Introduction

This report studies the data from the file diabetes_5050.csv to find the best model for predicting the diabetes status of a patient based on some features. The classifiers to be used include Naïve Bayes, Decision Tree and Logistic Regression.

Response Variable

The response variable in this report will be Diabetes_binary, as represented by the first column of the data. Diabetes_binary = 0 represents a person having no diabetes and Diabetes_binary = 1 represents prediabetes or diabetes. For simplification, a person is said to have no diabetes if Diabetes_binary = 0 and said to have diabetes if Diabetes_binary = 1.

Features

There are 70,692 rows of observations in the data and 21 columns of input variables, excluding the response variable Diabetes_binary which is the first column. Below is a short description of the 21 features.

HighBP: having high blood pressure: 0 = No; 1 = Yes

HighChol: having high cholesterol: 0 = No; 1 = Yes

CholCheck: have cholesterol checks in 5 years: 0 = No; 1 = Yes

BMI: Body mass index

Smoker: smoke at least 5 packs in life; 0 = No; 1 = Yes

Stroke: ever told to have a stroke: 0 = No; 1 = Yes

HeartDiseaseorAttack: have coronary heart disease: 0 = No; 1 = Yes

PhysActivity: done physical activity in the past 30 days excluding jobs: 0 = No: 1 = Yes

Fruits: consume fruits one or more times per day: 0 = No; 1 = Yes

Veggies: consume vegetables 1 or more times per day: 0 = No; 1 = Yes.
HvyAlcoholConsump: heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week): 0 = No; 1 = Yes.

AnyHealthcare: Do you have any kind of health care coverage (including health insurance, prepaid plans or government plans): 0 = No; 1 = Yes

NoDocbcCost: was there a time in the past 12 months when you needed to see a doctor but could not because of cost? 0 = No; 1= Yes.

GenHlth: would you say that in general, your health is: 1 = excellent; 2 = very good; 3 = good; 4 = fair; 5 = poor.

MentHlth: Thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good?

PhysHlth: Thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good?

DiffWalk: Do you have serious difficulty walking or climbing stairs? 0 = No, 1 = Yes.
Sex: 0 = Female; 1 = Male.

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.

Choosing Appropriate Features

Firstly, we would like to check the association between each feature and the response variable to identify the significant features that we should include in our model. This can help us to simplify the model by excluding unnecessary features.

The following features have two outcomes, being 1 = Yes and 0 = No.

- HighBP, HighChol, CholCheck, Smoker, Stroke, HeartDiseaseorAttack, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, AnyHealthcare, NoDocbcCos, GenHlth, DiffWalk, Sex

Odds ratio can be used to determine the association between the response and each of the variables above. Odds ratio is the ratio between two odds of success. In this report, we will consider having diabetes (Diabetes_binary = 1) as success. The further the odds ratio is from the value 1, the stronger the association. When odds ratio is 1, there is no difference in the odds of having diabetes between the two groups. When odds ratio is significantly lesser or greater than 1, the odds of having diabetes in one group is substantially lower or higher respectively. We will set the threshold to 0.5 for simplicity, so a variable is considered to be significant in the model if the odds ratio is <= 0.5 or >= 1.5.

For example, for the feature HighBP, the odds ratio of approximately 0.1965 indicates that the odds of individuals without high blood pressure getting diabetes are about 0.1965 the odds of individuals with high blood pressure getting diabetes. This means that HighBP is a significant feature since it significantly affects the odds of getting diabetes.

For the remaining non-binary features, boxplot will be used to check for any observable association.

- BMI, GenHlth, MentHlth, PhysHlth, Age, Edcation, Income

For example, the boxplot (Figure 1) of Age against Diabetes_binary below shows the median and interquartile range for Diabetes_binary = 1 higher than that of Diabetes_binary = 0.

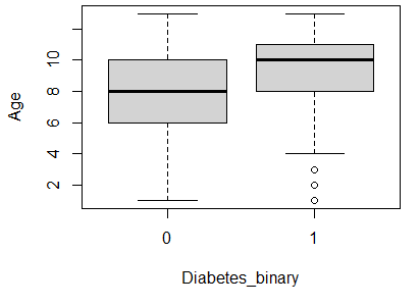

Figure 1: Box Plot of Age against Diabetes_binary

It can be inferred that those with diabetes are more likely to occur among larger age groups, thus Age is significant feature for determining the diabetes status of an individual.

After analysis, we will choose to include the following features for building the model.

- HighBP, HighChol, CholCheck, Stroke, HeartDiseaseorAttack, DiffWalk, PhysActivity, HvyAlcoholConsump, GenHlth, BMI, PhysHlth, Age, Income

Building Model

For the models, accuracy and Area Under the ROC Curve (AUC) will be used to measure its goodness of fit. Accuracy measures the ability of the model to predict an outcome correctly. It is the ratio of correctly predicted responses over the total number of responses. The AUC evaluates the model's performance in binary classification. It considers the True Positive Rate (TPR) and False Positive Rate (FPR), giving an idea about the model's ability to distinguish between positive and negative classes across different threshold values. It ranges from 0 to 1, where a higher value indicates better performance.

Model 1, M1 : Naïve Bayes

The first model, M1, will be built using Naïve Bayes classifier. We will randomly pick out 20% of the full data set to form our initial test set. The remaining data will be used to train the model. This ensures that the ratio of training data to testing data is 8:2. With initial M1, we will predict the response of the test data and compare the predicted responses against the actual responses. We then find the accuracy by computing the fraction of correctly predicted responses over the total responses. Next, the Receiver Operating Characteristic Curve (ROC) can be plotted and AUC can be extracted. Below is an example of the ROC curve for the initial model (Figure 2).

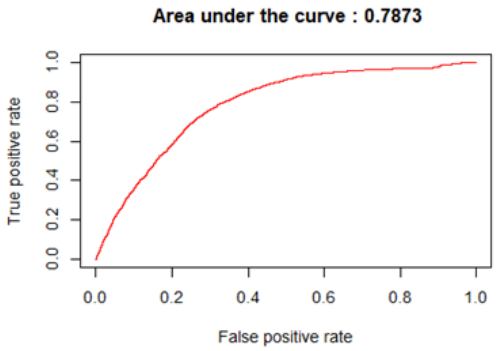

Figure 2: ROC and AUC for initial model M1

The process is repeated using a 5-fold cross-validation. The data is split into 5 sets, with each set taking turns to be the test set. With each loop, we compute the accuracy and AUC values and store them in a vector. The overall accuracy and AUC is given by the mean of the respective vectors.

Model 2, M2: Decision Tree

The second model, M2, will be built using Decision Tree classifier. We want to find the best 'cp' value for the algorithm such that the model will give us the highest accuracy. To do this, we will let 'cp' values take on a range of values from 10^-10 to 1. For each 'cp' value, we will do a 5-fold cross-validation and compute the average accuracy. The graph of accuracy against -log(cp) is given below (Figure 3).

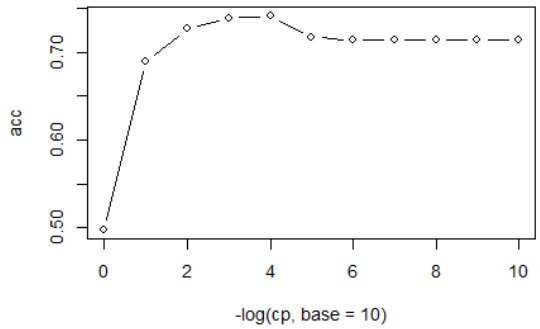

Figure 3: Graph of accuracy against -log(cp)

After analysis, accuracy is found to be highest when cp = 10^-4. The final model M2 will be built using the chosen 'cp' value which gives us a complex decision tree. A 5-fold cross-validation is then used to compute the mean AUC for M2.

Model 3, M3: Logistic Regression

The third model, M3, will be built using logistic regression. We will form an initial model using all the chosen features and check the significance of each regressor again using p-value. If p-value is small (<0.05), the regressor is considered significant, and vice versa. Using the

summary function in R, we can see that the p-value for PhysActivity is 0.0513 (3 s.f). PhysActivity is not very significant hence we will remove it and rebuild M3 without it.

M3 can be approximated as:

Log-odds = - 7.02 + 0.748 * I(HighBP=1) + 0.583 * I(HighChol=1)
+ 1.34 * I(CholCheck=1) + 0.163 * I(Stroke=1)
+ 0.301 * I(HeartDiseaseorAttack=1) + 0.0965 * I(DiffWalk=1)
- 0.737 * I(HvyAlcoholConsump=1) + 0.0763 * BMI
+ 0.592 * GenHlth - 0.00958 * PhysHlth + 0.153 *Age
- 0.0551 * Income

, where Log-odds represent the log of the odds of success. All coefficients are estimated to the nearest 3 significant figures (3 s.f.). To check the accuracy and AUC of the model, a 5-fold cross-validation is then used to compute the average accuracy and AUC.

Comparing Models

The following table gives the approximate summary of accuracy and AUC for each model (Figure 4).

Model	**Accuracy (3 s.f.)**	**AUC (3 s.f.)**
M1, Naïve Bayes	0.728	0.790
M2, Decision Tree	0.741	0.810
M3, Logistic Regression	0.747	0.824

Figure 4: Summary of Accuracy and AUC

Across all three models, the accuracy and AUC seem reasonably high, indicating a good fit of the models to the data. The combination of high accuracy and AUC suggests that the models effectively predict outcomes and discriminate between classes.

M3 has both the highest accuracy and highest AUC among the 3 models. In this case, we can easily conclude that M3 performs the best overall. It is most likely to make a prediction correctly and has the best discrimination ability among the other two models. Thus, we will choose M3 as the best model.

Final Model and Conclusion

A final analysis is performed on the chosen model, M3. We explore how its True Positive Rate (TPR) changes with False Positive Rate (FPR) at various threshold values. The threshold value determines the point at which a predicted probability is classified as positive (having diabetes) or negative (not having diabetes). The figure below shows the relation between TPR, FPR, and the threshold value (Figure 5).

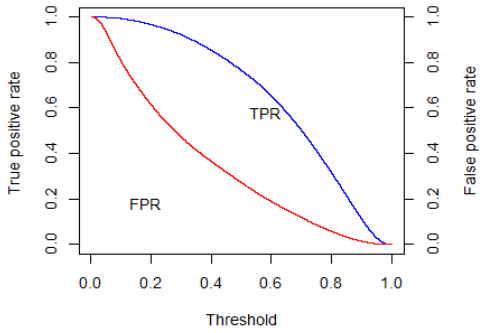

Figure 5: TPR, FPR and Threshold

The distance between the two curves shows us the difference between TPR and FPR. In context, TPR represents the probability that an individual who actually has diabetes is correctly predicted positive for diabetes. A high TPR is important so that we do not miss out on patients who require relevant medical treatment before it is too late. On the other hand, FPR represents the probability that an individual who does not have diabetes is incorrectly predicted to have diabetes. A low FPR is important so that individuals do not undergo unnecessary treatments which might be dangerous if they are non-diabetic.

In short, both TPR and FPR are equally crucial when predicting diabetes status and we should not compromise one for the other. Hence, the optimum threshold value should be one that maximizes TPR and minimizes FPR. In R, this can be done by comparing the TPR alongside FPR to find the threshold that gives the biggest difference between TPR and FPR (where TPR > FPR). After analysis, we find that the threshold value that will give the best fit, rounded to 3 s.f., is 0.470.

In conclusion, the logistic regression model, M3, with a threshold value of 0.470 is the best model for predicting the diabetes status of an individual.