STATISTICAL REPORT

Assignment 2 for DSA1101: Introduction to Data Science

Prepared by: Brian Bong Neng Ye, A0276127B

***Abstract* – This report aims to select an optimal classifier model to predict diabetes status based on the dataset titled "diabetes-dataset.csv". The dataset is explored with the help of visualization libraries. We fit four classifier models, including Logistic Regression, Naïve Bayes, Decision Tree and kNN. Using recall as a primary metric, this report finds the Logistic Regression classifier to be the most optimal model due to its high performance and interpretability.**

PART ONE: DATASET EXPLORATION

The dataset analysed in this assignment, titled "diabetes-dataset.csv", records the status of people with and without diabetes. The dataset, prepared by Mohammed Mustafa, is of format .csv (comma-separated values) with header included [1]. The purpose of this report is to explore the dataset, fit classifier models on the dataset, and thus select the best classifier model that best identifies people with diabetes given their status.

The dataset is first explored with R, utilizing "ggplot2" library as a visualization tool. The dataset consists of nine variables in total, including one categorical response variable "diabetes" and eight explanatory variables, including four categorical variables and four continuous variables. These variables are concluded in **Table 1** as follows.

TABLE 1

DESCRIPTION OF VARIABLES IN DATASET

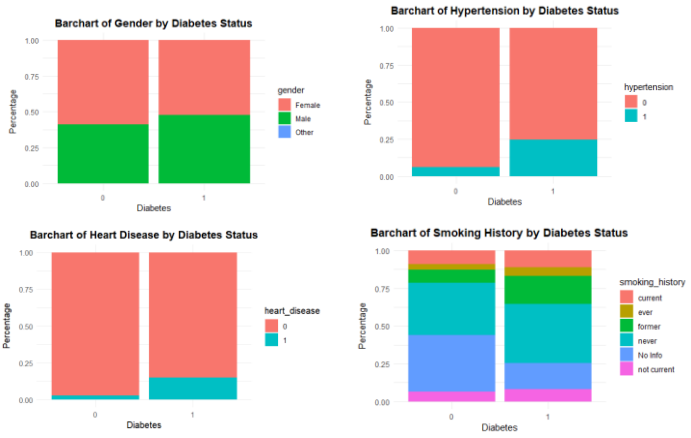
Variables	Description	Factors | Range
diabetes	Categorical Response	**1**: positive, **0:** negative
gender	Categorical	**"Male", "Female"**
age	Continuous	Range: $\mathbf{0.08 \leq x \leq 80}$
hypertension	Categorical	**1**: positive, **0:** negative
heart_disease	Categorical	**1**: positive, **0:** negative
bmi	Continuous	Range: $\mathbf{10.0 \leq x \leq 95.7}$
hbA1c_level	Continuous	Range: $\mathbf{3.5 \leq x \leq 9.0}$
blood_glucose_level	Continuous	Range: $\mathbf{80 \leq x \leq 300}$
smoking_history	Categorical	**"current", "ever", "former", "never", "not current", "No Info"**

The dataset consists of 100,000 clean samples (complete with no missing records). For the response variable "diabetes", we assume positive class for observations of "1" and negative class for observations of "0". A preliminary count reveals 8,500 samples of positive class samples, which is 8.5% of the total samples. This indicates a highly imbalanced dataset, which would be taken into account during the selection of metrics to evaluate the models fitted on the dataset.

Exploratory analysis is performed separately on continuous variables and categorical variables.

I. Categorical Variables

Categorical variables are visualized with proportional stacked bar charts grouped by the response variable "diabetes". The bar charts of the four categorical variables are represented in **Figure 1-4.**



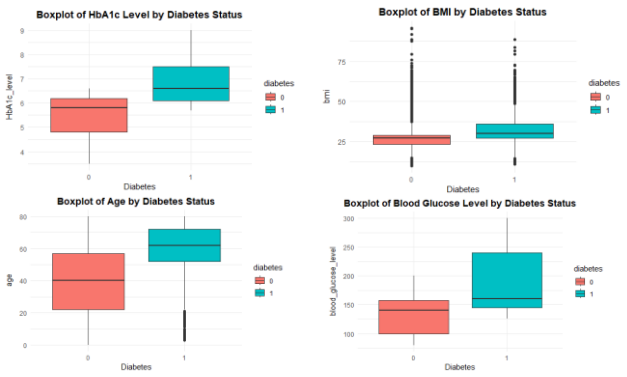
[Clockwise from top left] ***Figure 1****: Bar chart of gender by diabetes status;* ***Figure 2****: Bar chart of hypertension by diabetes status;* ***Figure 3****: Bar chart of heart disease by diabetes status;* ***Figure 4:*** *Bar chart of smoking history by diabetes status. Please generate via attached R file for figures of clearer resolution.*

According to the bar chart, interpretable trends are observed for the four categorical variables according to diabetes status. Specifically, people with diabetes are observed to be proportionally more likely to (1) be of male gender; (2) have hypertension; (3) have heart diseases than people without diabetes.[1]

For **Figure 4**, there is no significant trends due to multiple factors present; however, some inferences can still be made, particularly there are more current smokers and former smokers in the diabetes group and in the non-diabetes group. Less samples in the non-diabetes group have disclosed their smoking status than the diabetes group, thus the higher number of sample counts with "No Info". No observable differences can be noted for non-smokers and not current smokers.

II. Continuous Variables

Continuous variables are visualized with boxplots grouped by the response variable "diabetes". The boxplots of the four continuous variables are represented in **Figure 5-8.**



[Clockwise from top left] ***Figure 5:*** *Boxplot of HbA1c level by diabetes status;* ***Figure 6:*** *Boxplot of BMI by diabetes status;* ***Figure 7:*** *Boxplot of blood glucose status by diabetes status;* ***Figure 8:*** *Boxplot of age by diabetes status. Please generate via attached R file for figures of clearer resolution.*

There are observable trends for the four boxplots here, specifically people with diabetes are observed to be of (1) higher HbA1c level, with no significant outliers; (2) slightly higher BMI, although both classes have significantly intersecting higher and lower outlier; (3) older age, although there is significant amount of lower outliers for positive class; (4) higher blood sugar level.

PART TWO: METHODOLOGY, RESULTS AND DISCUSSIONS

I. Preprocessing and Metric Selection

The dataset is split by random in training and testing datasets by ratio 8:2. The same training set and testing set would be used to evaluate the models.

For each model, we gathered multiple metrics for evaluation, including ROC-AUC score, accuracy, recall and precision. However, since our data is highly imbalanced, overrepresentation of negative classes may affect the usefulness of ROC-AUC score and accuracy scores to capture the ability of the model to identify positive classes [2]. Thus, recall and precision are better metrics for representations of imbalanced datasets, specifically recall in our case of diabetes classification as we would like to minimize false negatives (people who have diabetes identified as non-diabetes).

To evaluate the models better, we have introduced a new metric, F1-score, defined as follows.

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

This metric combines both recall and precision and is hence more representative of the performance of models in identifying classes while minimizing errors, especially in imbalanced datasets.

II. Logistic Regression

Before training our logistic regression model we performed hyperparameter tuning on the value of the threshold δ. We iterate through the values of the threshold, and plotted the metrics of the performance of the models trained with each of the δ value, as shown in **Figure 9**. The optimal

[1] Note that this only compares the proportion between two groups (diabetes and non-diabetes) and does not imply the proportions within each group or the absolute count of the occurrences.

threshold for F1-score, 0.59, is used to train the model. The model is tested on the test dataset.

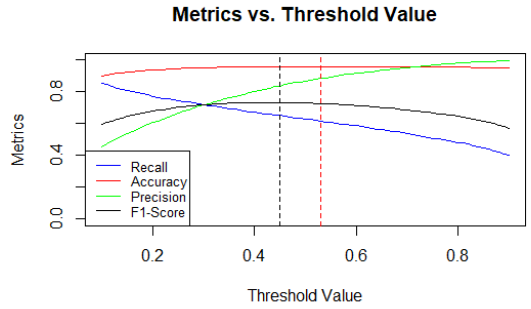


Figure 9: *The metrics of the model in different threshold values. The optimal thresholds to maximize F1-score and accuracy is marked on black and red vertical lines, respectively.*

The metrics and ROC-AUC curve of the model trained are presented in **Table 2** and **Figure 10. Figure 11** shows a summary of the regression model. The summary largely fits our expectations as outlined in Part One.

TABLE 2

METRICS FOR LOGISTIC REGRESSION

Accuracy	Precision	Recall	F1-Score	ROC-AUC
0.9591	0.8294	**0.6485**	**0.7279**	0.9625

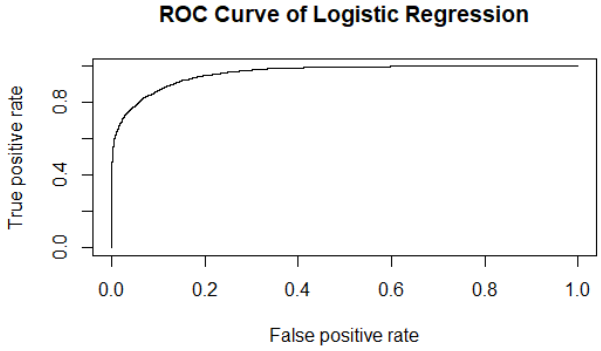


Figure 10: *ROC-AUC curve for the logistic regression model.*

```
Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -27.037684   0.325708 -83.012  < 2e-16 ***
genderMale                    0.302660   0.040352   7.500 6.36e-14 ***
genderOther                  -9.551811 105.538580  -0.091   0.9279
age                           0.046584   0.001258  37.041  < 2e-16 ***
hypertension1                 0.725065   0.052446  13.825  < 2e-16 ***
heart_disease1                0.719422   0.067873  10.600  < 2e-16 ***
smoking_historyever          -0.007592   0.102469  -0.074   0.9409
smoking_historyformer        -0.108327   0.078561  -1.379   0.1679
smoking_historynever         -0.155613   0.068127  -2.284   0.0224 *
smoking_historyNo Info       -0.748426   0.074794 -10.006  < 2e-16 ***
smoking_historynot current   -0.153042   0.092806  -1.649   0.0991 .
bmi                           0.088256   0.002852  30.940  < 2e-16 ***
HbA1c_level                   2.336978   0.039743  58.803  < 2e-16 ***
blood_glucose_level           0.033198   0.000537  61.818  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 11: *A summary of the parameters of the logistic regression model.*

The metrics indicate a relatively well performance of the model, with high performance in identifying negative classes and relatively lower performance in identifying positive classes (about 65% of the positive classes are classified correctly). Logistic regression models are also interpretable by observing threshold parameters for every variable.

III. Naïve Bayes

As there is not a suitable hyperparameter for tuning in Naïve Bayes models, the model is directly trained and tested on our datasets. The metrics and ROC-AUC curve of the model trained are presented in **Table 3** and **Figure 12.**

TABLE 3

METRICS FOR LOGISTIC REGRESSION

Accuracy	Precision	Recall	F1-Score	ROC-AUC
0.9543	0.7718	**0.6497**	**0.7055**	0.9507

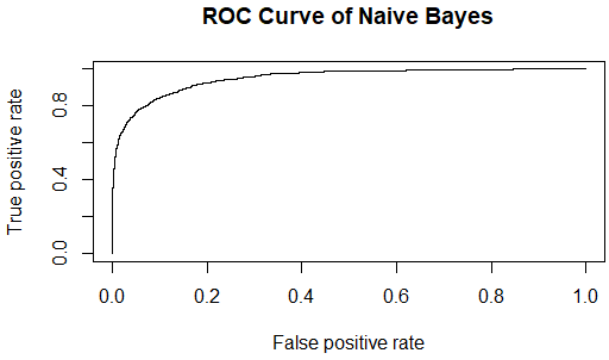


Figure 12: *ROC-AUC curve for the Naïve Bayes model.*

The performance of the Naïve Bayes model is comparable to the Logistic Regression, with a minute improvement in recall score and decrease in the rest of the metric scores. However, Naïve Bayes model suffers from some fundamental limitations due to its lack of hyperparameters for performance tuning and the lack of interpretability, which is important in diabetes prediction such that medical practitioners can diagnose the disease based on interpretable and observable factors. As such, in similar performance, Logistic Regression model is preferred over Naïve Bayes model.

IV. Decision Trees

Hyperparameter tuning of cp (complexity parameter) for the decision tree model is done, with the default value of 0.1 identified as the best value for the hyperparameter. An increase in cp value to 1 does not change the tree, and an increase above 1 reduces the tree to a node. Conversely, a decrease in cp value until 0.0001 does not change the tree, and any value below 0.0001, and any value below that produces complex trees that are at risk of overfitting. Experimenting of other parameters like max_depth has also concluded that default parameters are the most optimal.

The model is trained and tested based on default values of hyperparameters. The metrics and ROC-AUC curve of the model trained are presented in **Table 4** and **Figure 13. Figure 14** shows a plot of the decision tree model implemented.

TABLE 4
METRICS FOR DECISION TREE

Accuracy	Precision	Recall	F1-Score	ROC-AUC
0.9723	1.0000	**0.6710**	**0.8031**	0.8355

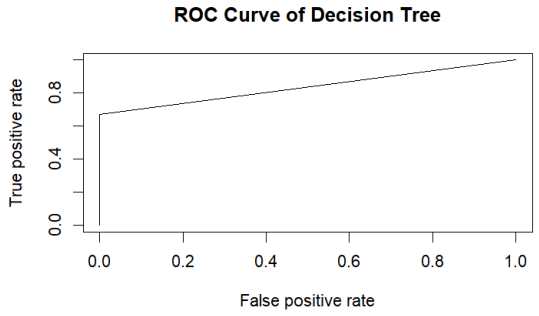


Figure 13: *ROC-AUC curve for the decision tree model.*

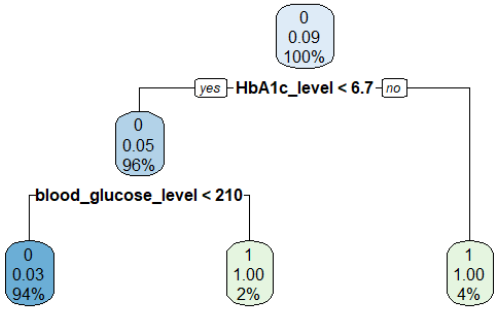


Figure 14: *A plot of the decision tree model.*

There are some interesting insights here. First, notice that the precision score is 1.000, indicating no false positives identified (all non-patients are identified correctly). Besides, the decision tree is relatively simple, employing only two variables to differentiate between positive and negative classes. Observing these two variables in **Figure 5** and **Figure 7** reveals a relatively clear threshold between the two groups. The precision and recall score of the model is the highest of all the models, despite lower performance in the rest of the metric.

The decision tree model provides an elegant and simple procedure to identify possible diabetes patients, with a significantly high recall score. However, there remains drawbacks in this model. The model does not use other variables in the data, neglecting other possible variables that may significantly affect the result. Furthermore, as demonstrated by the hyperparameter tuning, the entropy loss/Gini score calculation for the model makes it less flexible to changes in the dataset.

V. kNN Algorithm

Categorical variables are dropped, leaving four continuous variables. The variables are standardised by columns for better representation.

To select the optimal k value for the dataset, the elbow method is used. Due to the high demand of computational power and time for the algorithm, only certain values of k are chosen to gauge the performance of the model with different values of k. The points are plotted as follows in **Figure 15.**

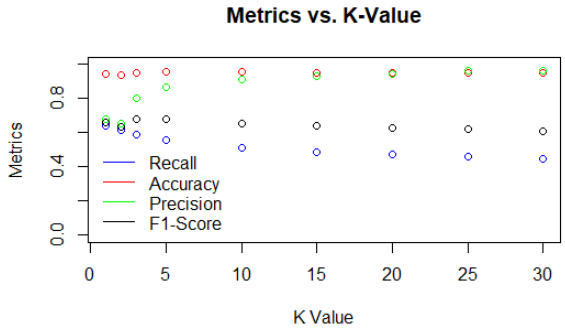


Figure 15: *A plot of the metrics associated with the kNN model trained on different k-values.*

As the k-value increases, there is a significant increase in precision and a slight decrease in recall, indicating an increase in negative class classifications. This is as predicted in imbalanced data, where higher values of k favour the majority class over the minority [3]. Using the elbow method, the metrics stabilise when k-value is approximately 15, thus this value is used to train the data.

The model is trained with a k-value of 15 and tested on the dataset. As kNN lacks a suitable method for the generation of ROC-AUC curves and ROC score, we omit this metric for the model. The other metrics are as follow in **Table 5**.

TABLE 5
METRICS FOR KNN

Accuracy	Precision	Recall	F1-Score
0.954	0.9333	**0.4896**	**0.6424**

As expected, the model does relatively poorly for minority classes (positive class in our case), which is in agreement with **Figure 15**. Thus, our model performance shows that kNN is prone to predicting majority classes in imbalanced data, therefore it may perform less well in our case where positive classes are underrepresented in the dataset. Furthermore, the model also cannot take use of categorical variables, which may be equally significant in predicting diabetes status.

PART THREE: CONCLUSION

Taking the four models and their performance into account, this report comes to the conclusion that the Logistic Regression model is the most optimal for classification tasks for our dataset, given its good performance in terms of recall and F1-score, besides being highly interpretable with different weights of parameters observable for the users.

While the Naïve Bayes model performed relatively well, it suffers from a limited degree of interpretability, which is important in our case as prediction of the diseases should ideally be done based on an observable sets of criteria, which may be difficult to achieve with the Naïve Bayes model.

The decision tree model outperforms the Logistic Regression model in terms of recall, indicating a better ability to correctly identify diabetes patients. However, the model is observed to be overly simple, which is ideal for interpretation but does not include many other variables in the dataset. Thus, the model may neglect changes in other variables that may affect the chances of one's cancer status prediction.

The kNN model reports significantly lower performance. This can be due to its intrinsic susceptibility of predicting majority classes, which may be problematic in highly imbalanced data like in our case. Besides, the model also relies only on continuous variables, thus it may disregards categorical variables that may be playing an important factor in cancer status prediction.

Thus, we conclude that the Logistic Regression performs best in our dataset. However, there are some intrinsic drawbacks to the method, namely its assumption of a linear relationship between factors and diabetes status, which may not be the true case. Hence, future directions can potentially focus on a modified version of the model to represent the dataset more accurately.

REFERENCES

1. Mustafa, M. (n.d.) Diabetes prediction dataset. Retrieved April 10, 2024 from https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset/data
2. Bekkar, Mohamed, et al. "Evaluation Measures for Models Assessment over Imbalanced Data Sets." *Journal of Information Engineering and Applications* , Vol. 3, No. 10, 1 Jan. 2013, pp. 27–38.
3. Beckmann, Marcelo, et al. "A KNN Undersampling Approach for Data Balancing." *Journal of Intelligent Learning Systems and Applications*, Vol. 07, No. 04, 2015, pp. 104–116. https://doi.org/10.4236/jilsa.2015.74010