Mini Project Report On

“Diabetes Prediction Using Data Mining”

BACHELOR OF COMPUTER ENGINEERING

SUBMITTED BY

Dnyanal Kumar Vedpathak	BCO18F67
Meghsham Vinayak Kapure	BCO19D23
Soham Santosh Solat	BCO18F63

UNDER THE GUIDANCE OF

Prof. S.B.Shirke

DEPARTMENT OF COMPUTER ENGINEERING

SAVITRIBAI PHULE PUNE UNIVERSITY

Batch of 2021-22

CERTIFICATE

This is to certify that the seminar report entitled

“Diabetes Prediction Using Data Mining”

Submitted by

Dnyanal Kumar Vedpathak	BCO18F67
Meghsham Vinayak Kapure	BCO19D23
Soham Santosh Solat	BCO18F63

Are bonafide student of this institute and the work has been carried out by her under the supervision of **Prof.S.B.Shirke** and it is approved for the partial fulfilment of the requirement of Savitribai Phule Pune University Computer Engineering.

Prof.S.B.Shirke
Guide

Prof.B.D.Thorat
HOD

Dr.S.B.Patil
Principal

Shri Chhatrapati Shivajiraje college of Engineering Pune

Place: Dhangwadi

Date:

ACKNOWLEDGEMENT

We extend our sincere and heartfelt thanks to our esteemed guide, **Prof.S.B.Shirke** for this exemplary guidance, monitoring and constant encouragement throughout the course at crucial junctures and for showing us the write way.

We would like to extends thanks to our respected HEAD of the division **Prof. B.D.Thorat** for following us to use to the facilities available. We would like to thanks faculty members also Last but not the least, We would like to thanks our friends and family for the support and encouragement they have give us during the course of our work

ABSTRACT

Diabetes is one of the major international health problems. World Health Organization reports says that around 422 million people have diabetes worldwide. Data mining plays a huge role in predicting diabetes in the healthcare industry. There are many algorithms developed for prediction of diabetes. But most of the algorithms failed in case of the accuracy estimation.

Introduction

Objective:

•To predict diabetes in healthcare industry using data mining.

Project Overview:

Diabetes is one of the major international health problems. World Health Organization reports says that around 422 million people have diabetes worldwide. Data mining plays a huge role in predicting diabetes in the healthcare industry. There are many algorithms developed for prediction of diabetes. But most of the algorithms failed in case of the accuracy estimation. Also, there is a need to automate the overall process of diabetes prediction. This automation of diabetic database helps in identification of impact of diabetes on various human organs. More the accuracy of prediction, more the chances of accurate severity estimation. Therefore this project concentrated on providing different prediction methods of diabetes.

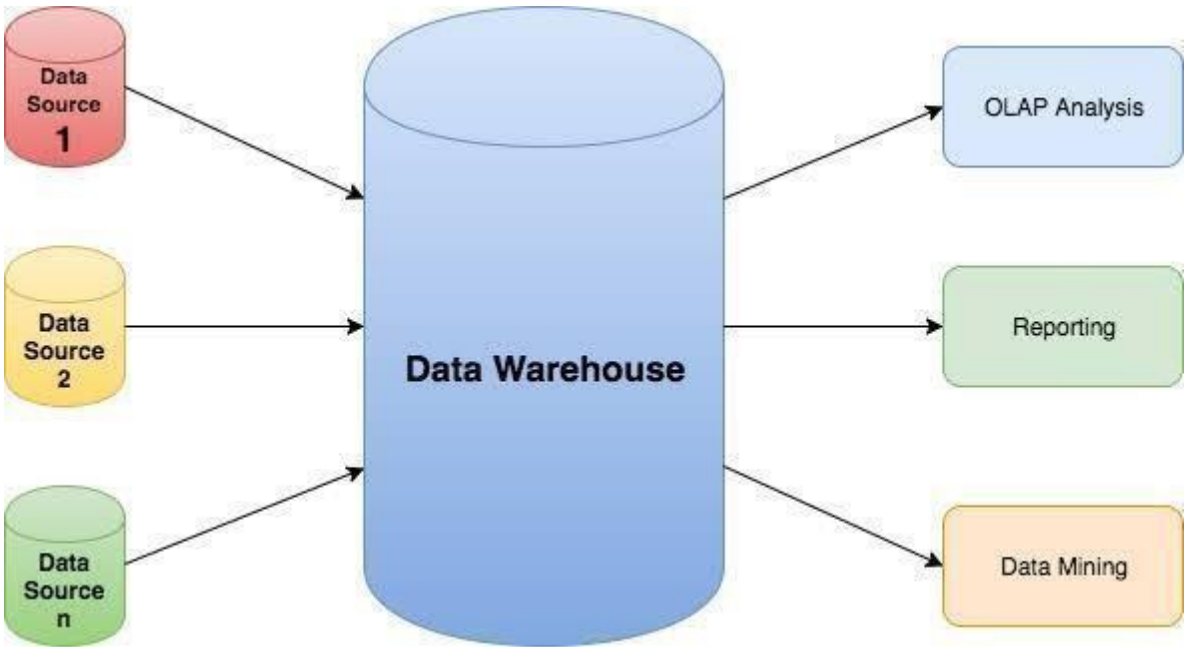
Propose system

Dataset:

Here PIMA Indian diabetes data set is considered. The data set is taken from UCI machine learning repository. The data set consists of 9 attributes: number of times pregnant, plasma glucose concentration, diastolic blood pressure, triceps skin folds thickness, serum insulin, body mass index, pedigree type, age, and class. Here, the class label is binary classification. It has two values

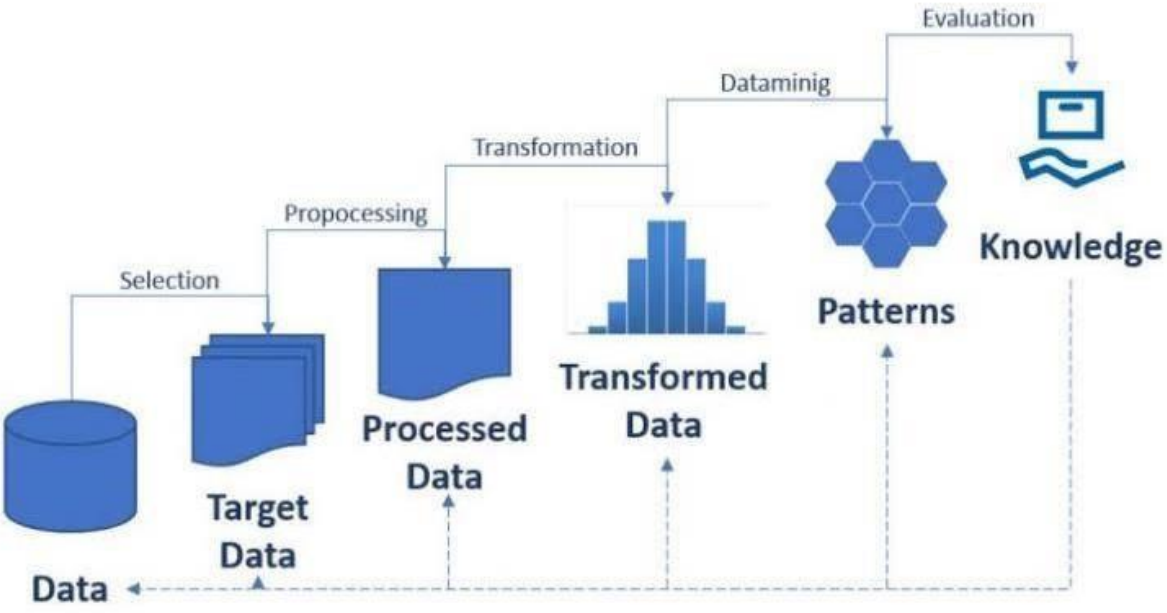
• Tested positive (1) which means diabetic

• Tested negative (0) which says nondiabetic



Methodology

Data preprocessing and data mining algorithms are used for the further process in the project. Data preprocessing technique data transformation is applied to the data set before applying data mining algorithms. The decision tree and regression models are built. Decision trees and Regression models are used to predict the final binary target variable. After running different types of models, model comparison needed to select the best algorithm. The best algorithm and best model is selected based on the high accuracy rate.



HARDWARE AND SOFTWARE REQUIREMENTS

Software Requirements:

Windows OS

Weka

Hardware Requirements:

Hard Disk – 1 TB or Above

RAM required – 8 GB or Above

Processor – Core i3 or Above

Technology Used:

Data Mining

Data Visualization

Coding:

```
opetion(repos = c(CRAN = "http://cran.rstudio.com"))

library(RSQLite)
 library(DBI)
library(datasets)
 library(caTools)
 library(e1071)

f<-file.choose("diab1.csv")
 mydata<-read.csv(f)
 datasetss<-read.csv(f)
#mydata<-read.csv(file = "/home/spllab01/Diabetes/diab1.csv",header = TRUE,sep = ",")
#datasetss<-read.csv(file = "/home/spllab01/Diabetes/diab1.csv",header = TRUE,sep = ",")

View(mydata)

View(datasetss)

#dataset<-read.table('diab.csv',header = T)

datasetss$SkinThickness = ifelse(is.na(datasetss$SkinThickness),
 ave(datasetss$SkinThickness, FUN = function(x) mean(x, na.rm = 'TRUE')),
datasetss$SkinThickness)
```

#Fills the NUll values with the average of that column values.

View(datasetss)

datasetss$Glucose = ifelse(is.na(datasetss$Glucose), ave(datasetss$Glucose, FUN = function(x) mean(x, na.rm = 'TRUE')), datasetss$Glucose)

View(datasetss)

h<-hist(datasetss$SkinThickness,main="Skin Thickness frquencies - histogram", xlab = "Skin Thickness", xlim = c(5,50),col = "blue")

h<-hist(datasetss$Glucose,main="Glucose frquencies - histogram", xlab = "Glucose Value",col = "blue",labels = TRUE, breaks = 8, border = "green",las=3)

train<-as.data.frame(datasetss[1:200,]) View(train)

test<-as.data.frame(datasetss[201:299,]) View(test)

head(train)

#Pregnancies Glucose BloodPressure SkinThickness Insulin BMI DiabetesPedigreeFunction Age

#201	0	113	80	16	0	31.0	0.874 21
#202	1	138	82	0	0	40.1	0.236 28
#203	0	108	68	20	0	27.3	0.787 32

#204	2	99	70	16	44 20.4	0.235 27
#205	6	103	72	32	190 37.7	0.324 55
#206	5	111	72	28	0 23.9	0.407 27

#Outcome

#201 0
#202 0
#203 0
#204 0
#205 0
#206 0

my_model<-naiveBayes(as.factor(train$Outcome)~.,train) pred1<-
predict(my_model,test[,-9])

pred1

#[1] 0 0 0 0 1 0 1 1 0 1 0 1 1 0 1 1 0 0 0 0 1 1 0 1 0 0 0 0 1 0 0 1 0 0 0
1 1 1 1 0 0 0 0 1 1 1

#[47] 1 1 1 0 0 0 0 0 1 0 0 0 1 1 1 0 0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 0 0 0 0
1 0 1 0 1 1 1 0 0 0 0

#[93] 1 1 1 1 1 0 1

#Levels: 0 1

#generate the confusion matrix

table(pred1,test$Outcome,dnn=c("predicted","actual"))

```
#actual

#predicted 0 1

#0 41 16

#1 19 23


#Build Classifier Models using Different Techniques.
#Cross Validation.


#Cross Validation K fold cross validation

library(caret) library(lattice)

 library(ggplot2)



# Define train control for k fold cross validation train_control
 <- trainControl(method="cv", number=10)




# Fit Naive Bayes Model model <- train(SkinThickness~., data=datasetss,
trControl=train_control, method="knn")
 # Summarise Results print(model)


#k-Nearest Neighbors


#299 samples

#8 predictor


#No pre-processing

#Resampling: Cross-Validated (10 fold)

#Summary of sample sizes: 269, 269, 269, 268, 270, 270, ...
```

#Resampling results across tuning parameters:

k RMSE Rsquared MAE

#5 11.87092 0.3645257 9.515780

#7 12.01004 0.3420931 9.699612

#9 11.83102 0.3558637 9.702612

#RMSE was used to select the optimal model using the smallest value.

#The final value used for the model was k = 9.

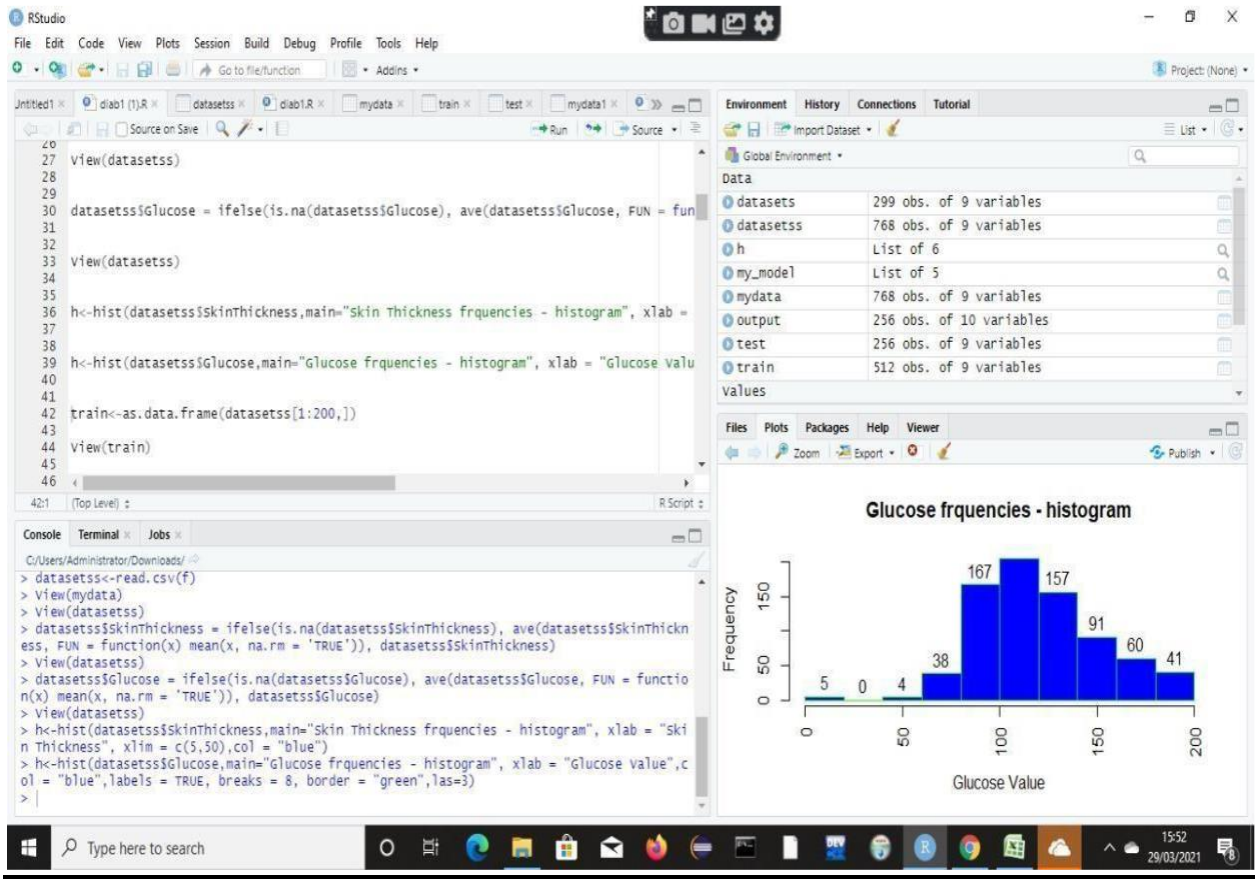
#scatterPolt matrix

```
head(datasetss)

pairs(datasetss[,1:4], pch = 19)

pairs(datasetss[,5:9], pch = 19)

pairs(datasetss[,1:4], pch = 19, lower.panel = NULL)

pairs(datasetss[,5:9], pch = 19, lower.panel = NULL)

View(datasetss)
```

```
#One more classifier model
 library(mlbench)
 library(caret)

# prepare training scheme control <-
 trainControl(method="repeatedcv", number=10, repeats=3)
# CART
set.seed(7) fit.cart <- train(SkinThickness~., data=datasetss,
 method="rpart", trControl=control)


# SVM
set.seed(7) fit.svm <- train(SkinThickness~., data=datasetss,
method="svmRadial", trControl=control)
# kNN
set.seed(7)
fit.knn <- train(SkinThickness~., data=datasetss, method="knn", trControl=control)
# Random Forest set.seed(7) fit.rf <- train(SkinThickness~., data=datasetss, method="rf",
trControl=control)
# collect resamples
results <- resamples(list(CART=fit.cart, SVM=fit.svm, KNN=fit.knn, RF=fit.rf))

summary(results)
```

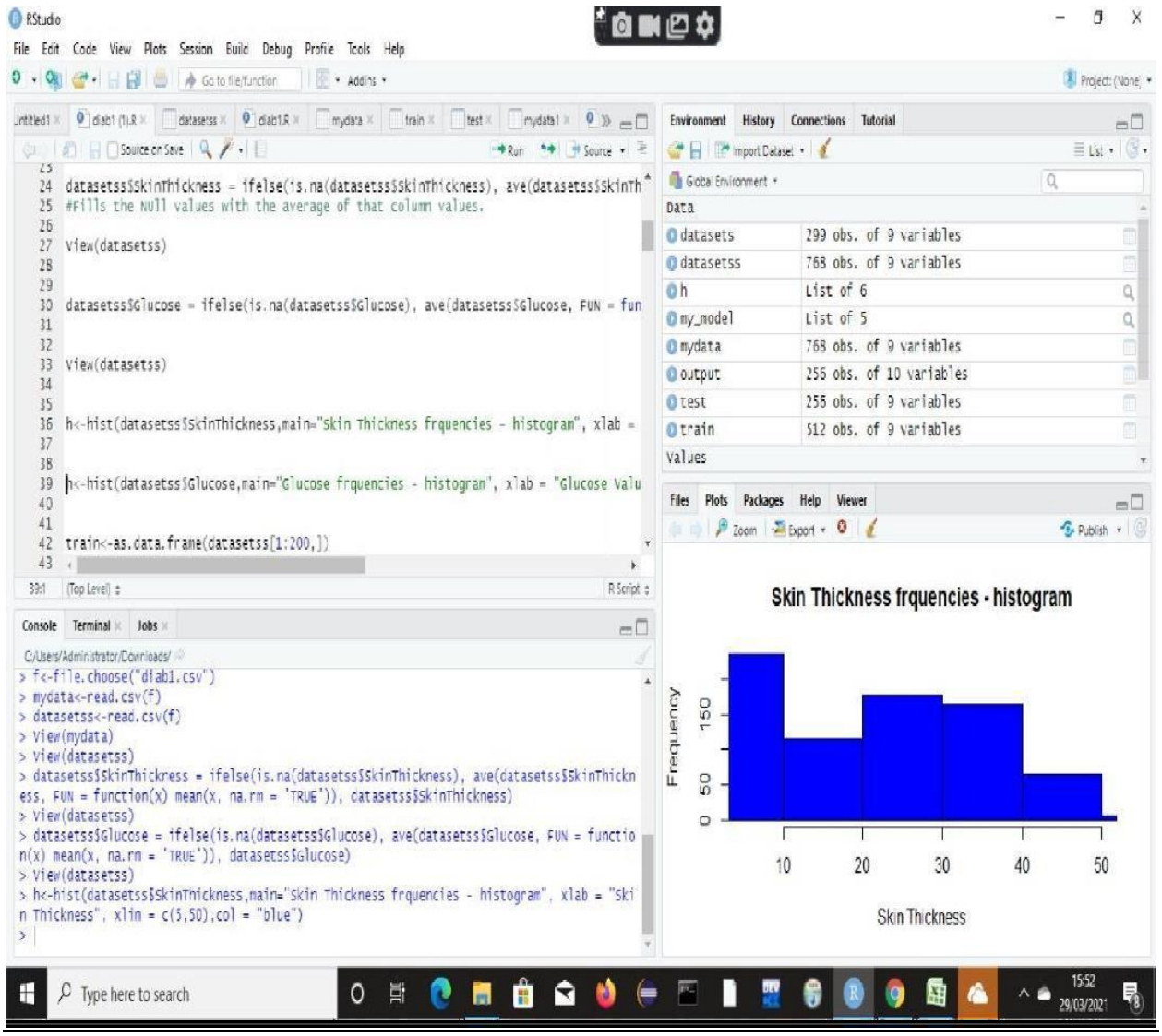
```
#box wisker scales <- list(x=list(relation="free"),
 y=list(relation="free")) bwplot(results, scales=scales)



#density plots scales <- list(x=list(relation="free"),
 y=list(relation="free")) densityplot(results, scales=scales,
 pch = "|")
#dot plots scales <- list(x=list(relation="free"),
 y=list(relation="free")) dotplot(results, scales=scales)
#parallel plots parallelplot(results)

#scatter plot splom(results)



#pair wise x and y plots


xyplot(results, models=c("KNN", "SVM"))

#statisticall significance test #
 difference in model predictions
 diffs <- diff(results)

# summarize p-values for pair-wise comparisons summary(diffs)
```

Glucose Frequencies-histogram:

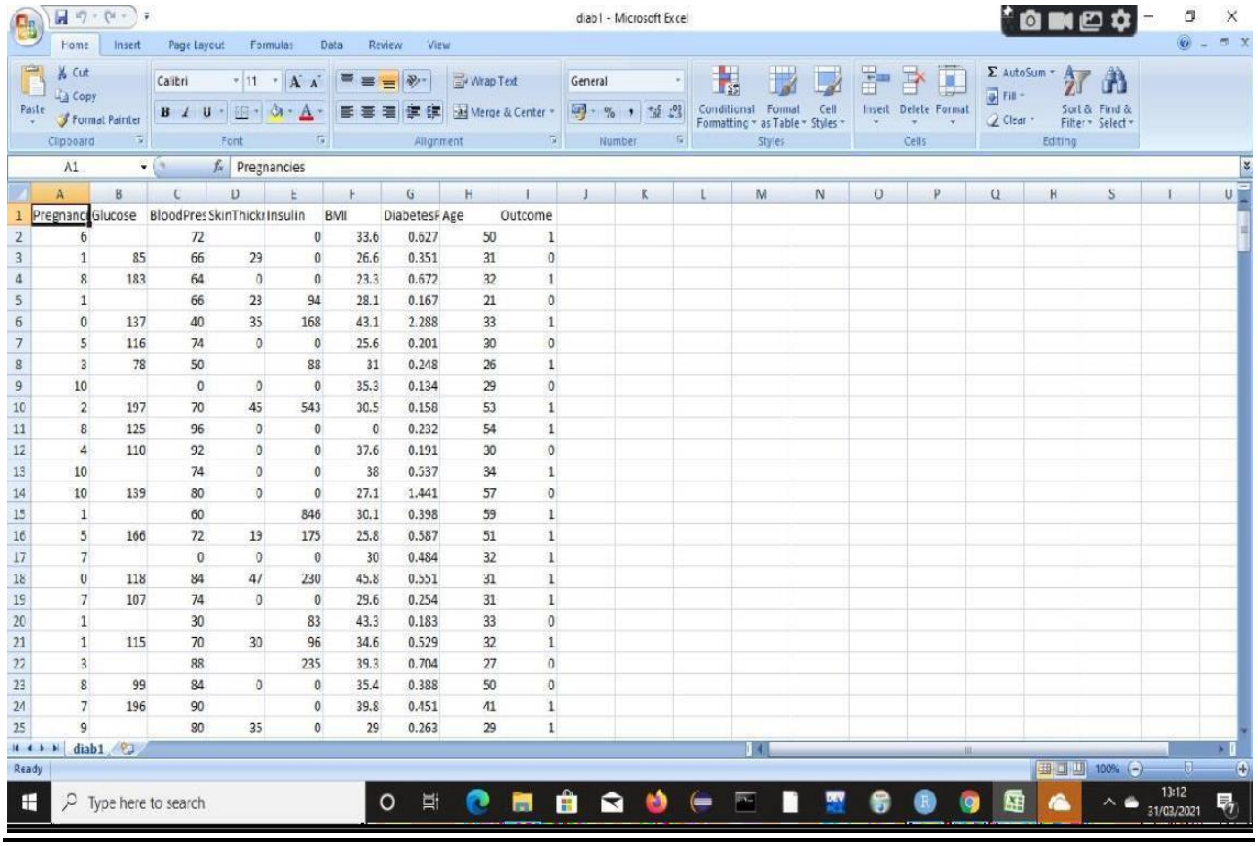


Outputs:

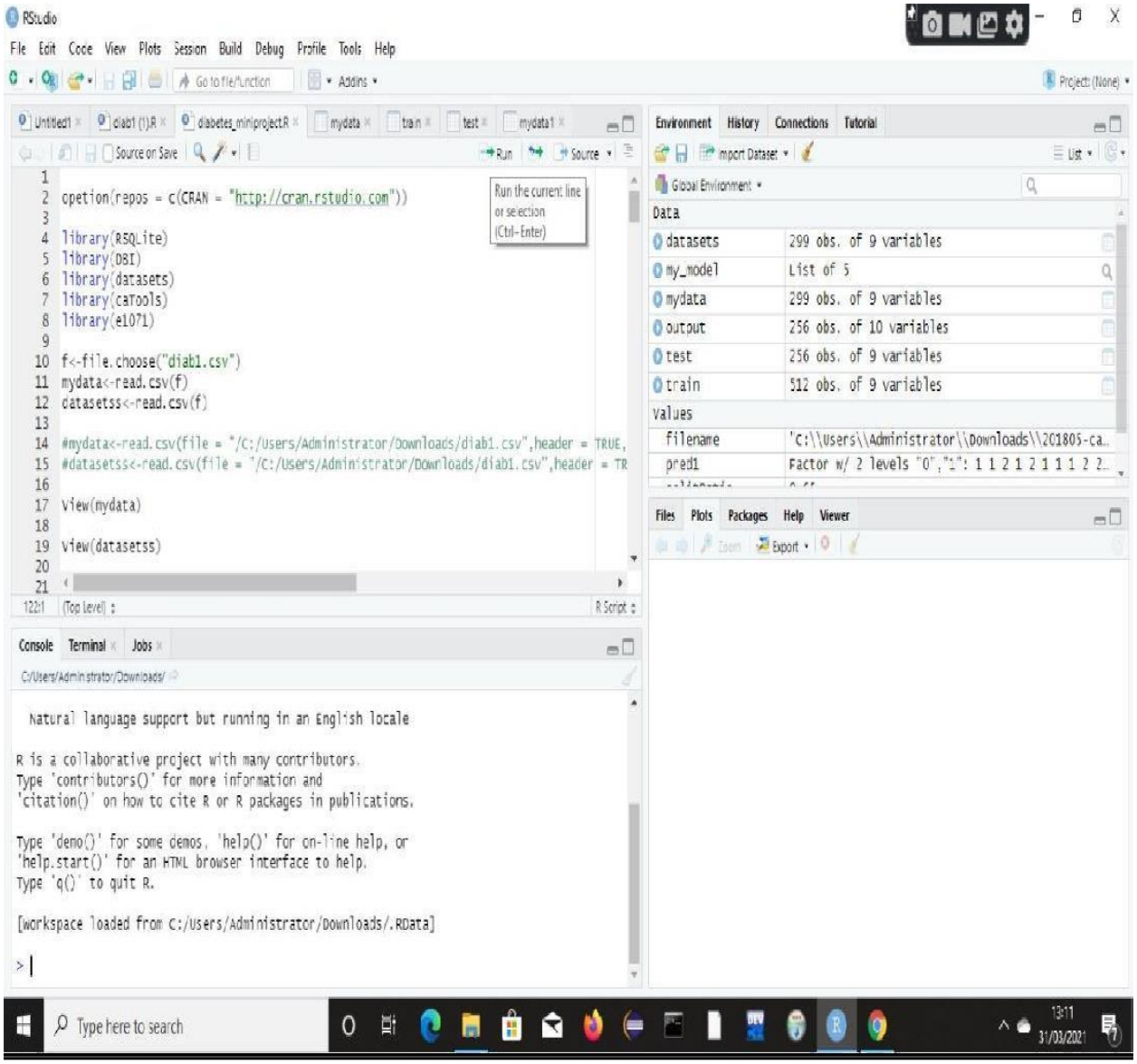
Skin Thickness Frequencies:



Database:



	A	B	C	D	E	F	G	H	I
1	Pregnanc	Glucose	BloodPres	SkinThickr	Insulin	BMI	DiabetesF	Age	Outcome
2	6		72		0	33.6	0.627	50	1
3	1	85	66	29	0	26.6	0.351	31	0
4	8	183	64	0	0	23.3	0.672	32	1
5	1		66	23	94	28.1	0.167	21	0
6	0	137	40	35	168	43.1	2.288	33	1
7	5	116	74	0	0	25.6	0.201	30	0
8	3	78	50		88	31	0.248	26	1
9	10		0	0	0	35.3	0.134	29	0
10	2	197	70	45	543	30.5	0.158	53	1
11	8	125	96	0	0	0	0.232	54	1
12	4	110	92	0	0	37.6	0.191	30	0
13	10		74	0	0	38	0.537	34	1
14	10	139	80	0	0	27.1	1.441	57	0
15	1		60		846	30.1	0.398	59	1
16	5	166	72	19	175	25.8	0.587	51	1
17	7		0	0	0	30	0.484	32	1
18	0	118	84	47	230	45.8	0.551	31	1
19	7	107	74	0	0	29.6	0.254	31	1
20	1		30		83	43.3	0.183	33	0
21	1	115	70	30	96	34.6	0.529	32	1
22	3		88		235	39.3	0.704	27	0
23	8	99	84	0	0	35.4	0.388	50	0
24	7	196	90		0	39.8	0.451	41	1
25	9		80	35	0	29	0.263	29	1


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Untitled1
diab1 (1).R
diabetes_miniprojectR
mydata
train
test
mydata1
Source on Save
Run
Source
Run the current line or selection (Ctrl-Enter)
1 cpetion(repos = c(CRAN = "http://cran.rstudio.com"))
3 library(RSQLite)
4 library(DBI)
5 library(datasets)
6 library(caTools)
7 library(e1071)
9 f<-file.choose("diab1.csv")
10 mydata<-read.csv(f)
11 datasetss<-read.csv(f)
13 #mydata<-read.csv(file = "/C:/Users/Administrator/Downloads/diab1.csv",header = TRUE,
14 #datasetss<-read.csv(file = "/C:/Users/Administrator/Downloads/diab1.csv",header = TR
16 view(mydata)
18 view(datasetss)
20
21
12:1 (Top Level) R Script
Console Terminal Jobs
C:/Users/Administrator/Downloads/
Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
[workspace loaded from C:/Users/Administrator/Downloads/.RData]
>
Environment History Connections Tutorial
Import Dataset
List
Global Environment
Data
datasets 299 obs. of 9 variables
my_model List of 5
mydata 299 obs. of 9 variables
output 256 obs. of 10 variables
test 256 obs. of 9 variables
train 512 obs. of 9 variables
Values
filename 'C:\\Users\\Administrator\\Downloads\\2C1805-ca...
pred1 Factor w/ 2 levels "0","1": 1 1 2 1 2 1 1 1 2 2...
Files Plots Packages Help Viewer
Zoom Export
Type here to search
13:11
31/03/2021

Conclusion:

Finally, decision is built using c4.5 decision tree algorithm. All the results are displayed to the end user using weka data visualization, regression provides the predicted outcome to the end user.