WEB MINING

Project Report

UNDER THE GUIDANCE OF
Prof. Hai Nhat Phan
Team Members
Priya Sangale(ps739)
Pratishtha Ashok(pa272)

DATA ANALYSIS ON DIABETES

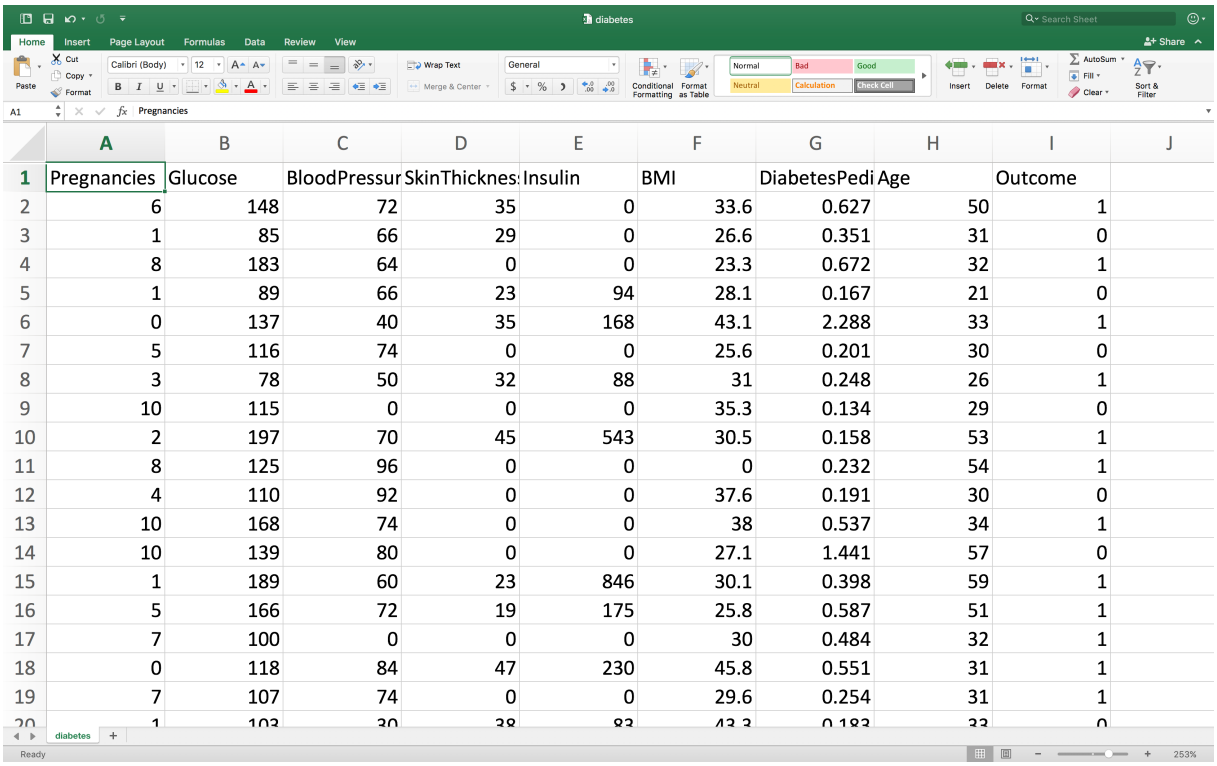
3.1 Data Set:

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective is to predict based on diagnostic measurements whether a patient has diabetes.

3.2 Objective:

The purpose of this data set mining implementation is to extrapolate the factors triggering diabetes from 'number of times pregnant' and other attributes. The analysis will also explain which variables leads to diabetes.

3.3 Data sheet snapshot:



	A	B	C	D	E	F	G	H	I	J
1	Pregnancies	Glucose	BloodPressur	SkinThickness	Insulin	BMI	DiabetesPedi	Age	Outcome	
2	6	148	72	35	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	89	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	32	88	31	0.248	26	1	
9	10	115	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	168	74	0	0	38	0.537	34	1	
14	10	139	80	0	0	27.1	1.441	57	0	
15	1	189	60	23	846	30.1	0.398	59	1	
16	5	166	72	19	175	25.8	0.587	51	1	
17	7	100	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	103	30	38	83	43.3	0.183	33	0	

3.4 Attributes and Source:

- **Pregnancies**: Number of times pregnant
- **Glucose:** Plasma glucose concentration a 2 hour in an oral glucose tolerance test
- **Blood Pressure:** Diastolic blood pressure (mm Hg)
- **Skin Thickness:** Triceps skin fold thickness (mm)
- **Insulin:** 2-hour Serum insulin (mu U/ml)
- **BMI:** Body mass index (weight in kg/ (height in m) ^2)
- **Diabetes Pedigree Function:** Diabetes Pedigree function
- **Age:** Age in years
- **Outcome:** Class variable (0 or 1)

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

3.6 Data Analytical Model:

The Data Analytical technique used for this model is Logistic Regression. The reason to choose this technique was the target variable being the type 'binomial', so Logistic regression will be the best fit. Furthermore, this type of regression will give a better view of the driver variables driving the target with high approximate value.

3.7 Result and Analysis:

R code for Analysis:

```
> library(ggplot2)
> library(visreg)
> data <- read.csv("/Users/pri0732/Downloads/diabetes.csv")
> str(data)
> head(data)
> summary(data)
> pairs(data)
> train <- data[1:500,]
> test <- data[501:768,]
> trainmodel <- lm(Outcome ~ ., data = train)
> result <- predict(trainmodel,test)
> summary(result)
```

R Code Snapshots for analysis:

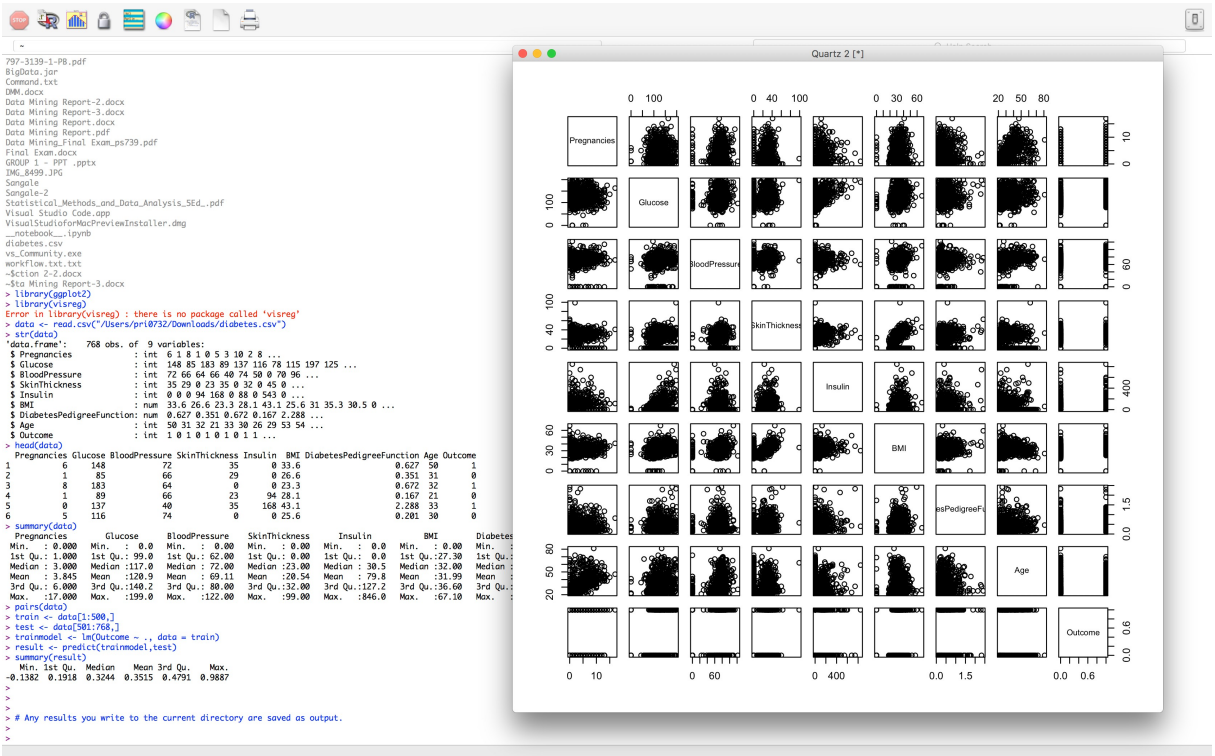
```
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

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.

[R.app GUI 1.68 (7288) x86_64-apple-darwin13.4.0]

[History restored from /Users/pri0732/.Rapp.history]

> diabetes <- read.csv("", header = TRUE)
Error in read.table(file = file, header = header, sep = sep, quote = quote,  :
  no lines available in input
In addition: Warning message:
In file(file, "rt") :
  file("") only supports open = "w+" and open = "w+b": using the former
> diabetes <- read.csv("../Downloads/diabetes.csv", header = TRUE)
Error in file(file, "rt") : cannot open the connection
In addition: Warning message:
In file(file, "rt") :
  cannot open file '../Downloads/diabetes.csv': No such file or directory
> diabetes <- read.csv("../Downloads/diabetes.csv", header = TRUE)
Error in file(file, "rt") : cannot open the connection
In addition: Warning message:
In file(file, "rt") :
  cannot open file '../Downloads/diabetes.csv': No such file or directory
> library(ggplot2)
> library(readr)
Error in library(readr) : there is no package called 'readr'
> source("/Users/pri0732/Downloads/diabetes.csv")
Error in source("/Users/pri0732/Downloads/diabetes.csv") :
  /Users/pri0732/Downloads/diabetes.csv:1:12: unexpected ','
1: Pregnancies,
               ^
> orig_data <- read.csv("/Users/pri0732/Downloads/diabetes.csv")
> orig_data$Outcome <- factor(orig_data$Outcome)
> str(orig_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                 : Factor w/ 2 levels "0","1": 2 1 2 1 2 1 2 1 2 2 ...
> summary(orig_data)
  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
 0:500
 1:268
```



```
797-3139-1-PB.pdf
BigData.jar
Command.txt
DMM.docx
Data Mining Report-2.docx
Data Mining Report-3.docx
Data Mining Report.docx
Data Mining Report.pdf
Data Mining_Final Exam_ps739.pdf
Final Exam.docx
GROUP 1 - PPT .pptx
IMG_8499.JPG
Sangale
Sangale-2
Statistical_Methods_and_Data_Analysis_SEd_.pdf
Visual Studio Code.app
VisualStudioforMacPreviewInstaller.dmg
__notebook__.ipynb
diabetes.csv
vs_Community.exe
workflow.txt.txt
~$ction 2-2.docx
~$ta Mining Report-3.docx
> library(ggplot2)
> library(visreg)
Error in library(visreg) : there is no package called 'visreg'
> data <- read.csv("/Users/pri0732/Downloads/diabetes.csv")
> 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 ...
> head(data)
  Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI DiabetesPedigreeFunction Age Outcome
1           6     148            72            35       0 33.6                    0.627  50       1
2           1      85            66            29       0 26.6                    0.351  31       0
3           8     183            64             0       0 23.3                    0.672  32       1
4           1      89            66            23      94 28.1                    0.167  21       0
5           0     137            40            35     168 43.1                    2.288  33       1
6           5     116            74             0       0 25.6                    0.201  30       0
> summary(data)
  Pregnancies        Glucose      BloodPressure    SkinThickness      Insulin          BMI        Diabetes
 Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00   Min.   :  0.0   Min.   : 0.00   Min.   :
 1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00   1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:
 Median : 3.000   Median :117.0   Median : 72.00   Median :23.00   Median : 30.5   Median :32.00   Median :
 Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54   Mean   : 79.8   Mean   :31.99   Mean   :
 3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00   3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:
 Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00   Max.   :846.0   Max.   :67.10   Max.   :
> pairs(data)
> train <- data[1:500,]
> test <- data[501:768,]
> trainmodel <- lm(Outcome ~ ., data = train)
> result <- predict(trainmodel,test)
> summary(result)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-0.1382  0.1918  0.3244  0.3515  0.4791  0.9887
>
>
>
> # Any results you write to the current directory are saved as output.
>
>
```

Output:

>library(ggplot2)

> library(visreg)

> data <- read.csv("/Users/pri0732/Downloads/diabetes.csv")

> **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 ...

> **head(data)**

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1
4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1
6	5	116	74	0	0	25.6	0.201	30	0

> **summary(data)**

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

> **pairs(data)**

> train <- data[1:500,]

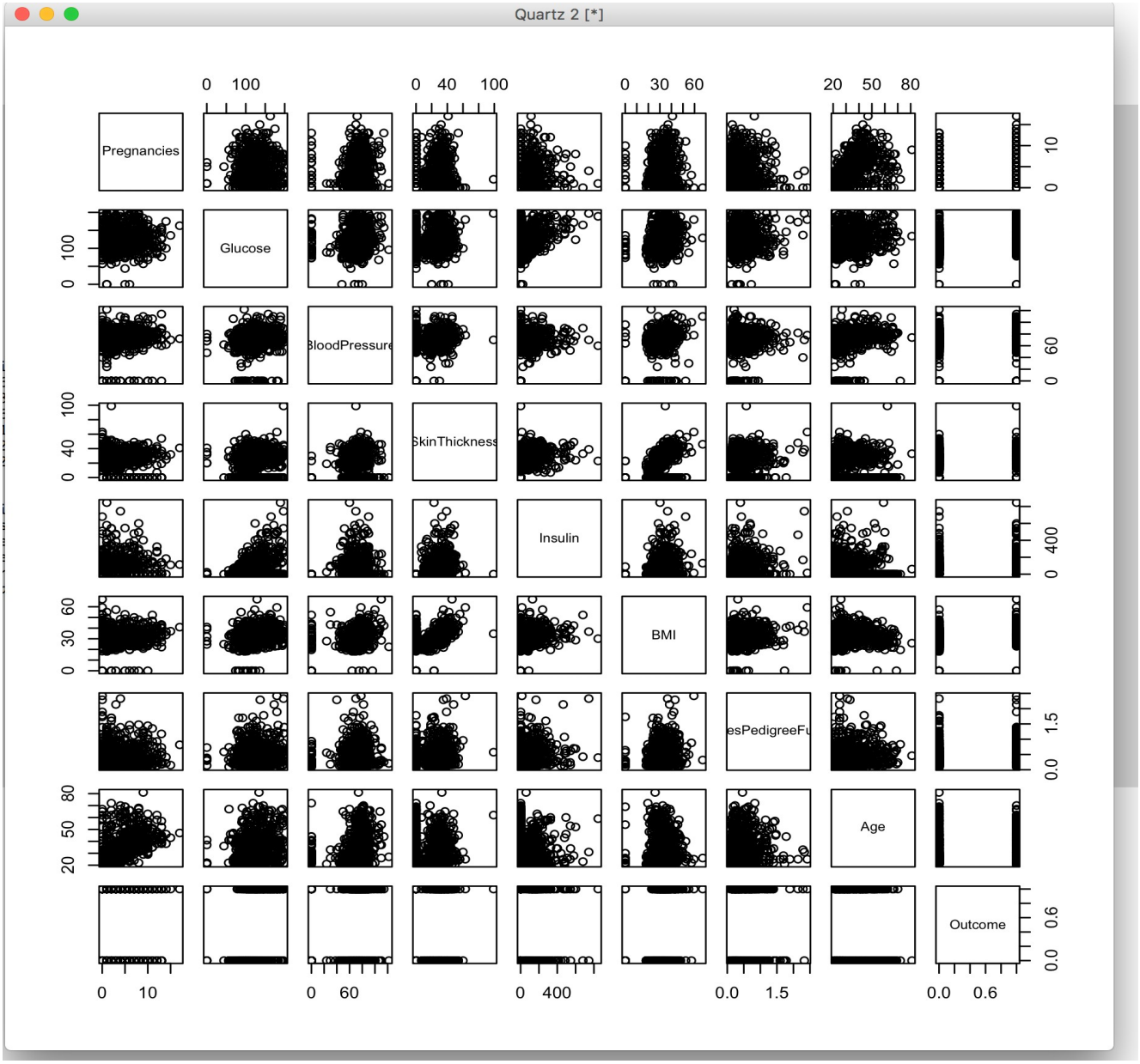
> test <- data[501:768,]

> trainmodel <- lm(Outcome ~ ., data = train)

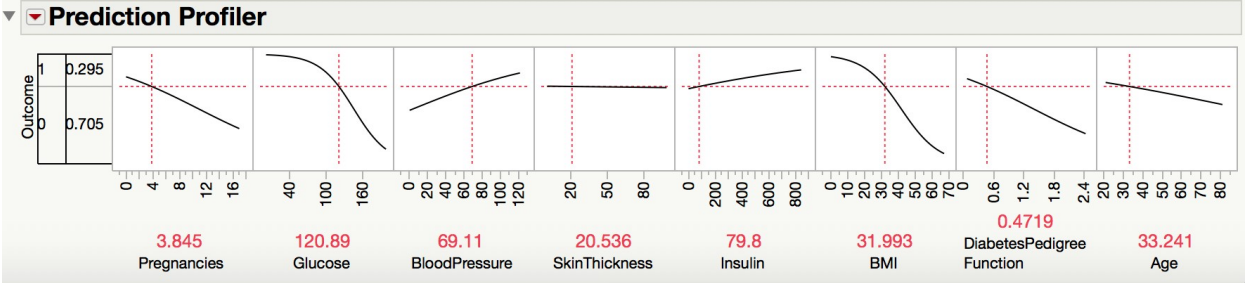
> result <- predict(trainmodel,test)

> summary(result)

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
-0.1382 0.1918 0.3244 0.3515 0.4791 0.9887 
>
>
>
> # Any results you write to the current directory are saved as output.
```



The above graph is converted in another form:



Solution Interpretation:

According to the above analysis,

- Pregnancy, glucose, Body mass index, diabetic pedigree function and age acts as driving variables for the person to be diabetic.
- Skin Thickness does not impact the target at all.
- Whereas, Insulin gives a reverse effect for the target outcome as being diabetic.