# 1. Introduction

The response variable is a binary variable indicating the presence of diabetes, where 0 is no and 1 is yes.

There is an equal distribution of people with diabetes and without diabetes, both at 35346. Hence, the mean and median binary are both 0.5. The total sample size is 70692.

There are a total of 21 input variables, which consist of some ordinal categorical variables such as GenHlth, representing General Health and Age. These include binary variables such as HighBP, representing High Blood Pressure, and Stroke. There is also a continuous variable, BMI, representing Body Mass Index.

```
Call:
glm(formula = Diabetes_binary ~ ., family = binomial(link = "logit"),
    data = diab)

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)           -6.865139   0.124479 -55.151  < 2e-16 ***
HighBP                 0.735319   0.019738  37.255  < 2e-16 ***
HighChol               0.587302   0.018861  31.139  < 2e-16 ***
CholCheck              1.360832   0.081318  16.735  < 2e-16 ***
BMI                    0.075617   0.001573  48.057  < 2e-16 ***
Smoker                -0.001680   0.018875  -0.089 0.929075
Stroke                 0.162029   0.040917   3.960 7.50e-05 ***
HeartDiseaseorAttack   0.252675   0.028435   8.886  < 2e-16 ***
PhysActivity          -0.033131   0.021294  -1.556 0.119726
Fruits                -0.034507   0.019592  -1.761 0.078185 .
Veggies               -0.061051   0.023333  -2.616 0.008884 **
HvyAlcoholConsump     -0.749713   0.048762 -15.375  < 2e-16 ***
AnyHealthcare          0.060813   0.047164   1.289 0.197264
NoDocbcCost            0.018988   0.034083   0.557 0.577452
GenHlth                0.584617   0.011447  51.071  < 2e-16 ***
MentHlth              -0.004360   0.001285  -3.394 0.000688 ***
PhysHlth              -0.008321   0.001192  -6.981 2.94e-12 ***
DiffWalk               0.114941   0.025861   4.445 8.81e-06 ***
Sex                    0.267261   0.019150  13.956  < 2e-16 ***
Age                    0.152242   0.003909  38.950  < 2e-16 ***
Education             -0.036815   0.010221  -3.602 0.000316 ***
Income                -0.058807   0.005193 -11.325  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

*Fig 1. Summary of Generalised Linear Model between Diabetes_binary and the input variables*

I performed logistic regression using the Generalised Linear Model function and found that some input variables have a high p-value, which means their contribution to the model is not significant.

The significant features are those with a "***" on the right, showing they have a p-value less than 0.001. Hence, I dropped the insignificant features: Smoker, PhysActivity, Fruits, Veggies, AnyHealthcare and NoDocbcCost.

For this report, I have randomly selected about 20% of the data for testing and the rest for training.

## 2. Models used and their goodness-of-fit

As the input variables are mostly categorical, with some being continuous, and the output is categorical, I propose to use 3 models for the prediction of diabetes. They are Decision Trees, Naive Bayes and Logistic Regression.

### 2.1 Decision Trees

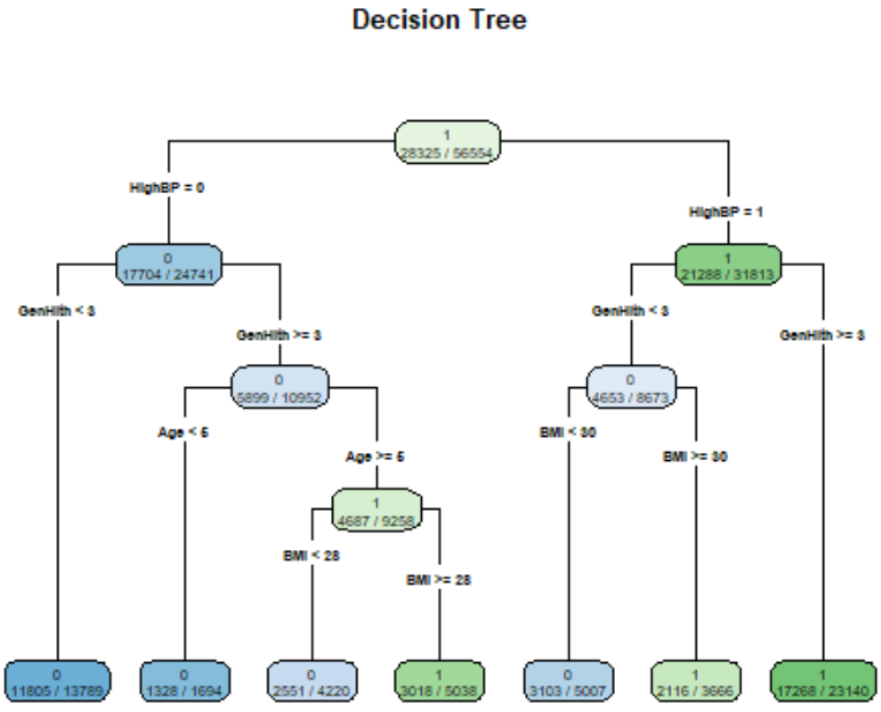


*Fig 2. Decision Trees model for predicting Diabetes_binary*

I chose a minsplit of 5000, which is about 10% of the sample size of the training set. From the Decision Trees model, we can see that High Blood Pressure, being the decision variable at the root node, is the most informative attribute, followed by General Health, and then other variables.

#### 2.1.1 Goodness-of-fit of the Decision Trees model

In this report, I will use 2 ways to calculate the goodness-of-fit of each model and thereafter determine which model is the best. The two tests I will be using are accuracy and Area Under the Receiver Operating Characteristic (ROC) Curve (AUC). Higher values for both tests will mean the model performs better.

Using the confusion matrix, I calculated the accuracy of the Decision Tree, which was 0.7254209.

The Decision Tree has an AUC value of 0.7258866. When calculated by getting probabilities instead of class for outcome, the AUC value is higher, at 0.7668546.

### 2.2 Naive Bayes

The Naive Bayes classifier is easy to implement and can be executed efficiently, however, it assumes the effects of inputs on the response are independent of each other. This might not be the case in the prediction of diabetes.

As the Naive Bayes classifier calculates probabilities and is not sensitive to the scale of the input features, I used the sample training set as the one I used for Decision Trees.

#### 2.2.1 Goodness-of-fit of Naive Bayes classifier

The accuracy value is 0.7218843, lower than that of the Decision Trees model.

The AUC value is 0.7862417, higher than both AUC values calculated in the Decision Trees model.

### 2.3 Logistic Regression

I scaled the input variables that were not binary in the Logistic Regression model to improve the numerical stability and for a more accurate calculation.

#### 2.3.1 Goodness-of-fit of Logistic Regression model

The accuracy of the model is 0.7435281, higher than the other models.
The AUC value calculated is 0.8219271, also the highest out of the various classifiers used.

The ROC plot for the classifiers is given below. As shown, Logistic regression gives the biggest area under the curve.

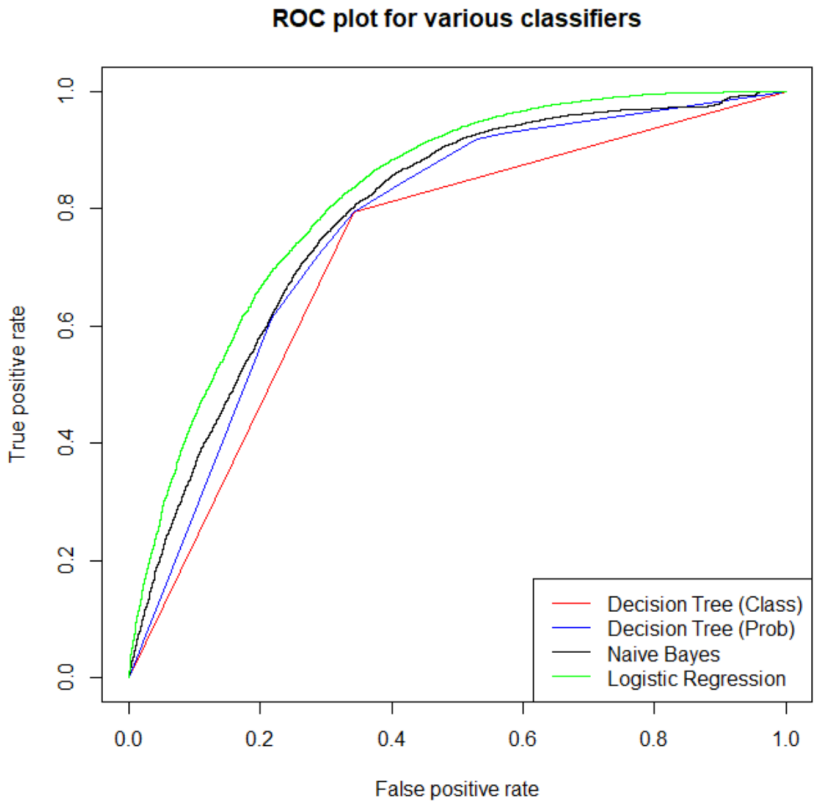


*Fig 3. ROC plot for the various classifiers*

## 3. Conclusion

Since Logistic Regression gives the highest AUC value as well as accuracy, I conclude that it is the best model for predicting the occurrence of diabetes.

To further evaluate the goodness-of-fit of the Logistic Regression model, I used N-folds cross validation with N = 5 to find the average accuracy for the model. It was calculated to be 0.748246, still higher than the other models used and higher than the accuracy calculated above using the confusion matrix.

The advantages of using the Logistic Regression model include its interpretability, it is easy to interpret the effect of each predictor, and its well-suitedness in producing binary outcomes.

However, the Logistic Regression model assumes a linear relationship between predictors and log odds of the outcome, which can limit its ability to capture complex relationships in the given data.