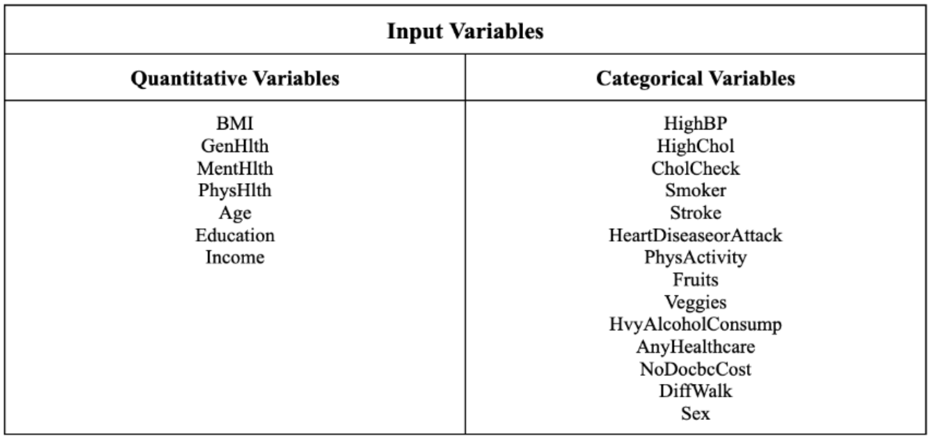
Statistical Report on Diabetes

1 Introduction

Diabetes is ranked as one of the most pervasive chronic diseases globally. The data used were collected through a survey conducted in the United States in 2015. It captured a diverse range of information related to diabetes, including demographic details, lifestyle factors, and crucial health indicators. The dataset has an equal 50-50 split of respondents with no diabetes and with either prediabetes or diabetes. This report will evaluate various classification methods aimed at identifying the most effective model by assessing its goodness of fit to predict diabetes status and detailing the rationale behind the choice of the best classifier.

2 Exploration of Data Set

There are a total of 21 input variables and 1 response variable, Diabetes_binary. The response variable, Diabetes_binary, is a categorical variable with 0 and 1 referring to non-diabetic and diabetic status respectively. By considering the ordinal input variable as a quantitative variable. The 21 input variables comprise of 7 quantitative variables and 14 categorical variables which can be found in Table 1 below.

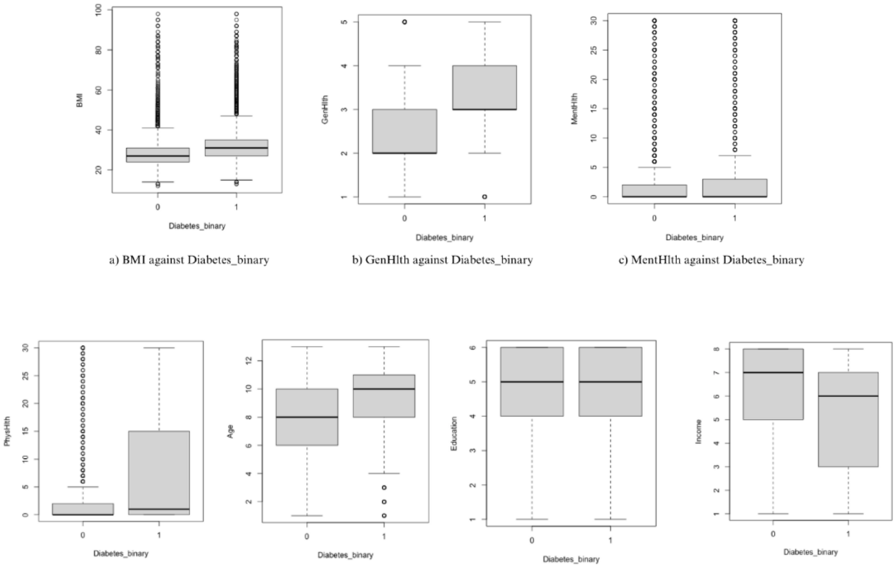


Input Variables	
Quantitative Variables	**Categorical Variables**
BMI GenHlth MentHlth PhysHlth Age Education Income	HighBP HighChol CholCheck Smoker Stroke HeartDiseaseorAttack PhysActivity Fruits Veggies HvyAlcoholConsump AnyHealthcare NoDocbcCost DiffWalk Sex

Table 1: Classification of Input Variables

2.1 Association between Quantitative Variables and Response Variable

To compare the association between the quantitative variables and the categorical response variable, Diabetic_Binary, the quantitative variables were plotted against the categorical response variable using boxplots. To determine if the quantitative variable affects the diabetic status, the Median and Interquartile Range (IQR) between both diabetics and non-diabetics were analysed.



a) BMI against Diabetes_binary

b) GenHlth against Diabetes_binary

c) MentHlth against Diabetes_binary

d) PhysHlth against Diabetes_binary

e) Age against Diabetes_binary

f) Education against Diabetes_binary

h) Income against Diabetes_binary

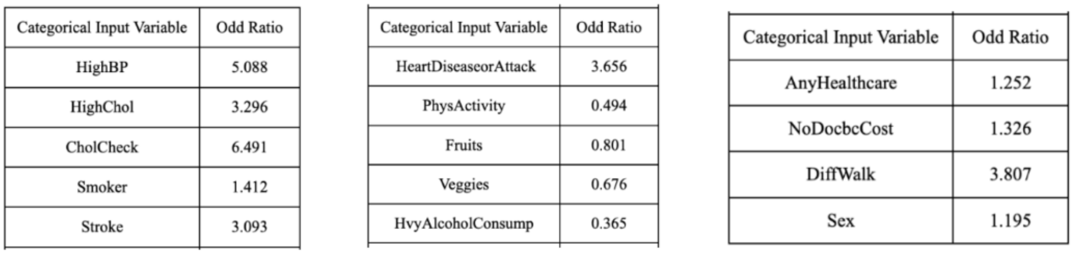
Figure 1: Boxplots of quantitative variables against response variable

2.2 Analysis of Association

With reference to subfigures in Figure 1 in alphabetical order. The median BMI value and interquartile range of a diabetic individual is slightly larger than a non-diabetic individual, therefore BMI affects diabetic status. The median general health value and interquartile range of a diabetic individual are larger than a non-diabetic individual, therefore general health affects diabetic status. The median values and interquartile range of mental health are similar with multiple outliers, therefore mental health does not affect the diabetic status. The median physical health is larger and has a wider interquartile range for a diabetic individual than a non-diabetic individual,therefore physical health affects diabetic status. The median age of a diabetic individual is larger than a non-diabetic individual, and the interquartile range for a diabetic individual is smaller than a non-diabetic individual, therefore age affects diabetic status. The median values and interquartile range for education are similiar, therefore education does not affect diabetic status. The median income of a diabetic individual is lower than non-diabetic individual, and the interquartile range for a diabetic person is larger than a non diabetic person therefore income affects diabetic status. Therefore it can be concluded that BMI, GenHlth, PhysHlth, Age, and Income are all significant input variables that will be included in fitting the model. MentHlth and Education are insignificant input variables that will be excluded from fitting the model.

2.3 Association between Categorical Variables and Response Variable

To compare the association between the categorical input variables and the categorical response variable, Diabetic_Binary, a contingency table was utilised to calculate the odd ratio between a diabetic individual over a non-diabetic individual. The odds ratio represents the ratio between the likelihood of having diabetes with a categorical variable compared to the likelihood of not having diabetes with the same categorical variable. The strength of association is directly proportional to the magnitude of the odds ratio: the farther the ratio is from 1, the more significantly the categorical variable affects the diabetic status.



Categorical Input Variable	Odd Ratio
HighBP	5.088
HighChol	3.296
CholCheck	6.491
Smoker	1.412
Stroke	3.093

Categorical Input Variable	Odd Ratio
HeartDiseaseorAttack	3.656
PhysActivity	0.494
Fruits	0.801
Veggies	0.676
HvyAlcoholConsump	0.365

Categorical Input Variable	Odd Ratio
AnyHealthcare	1.252
NoDocbcCost	1.326
DiffWalk	3.807
Sex	1.195

Figure 2: Odd ratio tables for each Categorical Variable

2.4 Analysis of Association

With reference to Figure 2, a categorical input variable with an odd ratio having an absolute difference of more than 0.5 from 1 larger is considered to have a significant effect on the diabetic status. Therefore, it can be concluded that HighBP, HighChol, CholCheck, Stroke, HeartDiseaseorAttack, HvyAlcoholConsump, and DiffWalk are considered to be significant input variables that will be included in fitting the model. Smoker, PhysActivity, Fruits, Veggies, AnyHealthcare, NoDocbcCost, and Sex are insignificant input variables that will be excluded from fitting the model.

2.5 Dropping of insignificant variables and separating of data

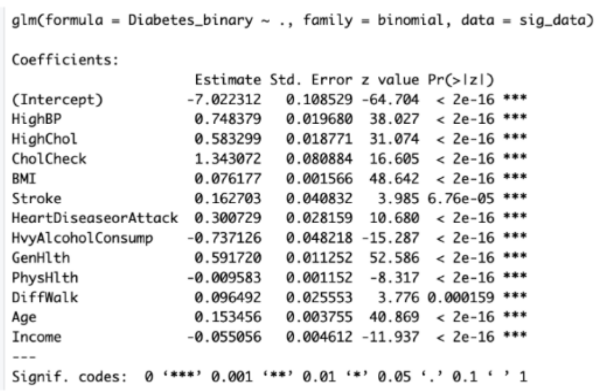
To ensure a better fit for the models, the insignificant variables were dropped from the dataset. In order to maintain an equal spread of diabetic and non-diabetic data in both the train and test data to prevent overfitting, the data was split into diabetic and non-diabetic. For each of these datasets, 5-fold cross-validation was performed to ensure an 80:20 ratio for the training and testing data respectively. The train data set for both diabetic and non-diabetic was then combined to form the overall train data set while the test data set for both diabetic and non-diabetic was then combined to form the overall test data set. This methodology was applied to all models.

3 Building of Classifiers

Since the response variable is a categorical variable, the classifiers proposed are K-Nearest Neighbors, Decision Tree, Naive Bayes, and Logistic Regression. The proposed diagnostics for the classifiers are precision, accuracy, and False Negative Rate (FNR) with FNR being the most significant diagnostics as it is detrimental if a diabetic individual is misclassified. Utilising the methodology in subsection 2.6, 5-fold cross-validation was performed for all the classifiers. For each iteration of the fold, a confusion matrix was utilised to calculate the diagnostics. After 5 folds, the mean for each diagnostics was calculated in order to compare between the classifiers.

3.1 Logistic Regression

There is a need to check the significance of the input variables in the fitted model, by extracting the P-value. If the P-value is below the significant value of 0.5, the input variable is considered significant otherwise it is removed from the model. Since the model returns raw probability, there is a need to find the threshold to classify the probability into the diabetic status, this can done by extracting the threshold (alpha value), False Positive Rate (FPR), and True Positive Rate (TPR) values and plotting the curve to determine the most optimal threshold with the highest TPR and lowest FPR. The 5-fold cross-validation mentioned in section 3 was performed on the training and test data for Logistic Regression with the optimal threshold value.



```
glm(formula = Diabetes_binary ~ ., family = binomial, data = sig_data)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)    
(Intercept)          -7.022312   0.108529 -64.704  < 2e-16 ***
HighBP                0.748379   0.019680  38.027  < 2e-16 ***
HighChol              0.583299   0.018771  31.074  < 2e-16 ***
CholCheck             1.343072   0.080884  16.605  < 2e-16 ***
BMI                   0.076177   0.001566  48.642  < 2e-16 ***
Stroke                0.162703   0.040832   3.985 6.76e-05 ***
HeartDiseaseorAttack  0.300729   0.028159  10.680  < 2e-16 ***
HvyAlcoholConsump    -0.737126   0.048218 -15.287  < 2e-16 ***
GenHlth               0.591720   0.011252  52.586  < 2e-16 ***
PhysHlth             -0.009583   0.001152  -8.317  < 2e-16 ***
DiffWalk              0.096492   0.025553   3.776 0.000159 ***
Age                   0.153456   0.003755  40.869  < 2e-16 ***
Income               -0.055056   0.004612 -11.937  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 3: P-values of Logistic Regression

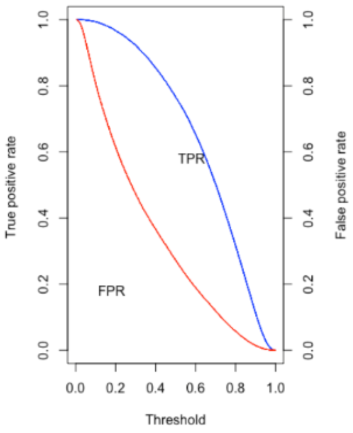


Figure 4: Logistic Regression Threshold plot

Average Accuracy	Average Precision	Average FNR
0.744	0.854	0.300

Table 2: Logistic Regression Diagnostic values

3.2 Logistic Regression Goodness of Fit

Referring to Figure 3, the P-value for all the input variables used to fit the model is very small and less than 0.5 hence it can be concluded that all the input variables are significant. Referring to Figure 4, the optimal threshold was chosen at 0.4 as it has a high TPR and a low FPR. The goodness of fit for the Logistic Regresion model can be observed in Table 2, with a relatively high accuracy and precision of 0.744 and 0.854 respectively.

3.3 K-Nearest Neighbors

On top of the 5-fold cross-validation, there is a need to find the K value that produces the lowest mean FNR. In consideration of the Bias-variance tradeoff, the range of K value was capped at 10 to determine the most optimal K value.

	Average Accuracy	Average Precision	Average FNR
K = 9	0.730	0.772	0.288

Table 3: KNN Diagnostic values at K = 9

3.4 K-Nearest Neighbors (KNN) Goodness of Fit

In the range of 10 K values, the most optimal K value is 9, it has the lowest average FNR and highest average accuracy and precision. The goodness of fit for the KNN model when K = 9 can be observed in Table 3, with a relatively low FNR at 0.288.

3.5 Decision Tree

The 5-fold cross-validation mentioned in section 3 was performed on the training and test data for the decision tree with a minimum split of 7000 (10% of the dataset) and split set to Gini.

Average Accuracy	Average Precision	Average FNR
0.726	0.800	0.303

Table 4: Decision Tree Diagnostic values

3.6 Decision Tree Goodness of Fit

The goodness of fit for the Decision Tree model can be observed in Table 4, with a relatively high precision at 0.800 and FNR at 0.303.

3.7 Naive Bayes

The 5-fold cross-validation mentioned in section 3 was performed on the training and test data for the Naive Bayes.

Average Accuracy	Average Precision	Average FNR
0.730	0.734	0.272

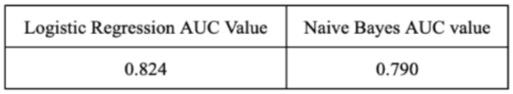
Table 5: Naive Bayes Diagnostic values

3.8 Naive Bayes Goodness of Fit

The goodness of fit for the Naive Bayes model can be observed in Table 5, with a relatively low FNR rate and relatively average accuracy and precision.

3.9 Comparison between Classifiers

By comparing all the diagnostics among all the proposed classifiers, Logistic Regression has the highest accuracy and precision and a relatively low FNR while Naive Bayes has the lowest FNR but a relatively average accuracy and precision. To determine the better classifier among the 2, there is a need to check on the Reciever Operating Characteristic (ROC) curve and Area under the Curve (AUC). The classifier with the larger ROC Curve and AUC value can be concluded to be the best classifier.



Logistic Regression AUC Value	Naive Bayes AUC value
0.824	0.790

Table 6: Logistic Regression and Naive Bayes AUC value

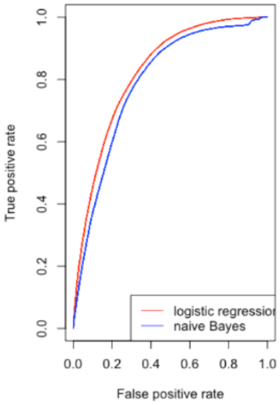


Figure 5: Logistic Regression and Naive Bayes ROC Curve

3.10 Best Classifier

Referring to Table 6 and Figure 5, Logistic regression has a larger AUC and a larger ROC curve. Therefore it can be concluded that Logistic Regression is a better classifier than Naive Bayes.

4 Conclusion

The best model to predict the diabetic status given the dataset would be Logistic Regression. With an average accuracy, precision and FNR of 0.744, 0.854 and 0.300 respectively, the model has the highest accuracy and precision. Similarly, the AUC value of the Logistic Regression model is the highest at 0.824, concluding that it has the highest ratio of TPR over FPR .However, it's worth noting that while the Logistic Regression model doesn't exhibit the lowest FNR compared to other classifiers, the difference in FNR between the Logistic Regression and the classifier with the lowest FNR is negligible.