# Predicting the Likelihood of Diabetes Using Common Signs and Symptoms

## Project Phase1 | MATH1298 Analysis of Categorical Data | RMIT University

Udeshika Dissanayake | s3400652 | Project Groups 60

September 23, 2020

# List of Contents



# Data Source and Description

The data set consists of signs and symptoms of 520 newly diabetic or would be diabetic patients, who presented at Sylhet Diabetes Hospital in Sylhet, Bangladesh. The data had been collected using direct questionnaires method at the hospital under the supervisor of Doctors. The Source for the data set is the UCI Machine Learning Repository (Dua, D. & Graff, C., 2017) at, archive.ics.uci.edu (https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset.) (Faniqul, M. M. I. et al., 2019). The data set has 16 descriptive features and one target feature.

## Descriptive Features

Below table explains the descriptive features in the data set that will be used in the model.

Table 1: Descriptive features

| Name | Data Type | Units | Description |
|---|---|---|---|
| Age | numerical | Years | Age of the patient |
| Gender | binary | Female, Male | Gender of the patient |
| Polyuria | binary | Yes, No | Body urinates more than usual and passes excessive or abnormally large amounts of urine each time |
| Polydipsia | binary | Yes, No | Dry mouth and excessive thirst |
| Sudden weight loss | binary | Yes, No | Unexplained sudden weight loss |
| Weakness | binary | Yes, No | Fatigue/ Feelings of exhaustion and lethargy |
| Polyphagia | binary | Yes, No | Excessive or extreme hunger |

| Name | Data Type | Units | Description |
|---|---|---|---|
| Genital thrush | binary | Yes, No | Itching, irritation and swelling around the Genital organs |
| Visual blurring | binary | Yes, No | Lack of sharpness of vision and inability to see fine details |
| Itching | binary | Yes, No | Irritating sensation that makes you want to scratch your skin |
| Irritability | binary | Yes, No | Feeling frustrated or getting upset easily |
| Delayed healing | binary | Yes, No | Delayed wound healing, recurrent or severe infections |
| Partial paresis | binary | Yes, No | Weakening of a muscle or group of muscles |
| Muscle stiffness | binary | Yes, No | Muscles feel tight and more difficult to move than you usually do |
| Alopecia | binary | Yes, No | Patches of hair loss on the head and on other parts of the body |
| Obesity | binary | Yes, No | Excessive amount of body fat |

## Target Feature

The name of the target feature is "Class" and it's labels are as follows,

$$\text{Class} = \begin{cases} Positive & \text{if the patient is diagnosed as a diabetic patient} \\ Negative & \text{if the patient is not diagnosed as a diabetic patient} \end{cases}$$

The target feature has two levels. Hence this can be classified as binomial target feature.

# Goals and Objectives

About one third of patients with diabetes do not know that they have diabetes according to the findings published by many diabetes institutes around the world (ASPE, 2017). Detecting and treating diabetes patients at early stages is critical in order to keep them healthy and to ensure their quality of life is not compromised. Early detection will also help to mitigate the risk of serious complications like heart disease & stroke, blindness, limb amputations, and kidney failures as a result of diabetes (ASPE, 2017).

This study intends to build a logistic regression model to predict the likelihood of having diabetes using common signs and symptoms presented by patients. A successful model will enable early detection of diabetes through signs and symptoms shown by possible patients.

This study consists with two phases: 1) Phase I - preprocess and explore the data set in order to make it ready to consume for model development. 2) Phase II - build a logistic regression model to predict the likelihood of having diabetes based on signs and symptoms.

All the activities have been performed in R package and the report has been compiled using R-Markdown. This report covers both narratives and R pseudocode for data preprocessing & exploration activities that have been performed under the phase I.

# Data Cleaning and Preprocessing

## Retrieving Data Set

The diabetes data set has been loaded in to R Studio using the *read_csv()* function in the *readr* package and then print the dimension of the data frame to check whether the data set has been loaded correctly.

```
diabetes<-read_csv("diabetes.csv")
dim(diabetes)
```

```
[1] 520  17
```

Random 5 rows have been printed using *sample_n()* function in *dplyr* package to inspect further and check whether the features and descriptions outlined in the source documentation are aligning with the data frame.

```
kbl(sample_n(diabetes,5), caption = "Table 2: Random 5 rows from data set") %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed"), full_width = F, position = "left",font_s
ize = 10)
```

Table 2: Random 5 rows from data set

| Age | Gender | Polyuria | Polydipsia | sudden weight loss | weakness | Polyphagia | Genital thrush | visual blurring | Itching | Irritability | delayed healing | partial paresis | muscle stiffness | Alopecia | Obesity | class |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 46 | Male | No | No | No | Yes | No | No | No | Yes | No | Yes | No | No | Yes | No | Negative |
| 44 | Male | No | Yes | No | No | No | Yes | No | No | No | Yes | No | No | No | Yes | Positive |
| 64 | Male | No | No | No | Yes | Yes | No | Yes | Yes | Yes | Yes | No | Yes | Yes | No | Negative |
| 61 | Male | No | No | No | Yes | No | Yes | No | Yes | No | Yes | No | No | Yes | No | Negative |
| 30 | Female | Yes | Yes | No | No | No | No | No | Yes | No | Yes | Yes | No | No | No | Positive |

As per the above R-outputs, the loaded data set is aligning with the data set description on the data source.

Data types in the original data frame are:

```
sapply(diabetes, class)
```

```
##               Age            Gender          Polyuria         Polydipsia
##          "numeric"       "character"       "character"       "character"
## sudden weight loss          weakness        Polyphagia     Genital thrush
##       "character"       "character"       "character"       "character"
##    visual blurring           Itching      Irritability    delayed healing
##       "character"       "character"       "character"       "character"
##    partial paresis   muscle stiffness          Alopecia           Obesity
##       "character"       "character"       "character"       "character"
##             class
##       "character"
```

As shown in the R-output above, the data type of the 'Age' feature is "numeric", whereas the data type for all the other descriptive features including target is "character".

# Data Type Conversion

All the variables except the 'Age' variable should be in factor data type. However in the data set they are defined as character variables. Using below code, variables with character data type have then been converted to "factor" type for this study.

```
diabetes[2:17] <- lapply(diabetes[2:17], as.factor)
```

After completing the data type conversion, the data types of the frame are as below:

```
#checking variable types in the data frame
sapply(diabetes, class)
```

```
##               Age            Gender          Polyuria         Polydipsia
##          "numeric"          "factor"          "factor"          "factor"
## sudden weight loss          weakness        Polyphagia     Genital thrush
##          "factor"          "factor"          "factor"          "factor"
##    visual blurring           Itching      Irritability    delayed healing
##          "factor"          "factor"          "factor"          "factor"
##    partial paresis   muscle stiffness          Alopecia           Obesity
##          "factor"          "factor"          "factor"          "factor"
##             class
##          "factor"
```

# Checking for Missing Values in the Data Set

Below codes have been executed to identify if there are any missing values in the data set. It is clearly evident that there are no missing values in the data set.

```
na_count <-sapply(diabetes, function(y) sum(length(which(is.na(y)))))
na_count <- data.frame(na_count)

kbl(na_count, caption = "Table 3: Count of missing values") %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed"), full_width = F, position = "left")
```

Table 3: Count of missing values

| | na_count |
|---|---|
| Age | 0 |
| Gender | 0 |
| Polyuria | 0 |
| Polydipsia | 0 |
| sudden weight loss | 0 |
| weakness | 0 |
| Polyphagia | 0 |
| Genital thrush | 0 |
| visual blurring | 0 |
| Itching | 0 |
| Irritability | 0 |
| delayed healing | 0 |
| partial paresis | 0 |
| muscle stiffness | 0 |
| Alopecia | 0 |
| Obesity | 0 |
| class | 0 |

# Checking for Typo in Categorical Features

Types of all categorical features, including the target feature in the data set has been checked by investigating the frequency tables using *summary()* function in *vcd* package. As can be seen below, there are no typos in the categorical features in the data set.

```
summary(diabetes[2:17])
```

```
##     Gender    Polyuria  Polydipsia sudden weight loss weakness  Polyphagia
##  Female:192   No :262   No :287    No :303            No :215   No :283
##  Male  :328   Yes:258   Yes:233    Yes:217            Yes:305   Yes:237
##  Genital thrush visual blurring Itching   Irritability delayed healing
##  No :404        No :287         No :267   No :394      No :281
##  Yes:116        Yes:233         Yes:253   Yes:126      Yes:239
##  partial paresis muscle stiffness Alopecia  Obesity        class
##  No :296         No :325          No :341   No :432   Negative:200
##  Yes:224         Yes:195          Yes:179   Yes: 88   Positive:320
```

# Checking Extra White-spaces & Capital Letter Mismatches in Categorical Features

Extra white-spaces & capital letter mismatches in the categorical data have already been checked while investigating the frequency tables in previous section ( Checking for typo in Categorical Features ).

# Checking for Impossible Numerical Values in Age Feature

Summary statistics has been checked using *summary()* function in the vcd package in order to check whether there are any impossible numerical values in 'Age' variable. As per the summary statistics, the 'Age' variable spans from 16 to 90. Therefore, this data set doesn't have any impossible values.

`

```
summary(diabetes$Age)
```

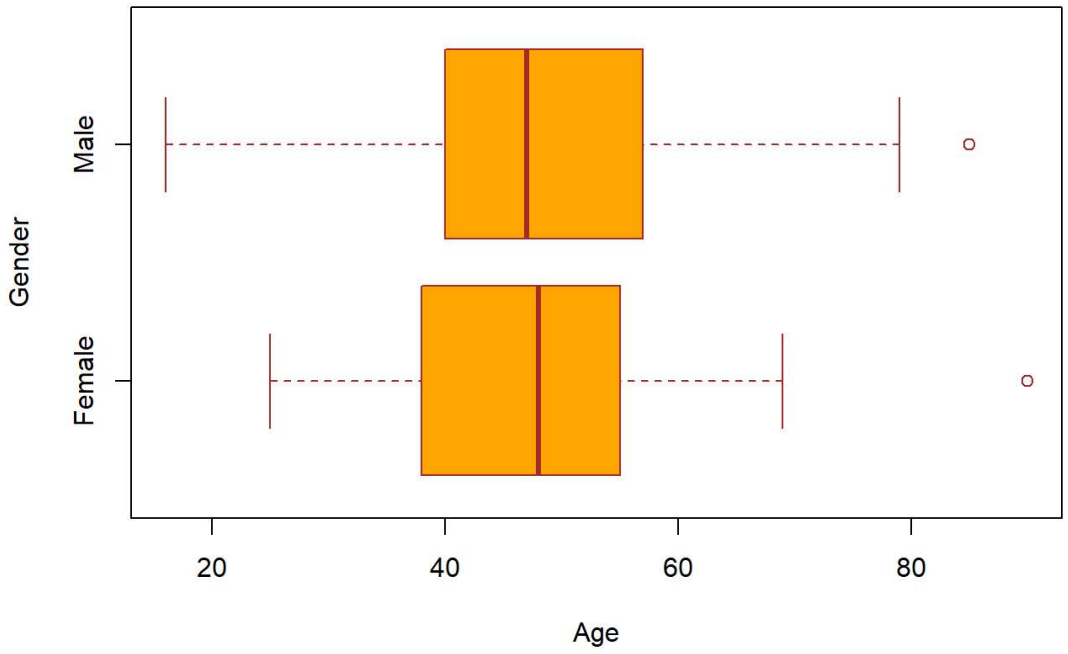
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   16.00   39.00   47.50   48.03   57.00   90.00
```

# Checking for Outliers in Age Feature

Box-plot is one of the best method to visualize outliers of numerical attributes. Any dots outside the whiskers are good candidates for outliers. The only numerical variable to be checked for outliers in the data set is 'Age' and as per the box-plot, few outliers can be seen:

```
boxplot(Age~Gender,data=diabetes, main = "Figure 1: Boxplot of Age Distribution Before Removing Outliers",
        xlab = "Age",
        col = "orange",
        border = "brown",
        horizontal = TRUE)
```

**Figure 1: Boxplot of Age Distribution Before Removing Outliers**




Then corresponding row numbers for these outliers are checked using the below R-Code.

```
# row number corresponding to these outliers
out <- boxplot.stats(diabetes$Age)$out
out_ind <- which(diabetes$Age %in% c(out))
out_ind
```

```
## [1] 102 103 186 187
```

Rows 102, 103, 186, and 187 are outliers as per the results above. It is better to investigate those rows before removing these outliers. As shown in the below table, two female and two male patients are found to be outliers and all of them are diagnosed as diabetes patients.

```
#Examining the relevant rows which are having outliers
diabetes[out_ind, ] %>%
  kbl(caption = "Table 4: Outliers in the data set") %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed"),
                font_size = 10, full_width = F, position = "left")
```

Table 4: Outliers in the data set

| Age | Gender | Polyuria | Polydipsia | sudden weight loss | weakness | Polyphagia | Genital thrush | visual blurring | Itching | Irritability | delayed healing | partial paresis | muscle stiffness | Alopecia | Obesity | class |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 85 | Male | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | No | Yes | Yes | Yes | No | No | Positive |
| 90 | Female | No | Yes | Yes | No | No | Yes | Yes | Yes | No | No | No | Yes | Yes | No | Positive |
| 85 | Male | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | No | Yes | Yes | Yes | No | No | Positive |
| 90 | Female | No | Yes | Yes | No | No | Yes | Yes | Yes | No | No | No | Yes | Yes | No | Positive |

Due to the fact that all four of these patients are above 85 years old, and assuming that they could have age related signs similar to that of diabetes symptoms, removing them from the data set is recommended to achieve the objective of the study of early detection of diabetes through its symptoms.

The z-score method has been used as below to remove those outliers from the data set.

```
#Summary statistics from z-score method
z.scores <- diabetes$Age %>% scores(type = "z")
z.scores %>% summary()
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -2.63580 -0.74303 -0.04352  0.00000  0.73828  3.45400
```

```
#Removing the Outliers
diabetes_new<- diabetes[which( abs(z.scores) <3 ),]
dim(diabetes_new)
```

```
## [1] 516  17
```

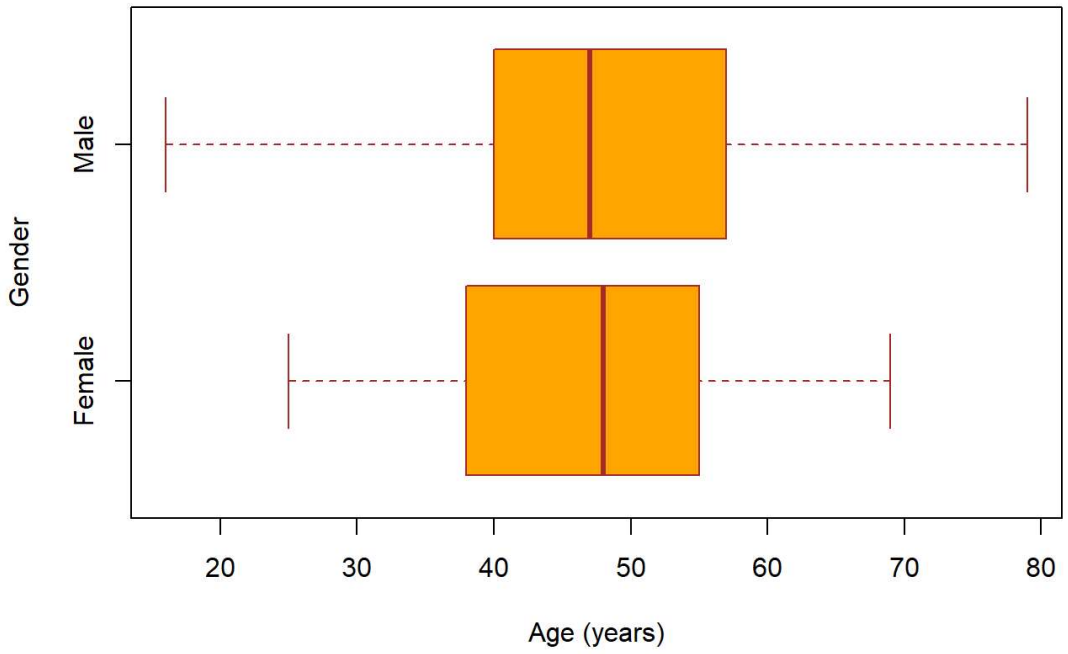
After removing the outliers, data set now contains information for 516 patients. As shown below, the Z-score test has again been executed to ensure that there are no further outliers.

```
z.scoresN <- diabetes_new$Age %>% scores(type = "z")
which( abs(z.scoresN) >3 )
```

```
## integer(0)
```

```
boxplot(Age~Gender,data=diabetes_new, main = "Figure 2: Boxplot of Age Distribution After Removing Outliers",
        xlab = "Age (years)",
        col = "orange",
        border = "brown",
        horizontal = TRUE)
```

**Figure 2: Boxplot of Age Distribution After Removing Outliers**




# Data Exploration and Visualization

## One-variable Plots

One-variable plots can be used to investigate the distribution and the characteristics of each attribute. The histogram has been used to explore the numerical feature, while frequency plots have been used to explore categorical features using *dplyr*, *ggplot2*, *tidyr* and *scales* packages.

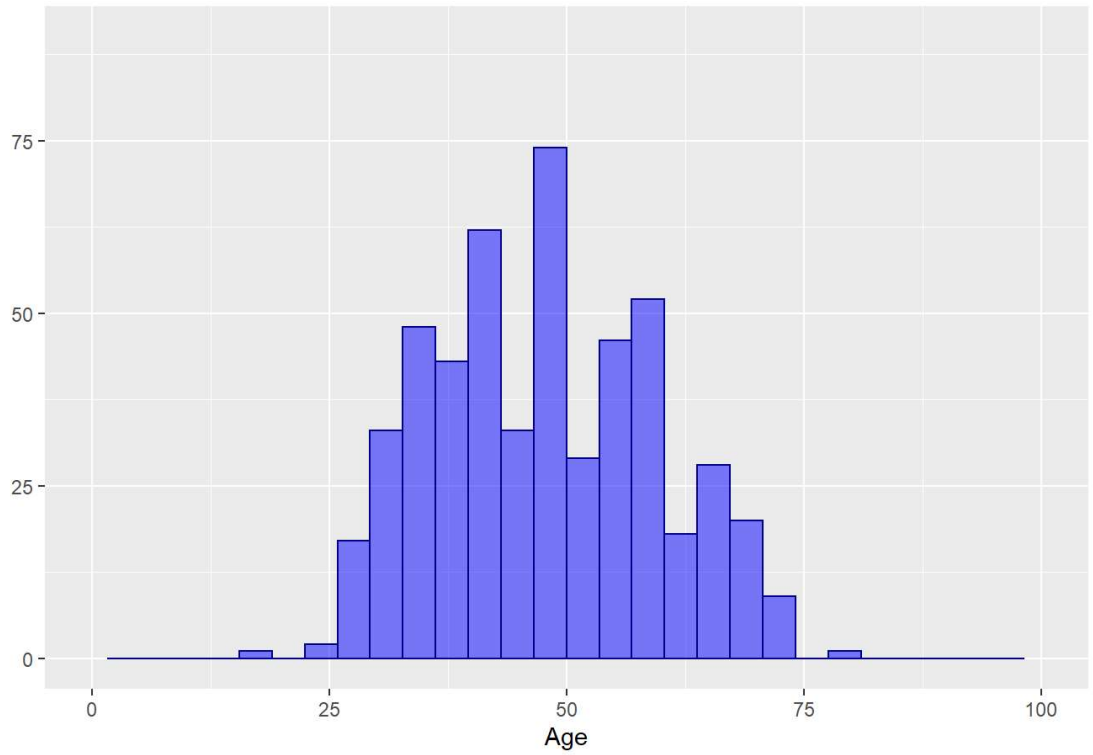
`

```
summary(diabetes_new$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   16.00   39.00   47.00   47.72   56.00   79.00
```

```
ggplot(data=diabetes_new, aes(x=Age)) +
  geom_histogram(col="dark blue",
                 fill="blue",
                 alpha = .5) +
  labs(title="Figure3: Histogram for Age", x="Age", y="") +
  xlim(c(0,100)) +
  ylim(c(0,90))
```

Figure3: Histogram for Age




Above figure shows the distribution of the 'Age' variable, which spans from around 16 years to almost 79 years. The middle 50% of the age resides between 39 years to 56 years as can be seen from summary statistics table. The shape of the histogram hints a slight right skewness with mean around 48 years. This suggests the higher proportion of the patients who visited this diabetes hospital are mid to older people.

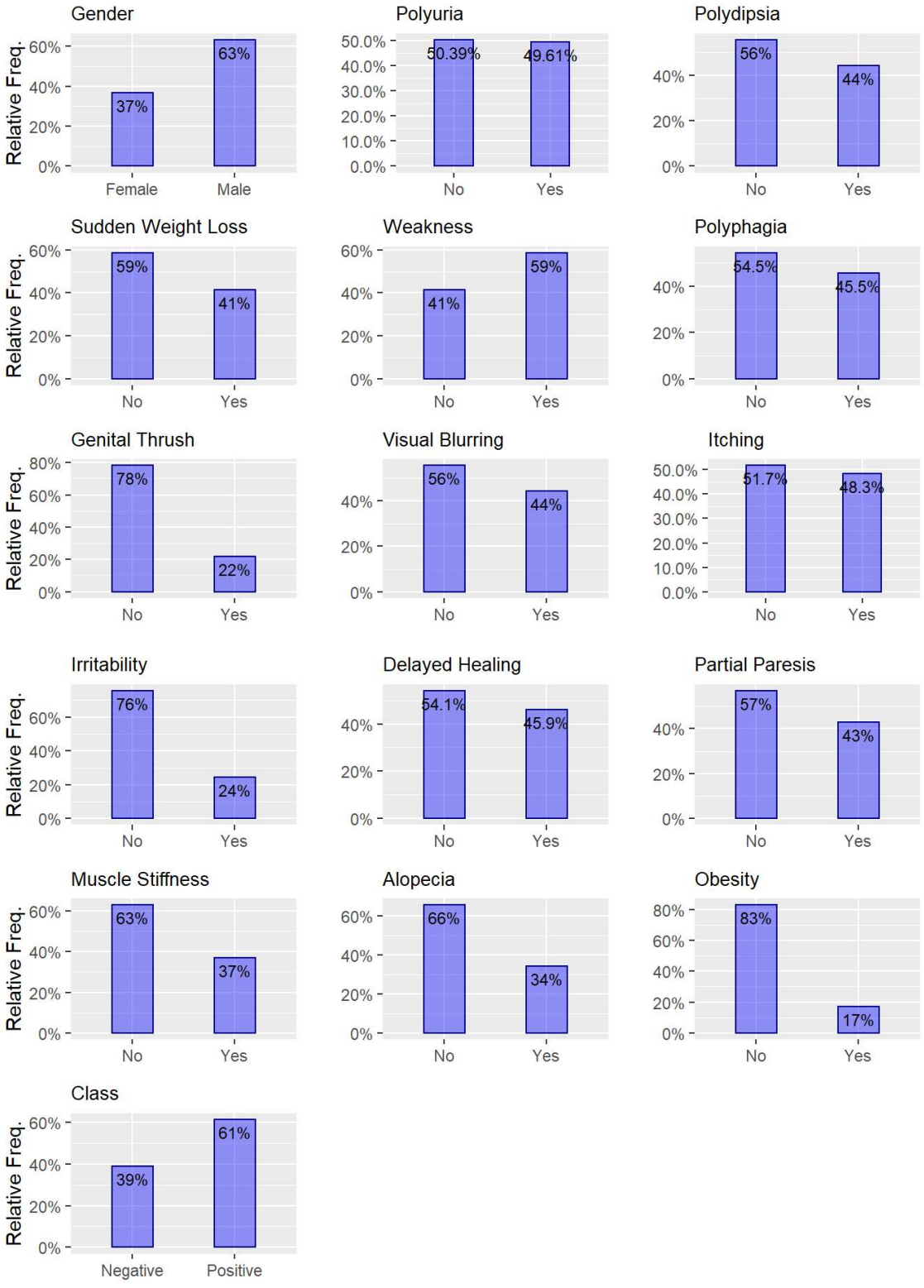
All other variables with factor data type have also been explored using relative frequency plots as shown below,

```
#Propotional Bar Charts for Gender

plot1 <- ggplot(diabetes_new, aes(Gender)) +
  geom_bar(aes(y = (..count..)/sum(..count..)),color="dark blue", fill="blue", alpha=0.4, width = 0.4) +
  scale_y_continuous(labels=scales::percent) +
  ylab("Relative Freq.")+
  geom_text(aes( label = scales::percent((..count..)/sum(..count..)),
                 y= (..count..)/sum(..count..) ), stat= "count", vjust = 1.5,colour="black",size=3)+
  labs(title="Gender")+
  theme(plot.title = element_text(size = 10),axis.title.x = element_blank())
```

```
grid.arrange(plot1, plot2, plot3, plot4,plot5,plot6,plot7,plot8,plot9,
             ncol=3, widths=c(2.6, 2.6, 2.6),
            top = grid::textGrob("Figure 4: Propotional Bar Charts for Categorical Features", x = 0, hjust = 0))
```

Figure 4: Propotional Bar Charts for Categorical Features



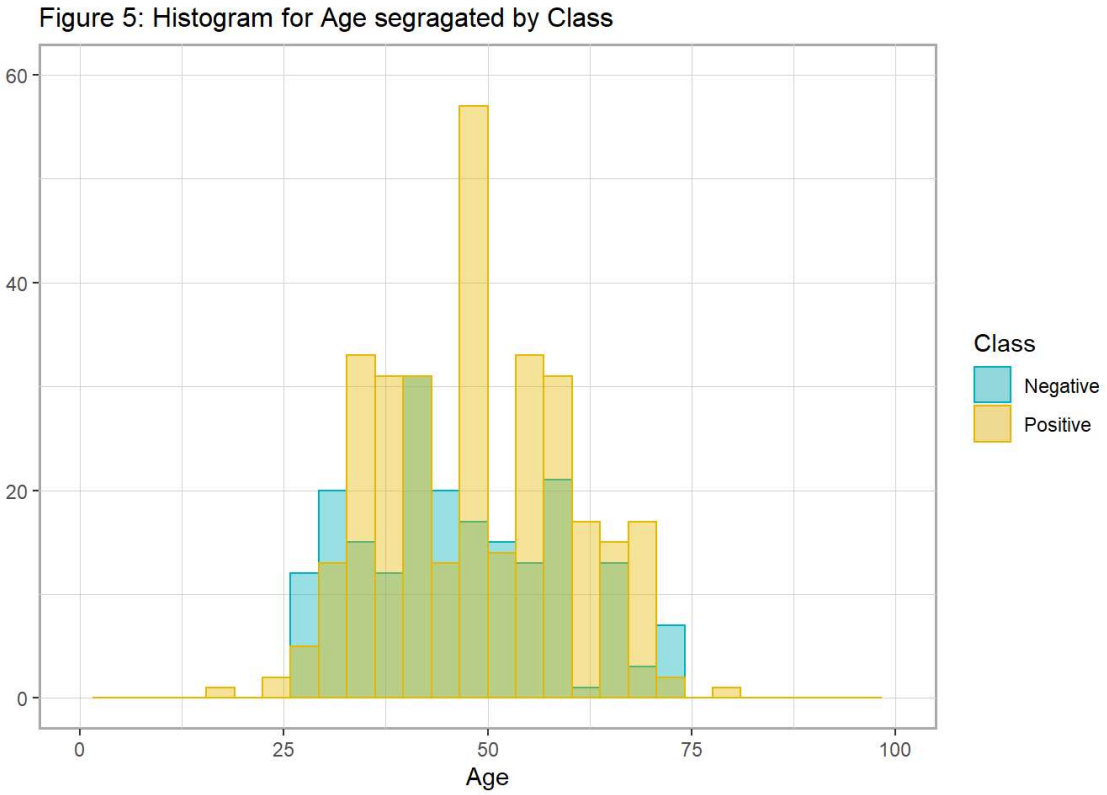

*Propotional Bar Charts for Categorical Features*

It is worth noting that the male population is dominating in the data set with 63%. As can be seen, there are fourteen sign and symptoms recorded in the data set and these signs and symptoms were presented within the sample patients ranging from 17% (least – Obesity) to 59% (most – Weakness). Finally, it is important to mention that only 61% of the patients in the data set are diabetes positive.

# Two-variable Plots

In order to obtain further insight on the data set, two-variable data exploration was performed. Below code plots the histograms for 'Age' feature segregated by Class (i.e. diabetes positive or negative).

```
# Histogram for Age segragated by Class
ggplot(diabetes_new, aes(x = Age)) +
  geom_histogram(aes(color = diabetes_new$class, fill = diabetes_new$class),
                 position = "identity", bins = 30, alpha = 0.4) +
  scale_color_manual(values = c("#00AFBB", "#E7B800"),name="Class") +
  scale_fill_manual(values = c("#00AFBB", "#E7B800"),name="Class")+
  labs(title="Figure 5: Histogram for Age segragated by Class") +
  xlim(c(0,100)) +
  ylim(c(0,60))+
  theme(plot.title = element_text(size = 12),axis.title.y = element_blank(),panel.background = element_rect(fill
= "white",colour = "dark gray",
                                size = 1, linetype = "solid"),
  panel.grid.major = element_line(size = 0.2, linetype = 'solid',
                                colour = "light gray"),
  panel.grid.minor = element_line(size = 0.1, linetype = 'solid',
                                colour = "light gray"))
```

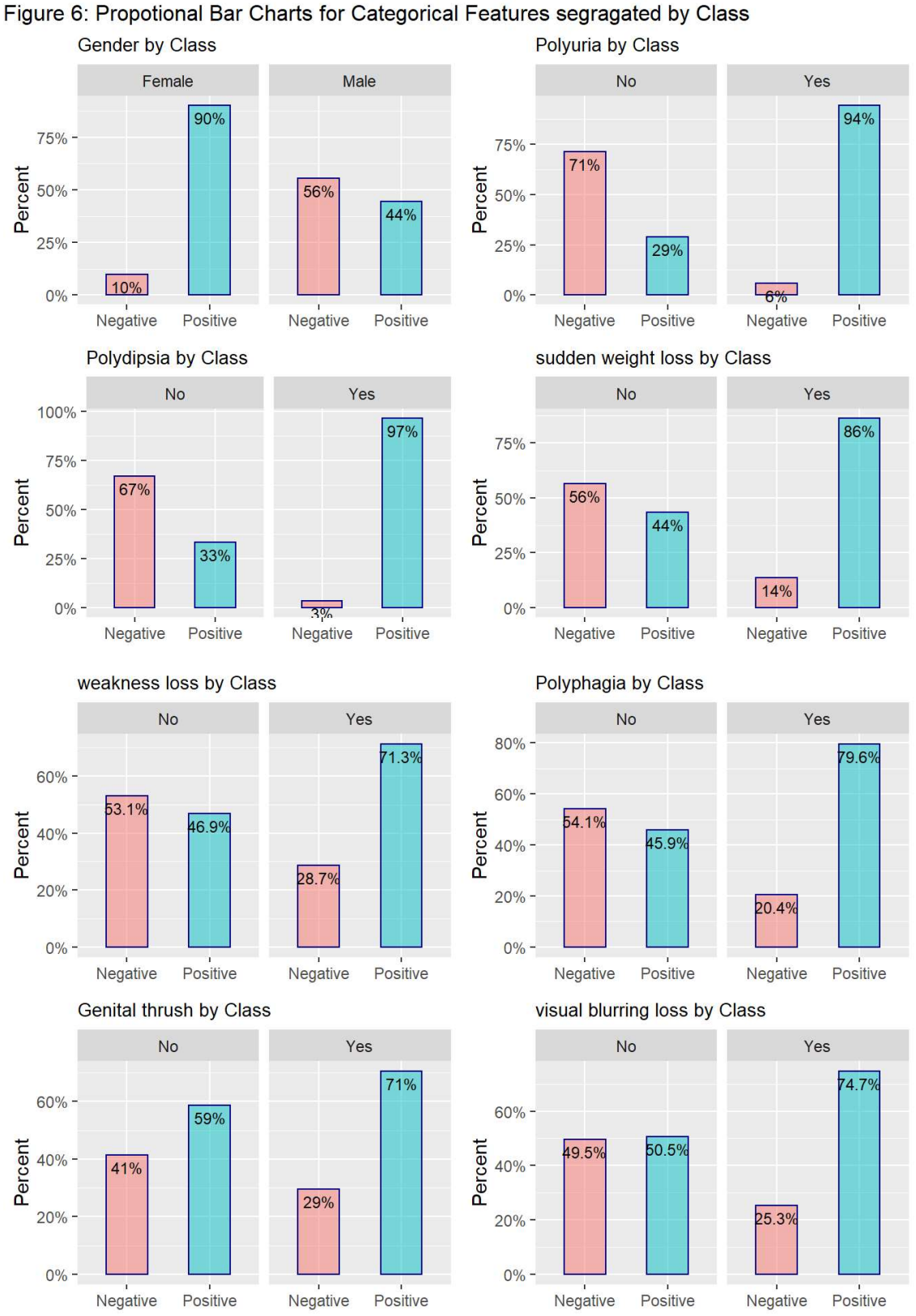


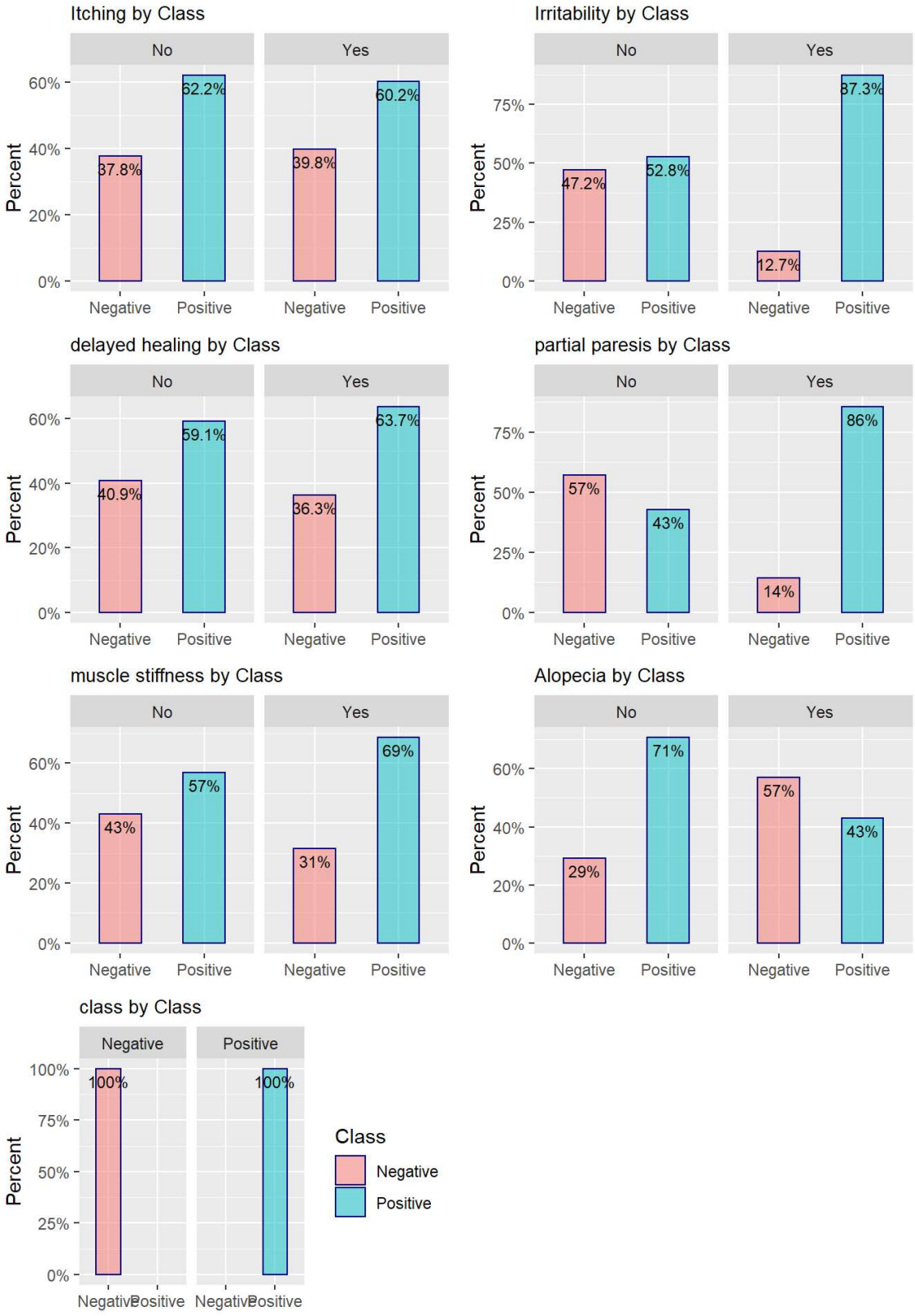

It has been noticed that most patients between age 25 to 32 years within the data set are proportionately diabetes negative, while majority of patients above 32 years are proportionately diabetes positive with the exception of 43 - 47 years, 50 – 53 years, and 71 - 74 years age groups, which shows slightly different results. Further, within the data set, it is observed that 47 - 50 and 62 - 65 age groups have shown a significantly high proportion of positive diabetes cases compared to other age groups. The shown variation of diabetes positive proportions across the age groups could be due to the small sample size and real trend (if any) with better intuition would be able to achieve by exploring larger data set.

The fourteen signs and symptoms which have categorical features, have been explored against target feature ‘Class’ (i.e. diabetes positive or negative) as shown below. Respective proportional bar plots segregated by ‘Class’ have been plotted in order to obtain better insight by comparing the normalized values instead of counts.

```
#Gender by Class
p1 <- ggplot(diabetes_new, aes(x= class,  group=Gender)) +
    geom_bar(aes(y = ..prop.., fill = factor(..x..)), stat="count", alpha=0.5,color="dark blue", width = 0.5,sho
w.legend = FALSE) +
    geom_text(aes( label = scales::percent(..prop..),
                   y= ..prop.. ), stat= "count", vjust = 1.5, colour="black",size=3) +
    labs(y = "Percent", title="Gender by Class") +
    facet_grid(~Gender) +
    scale_y_continuous(labels = scales::percent)+
  scale_fill_discrete(name="Class",labels=c("Negative", "Positive"))+
  theme(plot.title = element_text(size = 10),axis.title.x = element_blank())
```

```
grid.arrange(p1, p2, p3, p4,
             ncol=2, widths=c(2.6, 2.6),
              top = grid::textGrob("Figure 6: Propotional Bar Charts for Categorical Features segragated by Clas
s", x = 0, hjust = 0))
```






Surprisingly, the proportion of diabetes positive females in the data set is significant high (90%) compared to that of male (44%), despite the fact the female patients in the data set is noticeably low (37%) compared to male (67%). It will be interesting to conduct a study to investigate the reason behind this. Could this be due to females in Bangladesh are less likely to visit hospitals compared to males or could females be tolerating illnesses more compared to males. Such analysis is out of the scope of this study, therefore do not carry out further analysis on those lines in this study.

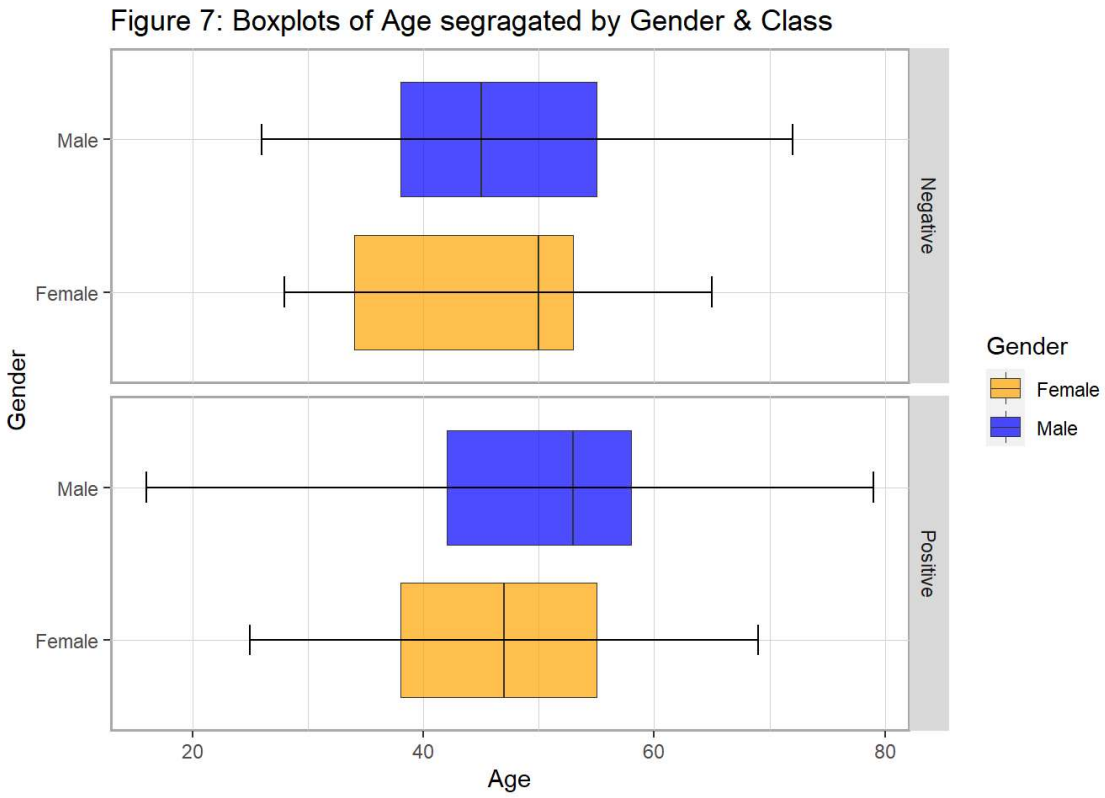
As can be seen from the above frequency plots, more than 70% of population with Polyuria, Ploydispia, Weight loss, Weakness, Ploypha gia, Genital thrush, Blurring, Irritability, and Partial paresis signs & symptoms, independently, have shown diabetes positive. On the other hand, more than 50% of population without Polyuria, Ploydispia, Weight loss, Weakness, Ploypha gia, and Partial paresis signs & symptoms, have shown diabetes negative.

This results at first sight tends someone to think that signs and symptoms like Polyuria, Ploydispia, Weight loss, Weakness, Ploypha gia, and Partial paresis would have high contributions to the logistic regression model that will be build in next phase.

# Three-variable Plots

Finally, features in the data set are explored taking three variables at a time and by plotting respective box plots as shown below,

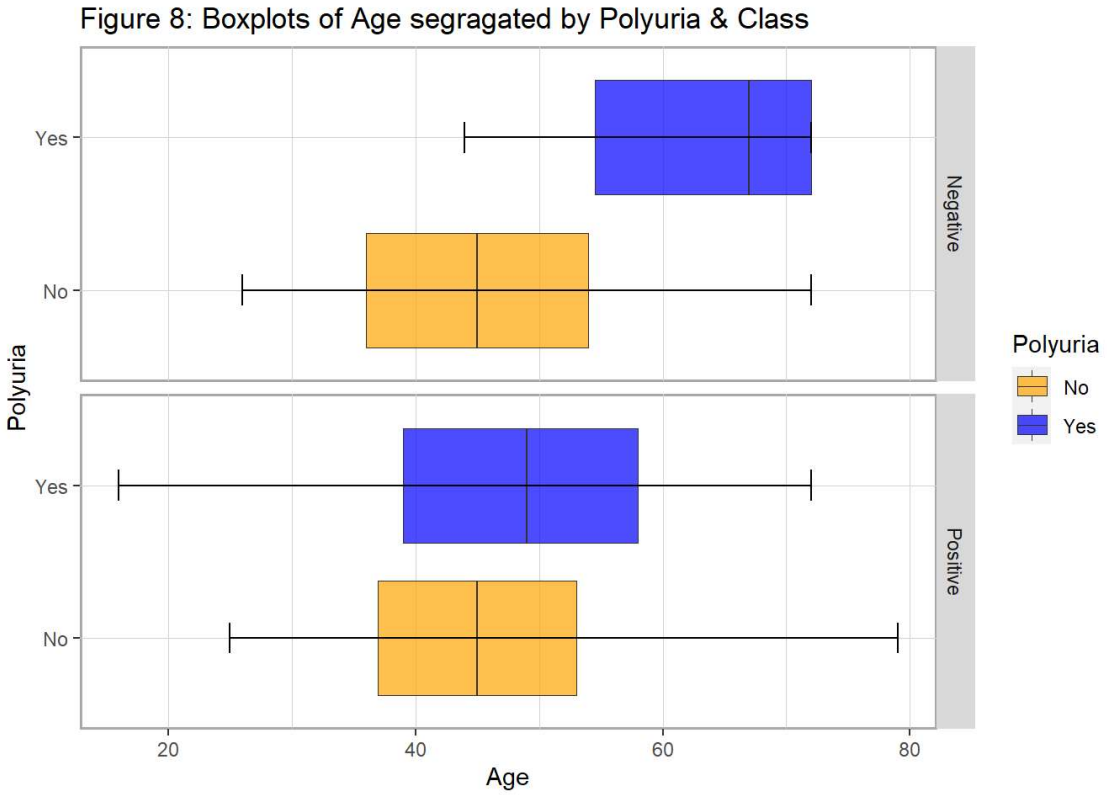
```
bp <- ggplot(data=diabetes_new, aes(x=Age, y=Gender, group=Gender)) +
  geom_boxplot(aes(fill=Gender), alpha=0.7,outlier.shape=NA,lwd=0.2)
bp + facet_grid(diabetes_new$class ~.)+ stat_boxplot(geom = 'errorbar', width = 0.2,coef = 3)+
  theme(
  panel.background = element_rect(fill = "white",colour = "dark gray",
                                  size = 1, linetype = "solid"),
  panel.grid.major = element_line(size = 0.2, linetype = 'solid',
                                  colour = "light gray"),
  panel.grid.minor = element_line(size = 0.1, linetype = 'solid',
                                  colour = "light gray"))+
  scale_fill_manual(name = "Gender", values = c("orange", "blue"))+
  labs(title="Figure 7: Boxplots of Age segragated by Gender & Class") +
  theme(plot.title = element_text(size = 13,colour = "black"))
```




It is clearly evident that the age distribution of diabetes positive male and female populations are higher compared to diabetes negative populations. However, the mean of diabetes negative females shows a fairly higher value, possibly due to the smaller sample size of diabetes negative female patients (count = 19).

Polyuria symptom that is believed to have high correlation to diabetes have been explored against respective ‘Gender’ and ‘Class’ as below,

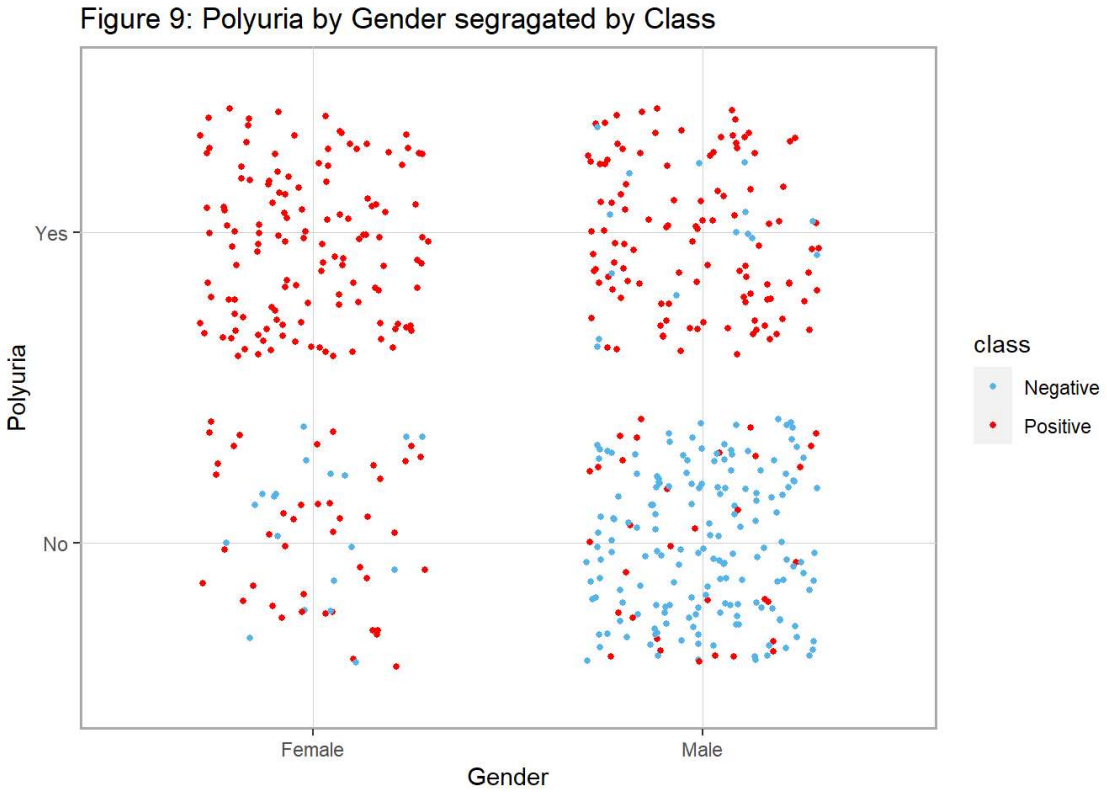
```
bp <- ggplot(data=diabetes_new, aes(x=Age, y=Polyuria, group=Polyuria)) +
  geom_boxplot(aes(fill=Polyuria), alpha=0.7,outlier.shape=NA,lwd=0.2)
bp + facet_grid(diabetes_new$class ~.)+ stat_boxplot(geom = 'errorbar', width = 0.2,coef = 3)+
  theme(
  panel.background = element_rect(fill = "white",colour = "dark gray",
                                  size = 1, linetype = "solid"),
  panel.grid.major = element_line(size = 0.2, linetype = 'solid',
                                  colour = "light gray"),
  panel.grid.minor = element_line(size = 0.1, linetype = 'solid',
                                  colour = "light gray"))+
  scale_fill_manual(name = "Polyuria", values = c("orange", "blue"))+
  labs(title="Figure 8: Boxplots of Age segragated by Polyuria & Class") +
  theme(plot.title = element_text(size = 13,colour = "black"))
```




The age distribution of Polyuria symptom segregated by 'Class' (i.e. diabetes positive or negative) is shown in above box plots. It is obvious from the diabetes negative plot (top) that Polyuria symptoms are present in older population; the age distributions of Polyuria "yes' and "no" show a clear separation of age (mean age of Polyuria 'no' is 45 years, while mean age of Polyuria "yes" is about 78 years). This suggests that Polyuria is an age-related sign in general community. However, this age separation between Polyuria "yes' and "no" populations are not prominent within diabetes positive population as shown in second plot. This supports someone to believe Polyuria is a diabetes related symptom at the first sight.

Finally, the colored scatter plots are used to visually show the grouping of 'Class' (i.e. diabetes positive in red and diabetes negative with blue) with respect to two other features. In the first plot below, "Gender' and 'Polyuria' have been used as features. It can be noticed that almost all the females with Polyuria symptom are diabetes positive, while majority (but let proportion compared to female) of males shows the similar pattern. On the other hand, the majority of the males without Polyuria symptoms are diabetes negative, and females are showing the similar pattern with less prominence.

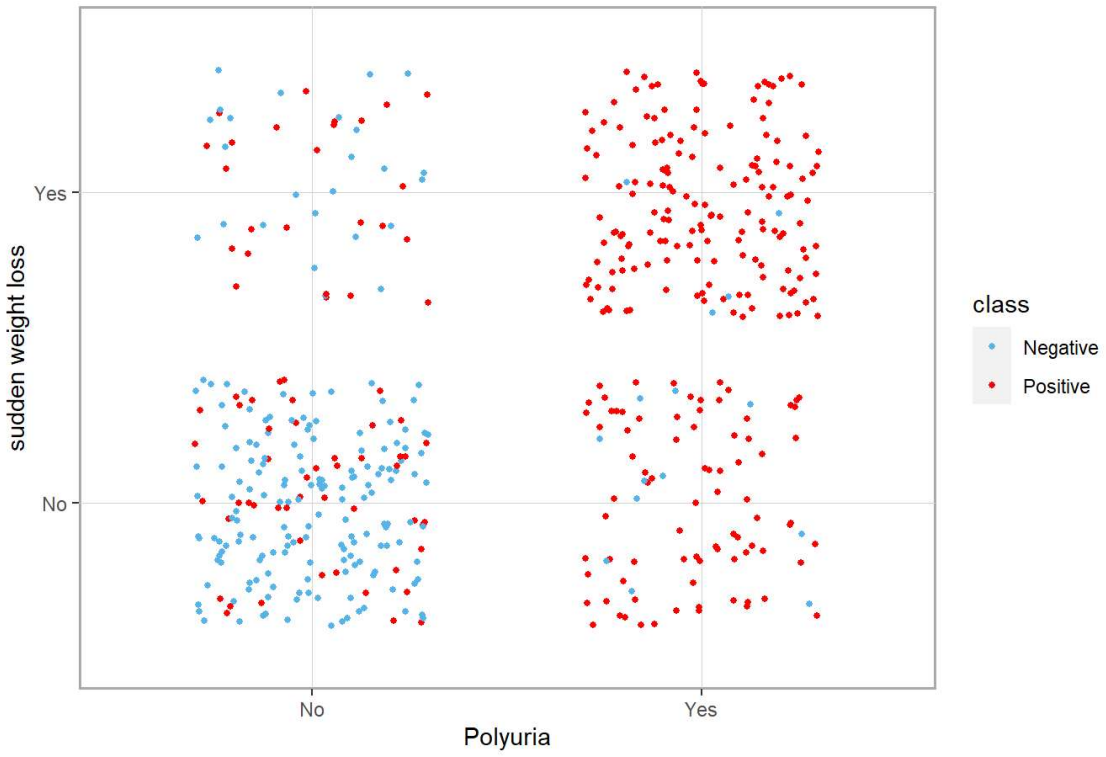
```
ggplot(diabetes_new, aes(Gender, Polyuria)) +
  geom_jitter(aes(color = class), size = 1,position=position_jitter(0.3))+
  theme(
  panel.background = element_rect(fill = "white",colour = "dark gray",
                                  size = 1, linetype = "solid"),
  panel.grid.major = element_line(size = 0.2, linetype = 'solid',
                                  colour = "light gray"))+
  scale_color_manual(values=c("#56B4E9", "red"))+
  labs(title="Figure 9: Polyuria by Gender segragated by Class") +
  theme(plot.title = element_text(size = 13,colour = "black"))
```




In the second plot, the grouping of 'Class' (i.e. diabetes positive in red and diabetes negative with blue) is shown again 'Sudden weight loss' and 'Polyuria' features. It is worth noting that very high proportion of the population that shows both of these symptoms are diabetes positive. In contrast, majority of the population that do not show either of these symptoms are diabetes negative.

```
ggplot(diabetes_new, aes(Polyuria, `sudden weight loss`)) +
  geom_jitter(aes(color = class), size = 1,position=position_jitter(0.3))+
  theme(
  panel.background = element_rect(fill = "white",colour = "dark gray",
                                  size = 1, linetype = "solid"),
  panel.grid.major = element_line(size = 0.2, linetype = 'solid',
                                  colour = "light gray"))+
  scale_color_manual(values=c("#56B4E9", "red"))+
  labs(title="Figure 10: Polyuria by `sudden weight loss` segragated by Class") +
  theme(plot.title = element_text(size = 13,colour = "black"))
```

Figure 10: Polyuria by `sudden weight loss` segragated by Class




# References


Aksakalli, V., Yenice, Z., Wong, Y. K., Ture, I., & Malekipirbazari, M. (2020). *Www.featureranking.com*. https://www.featureranking.com/ (https://www.featureranking.com/)

ASPE. (2017). *The importance of early diabetes detection*. https://aspe.hhs.gov/report/diabetes-national-plan-action/importance-early-diabetes-detection#:%7E:text=Early%20detection%20and%20treatment%20of,limb%20amputations%2C%20and%20kidney%20failure (https://aspe.hhs.gov/report/diabetes-national-plan-action/importance-early-diabetes-detection#:%7E:text=Early%20detection%20and%20treatment%20of,limb%20amputations%2C%20and%20kidney%20failure)

Dua, D., & Graff, C. (2017). *UCI machine learning repository*. University of California, Irvine, School of Information; Computer Sciences. http://archive.ics.uci.edu/ml (http://archive.ics.uci.edu/ml)

Faniqul, M. M. I., Ferdousi, R., Rahman, S., & Bushra, H. Y. (2019). Likelihood prediction of diabetes at early stage using data mining techniques. *Computer Vision and Machine Intelligence in Medical Image Analysis*, 113–125. https://doi.org/10.1007/978-981-13-8798-2_12 (https://doi.org/10.1007/978-981-13-8798-2_12)

kassambara. (2017). *Plot one variable: Frequency graph, density distribution and more*. http://www.sthda.com/english/articles/32-r-graphics-essentials/133-plot-one-variable-frequency-graph-density-distribution-and-more/#one-categorical-variable (http://www.sthda.com/english/articles/32-r-graphics-essentials/133-plot-one-variable-frequency-graph-density-distribution-and-more/#one-categorical-variable)

Sauer, S. (2016). *How to plot a 'percentage plot' with ggplot2*. https://sebastiansauer.github.io/percentage_plot_ggplot2_V2/ (https://sebastiansauer.github.io/percentage_plot_ggplot2_V2/)

Schneider, W. J. (2017). *Introduction to rmarkdown*. https://my.ilstu.edu/~wjschne/442/IntroductiontoRMarkdown.html (https://my.ilstu.edu/~wjschne/442/IntroductiontoRMarkdown.html)

Xie, Y. (2020). *R markdown cookbook*. https://bookdown.org/yihui/rmarkdown-cookbook/bibliography.html (https://bookdown.org/yihui/rmarkdown-cookbook/bibliography.html)

Zhu, H. (2020). *Create awesome html table with knitr::Kable and kableExtra*. https://cran.r-project.org/web/packages/kableExtra/vignettes/awesome_table_in_html.html (https://cran.r-project.org/web/packages/kableExtra/vignettes/awesome_table_in_html.html)