February 2022 | By: Isha Golakiya 002963001

Module 6: Project – Data Analytics

ALY 6000: Introduction to Analytics

Dr. Mohammad Shafiqul Islam

College of Professional Studies, Northeastern University

Introduction

Chronic condition is a disease or condition which is continuous or whose effect is long lasting. A major adverse effect is caused to the quality of life. One of the most acute diseases which are present all over the world is diabetes. Dealing with such disease is quite difficult. Most of the death all over the globe is caused due to this chronic disease. The statistics in the year 2013 of diabetes disclosed that approx. 380 million individual suffered with this chronic disease. In the year 2012, this disease was 8^{th} leading cause for death in men and 5^{th} leading cause among women. Around 693 million patients are predicted to have this disease in 2065. This chronic condition is also associated with high cost. Limited researches are available in such biological data. Here, 769 Indian patients of diabetes are analyzed wherein 350 female and 318 males are considered. Apart from this, it is perceived that gender has nothing to do with the chance of diabetes.

Methodology

Dataset:

Under the name Pima Indians Diabetes Database on kaggle, the dataset used here is and used. Here is the description provided on the website regarding dataset:

"This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict 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. A Predictor variable includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on."

The dataset has 9 attributes where 8 are independent of each other and 1 ("Outcome") is dependent, 769 observations. To predict whether a patient has diabetes or not based on few diagnostic measures included in the dataset, is the main aim of using this dataset. The best way to grasp the knowledge efficiently is by understanding it through visual tool like graphs, charts and plot. One can get a crystal clear idea regarding the data for further analysis through it.

Data Preprocessing:

The dataset is usually noisy and inconsistent due to the presence of missing values and unrealistic data. If the quality of dataset is low, it has adverse effect on the results. To ensure the best result, it is necessary to preprocess the dataset and then work on it. Cleaning, integrating, transformation, reduction, discretization of data is applied to process the dataset before using it for future use. With respect to time and cost, it is essential to create the data more accurate for further data mining and analysis.

Key Findings

The first and foremost part is to import all the essential libraries which is important for the further operations. Following are the packages and libraries:

- FSA (Fisheries Stock Assessment)
- FSAData
- Magrittr
- Dplyr
- Ggplot2
- Tidyverse
- Corrplot
- Formattable
- Naniar

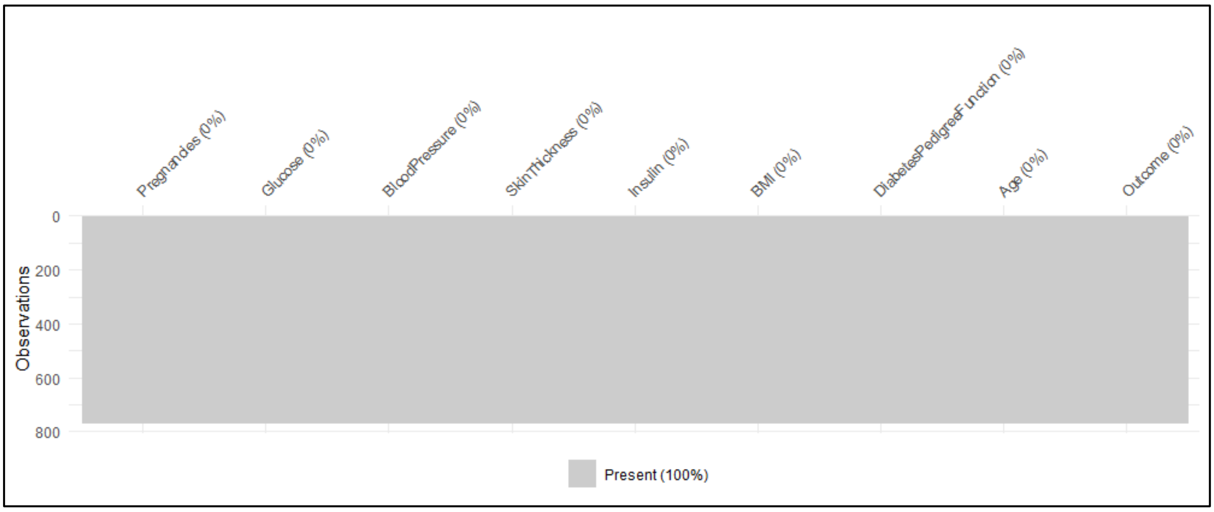
The dataset used here is diabetes.csv. For getting more knowledge regarding the dataset, it's better to display the first 5 and last 5 rows of the dataset. To get more insight regarding the dataset like what's the size of it, what the dataset comprises of, structure of the dataset is preferred to print.

```
> diabete <- read.csv("diabetes.csv")
> View(diabete)
> headtail(diabete)
    Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI DiabetesPedigreeFunction Age
1             6     148            72            35       0 33.6                    0.627  50
2             1      85            66            29       0 26.6                    0.351  31
3             8     183            64             0       0 23.3                    0.672  32
766           5     121            72            23     112 26.2                    0.245  30
767           1     126            60             0       0 30.1                    0.349  47
768           1      93            70            31       0 30.4                    0.315  23
    Outcome
1         1
2         0
3         1
766       0
767       1
768       0
> 
```

```
> #Information regarding dataset
> dim(diabete)
[1] 768   9
> str(diabete)
'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 ...
> 
```

Data cleaning: It comprises of filling up the missing values and datching the noisy data. Here, in the dataset, it can be observed from the screenshots, there is no missing values.

```
> #cleaning data
> # checking missing values
> cat("Number of missing value:", sum(is.na(diabete)), "\n")
Number of missing value: 0
> vis_miss(diabete)
> |
```



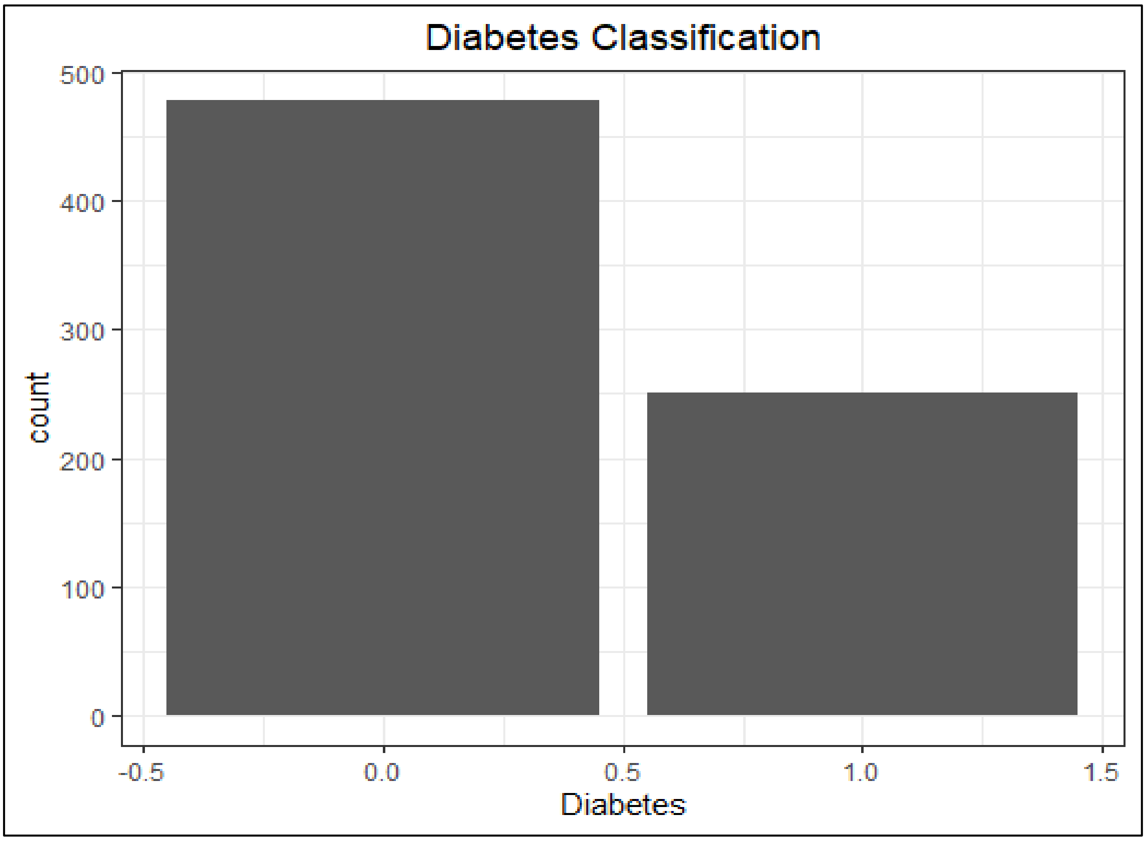
From the above graph, it is conspicuous that there's no missing values in the dataset. 100% data is present. The conversion of data type is essential and hence, changing the class of bloodpressure and BMI and age is been done. Although, the dataset does not have missing values, there are few unrealistic records present which is not needed. Below is the screenshot where a new object is created which does not have values less than 0 in bloodpressure nad BMI measure.

```
> #removing unrealistic values
> diab <- diabete[(diabete$BloodPressure > 0) & (diabete$BMI > 0),]
> diab
   Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI DiabetesPedigreeFunction Age
1            6     148            72            35       0 33.6                    0.627  50
2            1      85            66            29       0 26.6                    0.351  31
3            8     183            64             0       0 23.3                    0.672  32
4            1      89            66            23      94 28.1                    0.167  21
5            0     137            40            35     168 43.1                    2.288  33
6            5     116            74             0       0 25.6                    0.201  30
7            3      78            50            32      88 31.0                    0.248  26
9            2     197            70            45     543 30.5                    0.158  53
11           4     110            92             0       0 37.6                    0.191  30
12          10     168            74             0       0 38.0                    0.537  34
13          10     139            80             0       0 27.1                    1.441  57
14           1     189            60            23     846 30.1                    0.398  59
15           5     166            72            19     175 25.8                    0.587  51
17           0     118            84            47     230 45.8                    0.551  31
18           7     107            74             0       0 29.6                    0.254  31
```

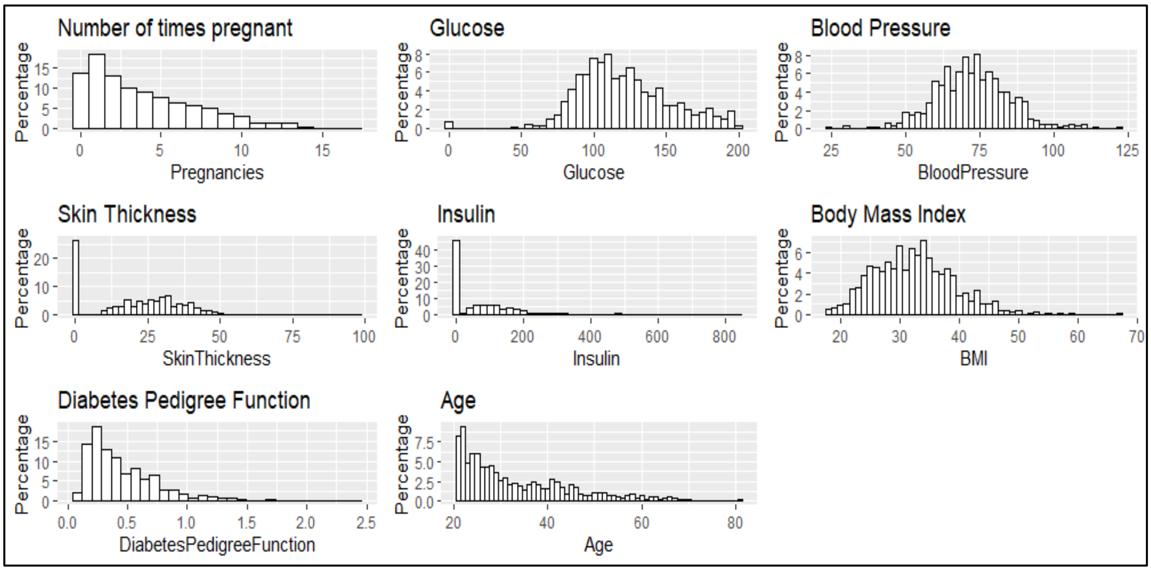
The table below, depicts the clear idea regarding dataset. It highlights the the values of daigonestic measures which is responsible for diabetes in a patient. For example, person having glucose level greater then 130 may have diabetes.

	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
7	3	78	50	32	88	31.0	0.248	26	✔ 1
9	2	197	70	45	543	30.5	0.158	53	✔ 1
11	4	110	92	0	0	37.6	0.191	30	✖ 0
12	10	168	74	0	0	38.0	0.537	34	✔ 1
13	10	139	80	0	0	27.1	1.441	57	✖ 0
14	1	189	60	23	846	30.1	0.398	59	✔ 1

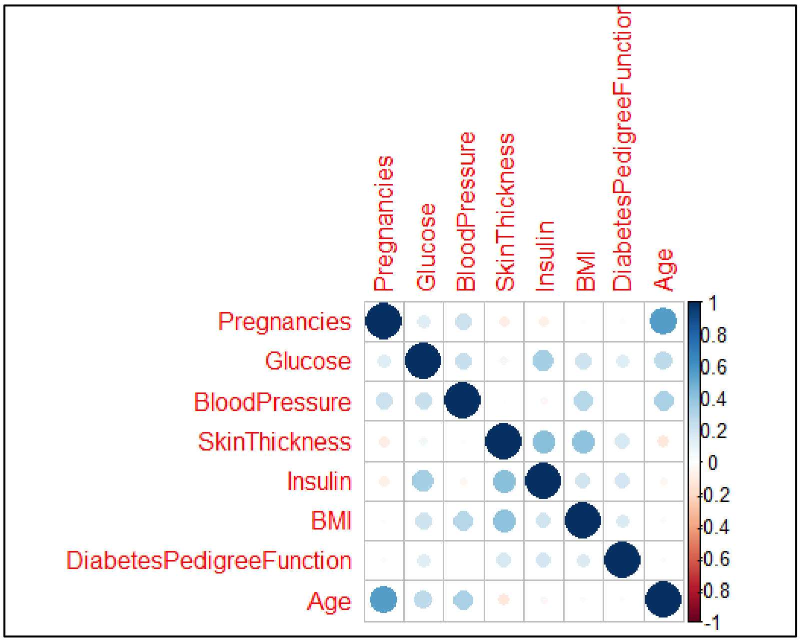
Here, outcome is dependent on others, where value value 1 with tick mark represents a patient has Diabetes and cross sign indicates no diabetic desease found. Below diagram provides a clear idea regarding number of patient suffers from diabetes.



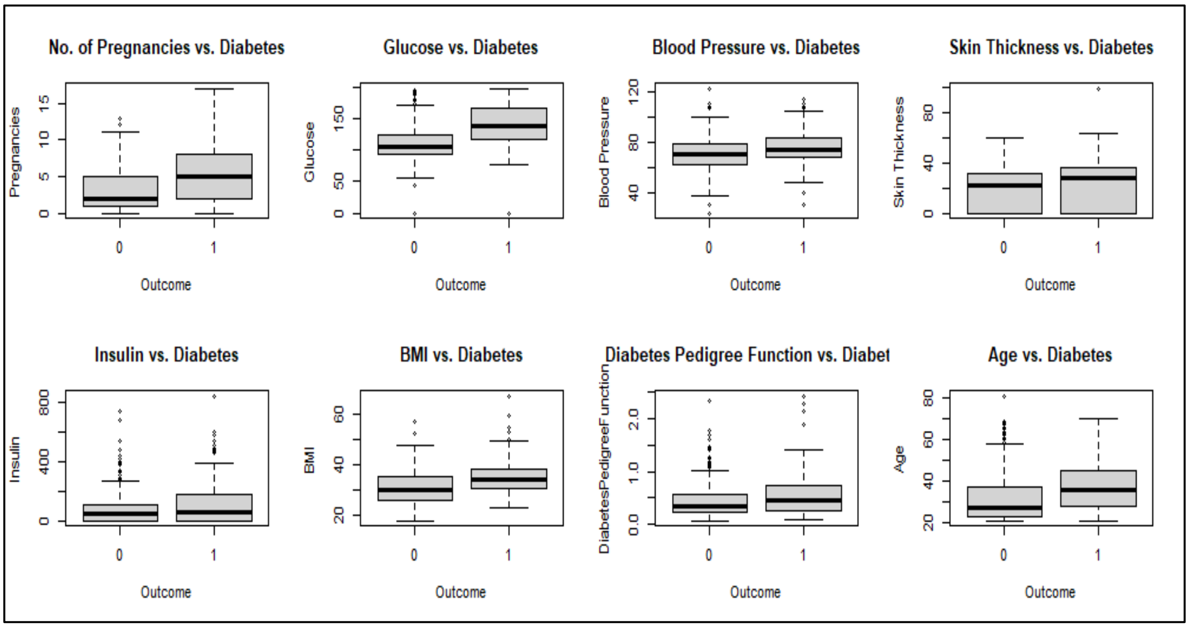
For getting more insight, it is necessary to study individual parameters like how many times patient is pregnant, glucose level of patients, blood pressure, etc. The histograms below depict the information regarding individual measures. All the parameters have a reasonable vast distribution and hence, would be kept for further use.



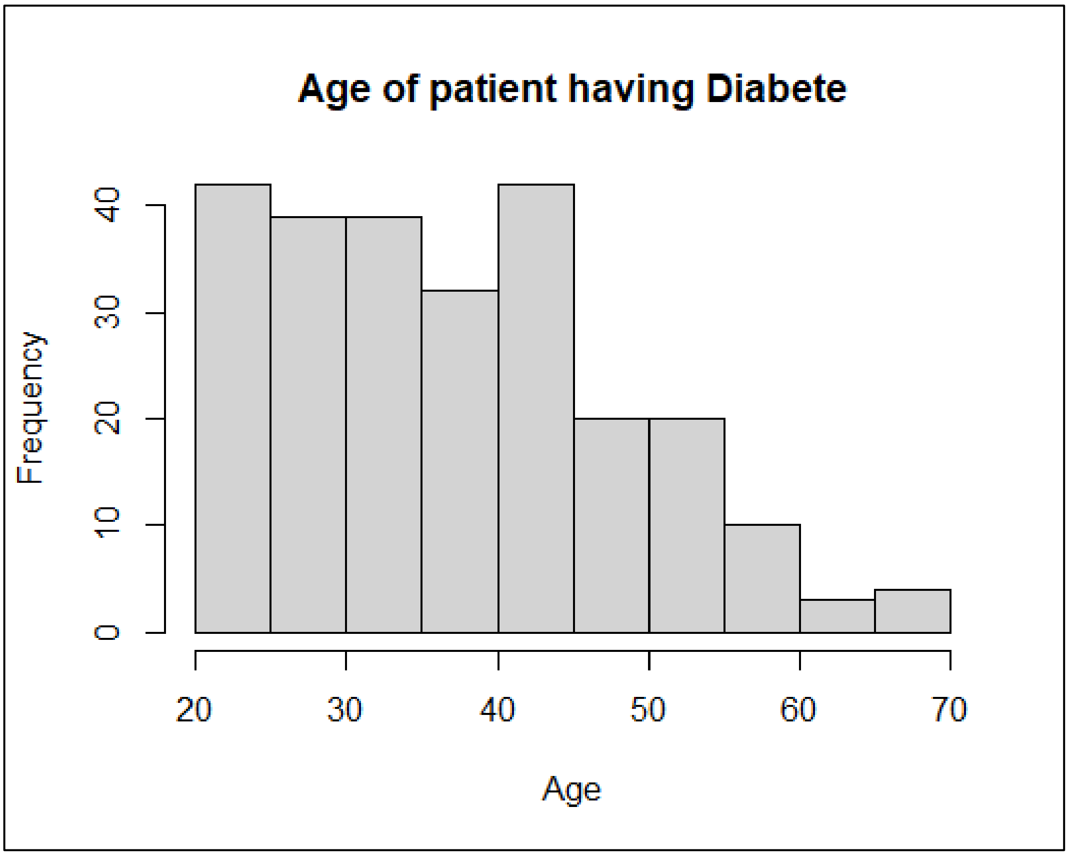
For predicting about a patient having diabetes with certain measure, it is necessary to correlate the entire numeric variables. Below is the correlation plot which depict that variables are almost not correlated.



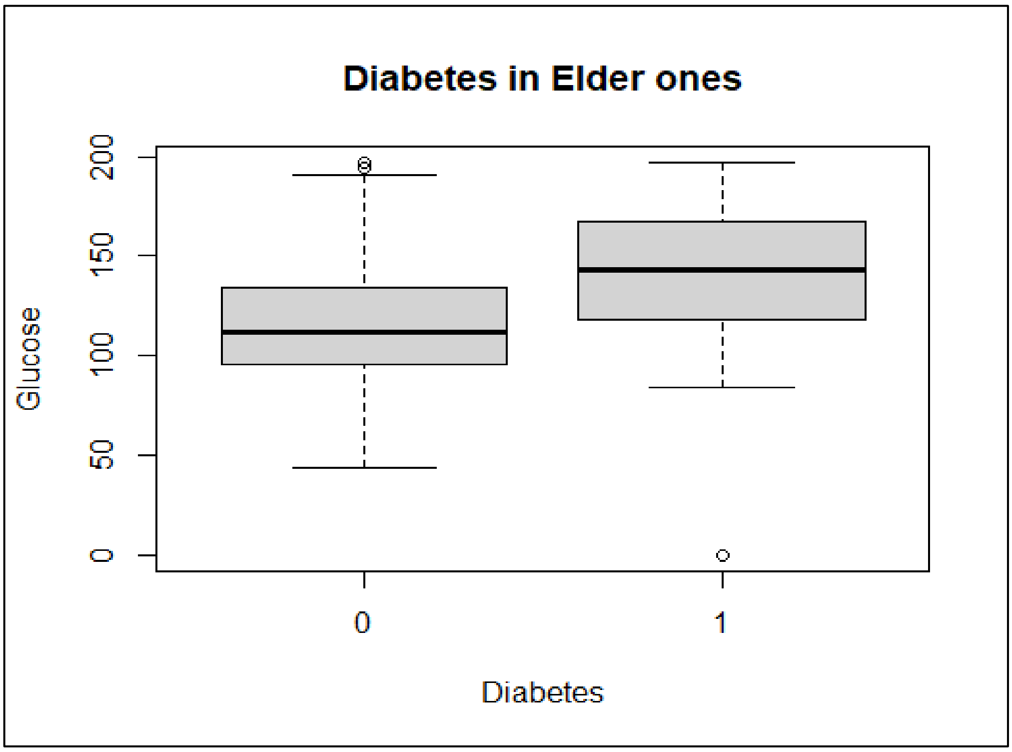
Hence, the only parameter with which all others are related is outcome. Below diagram shows the outliers as well as the individual relation with outcomes. To illustrate, below box plot depicts the patients who have glucose level more than 130 and is diabetic in single case. It can be noticed from the box plot that, blood pressure manifests slight discrepancy with diabetes.



Those patients who are suffering from diabetes are considered, then it can be predicted that the highest effect of this disease is among patient with age group 20-25 and 40-45 years.



However, while taking patient who are above 30 years into consideration, it can be noticed that the glucose level of patient who have diabetes in between 100 and 200.



Conclusion

The capability to predict diabetes early, assumes a vital role for the patient's appropriate treatment procedure. Diagnosis of diabetes is considered a challenging problem for quantitative research. The dataset consider here for prediction is diabetes.csv. It consists of 769 records of Indian patient and 8 diagnostic measures are considered. From the dataset it can be predicted that Elderly people whose glucose level is more than 130 are predicted to have diabetes. Apart from this, the age of patient between 20-25 and 40-45 suffers from diabetes. There is no relation among the diagnostic measures which are taken into consideration.

Bibliography

- Tutorial Point, Scatterplot, Retrieved from:
  Source: https://www.tutorialspoint.com/r/r_scatterplots.htm, Last accessed, February 16, 2022.
- Kaggle, Dataset, Retrieved from:
  Source: https://www.kaggle.com/uciml/pima-indians-diabetes-database, Last accessed, February 15, 2022.
- STHDA, Correlation matrix, Retrieved from:
  Source: http://www.sthda.com/english/wiki/correlation-matrix-a-quick-start-guide-to-analyze-format-and-visualize-a-correlation-matrix-using-r-software , Last accessed, February 17, 2022.
- R fortherestofus, Tables in r, Retrieved from:
  Source: https://rfortherestofus.com/2019/11/how-to-make-beautiful-tables-in-r/, Last accessed, February 18, 2022

Appendix

Here is the code:

```
print("Isha Golakiya")

print(" Diabetes Analysis")

library(FSA)

library(FSAdata)

library(magrittr)

library(dplyr)

library(ggplot2)

library(tidyverse)

library(corrplot)

library(formattable)

library(naniar)


#Reading dataset

setwd("/Users/HP/Downloads")

diabete <- read.csv("diabetes.csv")

View(diabete)

headtail(diabete)


#Information regarding dataset

dim(diabete)

str(diabete)


#cleaning data

# checking missing values
```

```
cat("Number of missing value:", sum(is.na(diabete)), "\n")
vis_miss(diabete)

class(diabete$BloodPressure)

#converting to numeric
diabete$BloodPressure <- as.numeric(diabete$BloodPressure)
diabete$BloodPressure
class(diabete$BloodPressure)
class(diabete$BMI)
diabete$BMI <- as.numeric(diabete$BMI)
diabete$BMI
class(diabete$Age)
diabete$Age <- as.numeric(diabete$Age)
class(diabete$Age)
diabete$Age
diabete$Pregnancies <- as.numeric(diabete$Pregnancies)
class(diabete$Pregnancies)
diabete$Glucose <- as.numeric(diabete$Glucose)
class(diabete$Glucose)
#removing unrealistic values
diab <- diabete[(diabete$BloodPressure > 0) & (diabete$BMI > 0),]
diab
summary(diab)
str(diab)

headtail(diab,5)
```

```
attach(diab)

formattable(diab,  list(

  Glucose = color_tile("white", "Orange"),

  BloodPressure = formatter("span", style = x ~ ifelse(x <= "130", style(color = "green",
font.weight = "bold"),NA)),

  Outcome = formatter("span",

                      style = x ~ style(color = ifelse(x, "green", "red")),

                      x ~ icontext(ifelse(x, "ok", "remove"), ifelse(x, "1", "0"))),

  Age = formatter("span", style = x ~ ifelse(x >= "45", style(color = "blue", font.weight =
"bold"),NA)),

  Pregnancies = formatter("span",

                          style = x ~ style(color = ifelse(x > "1", "red", "grey")))

  ))




#diabetes present or not

ggplot(diab, aes(diab$Outcome, fil= diab$Outcome)) + geom_bar() + theme_bw() +

  labs(title = "Diabetes Classification", x = "Diabetes") +

  theme(plot.title = element_text(hjust = 0.5))


# Histogram

p1 <- ggplot(diab, aes(x=Pregnancies)) + ggtitle("Number of times pregnant") +

  geom_histogram(aes(y = 100*(..count..)/sum(..count..)), binwidth = 1, colour="black",
fill="white") + ylab("Percentage")

p2 <- ggplot(diab, aes(x=Glucose)) + ggtitle("Glucose") +
```

```
  geom_histogram(aes(y = 100*(..count..)/sum(..count..)), binwidth = 5, colour="black", fill="white") + ylab("Percentage")

p3 <- ggplot(diab, aes(x=BloodPressure)) + ggtitle("Blood Pressure") +

  geom_histogram(aes(y = 100*(..count..)/sum(..count..)), binwidth = 2, colour="black", fill="white") + ylab("Percentage")

p4 <- ggplot(diab, aes(x=SkinThickness)) + ggtitle("Skin Thickness") +

  geom_histogram(aes(y = 100*(..count..)/sum(..count..)), binwidth = 2, colour="black", fill="white") + ylab("Percentage")

p5 <- ggplot(diab, aes(x=Insulin)) + ggtitle("Insulin") +

  geom_histogram(aes(y = 100*(..count..)/sum(..count..)), binwidth = 20, colour="black", fill="white") + ylab("Percentage")

p6 <- ggplot(diab, aes(x=BMI)) + ggtitle("Body Mass Index") +

  geom_histogram(aes(y = 100*(..count..)/sum(..count..)), binwidth = 1, colour="black", fill="white") + ylab("Percentage")

p7 <- ggplot(diab, aes(x=DiabetesPedigreeFunction)) + ggtitle("Diabetes Pedigree Function") +

  geom_histogram(aes(y = 100*(..count..)/sum(..count..)), colour="black", fill="white") + ylab("Percentage")

p8 <- ggplot(diab, aes(x=Age)) + ggtitle("Age") +

  geom_histogram(aes(y = 100*(..count..)/sum(..count..)), binwidth=1, colour="black", fill="white") + ylab("Percentage")

grid.arrange(p1, p2, p3, p4, p5, p6, p7, p8, ncol=2)


#Relation of all the numeric values (correlation matrix)

?corrplot

cor_data <- cor(diab[,setdiff(names(diab), 'Outcome')])

cor_data


corrplot(cor_data)
```

```
#Relation between Independent and dependent variable
attach(diab)
par(mfrow=c(2,4))
boxplot(Pregnancies ~ Outcome, main="No. of Pregnancies vs. Diabetes",
        xlab="Outcome", ylab="Pregnancies")
boxplot(Glucose~Outcome, main="Glucose vs. Diabetes",
        xlab="Outcome", ylab="Glucose")
boxplot(BloodPressure~Outcome, main="Blood Pressure vs. Diabetes",
        xlab="Outcome", ylab="Blood Pressure")
boxplot(SkinThickness~Outcome, main="Skin Thickness vs. Diabetes",
        xlab="Outcome", ylab="Skin Thickness")
boxplot(Insulin~Outcome, main="Insulin vs. Diabetes",
        xlab="Outcome", ylab="Insulin")
boxplot(BMI~Outcome, main="BMI vs. Diabetes",
        xlab="Outcome", ylab="BMI")
boxplot(DiabetesPedigreeFunction~Outcome, main="Diabetes Pedigree Function vs.
Diabetes", xlab="Outcome", ylab="DiabetesPedigreeFunction")
boxplot(Age~Outcome, main="Age vs. Diabetes",
        xlab="Outcome", ylab="Age")
#Adding a new column agegroup
attach(diab)
diab <- mutate(diab, Agegroup = if_else(Age > 30, "Elder", "Younger"))
diab
diagnosis <- diab[ (diab$Agegroup == "Elder"),]
diagnosis
```

```
boxplot(diagnosis$Glucose~diagnosis$Outcome, main= "Diabetes in Elder ones", xlab =
"Diabetes", ylab = "Glucose")


# Table with diabetic patient

out <- diab[(diab$Outcome == "1"),]

out

hist(out$Age, xlab = "Age", main = "Age of patient having Diabete")
```