STATISTICAL REPORT

I. Introduction

1) The research problem is diabetes - the most prevalence chronic diseases in the world is diabetes. This report will investigate 100,000 survey responses, provided by the author Mohammed Mustafa, to predict diabetes status.
2) The purpose of this statistical analysis is showing how I choose a classification method and propose the best classifier for predicting diabetes status.

II. Statistical Procedures Used

1) Summarize/describe variables:
   - Input variables: there are 8 input variables.
     - Gender: is a categorical variable, which consists of 58552 females, 41430 males and 18 others
     - Age: is a numerical variable. The highest age is 80, the lowest is 0.08. The median age is 43 and the mean of age variable is 41.89.
     - Hypertension: is a categorical variable, which consists of 92515 people doesn't have hypertension and 7485 people have hypertension
     - Heart disease: is a categorical variable, which consists of 96,058 people doesn't have heart disease and 3492 people have heart disease
     - Smoking History: is a categorical variable, which consists of 9286 people currently smoking, 4004 people smoke sometimes but not often, 9352 people smoked before but completely quitted, 35095 people have never smoked before and after, 6447 people have not smoked before but not sure in the future, and 35,816 people have no information about their smoking history.
     - BMI: is a numerical variable. The highest BMI is 95.69, while the lowest is 10.01. The median and mean BMI is similar at 27.32.
     - HbA1c Level: is a numerical variable. The highest level is 9, while lowest is 3.5. The median level is5.8 and the mean level is 5.528.
     - Blood Glucose Level: is a numerical variable. The highest level is 300, while the lowest is 80. The median level is 140 and the mean is 138.1.
   - Response variables is categorical variable, which is described as 0 = No and 1 = yes
     - Diabetes: There are 91500 people in data are not diabetes, while 8500 people are

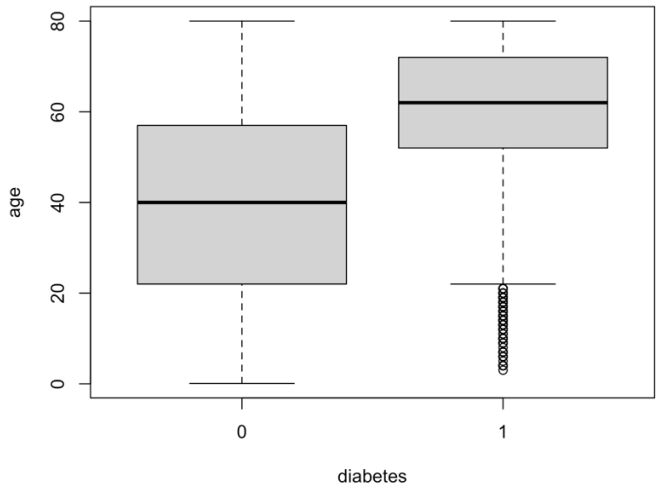
2) The association between the response and each input variable

- Diabetes and Gender:

diabetes	gender 2	Female	Male
0	0.00018	0.54091	0.37391
1	0.00000	0.04461	0.04039

  - For people are not diabetes (0): females have a higher proportion than males and others.
  - For people are diabetes (1): females have a slightly higher proportion than males.
    ⇨ Association strength is not strong.

- Diabetes and Age:



  - Older people (around 60) are more prone to diabetes.
    ⇨ Association strength is strong.

- Diabetes and Hypertension:

diabetes	hypertension 0	1
0	0.86103	0.05397
1	0.06412	0.02088

  - For people are not diabetes: people who do not have hypertension has a much higher proportion.
  - For people are diabetes: people who do not have hypertension has a slightly higher proportion.
    ⇨ Association strength is quite strong.

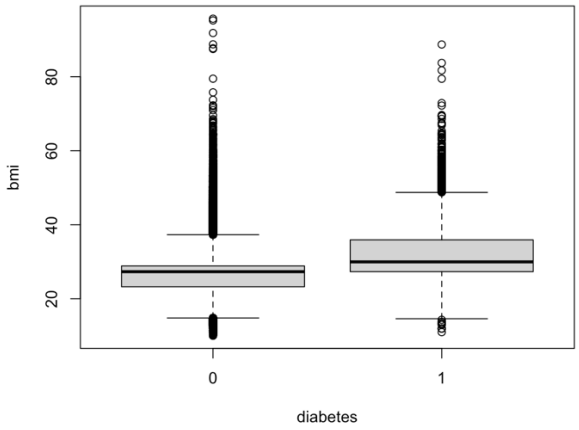
- Diabetes and Heart Disease

diabetes	heart_disease 0	1
0	0.88825	0.02675
1	0.07233	0.01267

  - For people are not diabetes: people who do not have heart disease has a significantly higher proportion.
  - For people are diabetes: people who do not have heart disease has a higher proportion.
    ⇨ Association strength is not strong
- Diabetes and Smoking History

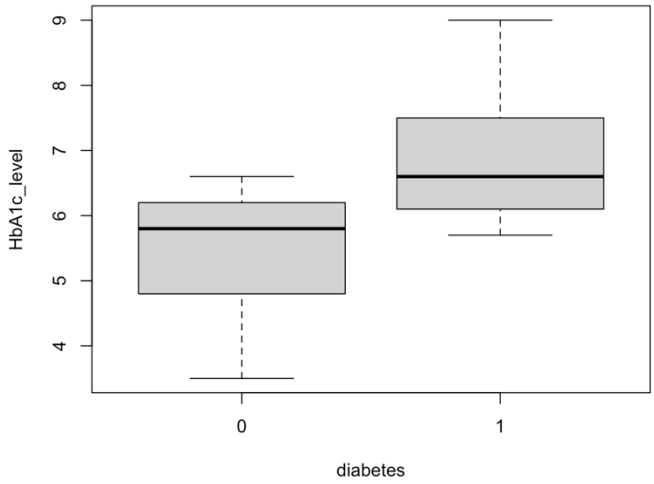
diabetes	smoking_history current	ever	former	never	No Info	not current
0	0.08338	0.03532	0.07762	0.31749	0.34362	0.05757
1	0.00948	0.00472	0.01590	0.03346	0.01454	0.00690

  - For people are not diabetes: people who never smoke before and after has the highest proportion of not being diabetes (significantly higher proportion than other people)
  - For people are diabetes: people who never smoke before and after has the highest proportion but not significantly higher proportion than others.
    ⇨ Association strength is not strong.
- Diabetes and BMI



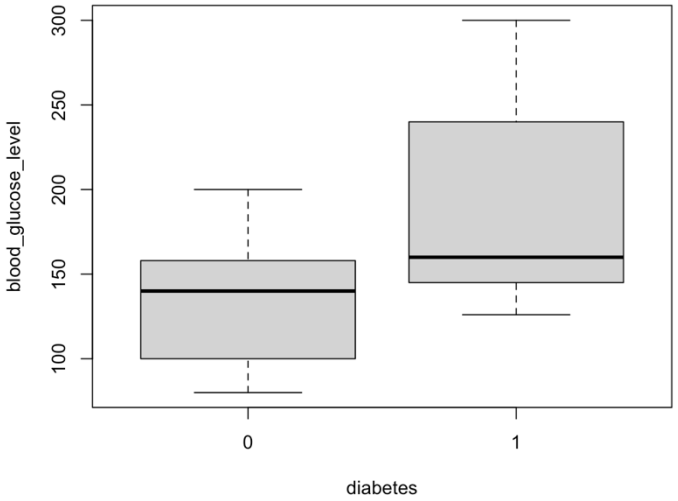
  - People who are diabetes have higher BMI than people who are not diabetes.
    ⇨ Association strength is strong.

- Diabetes and HbA1c



  - People who are diabetes have much higher HbA1c level than people who are not diabetes.
    - ⇨ Association strength is strong.

- Diabetes and Blood Glucose Level



  - People who are diabetes have higher blood glucose level than people who are not diabetes.
    - ⇨ Association strength is strong.

3) Propose some models/classifiers and examine its goodness of fit: by ROC, AUC, and accuracy.

	Average TPR	Average FPR	Average AUC
KNN	0.9739792	0.3658852	0.804047
Decision Tree	1	0.3307885	0.8346058
Naïve Bayes	0.9825582	0.3552465	0.9494347
Logistic Regression	0.9910277	0.3702855	0.9618196

III. Summary of statistical finding

	Pro	Cons
KNN	KNN has a quite high TPR	KNN is time consuming, although the data set has only 4 input variables, and it has the lowest AUC. Therefore, KNN is not a good classifier for this data
Decision Tree	DT processed the fastest, with the best result of TPR, FPR, and good AUC result. This is a good classifier for this data set	The AUC result of DT is lower than Naive Bayes and Logistic Regression
Naïve Bayes	Quite good result of TPR, FPR, and AUC, so Navies Bayes is also a good classifier for this data	Naïve Bayes is time-consuming
Logistic Regression	Logistic Regression is quite fast to process, has high TPR, highest AUC	Logistic Regression has the highest FPR with threshold = 0.5

- With the highest TPR, the lowest FPR, Decision Tree is the best classifier for this data set. Moreover, Decision Tree is also fast to run. Although Decision Tree has lower AUC result than Naïve Bayes and Logistic Regression but its AUC is quite high, which is acceptable.