Assignment 2 Statistical Report

Introduction

Diabetes is a chronic medical condition that occurs when the body cannot regulate blood glucose (sugar) levels. Common symptoms include increased thirst, frequent urination, unexplained weight loss, fatigue, blurred vision, and slow wound healing. blurred vision, and slow wound healing. Unfortunately, it is one of the most prevalent chronic diseases in the world. The diagnosis of diabetes involves a combination of clinical evaluation, laboratory tests, and assessment of symptoms. The purpose of this report is to propose the best classifier while showing its derivation, using a clean data set of 70,692 survey responses from a survey conducted in the US in 2015.

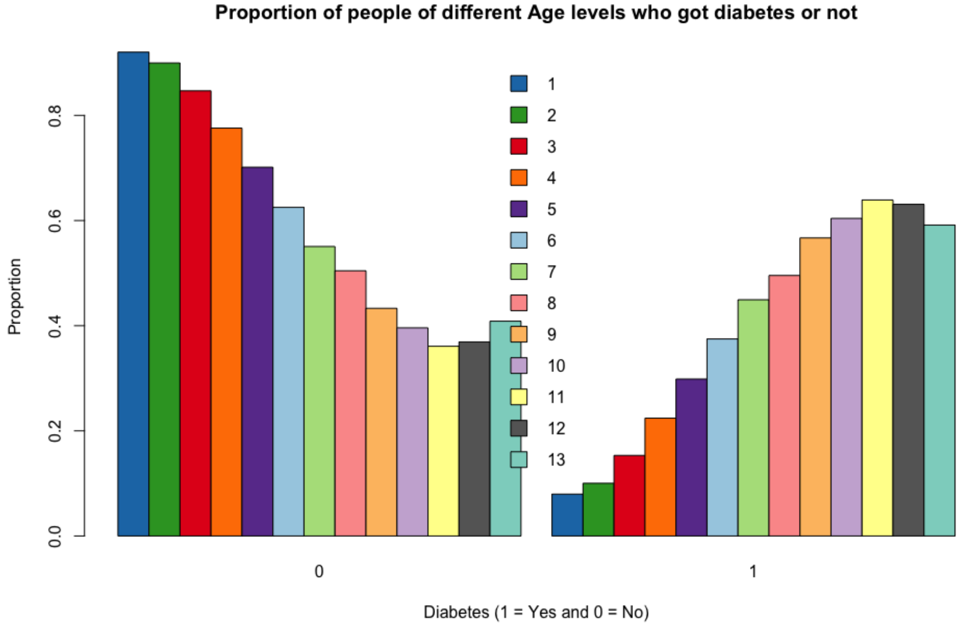
Statistical Procedures Used

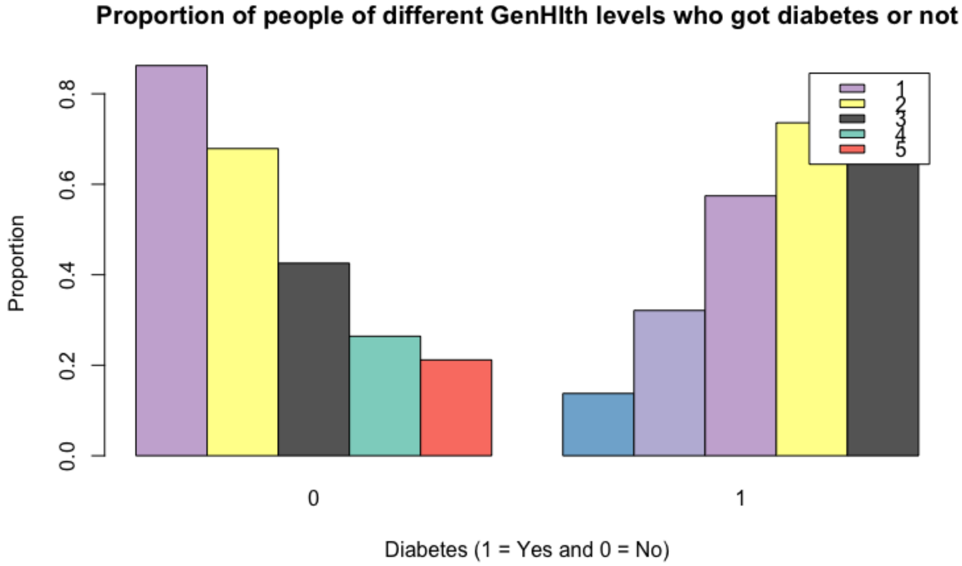
To simplify the model, reduce overfitting and increase computational efficiency in building a classification model from the large dataset, there is a need to first exclude some input variables from the dataset. Variables with no strong correlation with response variable (Diabetes_binary) and statistically insignificant will be excluded from the dataset.

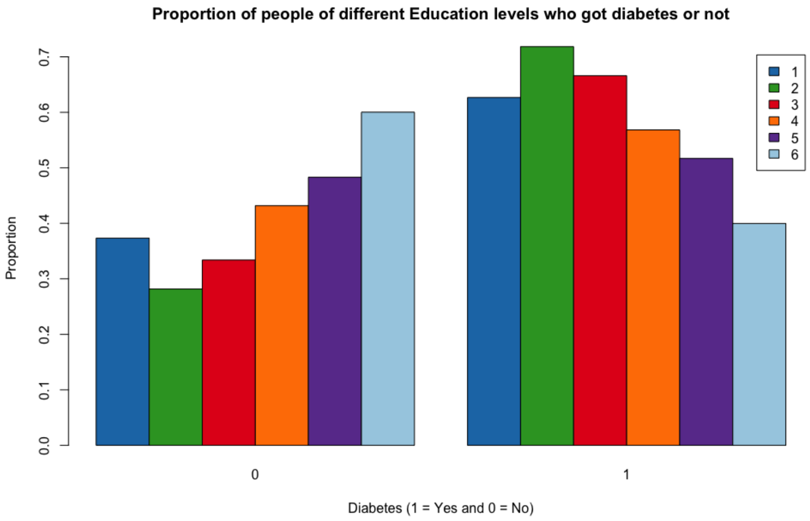
From the dataset, it can be seen that GenHlth, Age, Education and Income are ordinal input variables.

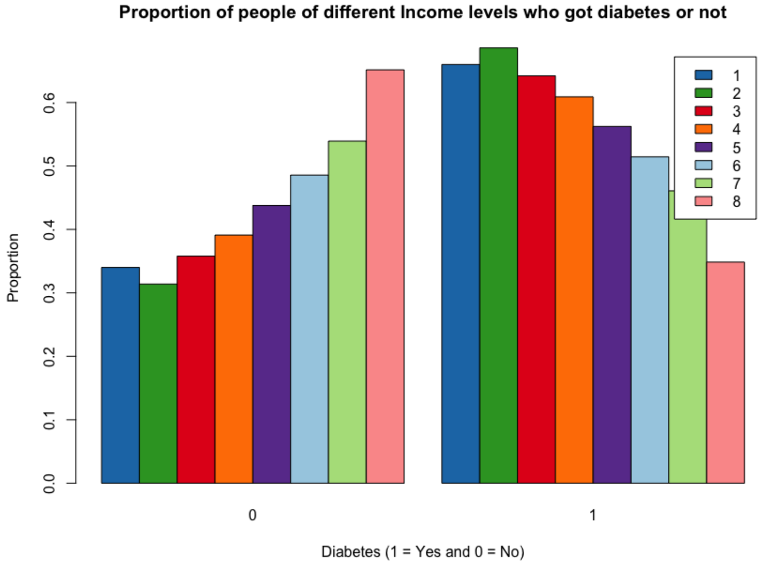
Variables	Description
GenHlth	Person’s judgement of his/her own health: 1 = excellent; 2 = very good; 3 = good; 4 = fair; 5 = poor.
Age	13 categories: 1 = age from 18 to 24; ….; 9 = age 60 to 64; 13 = age 80 or above
Education	Education level scale 1 to 6: 1 = never attended school or only kindergarten; 2 = elementary; ….
Income	income scale 1 to 8: 1 = less than 10k; …; 5 = less than 35k; …; 8 = 75k or more

These variables will be treated as quantitive instead of categorical as they have a clear ordering and possess many categories. Comparison of the relationship between the proportion of people with diabetes and the above variables using bar plots:









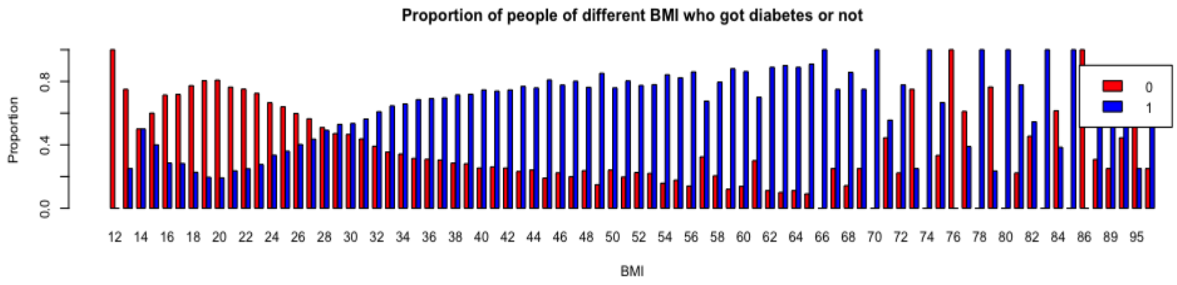
From the box plots above, it seems that there is a strong linear relationship between GenHlth, Age, Education, Income levels and one having diabetes (positive linear relationship for GenHlth and Age; negative linear relationship for Education and Age levels). This implies that these input factors appear to influence the probability of developing diabetes. Hence, these variables will be included in the modelling of classifiers.

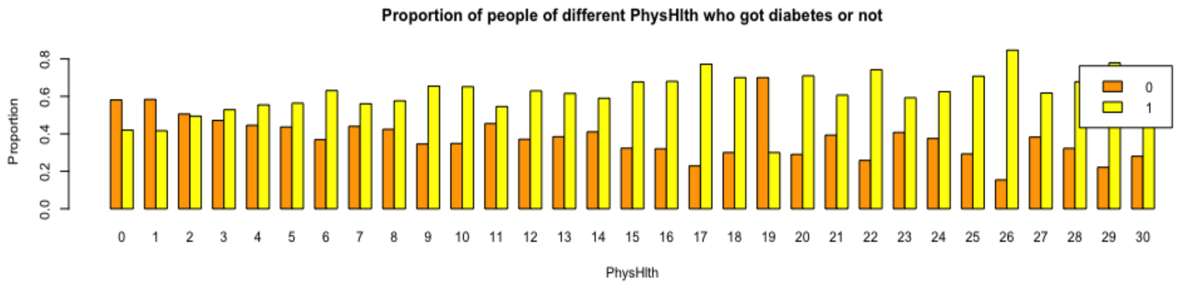
BMI, PhysHlth and MentHlth in the dataset are quantitive variables.

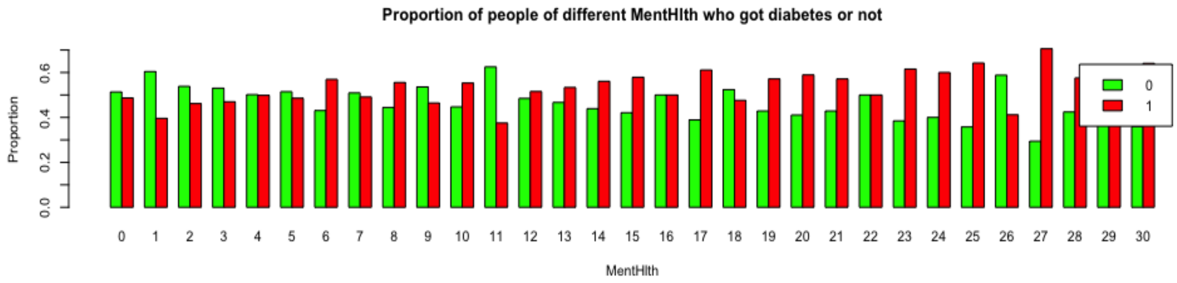
Variables	Description
BMI	Body mass index

PhysHlth	The number of days the person’s his/her physical health (physical illness and injury) not being good during the past 30 days
MentHlth	The number of days the person’s mental health (stress, depression, and problems with emotions) not being good during the past 30 days

Comparison of the relationship between the proportion of people with diabetes and the above variables using bar plots







From the box plots above, there is no clear correlation between the proportion of people having diabetes and their PhysHealth, MentHealth. However, there is a positive relationship between BMI and the proportion of people getting diabetes. Hence, PhysHealth, MentHealth will be excluded from the dataset in forming the classifiers.

Finally, a logistic regression model is performed on the dataset to determine which variables are statistically significant. Those with p-values > 0.05 are taken as statistically insignificant. These include AnyHealthcare, NoDocbcCost, Fruits, PhysActivity, Smoker variables. Hence, they will be excluded from the variable as there is insufficient evidence to conclude that the variable has a significant impact on the response variable (Diabetes_binary).

The dataset is then cut down from the initial 22 variables to 15 variables. These 15 variables will be used to build the classifiers. All classifiers will be modelled using 5-fold cross-validation and a 1-to-1 ratio of one having diabetes and one not having diabetes from the dataset. This is to ensure the model's performance is more reliable and does not become overfitting.

Decision trees, K-nearest-neighbors (KNN), Naïve Bayes and Logistic Regression are used as classification models. For each of them, the mean accuracy and type 2 error is computed. Type 2 error is computed as the assumption is that a false negative rate is more tolerable in the medical context (It is more severe to predict a negative result for a diabetic person). For KNN, 50 K values were tested to find which is the best K and the result is that K = 46 is the best K as it produces the highest mean accuracy and lowest type 2 error rate. Here are the findings for each models:

Classification Model	Mean Accuracy rate	Mean Type 2 Error Rate
Decision trees	0.726	0.233
KNN	0.744	0.230
Naïve Bayes	0.736	0.266
Logistic Regression	0.748	0.242

They all have very similar mean accuracy and type 2 error rate. However, Naïve Bayes will not be used as a classifier as it has the highest type 2 error rate and the second lowest mean accuracy rate. Additionally, Naïve Bayes assumes that all variables in the dataset are conditionally independent. However, this assumption is not true as many features are correlated to one another, such as Age and HighBp/HighChol.

To compare between Decision Trees, KNN and Logistic Regression in terms of their performance, AUC-ROC metric is used. A higher AUC value means the model has a higher true positive rate and a lower false negative rate. Here are the AUC values for each of the classifiers:

Classifier	AUC Value
Decision Trees	0.726
KNN	0.748

Logistic Regression	0.743

Conclusion

From the table above, KNN is the best model as it has the highest AUC value. While KNN does not require a training period which means that predictions are made based on the nearest neighbors during runtime and hence, advantageous in this context where the data evolves, it possesses computational complexity during prediction and requires a large memory to store the large dataset. However, one should not rely solely on the KNN algorithm to predict the patient’s health as it still does not have 100% accuracy and standard procedures should still be used in diagnosing whether one got diabetes or not.