

## EXPLORATORY STATISTICS USING R PROGRAMMING FOR DIABETES MELLITUS DATASET

Cheong Yih Jeng<sup>a</sup>, Khairani Abd. Majida<sup>a</sup>, Mohd Syazwan Mohamad Anuar<sup>b\*</sup>

<sup>a</sup> Department of Digital Health and Informatics, International Medical University Malaysia (IMU), Bukit Jalil, 57000 Kuala Lumpur, Malaysia

<sup>b</sup> Department of Mathematics, Center for Defence Foundation Studies, National Defence University of Malaysia, Sg. Besi Camp, 57000 Kuala Lumpur, Malaysia

### ARTICLE INFO

#### ARTICLE HISTORY

Received : 27-09-2024  
Revised : 01-11-2024  
Accepted : 02-01-2025  
Published : 31-05-2025

#### KEYWORDS

Exploratory Data Analysis  
Diabetes Mellitus  
R Programming  
Boxplot  
Correlation

### ABSTRACT

Exploratory data analysis (EDA) is a must in conducting every research. By going through the process we could organize the dataset, understand the variables, identify relationships between variables, choose the right model, and help find patterns in a dataset. In this research we choose a healthcare dataset concerning diabetes mellitus. We explored the dataset and came up with few conclusions. We identified the relationship between variables in the dataset with diabetes mellitus. The correlation between variables and the disease were presented in visual graphs. We developed R codes to assist analysis as well as graphical representations. While conducting the exploratory statistics we discovered the flaws of the data collection and proposed few steps to advance the research in a developing a predictive model.

## 1.0 INTRODUCTION

Diabetes mellitus is a chronic metabolic disease of raised blood glucose levels which may lead to many macrovascular and microvascular complications such as coronary artery diseases, diabetic nephropathy, chronic kidney disease, diabetic retinopathy, and stroke if the disease is not well controlled over time. According to the World Health Organization (WHO), around 422 million people worldwide are diagnosed with diabetes and the majority are living in low-income and middle-income nations [1]. Annually, about 1.5 million deaths are directly attributed to diabetes mellitus.

Malaysia is among the highest rate of diabetes in Western Pacific region [2]. The prevalence of diabetes in Malaysia has increased 68.3% from around 11.2% in 2011 to 18.3% in 2019. According to National Health and Morbidity Survey (NHMS) 2019, one (1) out of five (5) adults in Malaysia have diabetes mellitus and about 3.9 million adults aged 18 years and above had diabetes [3]. Diabetes mellitus is estimated to affect about 7 million Malaysian adults aged 18 and older by 2025. The increasing prevalence of diabetes in Malaysia is due to its multifactorial causality such as physical inactivity, increasing rates of obesity, expansion of population, population ageing and urbanization.

There is a globally agreed target to halt the rise in diabetes mellitus and obesity by 2025 according to WHO [1]. We should approach this global health threat as one where prevention is better than cure. As such, it is important for us to identify the risk factors (modifiable and non-modifiable) for diabetes mellitus and subsequently develop predictive models to identify individuals at risk of diabetes, given the alarming prevalence of diabetes mellitus in Malaysia and its potential serious complications which will increase Malaysia's national public healthcare expenditure tremendously in the future. Most datasets are not organized and therefore it is difficult to identify the patterns and relationship of variables with the predicted variables. Organized dataset also will help us understand the variables and choose a suitable model for it.

In this paper we explored the diabetes mellitus dataset and discussed the relationship between variables as well as predicted the variables which cause diabetes mellitus. We then analysed the characteristics of the dataset using statistical analysis and coded in R. With organized dataset we later can identify a proper model and relationship between predicted variable and explanatory variables. We recommended a few steps to be considered if further analysis and model development should take place. We discovered that the dataset can be improved by considering a proper design of data collections. In this paper we do not consider developing a model to predict diabetes mellitus but rather exploring the dataset.

## 2.0 RELATIONSHIP BETWEEN DIABETES MELLITUS AND VARIABLES

In this section, the review on pregnancies and diabetes mellitus, glucose and diabetes mellitus, body mass index (BMI) and diabetes mellitus, diabetes pedigree function and diabetes mellitus, and age, blood pressure, skin thickness, insulin level and diabetes mellitus.

### 2.1 Pregnancies and Diabetes Mellitus

It is important to identify women of reproductive age who are at risk of developing gestational diabetes mellitus. Liu et al. suggested that higher numbers of pregnancies are an independent risk factor of gestational diabetes mellitus [4]. The association between number of pregnancies and GDM was more prominent among women who were  $\geq 30$  years old or with a pre-pregnancy BMI  $< 24$  kg/m<sup>2</sup>. Besides that, Chengji et al. found that at least 4 pregnancies through childbearing age may be a potential risk factor for diabetes in postmenopausal women without a history of gestational diabetes [5].

### 2.2 Glucose and Diabetes Mellitus

According to the Malaysia's 6<sup>th</sup> edition of management of type 2 diabetes mellitus clinical practice guidelines (CPG) Kementerian Kesihatan Malaysia, risk-based screening for pre-diabetes and type 2 diabetes mellitus should be carried out in individuals aged more than 30 years ago and this screening should be done yearly [6]. Data from the National Health and Morbidity Survey (NHMS), around 49% of patients with diabetes were undiagnosed at time of screening [3]. It is due to this fact whereby up to 50% of patients with diabetes are not having any symptoms and thus it should be the standard of practice where screening should be done when specific risk factors are present.

Examples of risk factors are women with history of gestational diabetes mellitus, adults who are overweight or obese and all individuals with prediabetes. In Malaysia, an individual is diagnosed to be prediabetes if the fasting venous plasma glucose level is between 6.1 – 6.9 mmol/L and diabetes if the fasting venous plasma glucose level is more than or equals to 7.0 mmol/L Kementerian Kesihatan Malaysia [6]. Oral glucose tolerance test (OGTT) is also used to diagnose impaired fasting glucose, impaired glucose tolerance and type 2 diabetes mellitus (refers Table 1). HbA1c is used to diagnose diabetes mellitus as well (refers Table 2).

Table 1. Diagnostics value for glucose tolerance and T2DM based on OGTT [6]

OGTT Plasma Glucose values (mmol/L)		
Category	0-hour	2-hour
Normal	$< 6.1$	$< 7.8$
IFG	6.1-6.9	-
IGT	-	7.8-11.0
T2DM	$\geq 7.0$	$\geq 11.1$

Table 2. Diagnostic value for prediabetes and T2DM based on HbA1c [6]

	Normal	Prediabetes	T2DM
HbA1c	$< 5.7\%$	5.7% - $< 6.3\%$	$\geq 6.3\%$
	( $< 39$ mmol/mol)	(39-44 mmol/mol)	( $\geq 45$ mmol/mol)

### 2.3 BMI and Diabetes Mellitus

BMI is a person's weight in kilograms divided by the square of height in meters (kg/m<sup>2</sup>). Obesity is a predominant factor that contributes to the development of type 2 diabetes mellitus. According to the WHO, overweight and obesity are defined as abnormal or excessive fat accumulation that presents a risk to

health. Overweight is defined as a BMI over 25kg/m<sup>2</sup> while obesity is defined as BMI over 30kg/m<sup>2</sup>. Gupta & Bansal demonstrated that the likelihood of being both prediabetic and diabetic is higher among the overweight and obese individuals as compared to the non-overweight individuals [7].

## 2.4 Diabetes Pedigree Function and Diabetes Mellitus

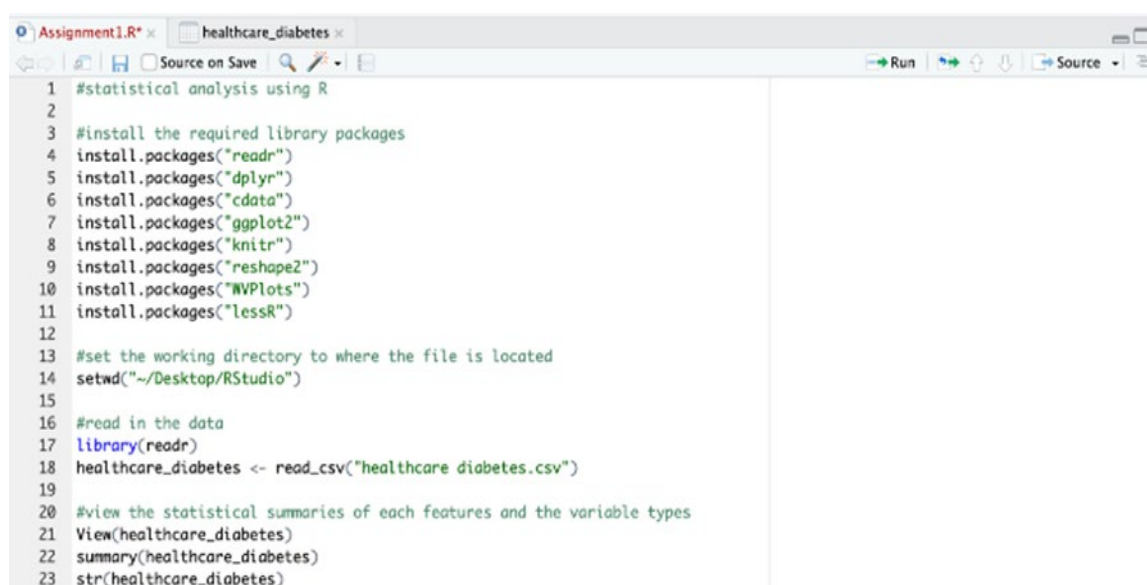
Diabetes pedigree function estimates diabetes probability depending on the individual's age and his family history of diabetes mellitus. Ghosh et al. showed a positive correlation between diabetes pedigree function, body mass index and glucose within a dataset collected to detect diabetes mellitus [8]. Besides that, Qayyum et al. estimates diabetes mellitus by using the diabetes pedigree function variable via a set of machine learning algorithms [9].

## 2.5 Age, Blood Pressure, Skin Thickness, Insulin Level and Diabetes Mellitus

There is a positive relationship between increasing life expectancy or aging and diabetes mellitus according to Chentli et al. [10]. On the other hand, Ruiz-Alejos et al. found a strong association between subscapular skinfold thickness and developing T2DM [11]. A study by Saxena et al. shows that "2-hour post-glucose insulin levels" appears to be a good indicator of insulin resistance [12].

## 3.0 METHODOLOGY

In our analysis we have taken the dataset in an open-data portal [13]. Before the analysis was done, we conducted the data preparation. We used R programming to perform descriptive statistics for the dataset. R is a programming language used for statistical computing and graphics. It is also open source and is platform independent. There are 10 columns in the dataset and each column is describe as follows. Figure 1 shows the steps involved for data preparation coded in R:



```

1 #statistical analysis using R
2
3 #install the required library packages
4 install.packages("readr")
5 install.packages("dplyr")
6 install.packages("cdata")
7 install.packages("ggplot2")
8 install.packages("knitr")
9 install.packages("reshape2")
10 install.packages("WVPlots")
11 install.packages("lessR")
12
13 #set the working directory to where the file is located
14 setwd("~/Desktop/RStudio")
15
16 #read in the data
17 library(readr)
18 healthcare_diabetes <- read_csv("healthcare diabetes.csv")
19
20 #view the statistical summaries of each features and the variable types
21 View(healthcare_diabetes)
22 summary(healthcare_diabetes)
23 str(healthcare_diabetes)

```

Figure 1. Data preparation

1. Id: Unique Identifier for Each Data Entry.
2. Pregnancies: Number Of Times Pregnant.
3. Glucose: Plasma Glucose Concentration Over 2 Hours in An Oral Glucose Tolerance Test.
4. Bloodpressure: Diastolic Blood Pressure (Mm Hg).
5. Skintickness: Triceps Skinfold Thickness (Mm).
6. Insulin: 2-Hour Serum Insulin (Mu U/MI).
7. BMI: Body Mass Index (Weight in Kg/Height in M<sup>2</sup>).
8. Diabetespedigreefunction: Diabetes Pedigree Function, A Genetic Score of Diabetes.
9. Age: Age In Years.
10. Outcome: Binary Classification Indicating the Presence (1) Or Absence (0) Of Diabetes Mellitus (Categorical Variable).

The data preparation steps involved are:

- Install The Required Library to Read the Csv File (Install.Packages("Readr")).
- Set The Working Directory to Where the File Is Located (Setwd("~/Desktop/Rstudio")).
- Use The Library (Readr) To Read in the Data (Healthcare\_Diabetes <- Read\_Csv("Healthcare\_Diabetes.Csv")).

As shown in Figure 2, the codes used in R to display the statistical summaries of the dataset and the variable types. it is noted that the number of variables is 10 and the number of rows (number of participants) is 2768.

```

20 #view the statistical summaries of each features and the variable types
21 View(healthcare_diabetes)
22 summary(healthcare_diabetes)
23 str(healthcare_diabetes)

> summary(healthcare_diabetes)
  Id      Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin    BMI  DiabetesPedigreeFunction  Age  Outcome
Min.   : 1.0   Min.   : 0.000   Min.   : 0.0   Min.   : 0.00   Min.   : 0.00   Min.   : 0.00   Min.   : 0.00   Min.   :0.0780   Min.   :21.00   0:1816
1st Qu.:692.8   1st Qu.: 1.000   1st Qu.:99.0   1st Qu.: 62.00   1st Qu.: 0.00   1st Qu.: 0.00   1st Qu.:27.30   1st Qu.:0.2440   1st Qu.:24.00   1: 952
Median :1384.5   Median : 3.000   Median :117.0   Median : 72.00   Median : 23.00   Median : 37.00   Median :32.20   Median :0.3750   Median :29.00
Mean   :1384.5   Mean   : 3.743   Mean   :121.1   Mean   : 69.13   Mean   : 20.82   Mean   : 80.13   Mean   :32.14   Mean   :0.4712   Mean   :33.13
3rd Qu.:2076.2   3rd Qu.: 6.000   3rd Qu.:141.0   3rd Qu.: 80.00   3rd Qu.: 32.00   3rd Qu.:1130.00   3rd Qu.:36.62   3rd Qu.:0.6240   3rd Qu.:40.00
Max.   :2768.0   Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :110.00   Max.   :1846.00   Max.   :80.60   Max.   :1.24200   Max.   :81.00

> str(healthcare_diabetes)
'spc.tbl.' [2,768 x 10] (S3: spec.tbl.df/tbl.df/tbl/data.frame)
 $ Id      : num [1:2768] 1 2 3 4 5 6 7 8 9 10 ...
 $ Pregnancies : num [1:2768] 6 1 8 1 0 5 3 10 2 8 ...
 $ Glucose    : num [1:2768] 148 85 183 89 137 116 78 115 197 125 ...
 $ BloodPressure : num [1:2768] 72 66 64 66 40 74 50 0 70 96 ...
 $ SkinThickness : num [1:2768] 35 29 0 23 35 0 32 0 45 0 ...
 $ Insulin     : num [1:2768] 0 0 0 94 168 0 88 0 543 0 ...
 $ BMI         : num [1:2768] 33.6 26.6 23.3 28.1 43.1 25.6 31 35.3 30.5 0 ...
 $ DiabetesPedigreeFunction: num [1:2768] 0.627 0.351 0.672 0.167 2.288 ...
 $ Age        : num [1:2768] 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 ...
 - attr(*, "spec")=
   .. cols(
   ..   Id = col_double(),
   ..   Pregnancies = col_double(),
   ..   Glucose = col_double(),
   ..   BloodPressure = col_double(),
   ..   SkinThickness = col_double(),
   ..   Insulin = col_double(),
   ..   BMI = col_double(),
   ..   DiabetesPedigreeFunction = col_double(),
   ..   Age = col_double(),
   ..   Outcome = col_double()
   .. )
 - attr(*, "problems")=externalptr=

```

Figure 2. Statistical summaries of the dataset and the variable types

## 4.0 RESULTS AND DISCUSSION

Figure 3 displays codes and results in R for the descriptive statistical analysis along with the measures of central tendency and measures of distribution (e.g. standard deviation). Pregnancies appear in an acceptable range from 0 to 17. However, it is important to take note that some variables in the data (Glucose, BloodPressure, SkinThickness, Insulin, BMI) include the value 0 and this is not possible in clinical practice.

```

> #the basic descriptive statistics
> library(readr)
> library(psych)
> healthcare_diabetes <- read_csv("healthcare diabetes.csv")
Rows: 2768 Columns: 10
— Column specification —
Delimiter: ","
dbl (10): Id, Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age, Outcome

i Use 'spec()' to retrieve the full column specification for this data.
i Specify the column types or set 'show_col_types = FALSE' to quiet this message.
> describe(healthcare_diabetes)

      vars  n  mean  sd median trimmed  mad  min  max  range  skew  kurtosis  se
1 2768 1384.50 799.20 1384.50 1384.50 1025.96 1.00 2768.00 2767.00 0.00 -1.20 15.19
2 2768   3.74   3.32   3.00   3.34   2.97 0.00 17.00 17.00 0.96  0.33  0.06
3 2768 121.10 32.04 117.00 119.62 29.65 0.00 199.00 199.00 0.16  0.57  0.61
4 2768  69.13 19.23  72.00  71.34 11.86 0.00 122.00 122.00 -1.85  5.26  0.37
5 2768  20.82 16.06  23.00  20.23 17.79 0.00 110.00 110.00 0.18 -0.03  0.31
6 2768  80.13 112.30  37.00  57.96 54.86 0.00 846.00 846.00 2.08  5.74  2.13
7 2768  32.14   8.08  32.20  32.04   6.97 0.00  80.60  80.60 -0.18  3.91  0.15
8 2768   0.47   0.33   0.38   0.42   0.25 0.08   2.42   2.34  1.84  5.16  0.01
9 2768  33.13 11.78  29.00  31.38 10.38 21.00  81.00  60.00  1.17  0.77  0.22
10 2768   0.34   0.48   0.00   0.31   0.00 0.00   1.00   1.00  0.66 -1.57  0.01

```

Figure 3. Basic statistical summary

As such, the 0 values need to be corrected. The impossible 0 values can be corrected by substituting them with the mean values at the data cleaning stage. The 'DiabetesPedigreeFunction' is a function that scores the probability of diabetes based on family history, with a realistic range of 0.08 to 2.42. Age has a realistic range from 21 to 81. The Outcome, "0" represents absence of diabetes mellitus, and "1" represents presence of diabetes mellitus.

#### 4.1 The Distribution of Data

From Figure 4, we can view the histograms of numerical values and its data distribution. Analysing the data distribution is extremely important to choose the right statistical test, identify outliers, check for normality, and visualize the data. In this case, it is noted that histograms of Glucose, BloodPressure, BMI, SkinThickness and Insulin show normal data distributions with outliers. The histograms of Pregnancies, DiabetesPedigreeFunction and Age display right-skewed distribution (positive-skew distributions) of data. In a skewed distribution, the mean and median become different (refers Figure 4).

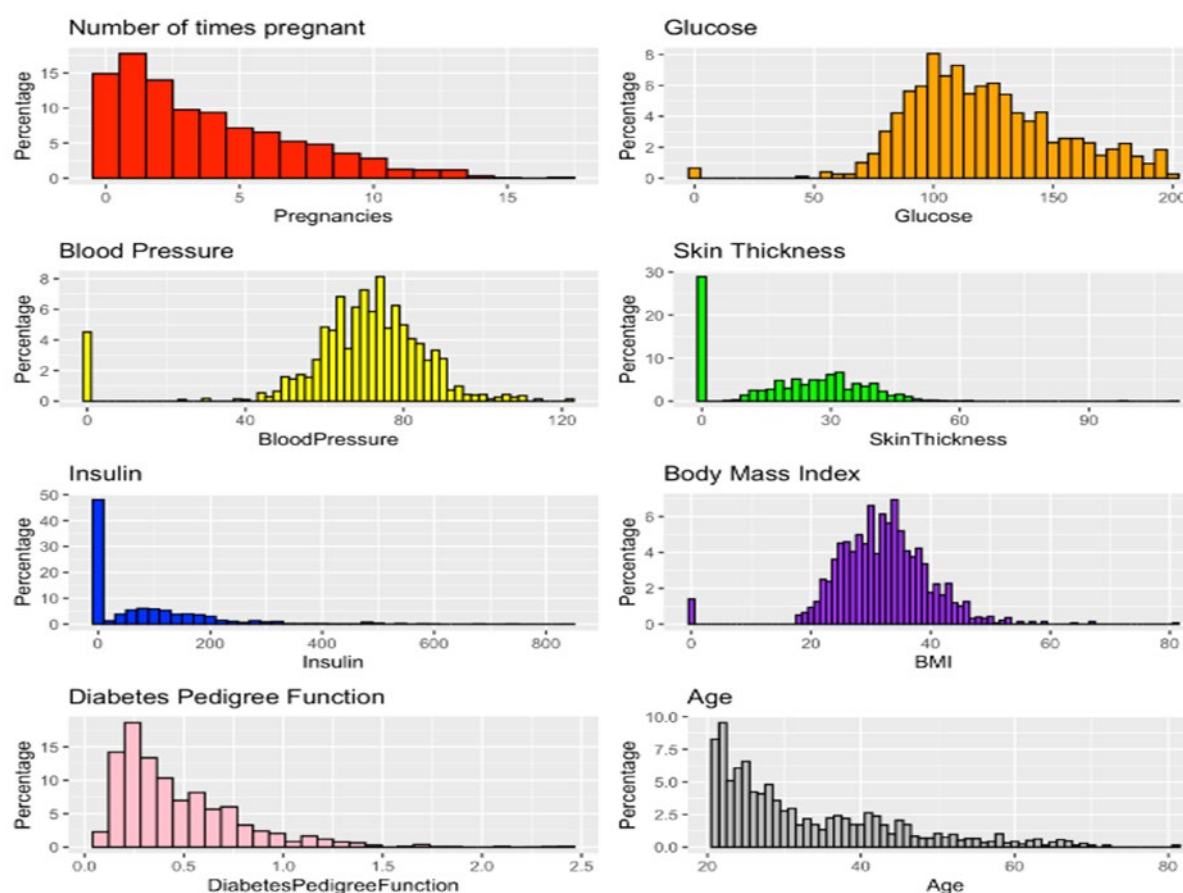


Figure 4. Histograms of different variables and their distribution

#### 4.2 Age and BMI Versus the Presence (1) or the Absence (0) of Diabetes Mellitus

As shown in Figure 5, the histogram on the left displays that at a younger age the absence of diabetes mellitus was very high. However, the histogram in the right-hand side showed that the prevalence of diabetes mellitus was more often for the group above 40 years old. The distributions for both histograms are skewed to the right because the ages of most of the patients are between 20 to 40. For Figure 6, the BMI attribute appears to be symmetrical in distribution but there are outliers in the dataset provided. There are 0 values in the dataset which is not possible at all. This suggests that there might be an error in the dataset.



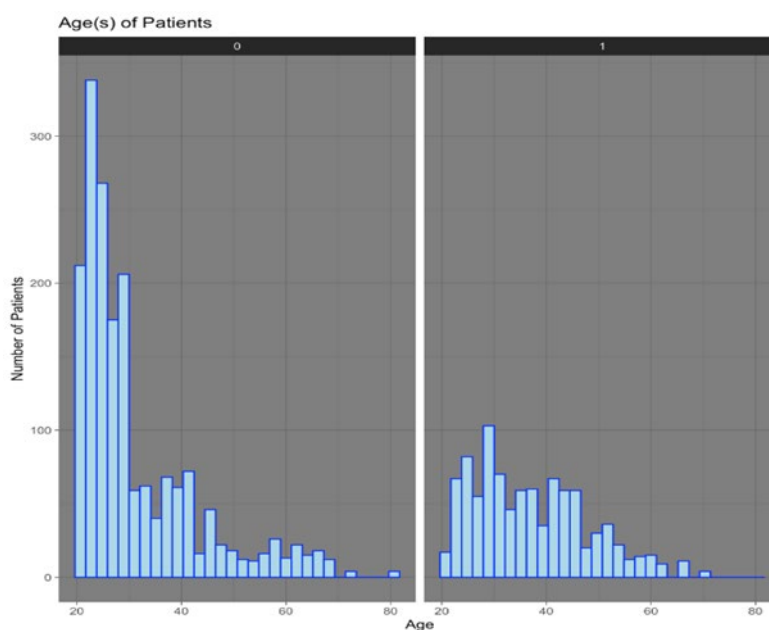


Figure 5. Histogram of age of patient without (left) and with (right) Diabetes Mellitus

### 4.3 Boxplot for the Association Between the Numerical Variables and Outcomes

From Figure 6, we can observe that in each variable there are outliers. The distribution of variables such as glucose, blood pressure, and BMI seems to be almost symmetrical, and these outcomes agreed with the outcomes in Figure 6. However, variable glucose shows that the outcome for the prevalence of diabetes mellitus is skewed to the left. Logically, we can agree that the higher the glucose level in the blood, the higher the chances that you will get diabetes. Variables such as number of pregnancies, skin thickness, insulin, age, and diabetes pedigree function are skewed to the right. This indicates that the dataset mostly consists of patients aged between 20 and 40 years old.

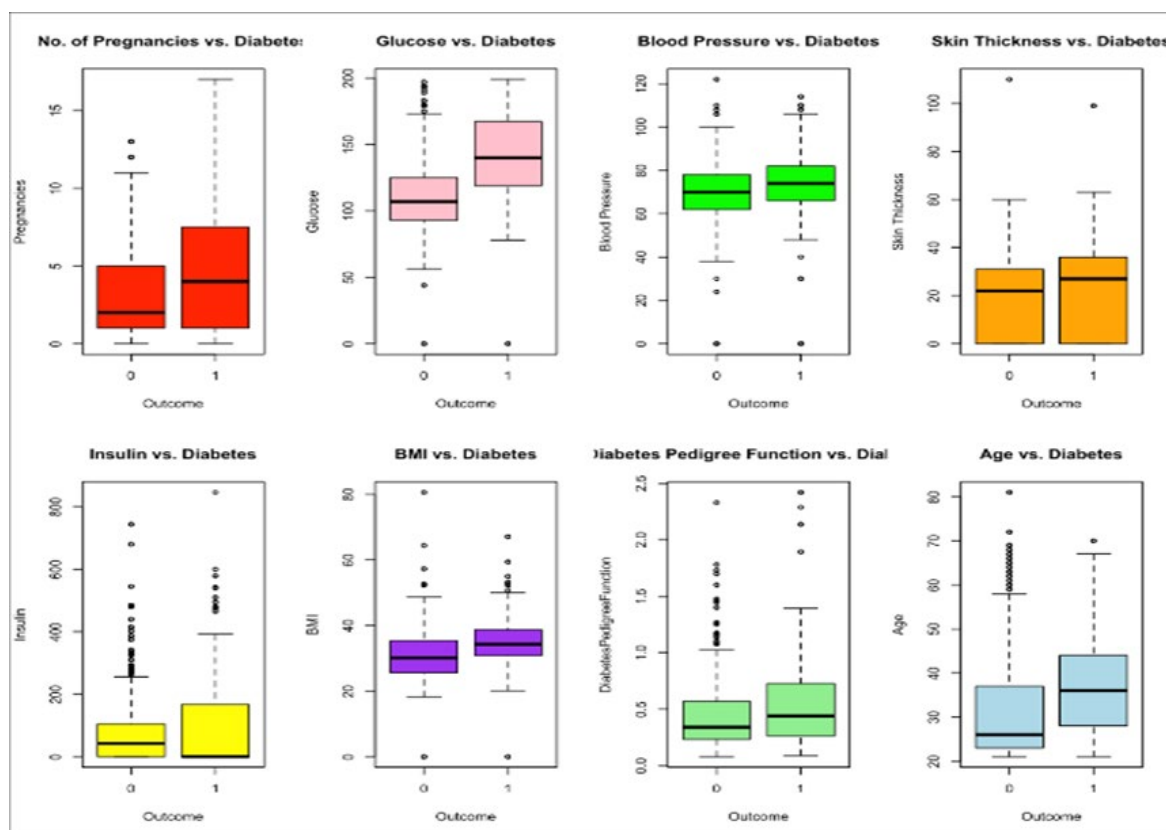


Figure 6. Boxplot for different variables with presence (1) and absence (0) Diabetes Mellitus

#### 4.4 Correlation for Numerical Variables

A correlation plot can be used to cross-examine multivariate data, show variance between variables, show whether any variables are like each other and detect whether there is a correlation between variables. In this correlation plot (Figure 7), almost all variables have weak linear correlations. Thus, this indicates that most of the attributes are more likely to have non-linear relationships instead. The correlation coefficients between all variables were less than 0.7, showing that multicollinearity was absent in the dataset.

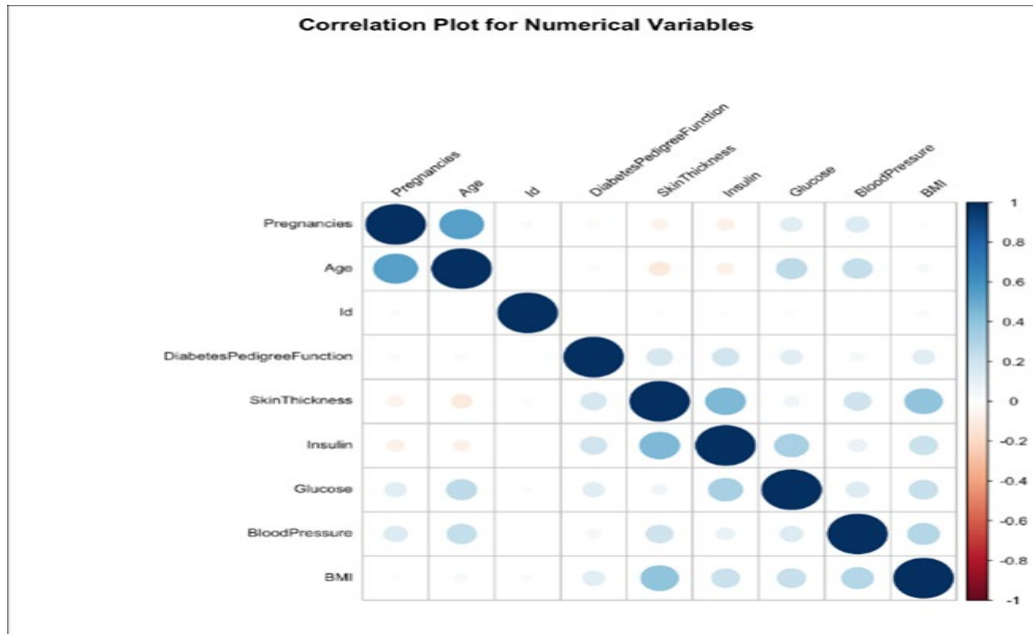


Figure 7. Correlation plot for numerical attributes in the dataset

#### 5.0 CONCLUSIONS

A total of 2768 number of participants participated in this study research. This dataset contains a diverse range of health-related attributes. The dataset was extracted from the Kaggle Dataset, Pore N. [13] website. In this broad analysis of the diabetes dataset, we aimed to explore the relationship between clinical variables and the presence of diabetes mellitus. The analysis resulted in several valuable findings which will help in understanding and addressing the diabetes risk factors. The data recorded in the database includes some limitations. Data cleaning can be performed to filter duplicate, incomplete and incorrect data to reduce error rates and enhance productivity and decision making. Attributes that are positively or negatively skewed can be improved to normal distribution through data cleaning methods. Central limit theorem is important in machine learning as it can be used as an assumption as the mean value of the sample is approximately equal to the mean of population regardless of the distribution.

#### 6.0 CONFLICT OF INTEREST

The authors declare no conflicts of interest.

#### 7.0 AUTHORS CONTRIBUTION

Cheong, Y. J. (Conceptualization; Writing - original draft; Data curation; Formal analysis)

Abd Majid, K. (Resources; Software; Writing - original draft; Project administration; Supervision)

Mohamad Anuar, M. S. (Validation; Writing - review & editing)

#### 8.0 ACKNOWLEDGEMENTS

The authors fully acknowledged Ministry of Higher Education (MOHE), International Medical University Malaysia (IMU) and National Defence University of Malaysia (NDUM) for their support in making this research feasible.

## List of Reference

- [1] World Health Organization, *Global Report on Diabetes*, WHO Press, ISBN 978 92 4 156525 7 (NLM classification: WK 810), 2016.
- [2] Chandran A., Zakaria N., *National Diabetes Registry Report*, Non-Communicable Disease Section, Disease Control Division, Ministry of Health, Malaysia, 2021.
- [3] National Institute of Health (NIH), *National Health and Morbidity Survey 2019, Non-Communicable Diseases, Healthcare Demand and Healthcare Literacy, Key Findings*, Ministry of Health Malaysia, 2020, ISBN 978-983-99320-6-5, 2020.
- [4] Liu, B., Song L., Zhang L., Wang L., Wu M., Xu S., Cao Z., Wang Y. (2020). Higher numbers of pregnancies associated with an increased prevalence of gestational diabetes mellitus: results from the healthy baby cohort study, *J. Epidemiol* 30(5), 208-212.
- [5] Chengjie, Lv., Chen C., Chen Q., Zhai H., Zhao L., Guo Y., Wang N. (2019). Multiple pregnancies and the risk of diabetes mellitus in postmenopausal women. *Menopause* 26(9), 1010-1015.
- [6] Kementerian Kesihatan Malaysia, *Management of Type 2 Diabetes Mellitus Clinical Practice Guidelines (CPG) (6<sup>th</sup> Edition)*, Ministry of Health Malaysia, 2021.
- [7] Gupta, S. & Bansal, S. (2020). Does a rise in BMI cause an increased risk of diabetes? Evidence from India, *PLoS ONE* 15(4), e0229716.
- [8] Ghosh, P., Azam, S., Karim, A., Hassan, M., Roy, K., & Jonkman, M. (2021). A comparative study of different machine learning tools in detecting diabetes. *Procedia Computer Science*, 192, 467-477.
- [9] Qayyum, A., Talpur, S., & Jawaid, M. (2021). Early Detection of Type 2 Diabetes using supervised machine learning. *Engineering Science and Technological International Research Journal*, 1(5).
- [10] Chentli, F., Azzoug, S., & Mahgoun S. (2015). Diabetes mellitus in elderly. *Indian J Endocrinol Metab.* 19(6), 744-752.
- [11] Ruiz-Alejos, A., Carrillo-Larco, R. M., Miranda, J. J., Gilman, R. H., Smeeth, L., & Bernabé-Ortiz, A. (2020). Skinfold thickness and the incidence of type 2 diabetes mellitus and hypertension: an analysis of the PERU MIGRANT study. *Public Health Nutr.* 23(1), 63-71.
- [12] Saxena, P., Prakash, A., Nigam, A. (2011). Efficacy of 2-hour post glucose insulin levels in predicting insulin resistance in polycystic ovarian syndrome with infertility. *J. Hum. Reprod. Sci.* 4(1), 20-22.
- [13] Pore N., *Healthcare Diabetes Dataset: A Comprehensive Dataset for Diabetes Risk Assessment*. National Institute of Diabetes and Digestive and Kidney Diseases, 2023.