Prediction of Chronic Diabetes

Pinakin Nimavat
Computer Science
Illinois Institute of Technology
Chicago, IL, USA.
pnimavat@hawk.iit.edu

Parth Rathod
Computer Science
Illinois Institute of Technology
Chicago, IL, USA
prathod@hawk.iit.edu

Purvaj Desai
Computer Science
Illinois Institute of Technology
Chicago, IL, USA
pdesai28@hawk.iit.edu

Illinois Institute of Technology
CSP 571 Data Preparation and Analysis
Professor: Jawahar Panchal

Table of Contents

1. Abstract

Diabetes mellitus is one of the most serious noncommunicable illnesses affecting people today. Many countries are currently seeing a rapid increase in the number of diabetics among their citizens. According to a World Health Organization (WHO) research, this figure will have risen to 552 million by 2030, implying that one in every ten adults will have diabetes by 2030 if no substantial action is done. Diabetes was predicted to affect 9 percent of individuals aged 18 and up worldwide in 2014. Furthermore, the goal of this study is to forecast/predict Diabetes based on diagnostic measurements. Also, to explore the readily available dataset from Kaggle and try to extract meaningful and interesting insights from it. There are 768 observations and 9 variables in this dataset. All patients here, in particular are women over the age of 21.

Keywords: Diabetes, statistical models, predictions, Exploratory Data Analysis.

2. Introduction

Diabetes is clearly the leading cause of blindness, amputation, and renal failure. Diabetes awareness, along with insufficient access to health services and life-saving medications, can cause a slew of problems. It is a global issue with enormous personal, societal, and economic ramifications, affecting about 300 million people globally.

There are a few different types of diabetes:

- Type 1 diabetes is an autoimmune disease. The immune system attacks and destroys cells in the pancreas, where insulin is made. It's unclear what causes this attack. About 10 percent of people with diabetes have this type.

- Type 2 diabetes occurs when your body becomes resistant to insulin, and sugar builds up in your blood.

- Prediabetes occurs when your blood sugar is higher than normal, but it's not high enough for a diagnosis of type 2 diabetes.

- Gestational diabetes is high blood sugar during pregnancy. Insulin-blocking hormones produced by the placenta cause this type of diabetes.

In this project we are trying to extract some insights from a readily available dataset of PIMA INDIA DIABETES from Kaggle. We also aim to build a statistical model that could make some predictions about whether a patient has diabetes or not.

3. Proposed Methodology

First we tried to look for missing data so that there aren't any inconsistencies. We checked by performing following code:-

```
colSums(is.na(data))
```

```
##              Pregnancies                  Glucose            BloodPressure
##                        0                        0                        0
##            SkinThickness                  Insulin                      BMI
##                        0                        0                        0
## DiabetesPedigreeFunction                      Age                  Outcome
##                        0                        0                        0
```

After that we performed Basic Exploratory Data Analysis, here we tried to add new columns in order to have some more visualizations. We added an AGE_CAT column which buckets users into AGE_GROUP. We also added a Correlation Matrix in order to show which columns are relevant.

Then, we studied our data in order to answer our problem statement, which is discussed in next section (4.3). After cleaning and wrangling the data and exploring different aspects of the data, we implemented various models for predictive analysis. As we came across different models during the semester, we have implemented 4 models namely, Decision Tree Model, Random Forest Model, Support Vector Machine and Logistic Regression. In order to check the validity and accuracy of our models we have used Confusion matrix, F1 score, and Precision and recall measures.

4. Data

4.1 Data Properties

Data we used in the project is taken from Kaggle https://www.kaggle.com/uciml/pima-indians-diabetes-database. The dataset's goal is to diagnose whether or not a patient has diabetes based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

The datasets consists of several medical predictor variables and one target variable, Outcome. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

```
str(data)
```

```
## 'data.frame':    768 obs. of  9 variables:
##  $ Pregnancies             : int  6 1 8 1 0 5 3 10 2 8 ...
##  $ Glucose                 : int  148 85 183 89 137 116 78 115 197 125 ...
##  $ BloodPressure           : int  72 66 64 66 40 74 50 0 70 96 ...
##  $ SkinThickness           : int  35 29 0 23 35 0 32 0 45 0 ...
##  $ Insulin                 : int  0 0 0 94 168 0 88 0 543 0 ...
##  $ BMI                     : num  33.6 26.6 23.3 28.1 43.1 25.6 31 35.3 30.5 0 ...
##  $ DiabetesPedigreeFunction: num  0.627 0.351 0.672 0.167 2.288 ...
##  $ Age                     : int  50 31 32 21 33 30 26 29 53 54 ...
##  $ Outcome                 : int  1 0 1 0 1 0 1 0 1 1 ...
```

4.2 Data Preprocessing, Cleaning and Wrangling

As part of our preprocessing steps, first we looked at the structure of the data.

```
str(data)
```

```
## 'data.frame':    768 obs. of  9 variables:
##  $ Pregnancies             : int  6 1 8 1 0 5 3 10 2 8 ...
##  $ Glucose                 : int  148 85 183 89 137 116 78 115 197 125 ...
##  $ BloodPressure           : int  72 66 64 66 40 74 50 0 70 96 ...
##  $ SkinThickness           : int  35 29 0 23 35 0 32 0 45 0 ...
##  $ Insulin                 : int  0 0 0 94 168 0 88 0 543 0 ...
##  $ BMI                     : num  33.6 26.6 23.3 28.1 43.1 25.6 31 35.3 30.5 0 ...
##  $ DiabetesPedigreeFunction: num  0.627 0.351 0.672 0.167 2.288 ...
##  $ Age                     : int  50 31 32 21 33 30 26 29 53 54 ...
##  $ Outcome                 : int  1 0 1 0 1 0 1 0 1 1 ...
```

We can see that there are no missing values in the dataset.

```
data$Outcome <- factor(make.names(data$Outcome))
biological_data <- data[,setdiff(names(data), c('Outcome', 'Pregnancies'))]
features_miss_num <- apply(biological_data, 2, function(x) sum(x<=0))
features_miss <- names(biological_data)[ features_miss_num > 0]
features_miss_num
```

```
##                 Glucose           BloodPressure           SkinThickness
##                       5                      35                     227
##                 Insulin                     BMI DiabetesPedigreeFunction
##                     374                      11                       0
##                     Age
##                       0
```

Upon careful consideration we noticed that many of the biological measurements such as glucose, blood pressure, BMI were 0.00 which do not make any sense so we replaced these values by taking the median of the dataset.

4.3 Problem Statement

What this project seeks to address:

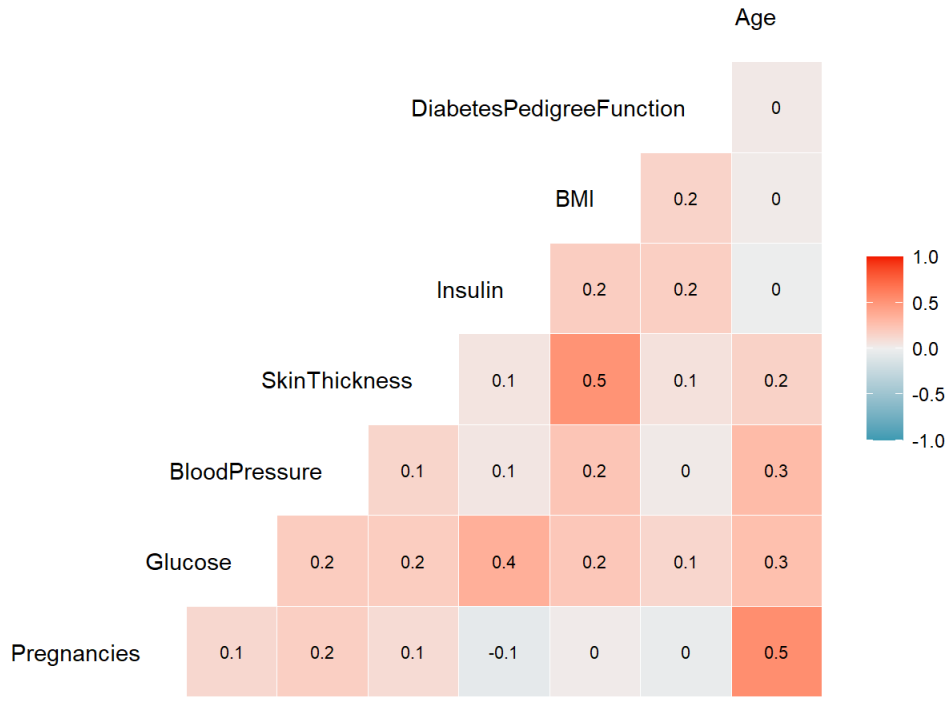
- Which features are most relevant?
- Is there any correlation between any features?
- Cross Model Performance Comparison.
- What is the effect of Data Cleaning on Model's Performance?
- What factors increase the chance of having Diabetes?
- Can we build a model to predict a customer's diabetic condition with sufficient accuracy?

- Can we rely on available customer data to predict chronic diseases?

5. EXPLORATORY DATA ANALYSIS

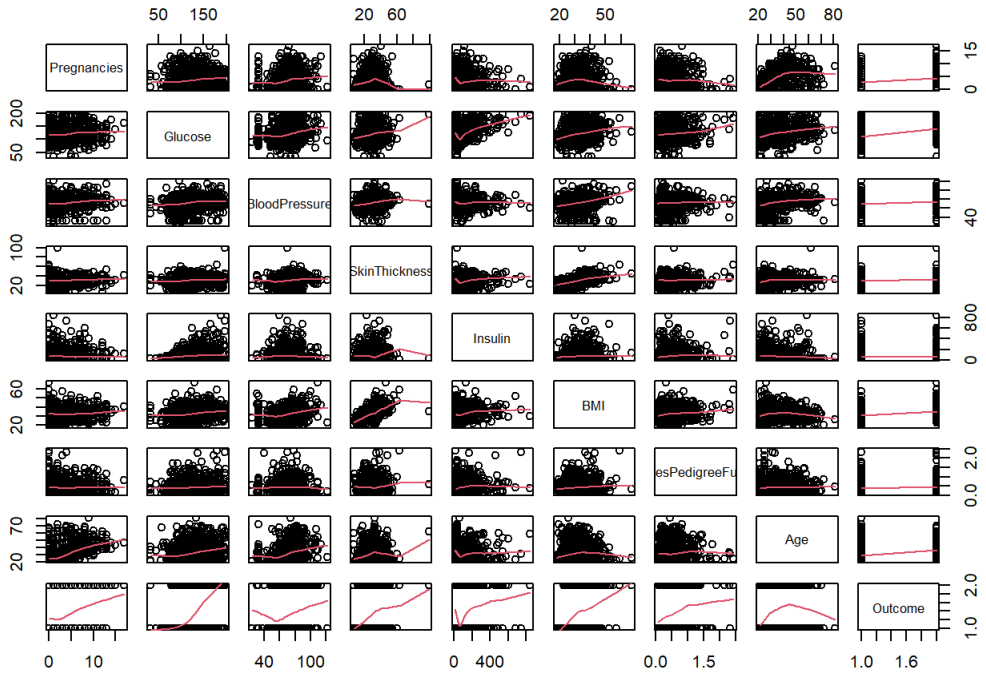
Correlation Visualization

First we tried to implement a correlation visualization and it states that there are no strong visualization among predictor variables

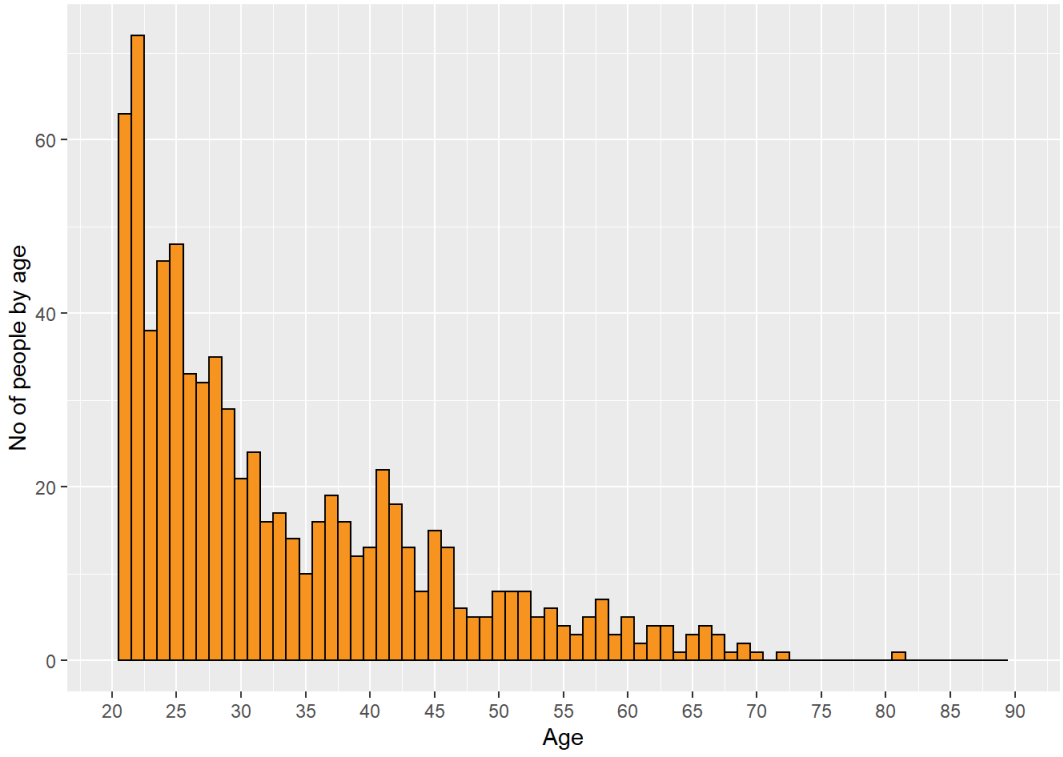


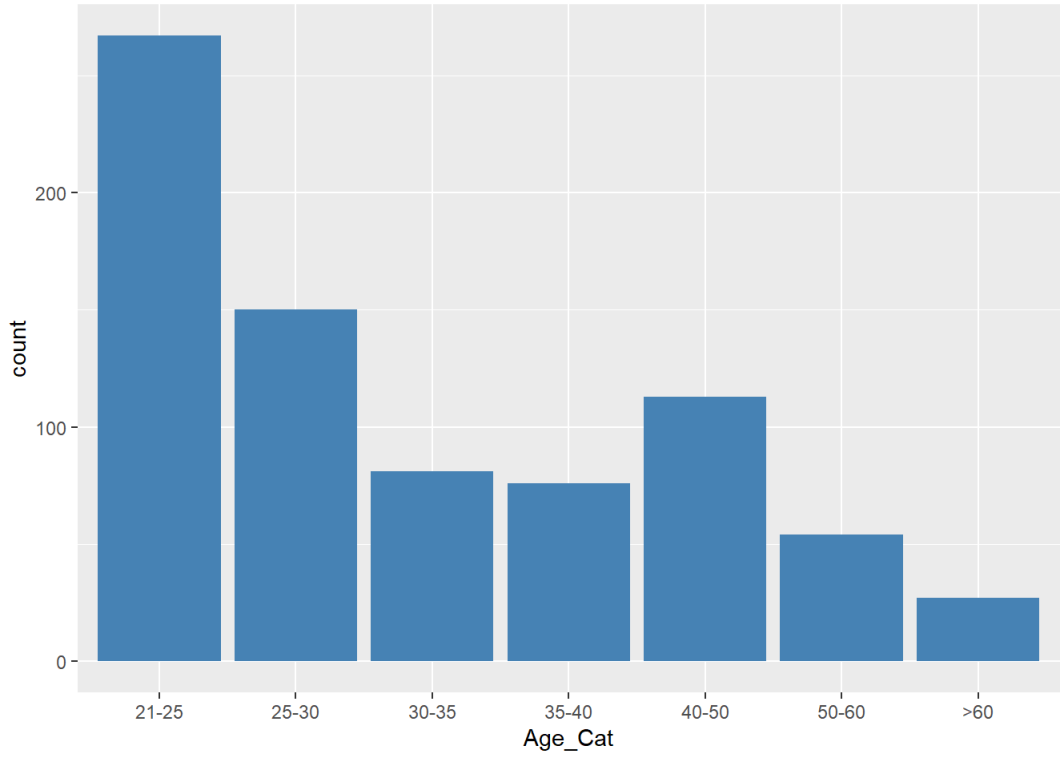
Scatter Plot

We created a scatter plot corresponding to each data frame. It gives a pairwise relationship between different variables in a dataset.

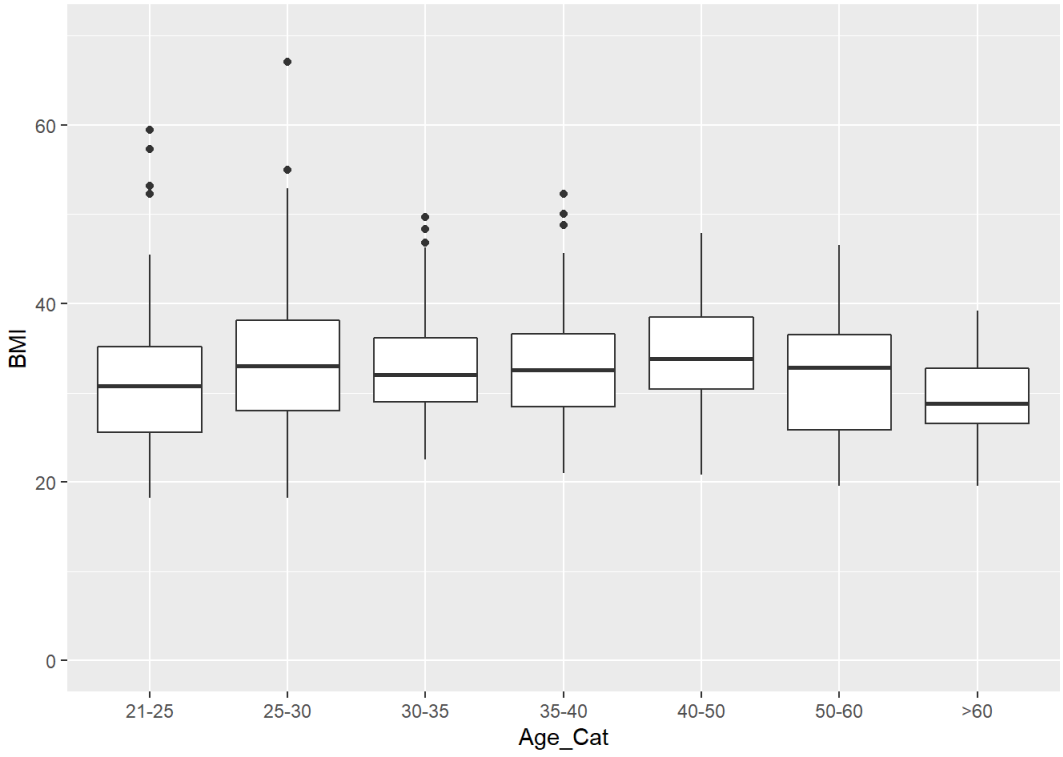


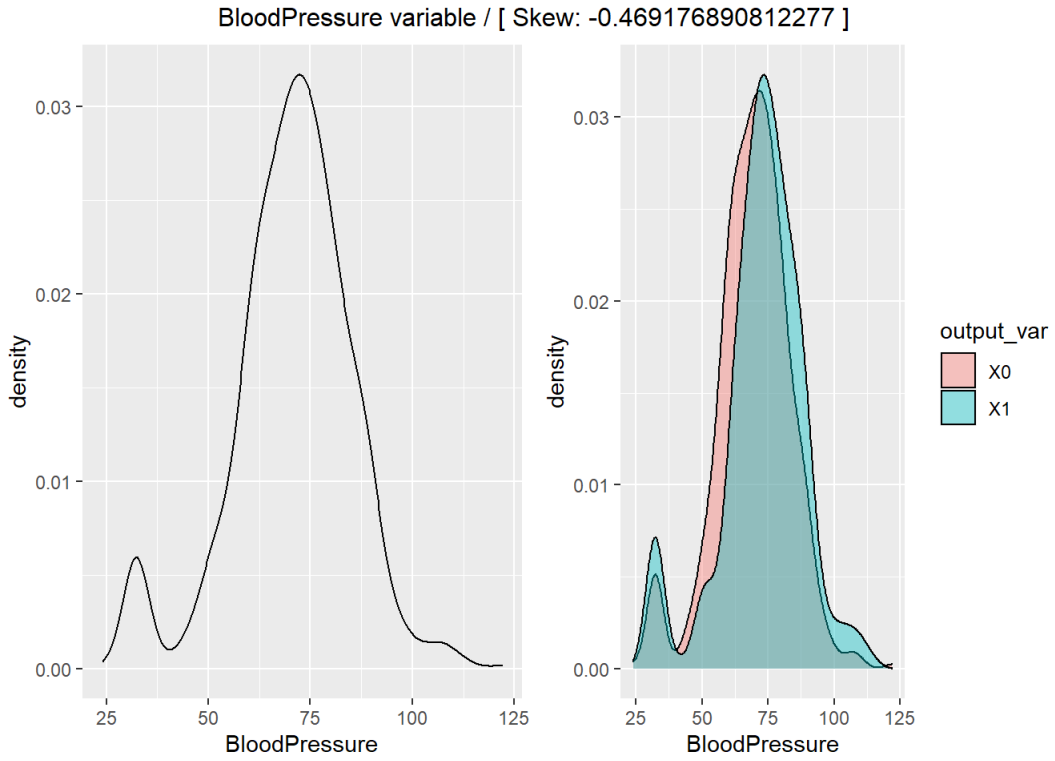
We also created a Histogram for Age vs No of People by Age, and it showed a lot of dataset consists of age between 20-30 age group.

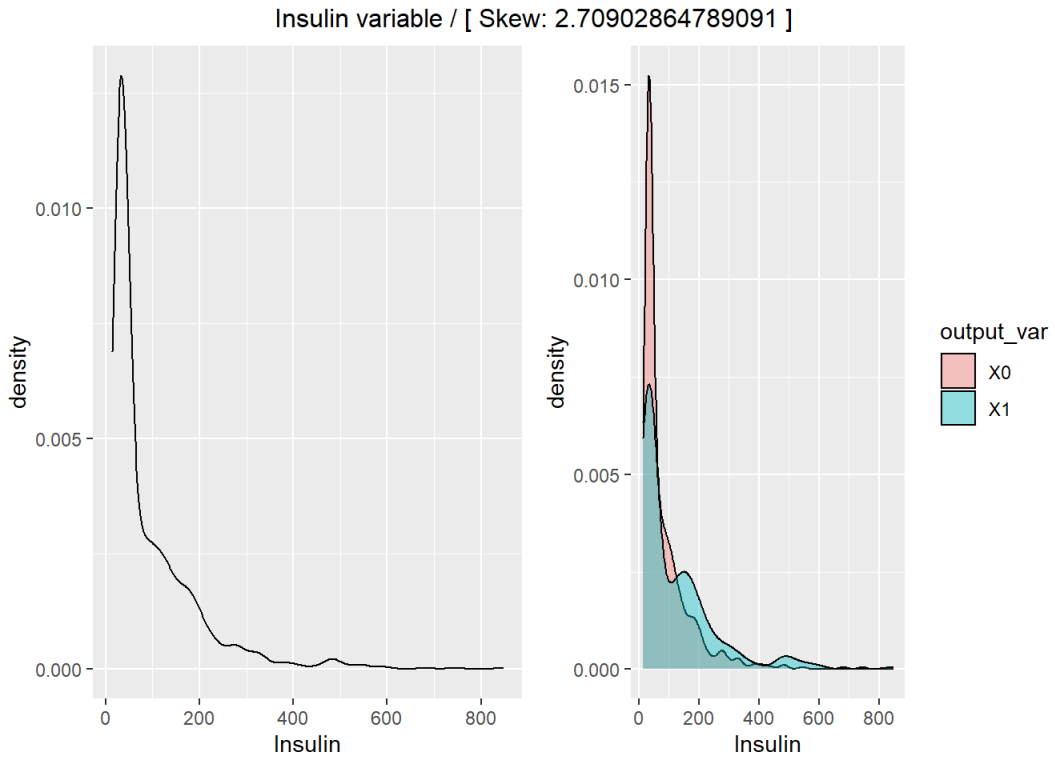




Box Plot of BMI vs Age_Cat





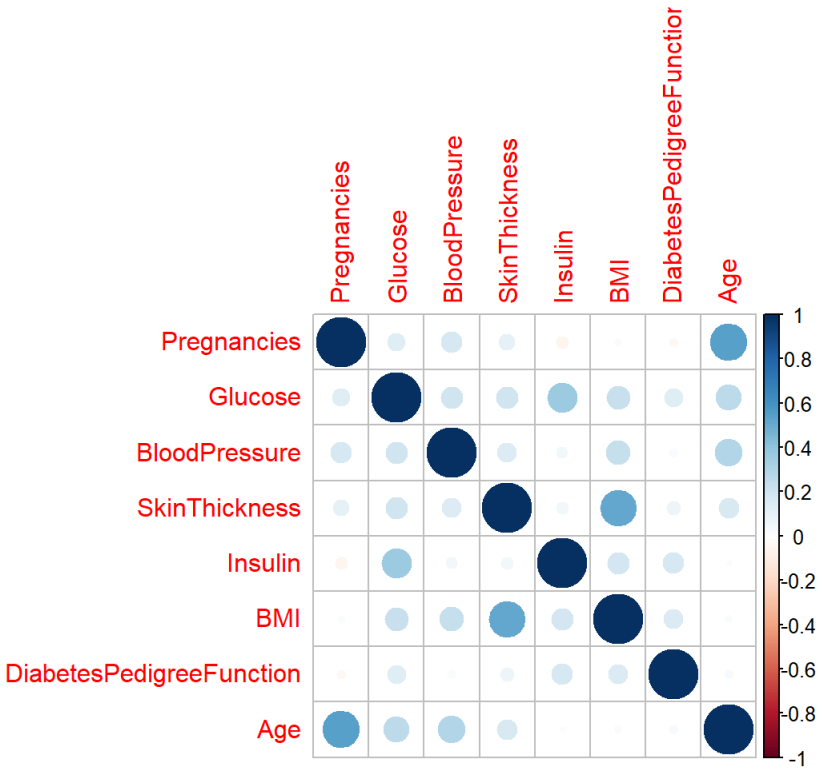


We performed Univariable analysis on each variable to find density and skewness. We were able to find that insulin, diabetesPedigreeFunction and age are highly skewed variables and blood pressure is a highly left skewed variable.

6. MODELING AND ANALYSIS

6.1 Correlations

A correlation plot among each variable which shows there are no strong correlation among themselves.



6.2 Train Test Split

In order to train our models and further test for measuring the accuracy, we first split dataset into 70-30% for training set and testing set respectively. This dataset contains all columns.

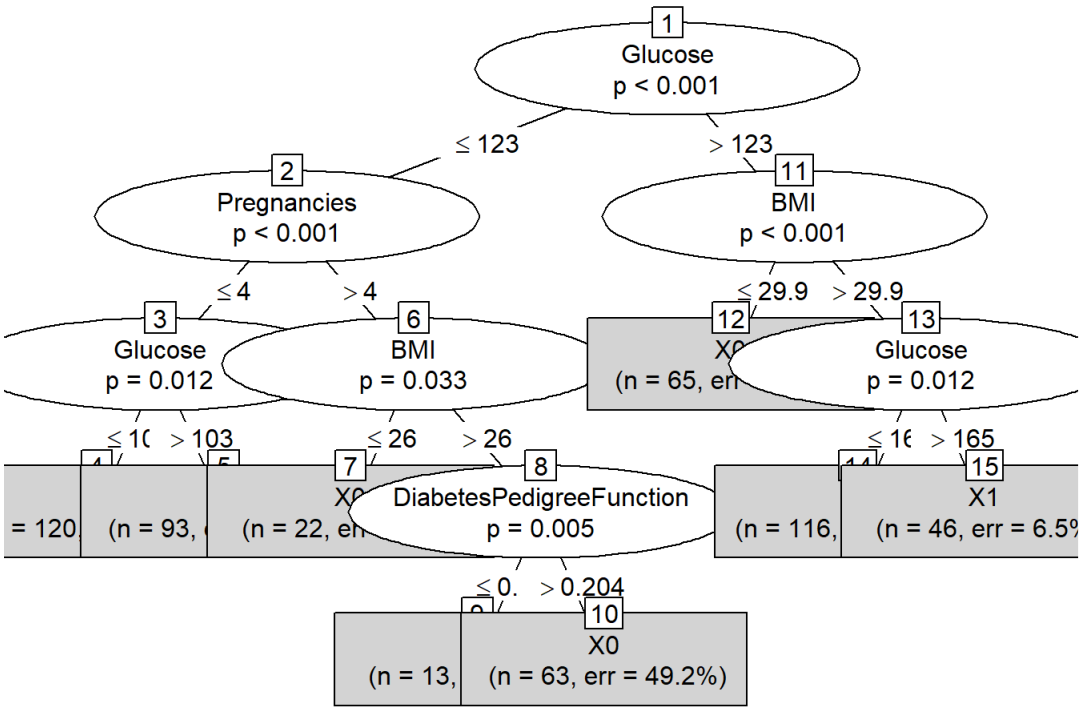
```
data = subset(data, select=-c(Age_Cat))
```

```
set.seed(7)
dindex <- createDataPartition(data$Outcome, p=0.7, list=FALSE)
data_train <- data[dindex,]
data_test <- data[-dindex,]
```

6.3 Modeling

6.3.1 Decision Tree

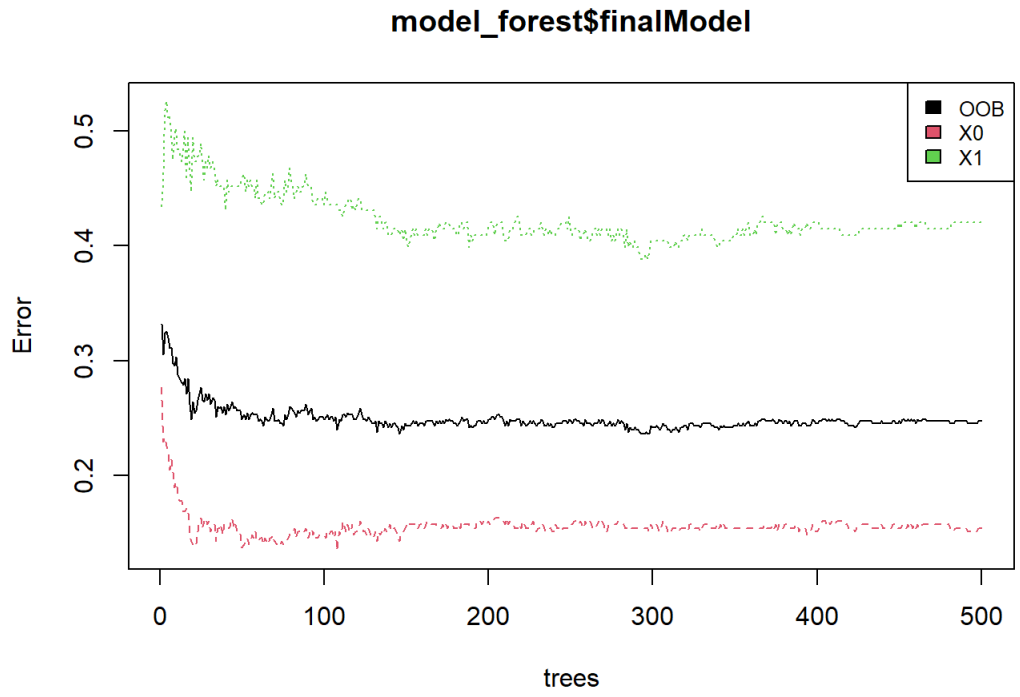
We created a Decision Tree model first with a train test split of 70% and 30%. We got accuracy of 73% with 95% Confidence Interval between (67% , 79%).



we can see the number of nodes and its distribution. In which :
[1] is root node
[2],[3],[6],[8], [11], and [13] are internal nodes or branch. Internal nodes shown by arrow pointing to/from them.
[4],[5],[7],[9], [10], [12], [14], [15] are leaf nodes or leaf.

6.3.2 Random Forest

We created Random Forest Model by using 5 Fold Cross Validation Technique, we also plotted an OOB error graph



Based on visualization above comparison of OOB and targeted variable. It depicts that from tree numbers around 100 the error of the model has been better, yet we can still use more than 400 trees to reduce our OOB.

We plotted importance variable required for prediction

```
## rf variable importance
##
##                           Overall
## Glucose                   100.000
## BMI                        51.783
## DiabetesPedigreeFunction  29.490
## Age                        27.784
## BloodPressure              7.253
## Pregnancies                5.932
## SkinThickness              4.346
## Insulin                    0.000
```

Based on result above, we know that glucose rate has the highest impact to the result while the other variables are only 50% or less than it.

The Accuracy of Random Forest Model is 74%

```
## Confusion Matrix and Statistics
##
##
## predict_forest  X0  X1
##             X0 124  33
##             X1  26  47
##
##                Accuracy : 0.7435
##                  95% CI : (0.6819, 0.7986)
##     No Information Rate : 0.6522
##     P-Value [Acc > NIR] : 0.001872
##
##                   Kappa : 0.4228
##
##  Mcnemar's Test P-Value : 0.434724
##
##             Sensitivity : 0.8267
##             Specificity : 0.5875
##          Pos Pred Value : 0.7898
##          Neg Pred Value : 0.6438
##              Prevalence : 0.6522
##          Detection Rate : 0.5391
##    Detection Prevalence : 0.6826
##       Balanced Accuracy : 0.7071
##
##        'Positive' Class : X0
##
```

6.3.3 Support Vector Machine

We also created SVM as one of our models. We have used the sigmoid as the kernel with default values of gamma and the cost. We can see the confusion matrix below.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  X0  X1
##         X0 126  33
##         X1  24  47
##
##                Accuracy : 0.7522
##                  95% CI : (0.6912, 0.8066)
##     No Information Rate : 0.6522
##     P-Value [Acc > NIR] : 0.0007075
##
##                   Kappa : 0.439
##
##  Mcnemar's Test P-Value : 0.2893148
##
##             Sensitivity : 0.5875
##             Specificity : 0.8400
##          Pos Pred Value : 0.6620
##          Neg Pred Value : 0.7925
##              Prevalence : 0.3478
##          Detection Rate : 0.2043
##    Detection Prevalence : 0.3087
##       Balanced Accuracy : 0.7137
##
##        'Positive' Class : X1
##
```

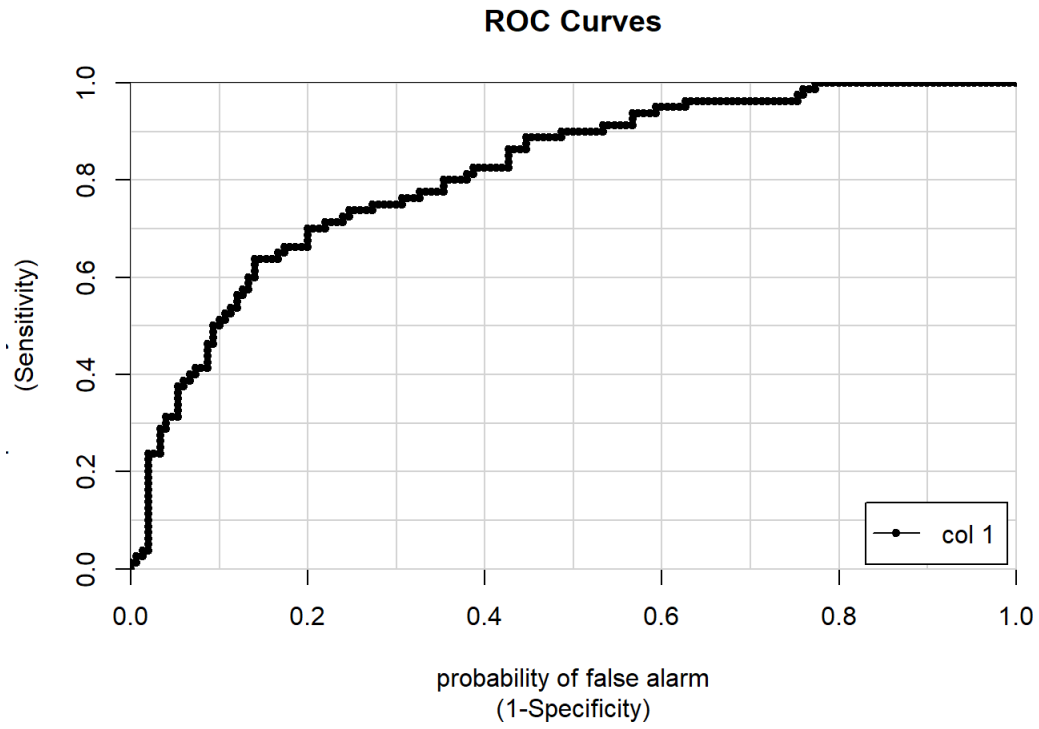
6.3.4 Logistic Regression

As part of modeling we created a Logistic Regression model. We trained the model using 70% of the dataset. We got an accuracy of 78.26%.

AUC has an 85.95% value.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  X0  X1
##         X0 129  29
##         X1  21  51
##
##                Accuracy : 0.7826
##                  95% CI : (0.7236, 0.8341)
##     No Information Rate : 0.6522
##     P-Value [Acc > NIR] : 1.156e-05
##
##                   Kappa : 0.5094
##
##  Mcnemar's Test P-Value : 0.3222
##
##             Sensitivity : 0.6375
##             Specificity : 0.8600
##          Pos Pred Value : 0.7083
##          Neg Pred Value : 0.8165
##              Prevalence : 0.3478
##          Detection Rate : 0.2217
##    Detection Prevalence : 0.3130
##       Balanced Accuracy : 0.7488
##
##        'Positive' Class : X1
##
```

Below is the ROC curve. This means that our model has the ability to correctly distinguish between the correct labels. Here 0 means the person is non-diabetic and 1 means the person has been predicted positive for diabetes.



6.4 Result Analysis

We see that logistic regression is the best model with accuracy of 78%. This accuracy is low since the dataset is imbalanced and we have less number of Positive Classes which is why it becomes difficult for the model to predict the classes.

Below table shows summary of all the models.

Sr No.	Model	Accuracy	F1 Score	Precision	Recall
1	Decision Tree Model	73%	80%	77%	83%
2	Random Forest	74%	80%	78%	82%
3	Support Vector Machine	75%	62%	66%	58%
4	Logistic Regression	78%	67%	70%	63%

7. CONCLUSION

In this project, we have successfully explored the diabetes dataset by extracting various insights. In addition to that we have also implemented and compared 4 models to perform prediction. The results analysis showed that Logistic Regression is the most accurate technique to predict diabetes with 78% accuracy as compared to other techniques. Logistic Regression clearly identified negative classes for diabetes as better predictors. When compared to other models, it is considered to the best and most suitable model for this dataset.

REFERENCES

[1] Asish Satpathya*, Satyajit Beharib
https://arxiv.org/pdf/2109.01863.pdf
[2] Félix Tena, Oscar Garnica, Juan Lanchares, J. Ignacio Hidalgo ∗†
https://arxiv.org/pdf/2109.02178.pdf
[3] Mingcheng Chen1∗ , Zhenghui Wang1∗ , Zhiyun Zhao2#, Weinan Zhang1# Xiawei Guo3, Jian Shen1 , Yanru Qu1 , Jieli Lu2 , Min Xu2 , Yu Xu2 Tiange Wang2 , Mian Li2 , Wei-Wei Tu3 , Yong Yu1 , Yufang Bi2 , Weiqing Wang2 , Guang Ning2
https://arxiv.org/pdf/2108.07107.pdf
[4] Talha Mahboob Alam, Muhammad Atif Iqbala, Yasir Alia, Abdul Wahabb, Safdar Ijazb, Talha Imtiaz Baigb, Ayaz Hussainc, Muhammad Awais Malikb, Muhammad Mehdi Razab, Salman Ibrarb, Zunish Abbasd
https://www.sciencedirect.com/science/article/pii/S2352914819300176
[5] Hang Lai, Huaxiong Huang, Karim Keshavjee, Aziz Guergachi & Xin Gao
https://bmcendocrdisord.biomedcentral.com/articles/10.1186/s12902-019-0436-6