so, it clearly explains that Diabetes(outcome) is related to Insulin as well.

Data Modelling

Data is modelled to know the actual reason for diabetes patient. We will use Decision Tree Model to get the reason for the diabetes patient. ROC curves are also used to know whether Decision Tree model is accurate or not.

After Modelling data, we get the decision tree as shown in Figure 8.

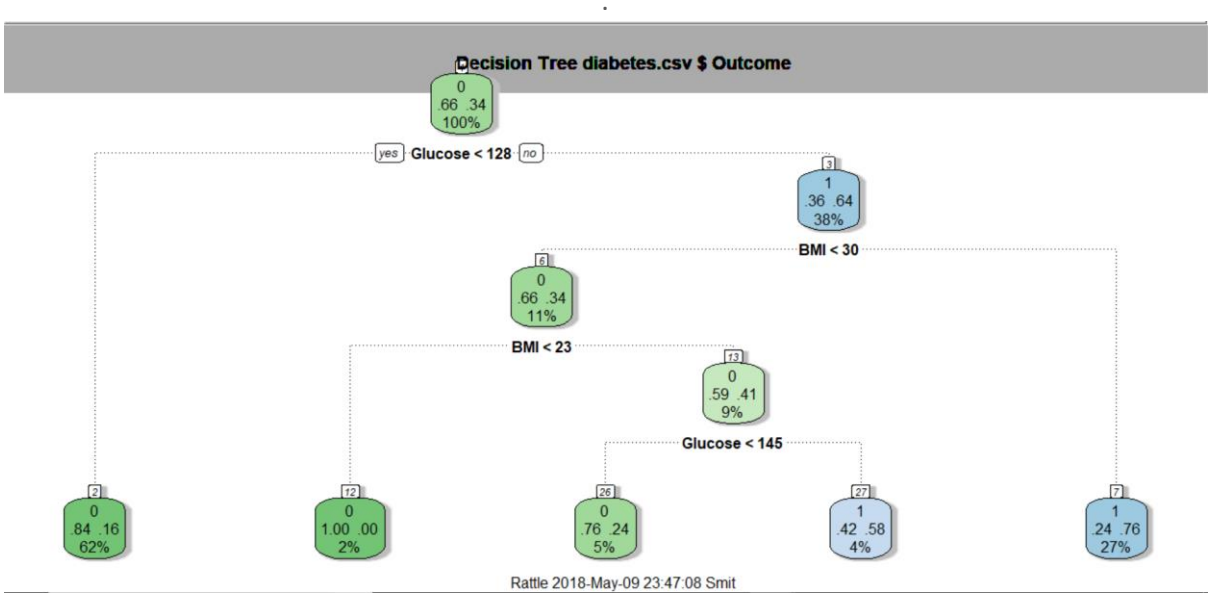
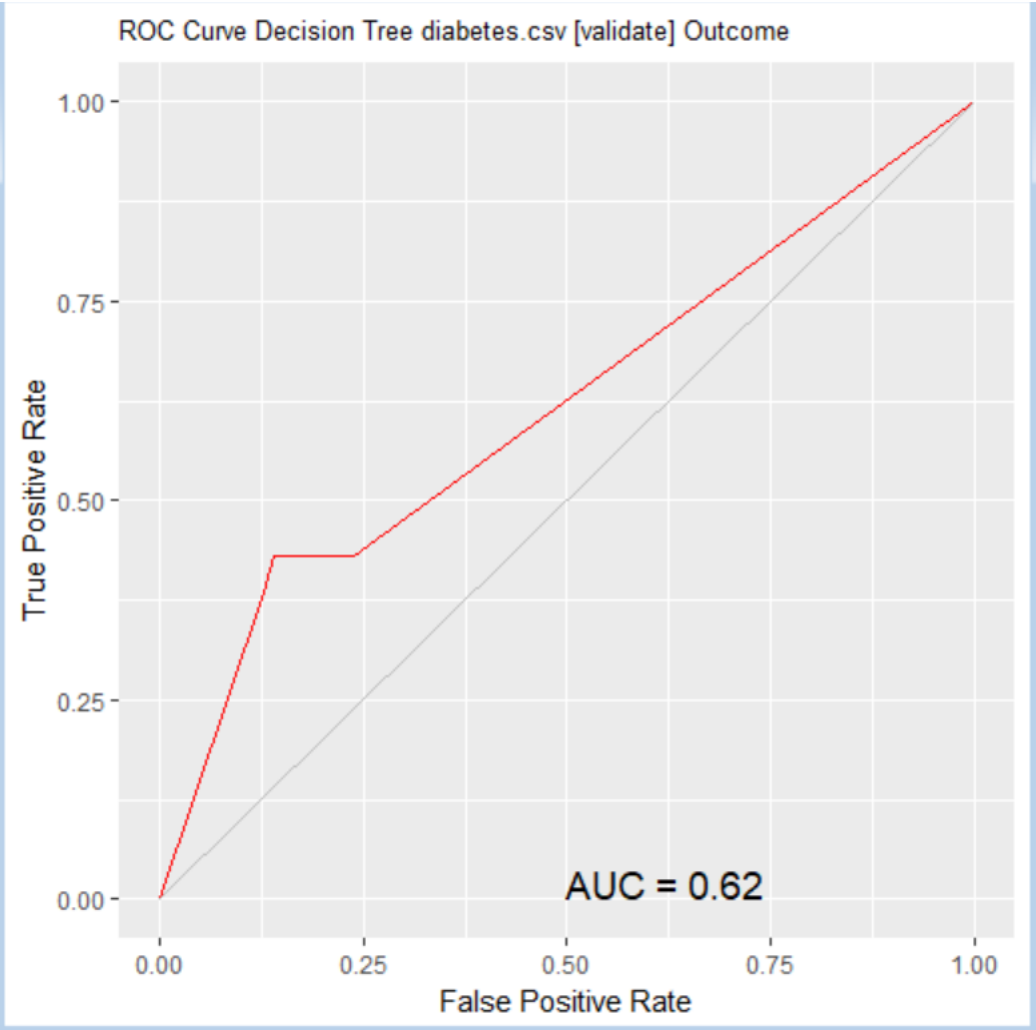


Figure-8

According to Figure-8 Glucose is the main variable that causes diabetes. People having glucose greater than 128 has probability of 38% of having Diabetes. People having Glucose greater than 128 and BMI > 30 has second highest probability of having 27%. So this Decision tree clearly shows the amount of patient getting diabetic.

Let’s see ROC curve to predict whether the decision tree is true or not.



ROC shows that decision tree is not very accurate as red line is not close to x and y axis but it has a better AUC = 0.62.

EVALUATION PHASE

Before deploying our model we need to check the quality of data. Number of data in the file is very limited for this vast analysis. We need to see that whether our analysis answered all questions and all the answer looking to the Business needs. More number of data can help to grow the Business understanding and help us to get more good result.

DEPLOYMENT PHASE

Right now, We are able to use all our modelling techniques to predict the right result of data. We can simply deploy the result for this limited amount of data and forecast future results using this data. The result of this data can be used to predict the future patients.

Conclusion

I came to conclusion that:

According to the results, Number of people suffering from Diabetes has high level of glucose and has BMI less than 30.