Introduction

In this report, we will be analysing the dataset called 'diabetes_5050.csv' and propose the best classifier for predicting diabetes status. This clean data set comprises 70,692 survey responses from a survey conducted in the US in 2015. Upon closer analysis, it was found to contain 22 distinct variables, including 1 response variable "Diabetes_binary" and 21 input variables. Examples of input variables include but are not limited to "HighBP", "BMI" and "AnyHealthcare".

Further calculation also revealed that the ratio of individuals with diabetes to those without is 50:50 as indicated by the equal number of '0 'and '1' under the response variable - 35346 each. However, it is important to note that this does not align with real-life statistics. The Centers for Disease Control and Prevention has reported that 37.3 million people in the US have diabetes, which is equivalent to 11.3% of the US population. Therefore, it is crucial to acknowledge that for the purpose of this study, this data has been artificially modified.

Analysing response and input variables

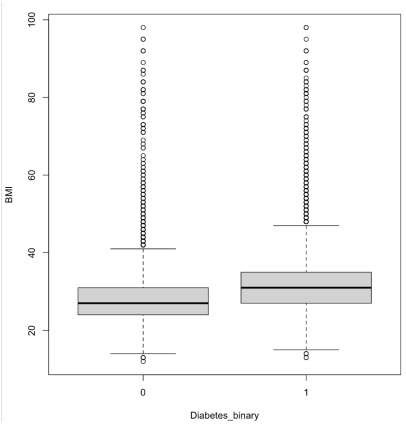
With the large amount of variables present in this study, it is vital for us to check the association between the response and input variables before fitting any classifiers or models. Including irrelevant variables and correlating them with the response variable could change the variance of the estimate, potentially making predictions less precise. To check for association, we will first categorize the response variables into quantitative, categorical or ordinal before subsequently employing the correct appropriate method to determine association between itself and the response variable.

Under ordinal variables, we have "GenHlth", "Age", "Education" and "Income".
For ordinal variables, chi-squared test has been adopted to measure the association. With the chi-squared test, we are able to obtain the p-value of all variables, which is surprisingly the same for all, at < 2.2e-16. Since the p-value for all variables is less than 0.05, it suggests that these variables are very significant to the response variable.

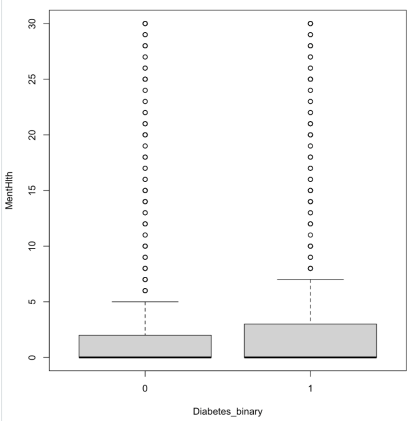
Under categorical variables, we have "HighBP", "HighChol", "CholCheck", "Smoker", "Stroke", "HeartDiseaseorAttack", "PhysActivity", "Fruits", "Veggies", "HvyAlcoholConsump", "AnyHealthcare", "NoDocbcCost", "DiffWalk" and "Sex".
For categorical variables, odd ratio has been adopted to measure the association. Since the process of finding odd ratios is the same for all 14 categorical variables, a for loop function has been written which produces all the odd ratios. The odd ratios for all 14 categorical variables are as follow:
Odds ratio for HighBP : **5.088477**
Odds ratio for HighChol : **3.296316**
Odds ratio for CholCheck : **6.491553**
Odds ratio for Smoker : **1.412383**
Odds ratio for Stroke : **3.093272**
Odds ratio for HeartDiseaseorAttack : **3.656197**
Odds ratio for PhysActivity : 0.4939616
Odds ratio for Fruits : 0.8007655
Odds ratio for Veggies : 0.6763796
Odds ratio for HvyAlcoholConsump : 0.3653163
Odds ratio for AnyHealthcare : **1.251644**
Odds ratio for NoDocbcCost : **1.326217**
Odds ratio for DiffWalk : **3.807365**
Odds ratio for Sex : **1.195343**

By interpreting odds ratio (OR), where OR=1 suggests that input variable is not associated to response variable, OR>1 suggests that input variable is positively associated to response variable and OR<1 suggests that input variable is negatively associated to response variable, it is known to us that only variables with the bolded OR have a possibility that they are significantly associated to the response variable. For further analysis, the phi coefficient (φ) for every variable which has a OR close to 1 is calculated and the phi coefficients are as follow:
φ of Smoker : 0.09
φ of AnyHealthcare : 0.02
φ of NoDocbcCost : 0.04
φ of Sex : 0.04
Since all the φ is less than 0.1, it indicates a small and weak association with the response variable. Hence, we will not be looking at these variables.

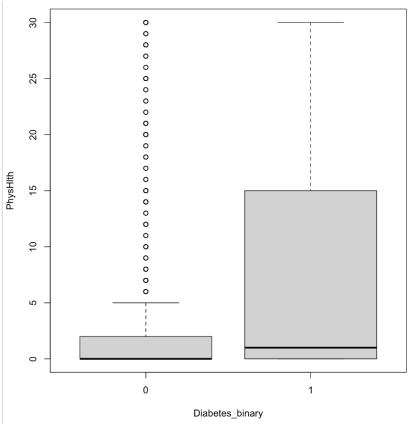
Under quantitative variables, we have "BMI", "MentHlth" and "PhysHlth".
For quantitative variables, we will be using boxplot to observe if there are any association before using coefficient correlation to confirm our observation.



"BMI"
From the figure, there is only a little overlapping of the two box plots which suggests a probable positive association between the two variables.
Further calculation revealed that the correlation coefficient is 0.2933727 which is greater than 0. Hence, there is a positive association between these two variables.



"MentHlth"
From the figure,the majority of the boxplot is overlapping which suggests a probable weak association between the two variables.
Further calculation revealed that the correlation coefficient is 0.08702877 which is too close to 0. Hence, there is a weak association between these two variables.



"PhysHlth"
From the figure, although there is some overlapping between the two boxplots, the majority of the boxplots are not overlapping which suggests a probable strong association between the two variables. Further calculation revealed that the correlation coefficient is 0.213081. Hence, there is a strong association between these two variables.

As such, we can conclude that only “PhysHlth” and “BMI” are the variables more closely associated with the response variable, “Diabetes_binary”.

After identifying the variables strongly associated with the response variable, data will be modified such that variables that aren’t closely related to the response variable will be dropped, leaving us with a data set that only has 12 input variables and 1 response variable.

Building different classifiers and analysing each of them

For this study, before building different classifiers, we have split the data into 80% for the training set and 20% for the test set to conduct n-cross validation of 5 folds.

In total, we have built a total of 4 classifiers, namely Logistic Regression, Decision Tree, Naive Bayes, and KNN. For each classifier, several performance estimators were calculated including accuracy, precision, recall value, f1 value as well as auc value.

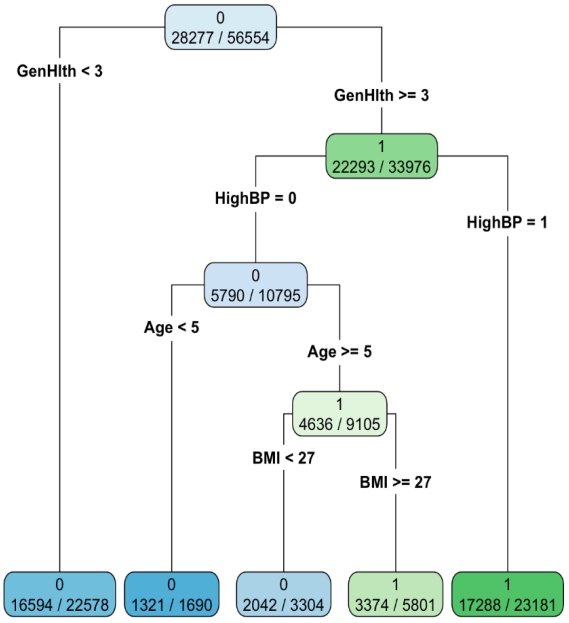
The table below shows the summary of every classifier with their performance estimator values:

Classifier	Accuracy	Precision	Recall value	F1 value	AUC value
Logistic Regression	0.746	0.761	0.731	0.751	0.822
Decision Tree	0.723	0.731	0.706	0.733	0.756
Naive Bayes	0.720	0.690	0.724	0.713	0.794
KNN	0.720	0.752	0.707	0.728	0.720

Analyzing each classifier as well as the pros and cons

Logistic regression

From one look of the table above, logistic regression can be considered to be the best classifier among all since it has the highest accuracy, precision, recall value, F1 value as well as AUC curve. However, it is important to exercise caution before hastily drawing any conclusions, as several limitations and assumptions need consideration. For instance, the underlying of logistic regression being a good classifier is that it assumes linearity between dependent and independent variables. Furthermore, due to the high dimensionality of data, it is susceptible to overfitting in the presence of a large number of predictor variables. Hence, this raises concerns about the generalizability of the model to new, unseen data.



Decision Tree

From the table, Decision Tree has the lowest recall value, which may indicate a high false negative due to specific variables in the data. Upon analysing the plotted diagram, it shows that GenHlth is identified as the most influential factor in predictions followed by HighBP, Age, BMI. This finding may be surprising, considering that GenHlth is a subjective factor as compared to factors such as BMI which is commonly emphasized by healthcare professionals for its importance in assessing our health. Hence, this raises the question of interchange of subjective and objective variables in influencing predictions not just in the Decision Tree but in other classifiers as well.

Naive Bayers

Naive Bayers had low accuracy value but a rather high AUC value in the table. This may be due to AUC being less sensitive to class imbalance than accuracy and hence AUC can still effectively discriminate between classes. However, if accuracy and precision is key for analysing the data, Naive Bayers should not be the classifier for it.

KNN

By comparing AUC value between each classifier, it seems that KNN is the least accurate classifier among all. This may be due to KNN being very sensitive to outliers and lead to misclassification. Since this data has many outliers, as seen from the box plots plotted for quantitative variables (Pg 2), hence, KNN algorithm may have misclassified many data, leading to it being the least suitable classifier. Furthermore, with the large amount of predictor variables, expected distance to nearest neighbour increases. However, KNN relies on the concept of proximity. As such, the nearest neighbours of a data point is most likely not going to be a true representative of local points and hence, leading to less accurate prediction.

Conclusion

Out of all the classifiers, logistic regression is still preferably the best classifier. Despite the assumption of linearity between variables, plotting a scatterplot allows for a quick assessment of the existence of a linear relationship between them. In addition, to address the problem of overfitting, further analysis of the variables can be conducted such as using standardised coefficients or the change in R-square for the last variable added to the model. This involves scrutinising the predictors more in depth and retaining only those of utmost significance. By addressing these assumptions and limitations, logistic regression still emerges as the optimal choice.