The Relationship between Medical Predictors and Diabetes Illness

Created by Group 7:
Allin Setiawan - 2602191301 - allin.setiawan@binus.ac.id- BINUS University - Data Science
Patricia Pepita - 2602174176 - patricia.pepita@binus.ac.id - BINUS University - Data Science
Putri Fatiha - 2602193042 - putri.nuzula@binus.ac.id - BINUS University - Data Science
Rachel Andrea - 2602179334 - rachel.sumaiku@binus.ac.id - BINUS University - Data Science

Problem Statement and Objectives

Based on WHO data, Diabetes illness tends to increase every year. Therefore, we chose the Diabetes data set that we found on Kaggle namely Diabetes Data Set. We want to know the relationship between health conditions and diabetes, so we can come up with the best solution.

- The number of people with diabetes rose from 108 million in 1980 to 422 million in 2014. Prevalence has been rising more rapidly in low- and middle-income countries than in high-income countries.
- Diabetes is a major cause of blindness, kidney failure, heart attacks, stroke and lower limb amputation.
- Between 2000 and 2019, there was a 3% increase in diabetes mortality rates by age.

Souce: WHO, 2023

Data Description

- Exploratory Data
- Data Understanding
- Modelling
- Create Prediction
- Data Result

Title: Diabetes Dataset — Author: Akshay Dattatray Khare

diabetes

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	627	50	1
1	85	66	29	0	26.6	351	31	0
8	183	64	0	0	23.3	672	32	1
1	89	66	23	94	28.1	167	21	0
0	137	40	35	168	43.1	2.288	33	1

There are 9 columns in the data set

- **Pregnancies:** the number of pregnancies
- **Glucose:** glucose level in blood
- **BloodPressure:** blood pressure measurement
- **SkinThickness:** the thickness of the skin
- **Insulin:** insulin level in the blood
- **BMI:** body mass index
- **DiabetesPedigreeFunction:** diabetes percentage
- **Age:** the age of the sample
- **Outcome:** the result (1 is yes, 0 is no)

Exploratory Data Analysis (EDA)

```
'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 ...
[1] 768   9
```

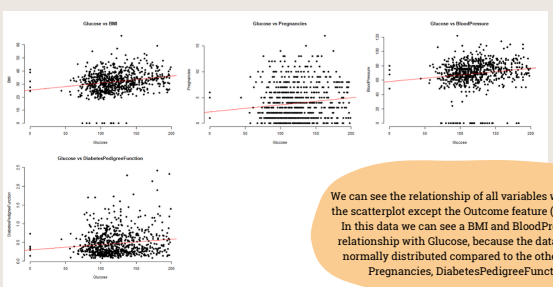
Diabetes Dataset has 768 records with 9 variables: Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age, and Outcome. All of these variables are numeric data types

When we checked the missing value we found that Insulin and SkinThickness variables have more than 20% of 0 value. Therefore, we removed them.

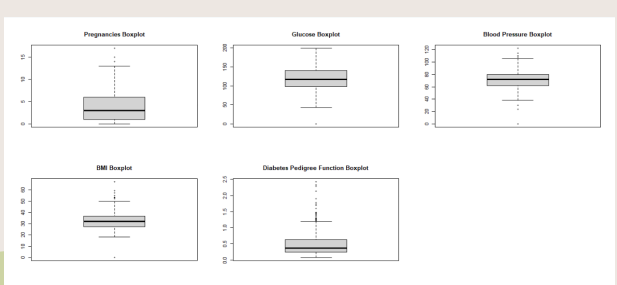
Description: df [7 × 8]

variable	type	levels	topLevel	topCount	topFrac	missFreq	missFrac
Pregnancies	integer	17	1	135	0.176	0	0
Glucose	integer	136	99	17	0.022	0	0
BloodPressure	integer	47	70	57	0.074	0	0
BMI	numeric	248	32	13	0.017	0	0
DiabetesPedigreeFunction	numeric	517	0.254	6	0.008	0	0
Age	integer	52	22	72	0.094	0	0
Outcome	integer	2	0	500	0.651	0	0

7 rows



We can see the relationship of all variables with Glucose through the scatterplot except the Outcome feature (because it is binary). In this data we can see a BMI and BloodPressure have a good relationship with Glucose, because the data that they have are normally distributed compared to the other features such as Pregnancies, DiabetesPedigreeFunction, and Age.



We use boxplot to detect outliers in each variable and we find:
- Pregnancies, Diabetes Pedigree, and Age variables have outliers that lie above max value
- Glucose variable has outliers that lie below min value
- Blood Pressure and BMI variables have outliers that lie both above max value and below min value

In addition, BMI, Diabetes Pedigree, and Age variables have many outliers.

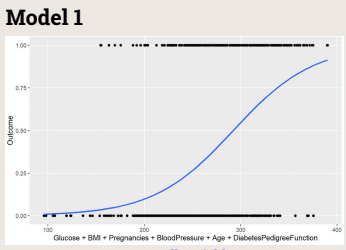
Predictive Model and Discussion

We use logistic regression to predict the outcome because logistic regression is known for its prediction accuracy which is why we think our model will defined clearly with this method to see the result of what we want to predict.

```
Call:
glm(formula = Outcome ~ ., family = "binomial", data = dataTrain)

Deviance Residuals:
    Min      1Q  Median      3Q     Max
-2.5471  -0.7341  -0.4367   0.7387   2.8788

Coefficients:
                          Estimate Std. Error z value
(Intercept)              -8.146149   0.811664 -10.036
Pregnancies               0.106812   0.037755   2.829
Glucose                   0.033063   0.004154   7.959
BloodPressure            -0.012747   0.005907  -2.158
BMI                       0.088221   0.017032   5.180
DiabetesPedigreeFunction  0.647896   0.331991   1.952
Age                       0.017916   0.010971   1.633
                          Pr(>|z|)
(Intercept)               < 2e-16 ***
Pregnancies               0.00467 **
Glucose                  1.73e-15 ***
BloodPressure             0.03093 *
BMI                      2.22e-07 ***
DiabetesPedigreeFunction  0.05099 .
Age                       0.10248
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 695.03  on 537  degrees of freedom
Residual deviance: 512.75  on 531  degrees of freedom
AIC: 526.75

Number of Fisher Scoring iterations: 5
```

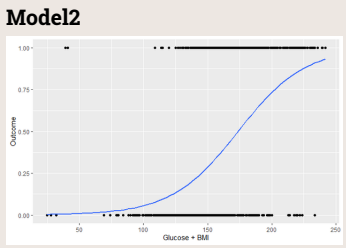
From the result of the fit model, we can see that Glucose and BMI features are statistically significant to the Outcome feature.

Model 1



outcome + all variables

Model2



outcome + glucose and BMI

From the models we created, we can conclude that the increase in medical predictors level resulting the outcome to 1 (having diabetes).

Check the accuracy

```
[1] "accuracy model 1: 0.77"
[1] "accuracy model 2: 0.76"
```

The regression that includes all of the variables has higher accuracy compared to the one that is not although the difference is not that much (only 0.01).

Check the sensitifity and specificity

```
[1] "Sensitivity model 1 is 0.785714285714286"
[1] "Specificity model 1 is 0.725806451612903"
[1] "Sensitivity model 2 is 0.76878612716763"
[1] "Specificity model 2 is 0.719298245614035"
```

Model 1 will likely to predict positive findings for people with Diabetes compared to model 2 since the sensitivity value is higher in model 1. Furthermore, model 1 will likely to predict people without Diabetes compared to the model 2 because the specificity value is higher in model 1. Therefore, model 1 has a better identification than model 2.

Conclusion

- From the regression, we see that the glucose and BMI value influence whether a person has diabetes or not.
- Higher glucose levels and BMI values resulting a higher chance to have diabetes.
- Moreover, we see that the regression model 1 is better if we compare it to the regression model 2. It is likely because model 1 (which includes all of the independent variables) is more accurate and sensitive in predicting the outcome.