Eng Cheng Xin matriculation no.: A0288099E

Assignment 2: Statistical Report

Part I: Exploring the data set

Association between response and input variables

The data "diabetes.csv" consists of 9 variables, namely "gender", "age", "hypertension", "heart_disease", "smoking_history", "bmi", "HbA1c_level", "blood_glucose_level" and "diabetes". Among these variables include categorical variables "gender", "hypertension", "heart disease" and "smoking_history". It also contains numerical variables such as "age", "bmi", "HbA1c_level", and "blood_glucose_level". The response variable is "diabetes", while the other variables are input variables. The response variable "diabetes" showed association with all input variables.

Association between diabetes and categorical variables within the data

1. **Gender and diabetes**

   By calculating the rate of diabetes given each gender, we obtain rate(diabetes|male)=0.076 and rate (diabetes|female)=0.097. The difference between both rates are relatively small, hence diabetes and gender is said to have a relatively weak association.

2. **Hypertension and diabetes**

   By calculating the rate of diabetes given the hypertension condition of a person, we obtain rate(diabetes|hypertension)=0.279 and rate (diabetes|no hypertension)=0.069. The difference between both rates are relatively significant, hence diabetes and hypertension is said to have a relatively strong association.

3. **Heart disease and diabetes**

   By calculating the rate of diabetes given the heart disease condition of a person, we obtain rate(diabetes|heart disease)=0.321 and rate (diabetes|no heart disease)=0.075. The difference between both rates are relatively significant, hence diabetes and heart disease is said to have a relatively strong association.

4. **Smoking history and diabetes**

   By calculating the rate of diabetes given the smoking history of a person, we obtain the following rates:

condition	rate
Rate(diabetes|current)	0.102
Rate(diabetes|ever)	0.118
Rate(diabetes|former)	0.170
Rate(diabetes|never)	0.095
Rate(diabetes|No Info)	0.041

Rate(diabetes|not current)	0.107

Excluding the 'No Info' feature, the rates of diabetes in other features remain unchanged. As each feature shows different rates of people with diabetes, we can say that smoking history is associated with diabetes. The difference in rates of the variables are relatively small, hence it is said to have a relatively weak association.

Association between diabetes and numerical variables within the data

1. **BMI and diabetes**

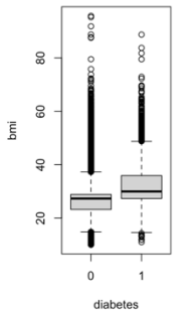


Diagram 1

Diagram 1 displays boxplots of BMI against diabetes. The positions of the medians within each boxplot and their interquartile range (IQR) are relatively similar. Both boxplots exhibit a large number of outliers, with the left boxplot containing relatively more outliers. BMI and diabetes have a correlation coefficient of approximately 0.214. Coupled with the mostly overlapping boxplots, this suggests that the association between BMI and diabetes is relatively weak.

2. **Age and diabetes**

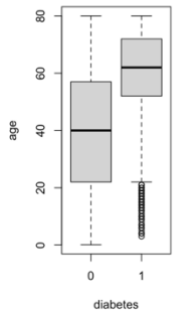


Diagram 2

Diagram 2 shows boxplots of age against diabetes. The positions of the medians within each boxplot are quite different. The left boxplot corresponding to individuals with diabetes exhibits larger interquartile range (IQR) compared to the right boxplot that corresponds to individuals without diabetes. The right boxplot has relatively more outliers. The correlation coefficient of age and diabetes is 0.258. Coupled with the mostly overlapping boxplots, this suggests that the association between BMI and diabetes is relatively weak.

3. HbA1c level and diabetes

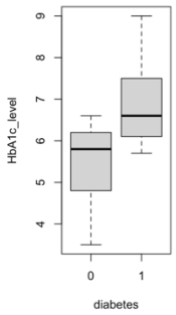


Diagram 3

Diagram 3 shows boxplots of HbA1c level against diabetes. The positions of the medians within each boxplot are moderately different. Both boxplots exhibit similar length of IQR and no outliers. The correlation coefficient of age and diabetes is 0.401. Coupled with the moderately overlapping boxplots, this suggests that the association between BMI and diabetes is moderately strong.

4. Blood glucose level and diabetes

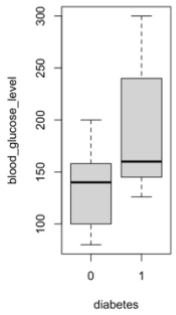


Diagram 4

Diagram 4 shows boxplots of blood glucose level against diabetes. The positions of the medians within each boxplot are moderately different. The right boxplot corresponding to individuals with diabetes The correlation coefficient of age and diabetes is 0.420. Coupled

with the moderately overlapping boxplots, this suggests that the association between BMI and diabetes is moderately strong.

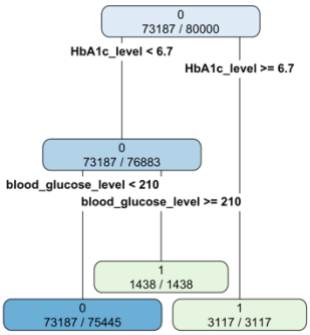
Part II: Building Model/Classifier and Conclusion

Methodology

Three classifiers were proposed for this set of data, which are decision tree, Naïve Bayes, and logistic regression classification methods. These models were trained on 80% of the dataset and evaluated on the remaining 20%. The model was trained to predict the presence or absence of diabetes based on features such as age, gender, body mass index (BMI), blood glucose level, smoking history, HbA1c level, and hypertension. The ROC curve was generated by varying the classification threshold and calculating sensitivity and specificity for each threshold. Furthermore, the accuracy of the model was manually calculated.

Results

1. **Decision Tree Model**



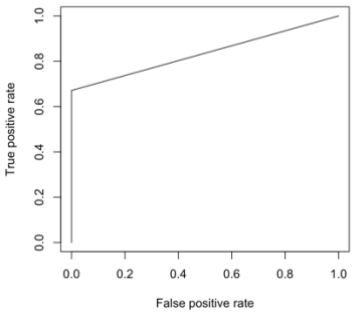


Diagram 5 Diagram 6

A decision tree (Diagram 5) was plotted using the training data. The decision tree model achieved an area under the ROC curve (AUC) of 0.836, indicating good discrimination ability between individuals with and without diabetes. The TPR and FPR of the model at the optimal threshold were approximately 0.68 and 0, respectively. The ROC curve (Diagram 6) demonstrates the trade-off between TPR and FPR across different threshold values. Furthermore, the accuracy of the model is calculated to be 0.972.

2. Naïve Bayes Model

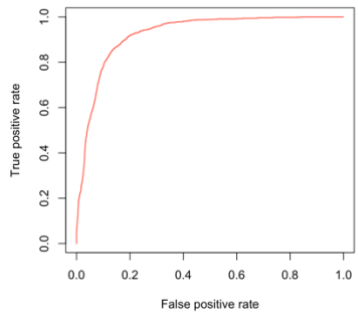


Diagram 7

The Naïve Bayes model achieved an area under the ROC curve (AUC) of 0.924, indicating good discrimination ability between individuals with and without diabetes. The TPR and FPR of the model at the optimal threshold were approximately 0.88 and 0.10, respectively. The ROC curve (Diagram 7) demonstrates the trade-off between TPR and FPR across different threshold values. Furthermore, the accuracy of the model is calculated to be 0.905.

3. Logistic Regression Model

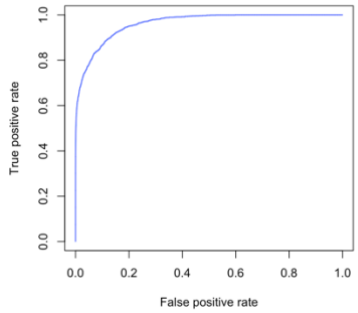


Diagram 8

The Logistic Regression model achieved an area under the ROC curve (AUC) of 0.962, indicating good discrimination ability between individuals with and without diabetes. The TPR and FPR of the model at the optimal threshold were approximately 0.90 and 0.12, respectively. The ROC curve (Diagram 8) demonstrates the trade-off between TPR and FPR across different threshold values. Furthermore, the accuracy of the model is calculated to be 0.960.

Comparison of Models

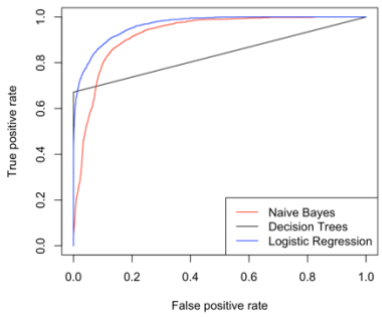


Diagram 9

Diagram 9 shows a comparison between the 3 models. The table below summarises the AUC values and accuracy of each model.

Model	AUC	Accuracy
Decision Tree	0.836	0.972
Naïve Bayes	0.924	0.905
Logistic Regression	0.962	0.960

The pros and cons of each model is as follows: Decision trees offer simplicity and interpretability, allowing for easy visualization of decision boundaries and feature importance. However, they are prone to overfitting and can be sensitive to small variations in the training data. Naive Bayes models, on the other hand, are computationally efficient and handle high-dimensional data well, but they make the strong assumption of feature independence, which may limit their predictive accuracy. Logistic regression models provide interpretable coefficients and output probabilities, making them suitable for binary classification tasks where class probabilities are needed. However, they assume a linear relationship between features and the log-odds of the target variable, which may not hold true in all cases.

Conclusion

In conclusion, the logistic regression model shows promising performance in predicting the risk of diabetes using demographic and clinical variables. With the highest AUC value of 0.962 and a relatively high accuracy of 0.960, it outperforms other models in terms of predictive power.