Pima Indians Diabetes

Xinning Chu

March 26, 2019

Abstract

Objective--To figure out which variables are significantly important for diabetes, predict the probability of diabetes for certain females, and provide some suggestions to prevent diabetes.

Methods-- Firstly, 650 observations are randomly picked up from the dataset and some descriptive statistics and coefficients were calculated. Then we implemented logistic regression models to predict the probability and find significant factors. Lastly, decision tree classifier was applied to detect subgroups of Pima Indian females that are more possible to have diabetes.

Results--Based on our model, the first given female has a relatively higher probability of having diabetes than the second one. Also, PRG, PLASMA, and BODY are most significant covariates.

Conclusion—Females with higher values of plasma glucose concentration in saliva, insulin, body mass index, diabetes pedigree function and age have a huge risk of having diabetes.

1. Introduction

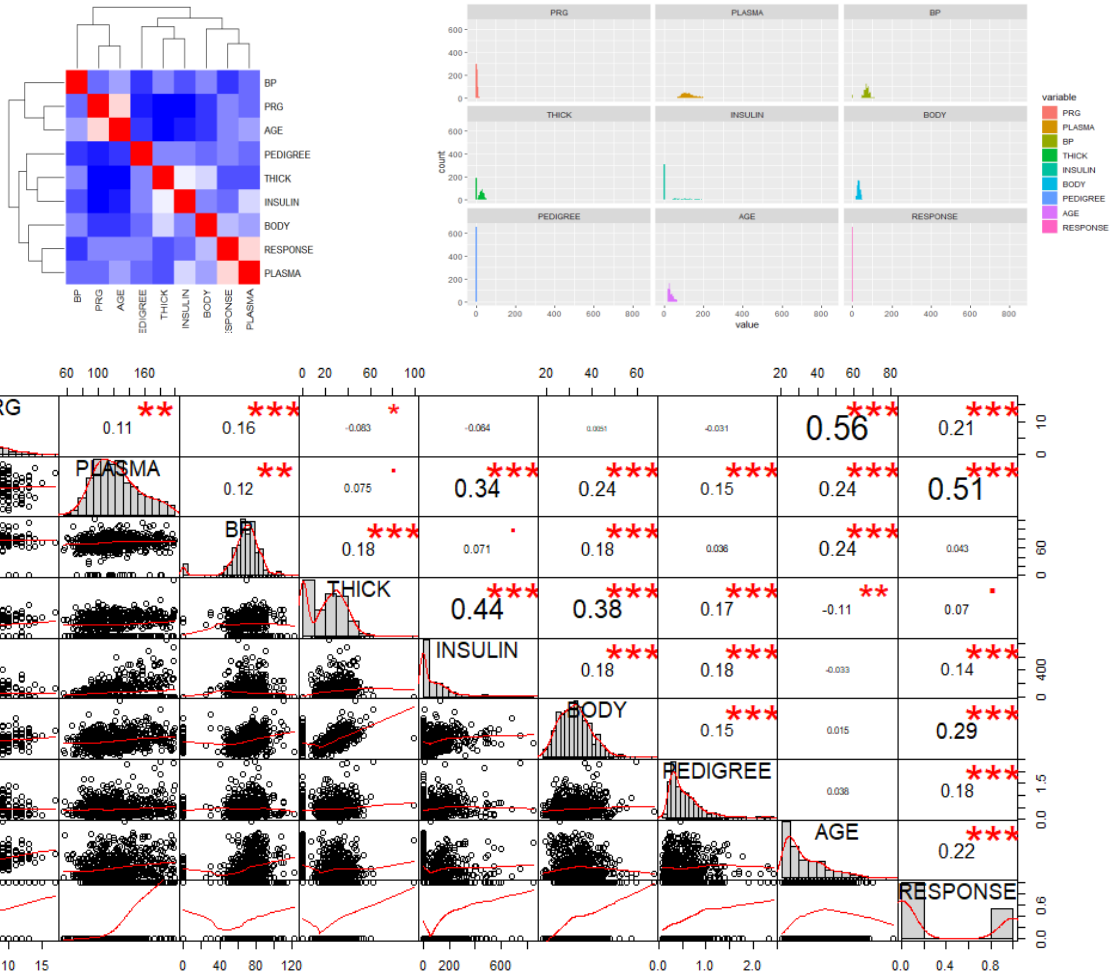
In this project, our purpose is to diagnostically predict whether a patient has diabetes, based on certain diagnostic measurements included in the dataset. Before building the model, we explored the data and got the correlations between variables. Then, we analyzed the data and detect the most relevant variables of diabetes and make suggestions to prevent diabetes.

2. Background and Dataset

The Pima Indians are the fattest population group in the fattest country on earth, and they have the highest prevalence of diabetes in the world. The problems have grown worse since first recognized in 1963, and the causes are multiple: environmental, genetic, cultural, and psychosocial. The real villain, however, is the disease, diabetes. Therefore, a solution is very necessary and essential.
We used data from the National Institute of Diabetes and Digestive and Kidney Diseases for all 750 patients that are females at least 21 years old of Pima Indian heritage. The dataset consists of 8 medical predictor variables and one target variable, RESPONSE. There are 768 observations in the dataset. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on. Details are shown in the Appendix I.

3. Data Analysis and Results

Initially, we picked up 650 total observations from 768 patients randomly. The data not selected can be used as testing dataset to assess the performance of models. The analysis of this problem was based on the chosen data. Then using R we calculated Spearman coefficients between variables.



From the figures above, PLASMA, BODY, AGE and PRG are highly correlated with RESPONSE. Additionally, the correlation coefficients of PRG and AGE is large, and it seems that these two variables seems to have a linear correlation, so we should delete one of them when building logistic regression models.

3.1 Logistic Regression

Since our target is a binary variable, logistic regression models could be performed. As it was mentioned before, AGE was deleted because it may have a linear correlation with PRG. From the correlation plot and results of ANOVA tests which can be seen in Appendix, BP is not statistically significant for the target, so probably it can be deleted.

Thus, we can get one model with BP and another model without BP. Bidirectional elimination was both applied in these two models to select variables. It's a combination of forward selection and backward elimination to testing at each step for variables to be included or excluded.

Table1. Comparison of Logistic Regression Models

	AIC	deviance	Hosmer-Lemeshow goodness of fit test

1: RESPONSE ~ PRG + PLASMA + BP + INSULIN + BODY + PEDIGREE	616.3704	602.3704	X-squared = 12.361, df = 8, p-value = 0.1358
2: RESPONSE ~ PRG + PLASMA + INSULIN + BODY + PEDIGREE	618.6296	606.6296	X-squared = 13.114, df = 8, p-value = 0.108

Before a model is relied upon to draw conclusions or predict outcomes, we should check, as far as possible, that the model we have assumed is correctly specified. The Hosmer-Lemeshow goodness of fit test is one approach to check the model which is based on dividing the sample up according to their predicted probabilities, or risks. According to Table 1, both results give us insignificant p-values, indicating good fits. We pick up the Model 1 as our final model, since the p-value of Model 1 is larger than Model 2. Another reason is that Model 1 has smaller AIC and deviance.

Table 2. Coefficients of Model 1

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-8.5994	0.7936	-10.836	< 2e-16 ***
PRG	0.1421	0.0305	4.657	3.20e-06 ***
PLASMA	0.0409	0.0041	9.966	< 2e-16 ***
BP	-0.0114	0.0056	-2.054	0.04000 *
INSULIN	-0.0013	0.0008	-1.532	0.12561
BODY	0.0821	0.0160	5.124	2.99e-07 ***
PEDIGREE	0.9208	0.3209	2.870	0.00411 **

From Table 2, BP is not statistically significant, but adding it to our model can make the fit stronger. In this model, the most significant variables are PLASMA, PRG and BODY. Clearly PEDIGREE and BP are also statistically significant. Though AGE was not included in this model, it's also statistically significant due to results of ANOVA tests.

Table 3. Values of VIF

Variable	PRG	PLASMA	BP	INSULIN	BODY	PEDIGREE
VIF	1.0755	1.1507	1.0907	1.1900	1.0878	1.0221

After fitting the model, diagnostics on model assumptions are required. There is an extreme situation, called multicollinearity, where collinearity exists between three or more variables even if no pair of variables has a particularly high correlation. One assumption of logistic regression mode is that no multicollinearity exists between explanatory variables. This means that there is redundancy between predictor variables. Therefore, VIF values of all explanatory were calculated. VIF is computed to access multicollinearity. All of them are smaller than 10, which means that there is no multicollinearity between explanatory variables. So it doesn't violate the model assumption.

Next, we used the Model 1 to predict the probability that two given females will have diabetes. The probability of diabetes for the female with (2,140,70,35,0,33.6,0.627,50) is 0.4888, while the probability for the female with (1,96,73,39,0,23.6,0.35,28) is 0.0431. The possible reason is that the first one is older and has higher Body Mass Index.

3.2 Decision Tree

Decision tree is a simple and widely used classification technique to build classification models form an input dataset, which applies a straightforward idea to solve the classification problem. It organizes a

series of test questions and conditions in a tree structure. Each time it receives an answer, a follow-up question is asked until a conclusion about the class label of the record is reached. The root and internal nodes contain attribute test conditions to separate records that have different characteristics. All the terminal node is assigned a class label Yes or No.

There are three main types of decision tree algorithms: ID3, C4.5 and CART. In this project, we applied CART to classify the outcome because explanatory variables are continuous. Gini is the default node split criteria of CART, and we used rpart function in R to automatically set the best values of parameters, such as max depths, min leaf samples and so on. The decision tree is displayed below.

According to the figure below, after excluding subgroups which contain too few samples, we can find three subgroups which have a high risk of getting diabetes: (1) 127.5 ≤ PLASMA < 157.5, BODY ≥ 29.95, AGE < 30.5, BP < 61; (2) 127.5 ≤ PLASMA < 157.5, BODY ≥ 41.8, AGE < 30.5, BP ≥ 61; (3) PLASMA ≥ 157.5

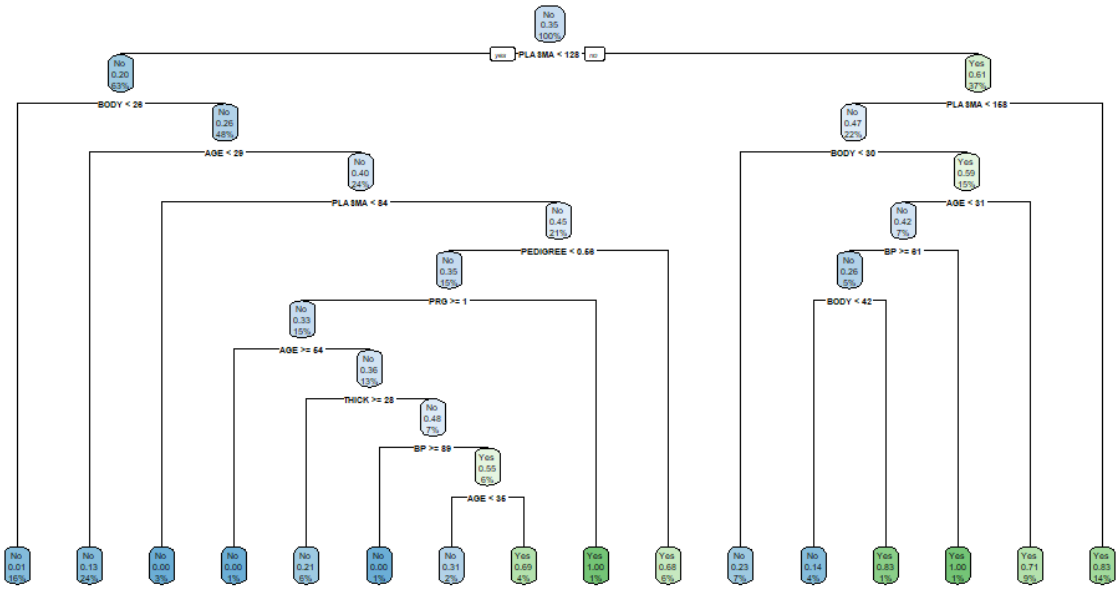
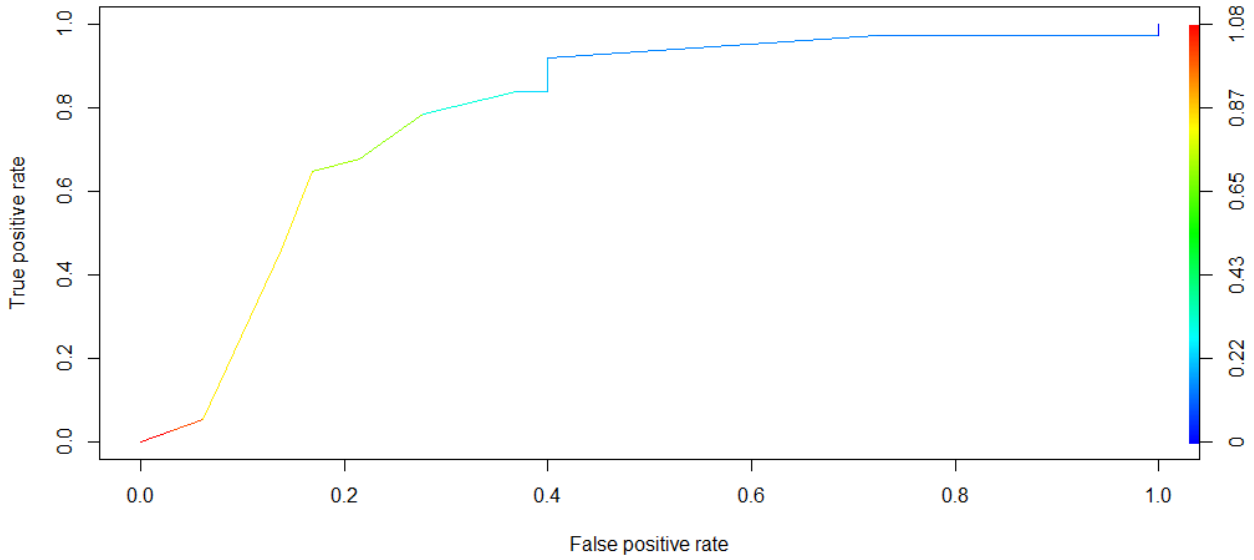


Table 4. Confusion Matrix and Statistics

	No (References)	Yes (References)	Accuracy: 0.7451
No (Predict)	48	9	95% CI: (0.6492, 0.8262)
Yes (Predict)	17	28	No Information Rate: 0.6373 P-Value [Acc > NIR]: 0.01366
Kappa: 0.4732		Mcnemar's Test P-Value: 0.16981	
Sensitivity: 0.7568		Specificity: 0.7385	
Pos Pred Value: 0.6222		Neg Pred Value: 0.8421	
Prevalence: 0.3627		Detection Rate: 0.2745	
Detection Prevalence: 0.4412		Balanced Accuracy: 0.7476	

Then the testing dataset was used to predict and evaluate the model. The confusion matrix was calculated. According to Table 4, accuracy of the model is 0.7451, which seems a god measure. At the same time, p-value is a small number, and this means that the model is performing very well. Sensitivity of 0.7568 indicated that nearly 76% of people with diabetes were predicted to be sick, and specificity of 0.7385 indicates that about 74% of people who were not sick, were truly predicted to be so.

To get deeper evaluation of this model, ROC curve and AUC score were used to visualize the performance of this classifier. From the plot, we saw a good “hump shape” curve and AUC=0.7896.So, this classifier was not so “bad”, it “learnt” something from training data.



4. Conclusion

Based on our research, there are 6 variables that can have statistically significant effects on diabetes. A female with higher values of PRG, plasma glucose concentration in saliva, insulin, BMI and diabetes pedigree function is more likely to have diabetes. Thus, here are our recommendation for preventing diabetes:

- Work out regularly;
- Drink water instead of other beverages, thereby control blood sugar and insulin levels
- Loss weight if necessary;
- Follow a very-low-carb and high-fiber diet;
- Optimize Vitamin D levels to control blood sugar;
- Take natural herbs, like Curcumin and Berberine.

5. References

https://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/
https://www.r-bloggers.com/how-to-perform-a-logistic-regression-in-r/
https://stats.idre.ucla.edu/r/dae/logit-regression/
https://cran.r-project.org/web/packages/plotROC/vignettes/examples.html
https://cran.r-project.org/web/packages/pROC/pROC.pdf

Appendix 1: Summary Statistics of the Training Dataset

```
      PRG              PLASMA             BP              THICK           INSULIN           BODY      
 Min.   : 0.000   Min.   : 44.0   Min.   :  0.00   Min.   : 0.00   Min.   :  0.00   Min.   :18.20  
 1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 64.00   1st Qu.: 0.00   1st Qu.:  0.00   1st Qu.:27.60  
 Median : 3.000   Median :117.0   Median : 72.00   Median :23.50   Median : 44.00   Median :32.40  
 Mean   : 3.885   Mean   :121.2   Mean   : 70.01   Mean   :21.19   Mean   : 83.99   Mean   :32.57  
 3rd Qu.: 6.000   3rd Qu.:139.8   3rd Qu.: 80.00   3rd Qu.:33.00   3rd Qu.:130.00   3rd Qu.:36.80  
 Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00   Max.   :846.00   Max.   :67.10  
    PEDIGREE           AGE        RESPONSE   
 Min.   :0.0780   Min.   :21.00   No :423  
 1st Qu.:0.2440   1st Qu.:24.00   Yes:227  
 Median :0.3705   Median :29.00            
 Mean   :0.4683   Mean   :33.48            
 3rd Qu.:0.6228   3rd Qu.:41.00            
 Max.   :2.3290   Max.   :81.00            
```

Appendix 2: Description of Variables in the Dataset

- PRG: Number of times pregnant
- PLASMA: Plasma glucose concentration in saliva
- BP: Diastolic blood pressure
- THICK: Triceps skin fold thickness
- INSULIN: Two hours serum insulin
- BODY: Body mass index (Weight/Height)
- PEDIGREE: Diabetes pedigree function
- AGE: In years
- RESPONSE: 1: Diabetes, 0:Not

Appendix 3: Results of ANOVA Tests

```
> fit1 <- aov(PRG ~ RESPONSE, data = Train)
> fit2 <- aov(PLASMA ~ RESPONSE, data = Train)
> fit3 <- aov(BP ~ RESPONSE,data = Train)
> fit4 <- aov(THICK ~ RESPONSE, data = Train)
> fit5 <- aov(INSULIN ~ RESPONSE, data = Train)
> fit6 <- aov(BODY ~ RESPONSE,data = Train)
> fit7 <- aov(PEDIGREE ~ RESPONSE, data = Train)
> fit8 <- aov(AGE ~ RESPONSE, data = Train)
> summary(fit1)
             Df Sum Sq Mean Sq F value   Pr(>F)    
RESPONSE      1    352   352.5   31.31 3.25e-08 ***
Residuals   648   7296    11.3                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(fit2)
             Df Sum Sq Mean Sq F value Pr(>F)    
RESPONSE      1 139844  139844   194.1 <2e-16 ***
Residuals   648 466958     721                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(fit3)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
RESPONSE      1   1039  1039.4   3.156 0.0761 .
Residuals   648 213407   329.3
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(fit4)
             Df Sum Sq Mean Sq F value Pr(>F)
RESPONSE      1    841   840.6   3.305 0.0696 .
Residuals   648 164842   254.4
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(fit5)
             Df  Sum Sq Mean Sq F value   Pr(>F)
RESPONSE      1  197087  197087   14.03 0.000196 ***
Residuals   648 9100713   14044
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(fit6)
             Df Sum Sq Mean Sq F value   Pr(>F)
RESPONSE      1   2989  2988.6   69.95 3.73e-16 ***
Residuals   648  27686    42.7
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(fit7)
             Df Sum Sq Mean Sq F value   Pr(>F)
RESPONSE      1   1.84  1.8416   18.33 2.14e-05 ***
Residuals   648  65.11  0.1005
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(fit8)
             Df Sum Sq Mean Sq F value   Pr(>F)
RESPONSE      1   5074    5074   39.48 6.08e-10 ***
Residuals   648  83277     129
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Appendix 4: Details of the Decision Tree

```
Rule number: 53 [RESPONSE=Yes cover=9 (1%) prob=1.00]
   PLASMA>=127.5
   PLASMA< 157.5
   BODY>=29.95
   AGE< 30.5
   BP< 61

 Rule number: 93 [RESPONSE=Yes cover=4 (1%) prob=1.00]
   PLASMA< 127.5
   BODY>=26.45
   AGE>=28.5
   PLASMA>=83.5
   PEDIGREE< 0.558
   PRG< 0.5
```

```
Rule number: 105 [RESPONSE=Yes cover=6 (1%) prob=0.83]
  PLASMA>=127.5
  PLASMA< 157.5
  BODY>=29.95
  AGE< 30.5
  BP>=61
  BODY>=41.8

Rule number: 7 [RESPONSE=Yes cover=92 (14%) prob=0.83]
  PLASMA>=127.5
  PLASMA>=157.5

Rule number: 27 [RESPONSE=Yes cover=56 (9%) prob=0.71]
  PLASMA>=127.5
  PLASMA< 157.5
  BODY>=29.95
  AGE>=30.5

Rule number: 1487 [RESPONSE=Yes cover=26 (4%) prob=0.69]
  PLASMA< 127.5
  BODY>=26.45
  AGE>=28.5
  PLASMA>=83.5
  PEDIGREE< 0.558
  PRG>=0.5
  AGE< 53.5
  THICK< 27.5
  BP< 89
  AGE>=34.5

Rule number: 47 [RESPONSE=Yes cover=38 (6%) prob=0.68]
  PLASMA< 127.5
  BODY>=26.45
  AGE>=28.5
  PLASMA>=83.5
  PEDIGREE>=0.558

Rule number: 1486 [RESPONSE=No cover=16 (2%) prob=0.31]
  PLASMA< 127.5
  BODY>=26.45
  AGE>=28.5
  PLASMA>=83.5
  PEDIGREE< 0.558
  PRG>=0.5
  AGE< 53.5
  THICK< 27.5
  BP< 89
  AGE< 34.5

Rule number: 12 [RESPONSE=No cover=47 (7%) prob=0.23]
  PLASMA>=127.5
  PLASMA< 157.5
  BODY< 29.95

Rule number: 370 [RESPONSE=No cover=39 (6%) prob=0.21]
  PLASMA< 127.5
  BODY>=26.45
```

```
  AGE>=28.5
  PLASMA>=83.5
  PEDIGREE< 0.558
  PRG>=0.5
  AGE< 53.5
  THICK>=27.5

Rule number: 104 [RESPONSE=No cover=28 (4%) prob=0.14]
  PLASMA>=127.5
  PLASMA< 157.5
  BODY>=29.95
  AGE< 30.5
  BP>=61
  BODY< 41.8

Rule number: 10 [RESPONSE=No cover=155 (24%) prob=0.13]
  PLASMA< 127.5
  BODY>=26.45
  AGE< 28.5

Rule number: 4 [RESPONSE=No cover=103 (16%) prob=0.01]
  PLASMA< 127.5
  BODY< 26.45

Rule number: 742 [RESPONSE=No cover=6 (1%) prob=0.00]
  PLASMA< 127.5
  BODY>=26.45
  AGE>=28.5
  PLASMA>=83.5
  PEDIGREE< 0.558
  PRG>=0.5
  AGE< 53.5
  THICK< 27.5
  BP>=89

Rule number: 184 [RESPONSE=No cover=8 (1%) prob=0.00]
  PLASMA< 127.5
  BODY>=26.45
  AGE>=28.5
  PLASMA>=83.5
  PEDIGREE< 0.558
  PRG>=0.5
  AGE>=53.5

Rule number: 22 [RESPONSE=No cover=17 (3%) prob=0.00]
  PLASMA< 127.5
  BODY>=26.45
  AGE>=28.5
  PLASMA< 83.5
```

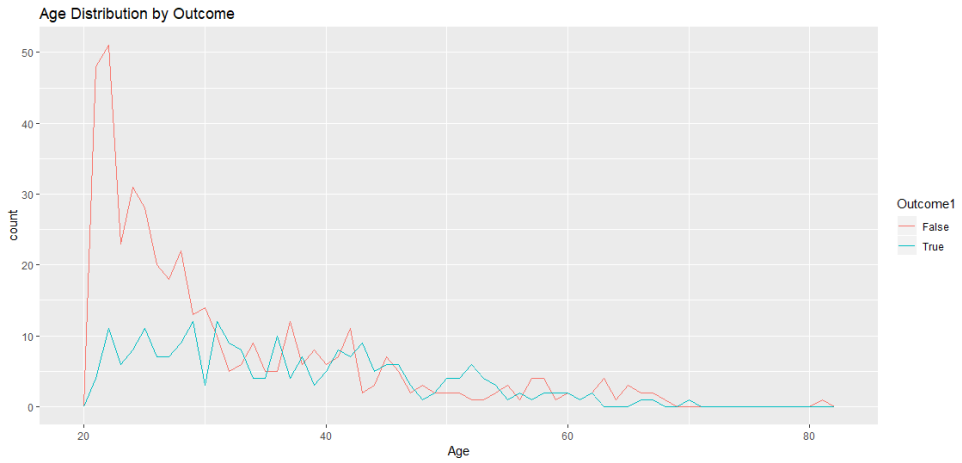
Appendix 5: R Codes Used in the Project

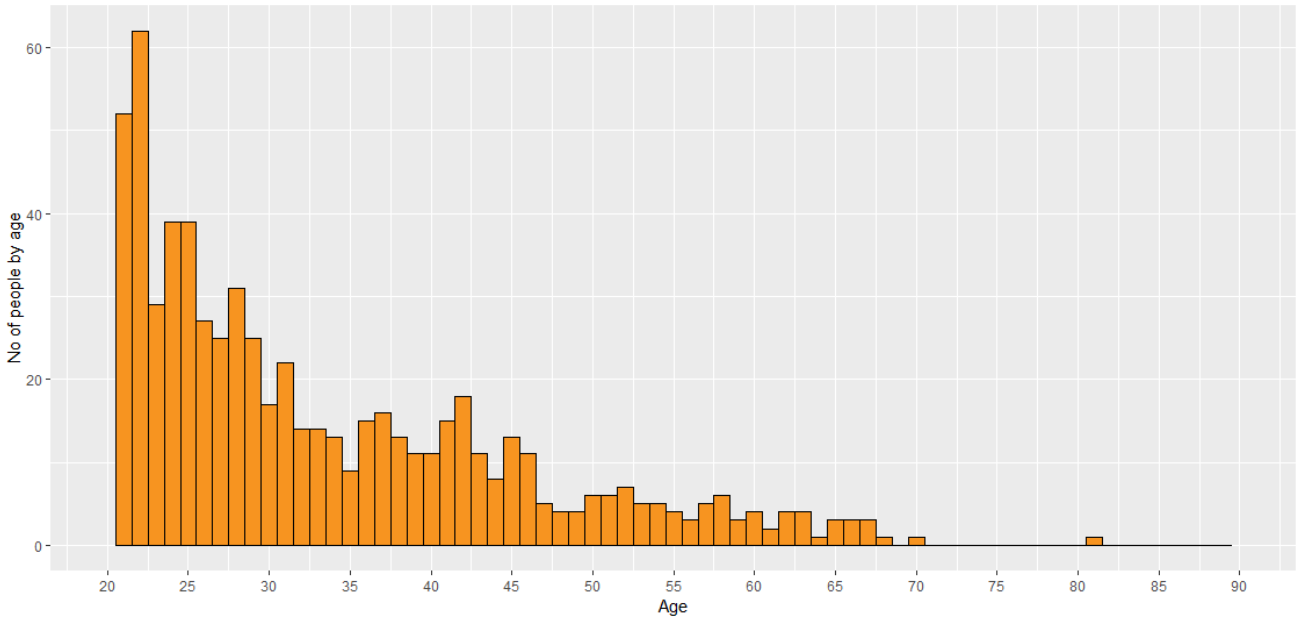
```
setwd("C:/Users/derri/Onedrive/r")
diabetes <- read.csv('diabetes.csv', header = TRUE)
str(diabetes)
summary(diabetes)
library(splitstackshape)
library(ggplot2)
library(xts)
library(zoo)
library(PerformanceAnalytics)
library(lattice)
library(DAAG)
library(randomForest)
library(ResourceSelection)
library(car)
library(pROC)
library(caret)
#random sampling-----------------------------------------------------------
set.seed(1)
index <- createDataPartition(diabetes$RESPONSE, p=0.86436, list=FALSE)
Train <- diabetes[index,]
Test <- diabetes[-index,]
summary(Train)
write.csv(Train,"C:/Users/derri/Onedrive/r/train.csv",row.names = FALSE)
write.csv(Test,"C:/Users/derri/Onedrive/r/test.csv",row.names = FALSE)
Train <- read.csv('train.csv',header=TRUE)
Test <- read.csv('test.csv',header=TRUE)
#EDA-----------------------------------------------------------------------------------
#correlation matrix
cormatr <- cor(Train)
symnum(cormatr)

col = colorRampPalette(c("blue", "white", "red"))(20)
heatmap(x = cormatr, col = col, symm = TRUE)
chart.Correlation(Train,histogram = TRUE,pch=19)

#Spearman Correlation Coeffients
chart.Correlation(Train,histogram = TRUE,pch=19,method = "spearman")

#ANOVA---------------------------------------------------------------------------
fit1 <- aov(PRG ~ RESPONSE, data = Train)
fit2 <- aov(PLASMA ~ RESPONSE, data = Train)
fit3 <- aov(BP ~ RESPONSE,data = Train)
fit4 <- aov(THICK ~ RESPONSE, data = Train)
fit5 <- aov(INSULIN ~ RESPONSE, data = Train)
fit6 <- aov(BODY ~ RESPONSE,data = Train)
```

```
fit7 <- aov(PEDIGREE ~ RESPONSE, data = Train)
fit8 <- aov(AGE ~ RESPONSE, data = Train)
summary(fit1)
summary(fit2)
summary(fit3)
summary(fit4)
summary(fit5)
summary(fit6)
summary(fit7)
summary(fit8)
#logistics regression
LR1 = glm(RESPONSE ~ PRG + PLASMA + BP +  THICK + INSULIN + BODY + PEDIGREE,
                  family=binomial(link='logit'),data = Train)
summary(LR1)
logit.step1 <- step(LR1, direction = "both")
summary(logit.step1)
logit.step1a <- step(LR1, direction = "back")
logit.step1b <- step(LR1, direction = "forward")
summary(LR2)
LR2 = glm(RESPONSE ~ PRG + PLASMA  +  THICK + INSULIN + BODY + PEDIGREE,
                family=binomial(link='logit'),data = Train)
logit.step2 <- step(LR2, direction = "both")
summary(logit.step2)
LR3 = glm(RESPONSE ~ PRG + PLASMA + BP +  THICK + INSULIN + BODY + PEDIGREE + AGE,
       family=binomial(link='logit'),data = Train)
logit.step3 <- step(LR3, direction = "both")
summary(logit.step3)
ht1 <- hoslem.test(Train$RESPONSE,fitted(logit.step1),g=10)
ht2 <- hoslem.test(Train$RESPONSE,fitted(logit.step2),g=10)
ht3 <- hoslem.test(Train$RESPONSE,fitted(logit.step3),g=10)
ht1
ht2
newdata <- read.csv('newdata.csv', header = TRUE)
predict(logit.step1,newdata,type="response")


OT2 = randomForest(RESPONSE~.,Train,ntree=1000,mtree=3,importance=TRUE)
print(OT)
importance(OT2)
varImpPlot(OT2)

vif(logit.step2)
vif(logit.step1)


library(reshape2)
gg <- melt(Train)
ggplot(gg, aes(x=value, fill=variable)) +   geom_histogram(binwidth=5) +   facet_wrap(~variable)
```

```
#decision tree------------------------------
library(rpart)
library(rpart.plot)
library(gplots)
library(ROCR)
library(rattle)
library(e1071)
Train$RESPONSE <- factor(Train$RESPONSE, levels=c(0,1), labels=c("No", "Yes"))
Test$RESPONSE <- factor(Test$RESPONSE, levels=c(0,1), labels=c("No", "Yes"))
fit=rpart(RESPONSE~.,method = "class",
        data=Train,control = rpart.control(minsplit = 1) , parms = list(split="information"))
rpart.plot(fit)
asRules(fit)

pred1 <- predict(fit, Test, type="class" )
pred1
confusionMatrix(pred1, reference = Test$RESPONSE, positive="Yes")
pred1_prob <- predict(fit, Test)
prediction1 <- prediction(pred1_prob[,2],Test$RESPONSE)
perf1 <- performance(prediction1, "tpr", "fpr")
plot(perf1, colorize=TRUE)
performance(prediction1, "auc")@y.values
fit.pru <- prune(fit, cp= fit$cptable[which.min(fit$cptable[,"xerror"]),"CP"])
rpart.plot(fit.pru,branch=1, extra=106, under=TRUE, faclen=0,
         cex=0.8, main="决策树")
```

Appendix 6: Graphs




No of people by age
60
40
20
0
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
Age