DSA1101 Statistical Report

Sandrina Agnes Natalie A0276244B

1 Background

Diabetes is a chronic disease prevalent in modern society. Methods of detecting and predicting the disease have been developed through the contributions of many fields of science. In the field of data science, predicting diabetes can be done by making use of classifiers to construct a model. This model will then be used to predict the response variable of whether one has diabetes or not through the pattern of the input variables. The question arises in regard to which method is the best to predict diabetes status. This report aims to resolve that and propose the best classifier to predict diabetes status.

2 Aim

The aim of this statistical report is to:

1. Analyze different classification methods for predicting diabetes status in order to propose the best classifier.
2. Investigate and explore the goodness-of-fit for the proposed classifier.

3 Introduction of Data

The data set given, ”diabetes_5050.csv” is a clean data set of 70,692 survey responses from a survey conducted in the US in 2015. In this report, the response variable that will be predicted is the diabetes status of a person based on multiple input variables. In the data, Diabetes_binary represents the status of diabetes, with 0 being no diabetes status and 1 being prediabetes or diabetes status. To determine whether a person has diabetes or not, there are input variables to be considered in predicting the status. The input variables are blood pressure (HighBP), cholesterol (HighChol), cholesterol check (CholCheck), body mass index (BMI), smoking status (Smoker), stroke (Stroke), heart disease or attack (HeartDiseaseorAttack), physical activity (PhysActivity), fruits (Fruits), vegetables (Veggies), alcohol consumption (HvyAlcoholConsump), healthcare coverage (AnyHealthcare), need of doctor but hindered due to cost (NoDocbcCost), general health status (GenHlth), mental health (MentHlth), days with bad physical health (PhysHlth), difficulty walking or climbing stairs (DiffWalk), sex (Sex), 13 categories of age (Age), 6 scales of education (Education), 8 scales of income (Income).

Based on the correlation with respect to the response variable, the input variables are filtered into significant variables by using only the variables with a correlation above 0.1, which are HighBP, HighChol, CholCheck, BMI, Stroke, HeartDiseaseorAttack, PhysActivity, GenHlth, PhysHlth, DiffWalk, Age, Education, and Income. In this case, the input variables with a correlation value above 0.1 are significant to determine the response. This is further proven by the p-value of less than 0.001, except for the variable Stroke with a p-value less than 0.01.

The binary variables are Diabetes_binary, HighBP, HighChol, CholCheck, Stroke, HeartDiseaseorAttack, PhysActivity, and DiffWalk, while the categorical variables are BMI, GenHlth, PhysHlth, Age, Education, and Income.

For HighBP, 0 indicates no high blood pressure and 1 indicates high blood pressure. For HighChol, 0 indicates no high cholesterol and 1 indicates high cholesterol. For CholCheck, 0 indicates no cholesterol checks in 5 years and 1 indicates there is a cholesterol check in 5 years. BMI indicates body mass index ranging between 12 to 98 (inclusive). For Stroke, 0 indicates no, and 1 indicates yes. For PhysActivity, 0 indicates no physical activity for the past 30 days, not including jobs and

1 indicates there is. GenHlth indicates general health status ranging from 1 to 5, with 1 indicating excellent to 5 indicating poor. PhysHlth indicates the number of days during the past 30 days that the physical health is not good. For DiffWalk, 0 indicates no difficulty in walking or climbing the stairs while 1 indicates difficulty in walking or climbing the stairs. Age has 13 categories, with 1 being age 18 to 24, then 9 being range 60 to 64, and 13 being age 80 or above. Education has 6 categories with 1 indicating never attended school, 2 indicating elementary, and so on. Income has 8 categories, 1 indicates less than 10k, 5 indicates less than 35k, and 8 indicates 75k or more.

The binary input variables' relation with the response variable is represented with contingency tables and the categorical input variables' relation with the response variable is represented with boxplots to visualize the spread of data. The columns represent the diabetes status.

HighBP	0	1
0	22118	8742
1	13228	26604

HighChol	0	1
0	21869	11660
1	13477	23686

CholCheck	0	1
0	1508	241
1	33838	35105

Stroke	0	1
0	34219	32078
1	1127	3268

HeartDiseaseorAttack	0	1
0	32775	27468
1	2571	7878

PhysActivity	0	1
0	7934	13059
1	27412	22287

DiffWalk	0	1
0	30601	22225
1	4745	13121

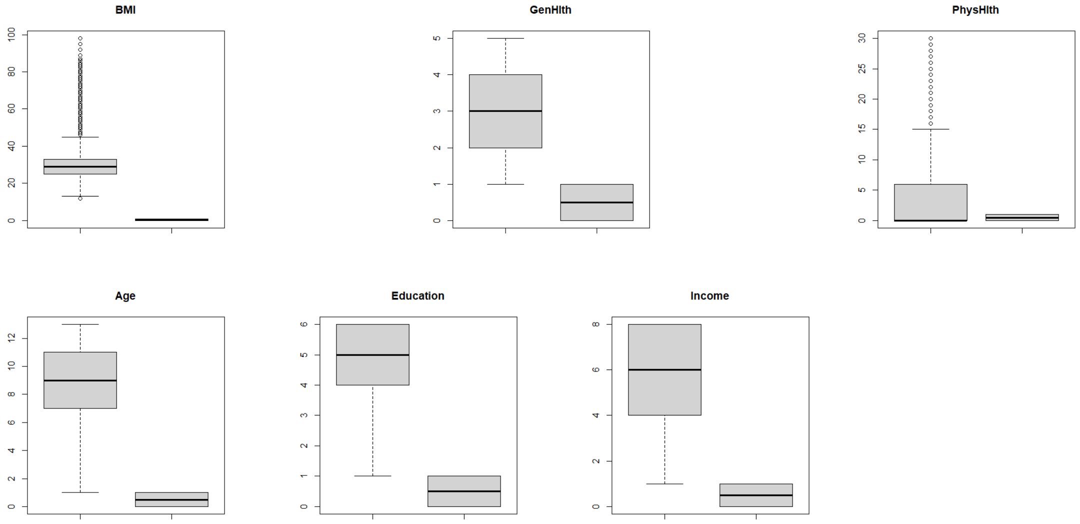

Figure 1: The illustrations for the comparison of the input variables and the response variable

Right away, the relation between the input variables and the response variable can be seen where most of the positive indication or the higher value of the input variables corresponds with the positive indication of the response variable.

Before processing the data with classifiers to form the model, the data "diabetes_5050.csv" will be separated into two parts, train data and test data, with a ratio of 80:20. The train data will then have 56554 observations, while the test data will have 14138 observations. The data separation will be done by ensuring both parts have the same proportion of positive and negative responses. The data separation is done by using the createDataPartition() function. The input variables of BMI, PhysHlth, and Age will be standardized.

4 Statistical Procedures

This data has a response variable, hence to categorize the data, the supervised learning method is the best way to do so. In the course DSA1101, there are five classifiers for supervised learning, which are linear regression, KNN (K-Nearest Neighbor), decision tree, naïve Bayes, and logistic regression.

To predict diabetes status based on the input variables, a few models could be constructed from the five classifiers of supervised learning. Each of these classifiers will be trained with the train data and tested with the test data which have been separated from the “diabetes_5050.csv” data.

4.1 Linear Regression

Linear regression is a statistical modeling technique used to examine the relationship between a dependent variable and one or more independent variables. The method assumes a linear association, aiming to find the best-fitting line that represents the overall trend in the data. The model estimates coefficients for each independent variable, allowing for the prediction of the dependent variable. Although linear regression is more suitable for predicting continuous response variables, this report will still test the classifier and compare it to another classifier by categorizing the values of the linear regression predictors based on the chosen threshold.

From the significance of the linear regression model’s coefficients, the variables can be further filtered to only the significant variables. The significant variables are HighBP, HighChol, CholCheck, BMI, Stroke, HeartDiseaseorAttack, GenHlth, PhysHlth, DiffWalk, Age, Education, and Income.

The formula is as follows:

$$\hat{y} = 0.1523 \cdot \text{HighBP} + 0.1044 \cdot \text{HighChol} + 0.1632 \cdot \text{CholCheck} + 0.0870 \cdot \text{BMI} + 0.0216 \cdot \text{Stroke}$$

$$+0.0612 \cdot \text{HeartDiseaseorAttack} + 0.1057 \cdot \text{GenHlth} - 0.0180 \cdot \text{PhysHlth} + 0.0283 \cdot \text{DiffWalk} + 0.0742 \cdot \text{Age}$$

$$-0.0075 \cdot \text{Education} - 0.0097 \cdot \text{Income}$$

4.2 KNN (k-Nearest Neighbors)

The k-Nearest Neighbors (KNN) classifier is a simple and effective machine learning algorithm for classification tasks. It assigns a class to a new data point based on the majority class among its k nearest neighbors in the feature space. KNN is non-parametric, requires no training phase, and is sensitive to the choice of the parameter k, which determines the number of neighbors considered. This is the reason why the formula will be iterated for different cases of k to find the best accuracy amongst the different values of k. While KNN is straightforward, its performance can be influenced by the curse of dimensionality in high-dimensional spaces. This is the reason why data with big variations must be standardized to increase its robustness.

From iterating the k values, we can conclude that the best value of k is 74 with an accuracy of 0.7513793. This k will be used for the model to predict the outcomes of the test data response.

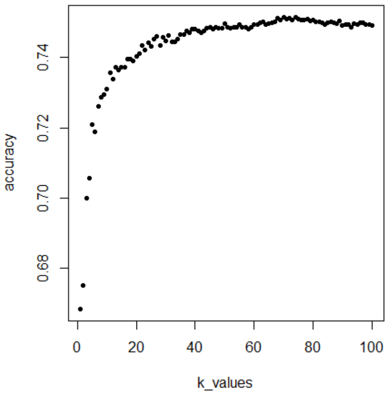

Figure 2: The plot for accuracies along the different values of k from 1 to 100

The confusion matrix for the test response that is compared to the prediction of the KNN is as follows:

KNN Confusion Matrix	0	1
0	4988	1427
1	2081	5642

4.3 Decision Tree

A decision tree is a versatile and interpretable machine-learning algorithm used for both classification and regression tasks. It recursively splits the dataset into subsets based on the most informative features, creating a tree-like structure of decisions. Each internal node represents a decision based on a feature, and each leaf node represents the predicted outcome. Decision trees are intuitive, allowing for visual interpretation of decision-making processes. They are sensitive to the choice of splitting criteria and depth, impacting model complexity. Decision trees are employed in various fields for their transparency, ease of use, and ability to capture complex relationships in data.

The model will be computed by the method class, minsplit 1, and parms list split information. The decision tree based on the model is as follows:

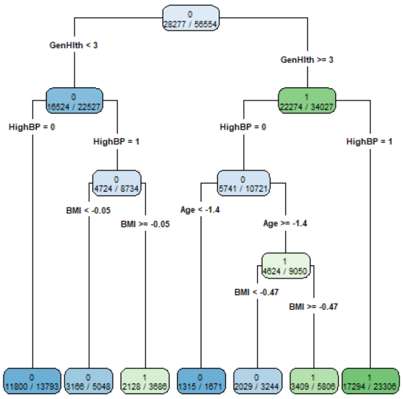

Figure 3: The decision tree to predict the diabetes status

Based on the decision tree, the input variables that are the most significant and is used to predict the response are GenHlth, HighBP, BMI, and Age.

The confusion matrix for the test response that is compared to the prediction of the decision tree is as follows:

DT Confusion Matrix	0	1
0	4623	1422
1	2446	5647

4.4 Naïve Bayes

Naïve Bayes is a probabilistic classification algorithm based on Bayes' theorem. It assumes that the features used to describe instances are conditionally independent, given the class label. Despite its "naïve" independence assumption, Naive Bayes has proven effective in practice, particularly for text classification tasks. It calculates the probability of each class for a given set of features and assigns the class with the highest probability to the instance. Naive Bayes is computationally efficient and requires a small amount of training data. While its assumptions may not always hold in reality, it remains a popular and accessible choice for classification problems, especially in the context of high-dimensional and sparse data.

The confusion matrix for the test response that is compared to the prediction of the naïve Bayes is as follows:

nB Confusion Matrix	0	1
0	5275	2258
1	1794	4811

4.5 Logistic Regression

Logistic Regression is a statistical method for binary classification that models the probability of an instance belonging to a particular class. It utilizes the logistic function to constrain the output between 0 and 1, representing the probability. The model estimates coefficients for each predictor, indicating their impact on the log odds of the outcome. Logistic Regression is widely used due to its simplicity, interpretability, and effectiveness in predicting binary outcomes. It is suitable for scenarios where the relationship between predictors and the log odds of the outcome is assumed to be linear. The model's performance is often assessed using metrics like accuracy, precision, recall, and the area under the Receiver Operating Characteristic (ROC) curve.

From the significance of the logistic regression model's coefficients, the variables can be further filtered to only the significant variables. The significant variables are HighBP, HighChol, CholCheck, BMI, Stroke, HeartDiseaseorAttack, GenHlth, PhysHlth, DiffWalk, Age, Education, and Income.

The formula is as follows:

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -3.274 + 0.7107 \cdot \text{HighBP} + 0.5557 \cdot \text{HighChol} + 1.3356 \cdot \text{CholCheck} + 0.5501 \cdot \text{BMI}$$
$$+0.1399 \cdot \text{Stroke} + 0.3268 \cdot \text{HeartDiseaseorAttack} + 0.5830 \cdot \text{GenHlth} - 0.0910 \cdot \text{PhysHlth}$$
$$+0.1140 \cdot \text{DiffWalk} + 0.4508 \cdot \text{Age} - 0.0414 \cdot \text{Education} - 0.0519 \cdot \text{Income}$$

4.6 Threshold Values for Linear Regression and Logistic Regression

The TPR (True Positive Rate) and FPR (False Positive Rate) comparison for the threshold values of the linear regression and logistic regression is as follows:

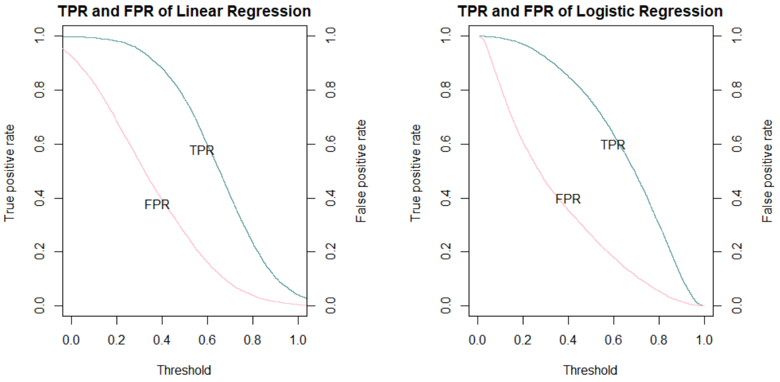

Figure 4: The TPR and FPR comparison for each threshold values of Linear Regression and Logistic Regression

The best value proposed for the threshold of the linear regression and logistic regression is both 0.5, as it has a significant difference between the TPR and FPR.

5 Result

5.1 Best Classifier

Using ROC (receiver operating characteristic curve), which is a graph showing the performance of a classification model at all classification thresholds, we can compare the models with the best fit to predict the data. The comparison is made by obtaining the AUC (Area Under the Curve) for each of the classifiers' ROC curves, which is also the accuracy of the model. Based on the calculation, logistic regression has the best accuracy with a value of 0.8264. Linear regression's accuracy came next closely with a value of 0.8258. As for the rest, KNN, decision tree, and naive Bayes have values of 0.7519, 0.7264, and 0.7134, respectively.

In a general case, accuracy is the determining factor in choosing the best model, and in turn, classifier. As there are no required parameters to choose the best classifier, accuracy will be the main factor to determine the best classifier. In this case, this report will also consider the suitability of the classifier with respect to the data. The response variable of diabetes status is a categorical binary response, hence, although linear regression has a similar accuracy as the logistic regression, it is not the best fit to classify the data. Based on the accuracy and suitability, the best classifier for the diabetes status data is the logistic regression classifier.

The ROC curve for the classifiers is as follows:

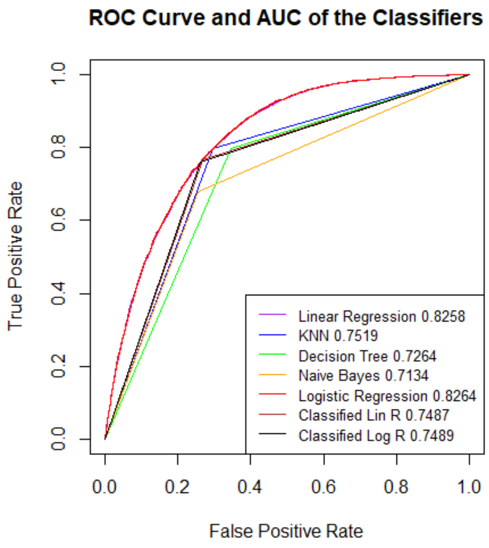

Figure 5: The ROC curve of the classifiers along with the AUC/accuracy of each classifier

5.2 Advantage and Disadvantages of Logistic Regression

The advantage of logistic regression are that it provides easily interpretable results and probabilities that allow intuitive interpretation of the likelihood of an event occurring and it is less prone to overfitting compared to more complex models. Logistic regression is suitable for binary outcomes. On the other side, the disadvantages of logistic regression are that it assumes a linear relationship between predictors and the log odds of the outcome, which may not always hold in practice, it may not capture complex relationships in the data as effectively as more flexible models like decision trees or neural networks, it is sensitive to outliers, highly influenced by sample size, not suitable for non-linear relationships, and it may struggle with missing data.

6 Conclusion

In the report, five different kinds of supervised learning classifiers have been tested to predict the response for diabetes status with the provided input variables. The process is done by splitting the data into two parts for train and test with a ratio of 80:20, respectively. The models are all trained with the train set and are tested by predicting the input variables in the test set and then compared to the response variable of the test set. The classifiers were then compared based on the ROC curve and AUC, which is also the accuracy for the classifiers. Out of all the supervised learning classifiers learned in the course DSA1101, the best fit for predicting diabetes status based on the high level of AUC/accuracy and suitability is logistic regression.

While logistic regression is a valuable and interpretable tool for binary classification, it has limitations, especially when facing complex, non-linear relationships or when dealing with small datasets. In this case, logistic regression can work due to the minimum effect of the data as this diabetes status data has a linear relationship and is considerably large data. To predict diabetes status, the conventional method is to do a medical checkup. As the accuracy of the logistic regression model is quite high, 0.8264, it can be used as the initial step to predict the diabetes status for a quick prediction. This would be useful for large-scale diabetes status prediction. In the situation of widespread diabetes cases, timely implementation of prevention and solutions can be made due to immediate awareness of the prediction result.

References

ttps://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset