# Diabetes Statistical Report

## Section I: Introduction

Diabetes stands as one of the most widespread chronic conditions. Mohammed Mustafa curated a collection of 100,000 survey responses on diabetes, incorporating eight features which are age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, and blood glucose level. The data is compiled in the file "diabetes-dataset.csv".

Description on the categorical features:

- Gender contains three categories: male, female, and other (LGBT)
- Hypertension contains two categories: the person does not have hypertension (0) and the person has hypertension (1)
- Heart disease contains two categories: the person does not suffer from heart disease (0) and the person suffers from heart disease (1)
- Smoking history contains *current* (the person is currently smoking); *ever* (the person smokes sometimes but not often); *former* (the person formerly smoked but has completely quit); *never* (the person has never smoked and will never smoke); *not current* (the person has never smoked but is uncertain if they will smoke); *no info* (no information from the person)

The objective of the report is to propose the best classifier to predict diabetes status. Therefore, the diabetes status in the dataset is the response variable, with the eight features as the input variables.

## Section II: Statistical Procedures Used

Before classification: Association between each input variable and response variable

The strength of association is assessed by using the contingency table of proportion, to identify the potential features for the classifier.

Input variable 1: Gender

Based on the contingency table of proportion by gender, 7.6% of females have diabetes. Among males, around 9.7% of males have diabetes. Lastly, none of the other genders has diabetes. Thus, there is a rather clear association between gender and diabetes; males have a slightly higher chance of getting diabetes compared to females, while other genders are not likely to get diabetes. However, further investigation on the association is warranted.

Input variable 2: Age

Age is divided into three groups based on Our World in Data[1]: young (<15 years old), middle (15-64 years old), and old (>=65 years old). Based on the contingency table of proportion by age, 0.3% of the young category have diabetes, 6.9% of the middle category have diabetes, and

[1] https://ourworldindata.org/age-structure

20.5% of the old category have diabetes. There is a strong and positive association between age and diabetes; the older the person is, the higher chance the person gets diabetes.

Input variable 3: Hypertension

The contingency table of proportion by hypertension reveals that among the people with no hypertension, 6.9% have diabetes while among the people with hypertension, 27.9% have diabetes. There is a strong and positive association between hypertension and diabetes; diabetes is more likely to those who have hypertension.

Input variable 4: Heart disease

The contingency table of proportion by heart disease reveals that among the people with no heart disease, 7.5% have diabetes and among those with heart disease, 32.1% have diabetes. There is a strong and positive association between heart disease and diabetes; diabetes is more susceptible to those who suffer from heart disease.

Input variable 5: Smoking history

Based on the contingency table of proportion by smoking history, it is obtained:

- Among those who are currently smoking, 10.2% have diabetes.
- Among those who smoke sometimes, 11.8% have diabetes.
- Among those who have quit smoking, 17% have diabetes.
- Among those who have not smoked and will never smoke, 9.5% have diabetes.
- Among those who have never smoked before but unknown in the future, 10.7% have diabetes.
- Among those who are unknown for their smoking history, 4% have diabetes.

There is a somewhat strong and positive association between smoking history and diabetes; people who have smoked tend to have a higher chance of diabetes.

Input variable 6: Body mass index

BMI is categorised into four groups: underweight (<18.5), healthy weight (18.5-24.9), overweight (25-30), and obese (>30) according to Britannica[2]. The contingency table of proportions by BMI displays the following diabetes rates:

- Among those who are underweight, 0.75% have diabetes.
- Among those who are of healthy weight, 3.9% have diabetes.
- Among those who are overweight, 7.3% have diabetes.
- Among those who are obese, 18% have diabetes.

There is a strong and positive association between BMI and diabetes; the larger the BMI, the more likely to get diabetes.

[2] https://www.britannica.com/science/body-mass-index

Input variable 7: HbA1c level

HbA1c levels are categorised into three groups as per the National Library of Medicine[3]: low (<5.7%), medium (5.7%-6.49%), and high (>=6.5%). Based on the contingency table of proportion by HbA1c level, it is obtained: none of those with low HbA1c level has diabetes; 8% of those with medium HbA1c level have diabetes; 24.96% of those with high HbA1c level have diabetes. There is a strong and positive association between HbA1c level and diabetes; the higher the HbA1c level, the more likely to get diabetes.

Input variable 8: Blood glucose level

Blood glucose levels are grouped into three categories based on Diabetes Singapore[4]: low (<140 mg/dL), medium (140-199 mg/dL), and high (>=200 mg/dL). Based on the contingency table of proportion by blood glucose level, it is obtained: 3% of the low category people have diabetes; 7.1% of the medium category have diabetes; 36% of the high category have diabetes. There is a strong and positive association between blood glucose level and diabetes; the higher the blood glucose level, the higher chance to get diabetes.

Conclusion: All the input variables manifest a seemingly strong association with the diabetes status.

Classification

The dataset is first allocated into a training set and testing set with the ratio of 8:2 respectively. The study employs Naïve Bayes, Decision Trees, and Logistic Regression classifiers, each getting probabilities rather than class labels to fetch the threshold (for converting probabilities into class labels). Model assessment is based on ROC curve, AUC value, and False Negative Rate (FNR). FNR, indicating the percentage of positives classified as negatives, is prioritised to minimise the risk of overlooking disease cases for early detection and treatment. A lower FNR signifies a fitter classifier.

Classifier 1: Naïve Bayes

Assessments of the goodness-of-fit of the classifier:

1. ROC curve

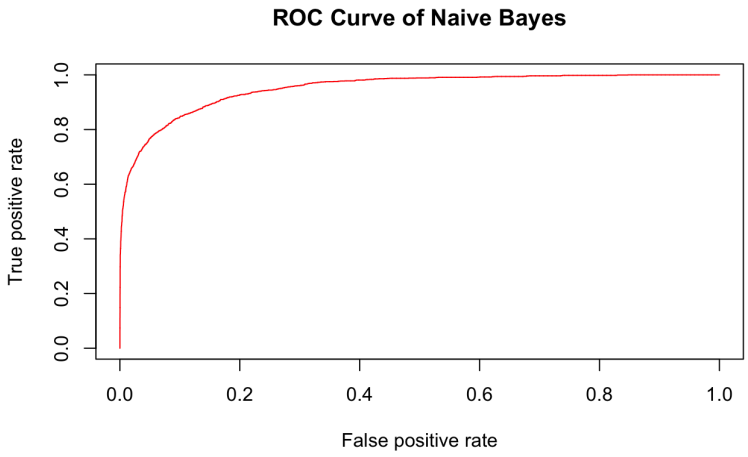


Figure 1. The ROC curve of the Naïve Bayes classifier.

There is a sharp rise in true positive rate (TPR) until around 0.6 in point 0.0 on the x-axis, suggesting that if one wants 0 false

[3] https://www.ncbi.nlm.nih.gov/books/NBK549816/
[4] https://www.diabetes.org.sg/resource/managing-diabetes/

positive rate (FPR), the highest TPR they can get is around 0.6. Subsequently, there is a more gradual rise in TPR from 0.1 to 0.3 on the x-axis, followed by a plateau. The optimal threshold for the classifier might be around a TPR of 0.8 and an FPR of 0.1.

2. AUC value: 0.9506937
   This value is considered very high and close to 1. The closer the AUC value to 1, the better the classifier's performance is.
3. FNR: 0.1351512 (when threshold = 0.0823)
   This rate is considered low, suggesting that the model performs great in accurately identifying positives without misclassifying them as negatives.

Classifier 2: Decision Trees

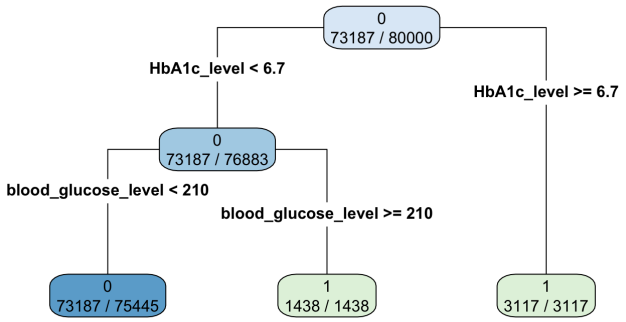


Figure 2. The output or decision tree.

Only two features, HbA1c and blood glucose level, are included which means they are the primary indicators. HbA1c acts as the root node, contributing significantly to predictions. The key outcome (73187/80000) shows that out of the people in the training data, 73187 of them have no diabetes. Besides, people who have HbA1c level >= 6.7 and people who have blood glucose level >=210 are definitely diabetic.

1. ROC curve

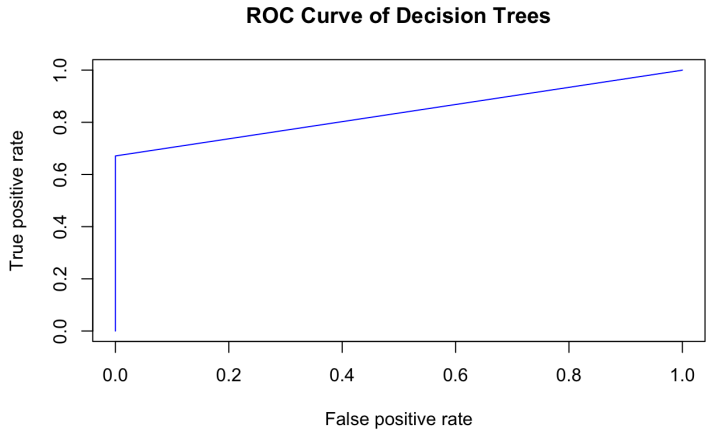


Figure 3. The ROC curve of Decision Trees.

The figure is less flexible or more rigid compared to other models. At first, there is a sharp ascent until approximately 0.7 in TPR at 0.0 FPR, showing that at 0 FPR, the highest TPR that can be obtained is around 0.7. Following this, there is a sudden dramatic increase in TPR, followed by a linear growth. The optimal threshold for the classifier might be around a TPR of 0.7 and an FPR of 0 or 0.1.

2. AUC value: 0.8355068
   This value is relatively high but it would be better if it is within the range of 0.9.
3. FNR: 1
   This is the maximum value of FNR; therefore, Decision Trees classifier is not preferred for prediction in this dataset. The maximum FNR can occur due to a high threshold value, which is 1.

```
      dt.threshold dt.fpr dt.tpr
[1,]          Inf      0  0.000
[2,]       1.0000      0  0.671
[3,]       0.0299      1  1.000
```

Figure 4. The thresholds for Decision Trees classifier; 0.0299 and 1.

Threshold value of 1 is selected because it shows minimum value of FPR and intermediate (although relatively low) value of TPR compared to others.

Alternatively, the author gets the class labels when predicting, instead of probabilities. An FNR of 0.3289864 is obtained. This value is still relatively high, it would be better to have a lower rate.

Classifier 3: Logistic Regression

```
Call:
glm(formula = diabetes ~ ., family = binomial(link = "logit"),
    data = train.data)

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   0.002853  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 5. The summary of the logistic model.

The p-values for gender (other) and smoking history (ever, former, never, not current) are large, suggesting that these categories do not contribute significantly to the model when predicting the response.

1. ROC curve

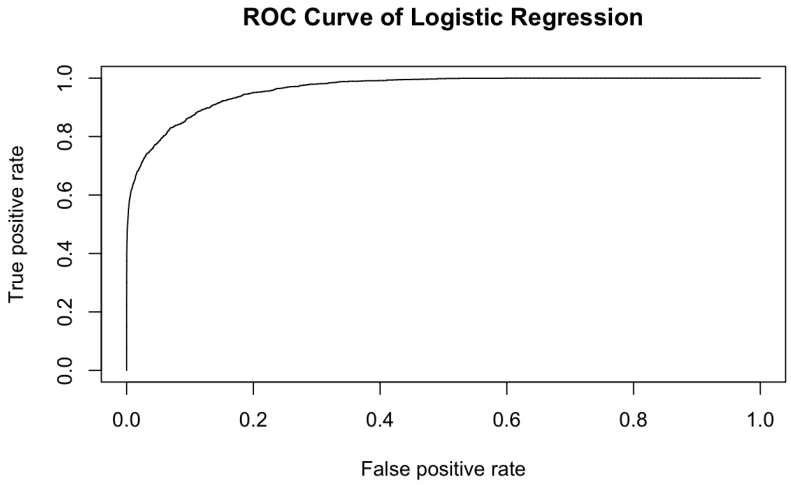


Figure 6. The ROC curve of Logistic Regression model.

Initially, there is a sharp incline in TPR until approximately

0.6, indicating that at 0 FPR, the highest TPR they can get is around 0.6. Afterwards, the increase in TPR becomes more gradual, spanning from points 0.1 to 0.2 on the x-axis, before reaching a plateau. The plateau suggests that as TPR approaches its maximum value of 1, the FPR steadily rises from approximately 0.3 to 1. The best threshold may be somewhere with 0.9 TPR and 0.1 FPR.

2. AUC value: 0.962497
   This value is very high and very close to 1, suggesting that the model performs great in classifying.
3. FNR: 0.1108477 (when threshold = 0.0823)
   This value is low enough, implying that fewer positive instances are being incorrectly classified as negative by the model.

## Section III: Summary of Statistical Findings

Considering the ROC curve, highest AUC value, and lowest FNR, the best classifier is the Logistic Regression classifier. It has the highest AUC value of 0.96 and lowest FNR of 0.1 compared to the other classifiers. In spite of that, the best classifier depends on the assessments and priorities the researcher concentrates on.

| Classifier | **Pros** | **Cons** |
|---|---|---|
| **Naïve Bayes** | Computationally efficient, especially for large datasets, due to its simplicity and independence assumption.<br>Easy and simple to implement.<br>Is robust to irrelevant features due to feature independence assumption. | Assumes feature independence, which may not hold true in real-world datasets.<br>Ignores feature dependency, potentially degrading model prediction performance.<br>The author assumes in determining the best threshold (select the threshold that is roughly similar to the percentage of people with diabetes). |
| **Decision Trees** | Produces models that are easy to interpret and visualise, which are useful for decision-making.<br>Performs automatic feature selection by selecting the most informative features. | N-fold cross-validation (CV) is needed to determine the optimal value of arguments like minsplit and cp for model fitness. However, the author uses default values instead of conducting N-fold CV.<br>In this dataset, it has the lowest AUC value and highest FNR. |
| **Logistic Regression** | In this dataset, it has the highest AUC value and lowest FNR.<br>Does not require installing packages (time-efficient). | The author makes assumptions on determining the best threshold.<br>Assumes linearity between the dependent variable and independent variables. |

Figure 7. Evaluation of all the classifiers used.