I. Introduction

i. Background of the research problem

Diabetes is among the most prevalence chronic diseases in the world. The dataset shows a total of 70,692 survey responses regarding Diabetes form a survey conducted in the US in 2015.

ii. Exploring the dataset

The dataset has an equal proportion of people diagnosed with diabetes and people who do not.

In the dataset, the response variable is “Diabetes binary”, which indicates whether the person has diabetes.

There are 21 input variables, of which some significant variables include “HighBP”, “GenHlth”, “DiffWalk”, “Age”, “Income”, “HeartDiseaseorAttack”, “HighChol”, “BMI” and “PhysHlth”.

1. “GenHlth” indicates the scale of general health level of the person (1 = excellent; 2 = very good; 3 = good; 4 = fair; 5 = poor) , and it has a correlation value of 0.408 with the response variable, showing the relative strong positive correlation between income and diabetes.
2. “HighBP” indicates that whether the person has a high BP (0 = no high BP; 1 = high BP), and it has a correlation value of 0.382 with the response variable, showing the relative strong positive correlation between income and diabetes.
3. “BMI” shows the body mass index, and it has a correlation value of 0.293 with the response variable, showing the relative strong positive correlation between income and diabetes.
4. “High Chol” indicate that whether the person has a high cholesterol (0 = no high cholesterol, 1= high cholesterol), and it has a correlation value of 0.289 with the response variable, showing the relative strong positive correlation between income

and diabetes.

5. “Age” has a correlation value of 0.279 with the response variable, showing the relative strong positive correlation between income and diabetes.
6. “DiffWalk” indicates whether the person has serious difficulty walking or climbing stairs (0 = no, 1 = yes) and it has a correlation value of 0.273 with the response variable, showing the weak positive correlation between income and diabetes.
7. “PhysHlth” shows the number of days in the past 30 days the person had physical illness and injury and it has a correlation value of 0.213 with the response variable, showing the weak positive correlation between income and diabetes.
8. “HeartDiseaseorAttack” shows whether the person has coronary heart disease (CHD) or myocardial infarction (MI) (0 = no; 1 = yes) and it has a correlation value of 0.212 with the response variable, showing the weak positive correlation between income and diabetes.
9. “Income” has a correlation value of -0.224 with the response variable, showing the weak negative correlation between income and diabetes.

In this report, I will be focusing on these 9 significant input variables mentioned above.

iii. Purpose of the Report

In this report, I will

1. choose a classification method for predicting diabetes status,
2. propose the best classifier, and
3. investigate the goodness of fit of that classifier.

II. Statistical Procedures Used

In this report, I will focus on two models which are Naïve Bayes and K-Nearest Neighbors and investigate their accuracy respectively to compare which one is a better classifier.

i. Method to Divide the Dataset

I will split the dataset into two parts for training and for testing with the ratio 4:1 while ensuring a same proportion of the binary responses in each part.

To achieve this, at the beginning, I split the dataset to two groups based on the outcome of “Diabetes_binary”. To divide each group to 5 subsets randomly, I apply the sample function twice to the two groups. I combine two subsets from each group and treat them as the test data while keeping the rest as the train data.

ii. Assessing Goodness-of-fit

a) Naïve Bayes Model

The response variable for Naïve Bayes model is usually categorical, so we change the “Diabetes_binary” to categorical variable first. Firstly, we initialize an empty vector for accuracy. We carry out N-fold Cross-Validation method to combine different subsets of the two groups as the testing data using a “for” loop. In each iteration, we fit the dataset to the naïve bayes model and calculate out the accuracy of the model and stored in the empty vector. The goodness-of-fit of the model can be attained by calculating the average value of accuracy stored in the empty vector, which is 0.710.

b) K-Nearest Neighbors Model (KNN)

In this model, we should treat the response variable “Diabetes_binary” as a quantitative variable. Similar to the naïve bayes model, we carry out N-fold Cross-Validation method to combine different subsets of the two groups as the testing data using a “for” loop and in each loop, we fit the dataset to the KNN model to assess its goodness-of-fit by calculating the average accuracy and storing it in an empty vector. However, we must choose the best value of k (the number of neighbors for the classifier) which gives highest value of accuracy. Hence, we add an outer loop to calculate and store the respective average accuracy for different values of k (eg. from 1 to 20). Lastly, we find out the highest value of the average accuracy, which is 0.724 when k = 20 (Fig.1).

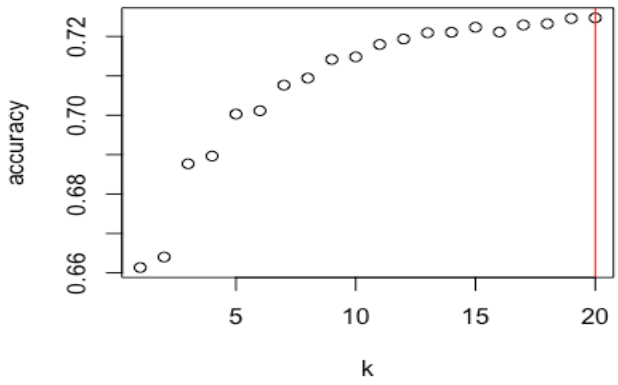

Fig.1

iii. Comparing Models

We compare the accuracy of the two models and find out that KNN model has a higher accuracy, indicating that it is a better model to fit for this dataset.

iv. Comments on the better model

Although KNN is more accurate in predicting the response variable for this dataset, it has a lower efficiency in processing data especially when the dataset is large. Moreover, we must test for different values of k which will make the algorithm slower. Furthermore, it is very sensitive to the outliers as it involves calculation of the distance between points.

III. Summary of Statistical Findings

After applying two different models to the dataset, I found that KNN is a better classifier as it has a higher accuracy although it has its limitations as mentioned above.