DSA1101: Introduction to Data Science

Assignment 2: Statistical Report

I. INTRODUCTION

Diabetes is one of the most prevalent chronic diseases all over the world. In this statistical report, I am going to propose some suitable classification methods for predicting diabetes status and then choose the best one with a good goodness-of-fit.

The models use the data set from diabetes-dataset.csv, which is a clean data set of 100,000 survey responses, provided by the author Mohammed Mustafa.

II. METHODS

1. Summary of the response variable and different input variables

1.1. The response variable: "diabetes"

- The response variable "diabetes" is a categorical variable represented by the numeric values "0" = No and "1" = Yes.
- With over 100,000 observations in the data, there are 91,500 people without diabetes and 8,500 people with diabetes.

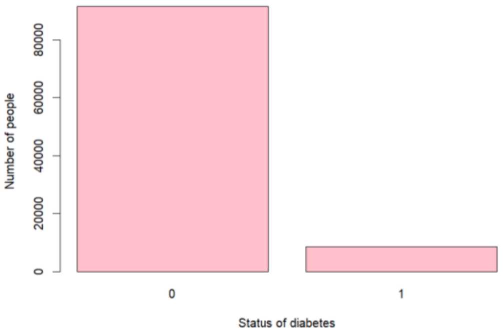


Figure 1. Bar plot of diabetes status

1.2. 8 input variables

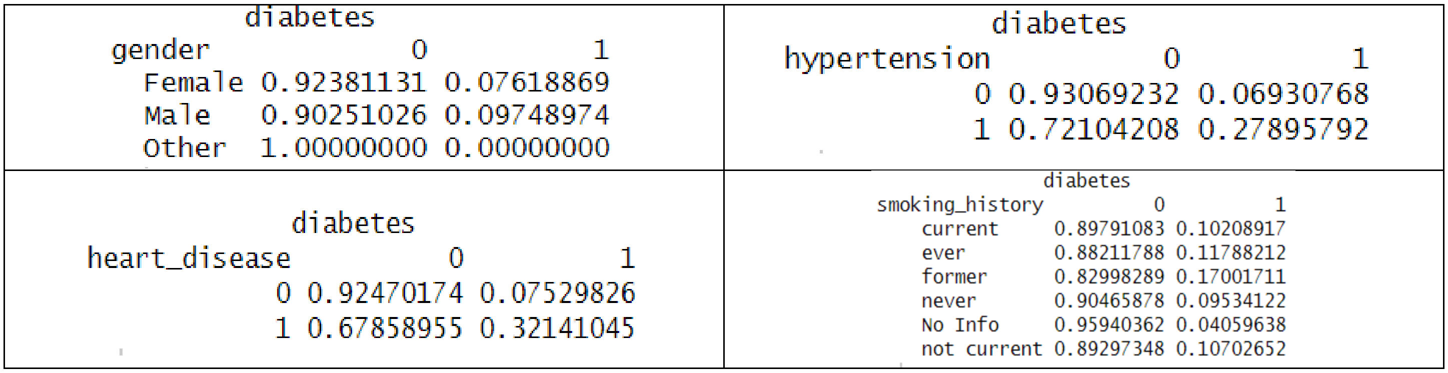
Categorical variables		Quantitative variables	
gender	3 categories: "Female," "Male," "Other." The number of females = 1.4 times the number of males.	**age**	Age of people in the data set. The smallest age is 0.08, the oldest age is 80 while the mean is 42.
hypertension	2 categories: "0" = No and "1" = Yes. Around 7.5% of people in the data set have hypertension.	**bmi**	BMI index of people in the data set. The smallest BMI is 10.01, the highest is 95.69 while the mean is 27.32.
heart disease	2 categories: "0" = No and "1" = Yes. Around 3.9% of people	**HbA1c level**	HbA1c level of people in the data set. The smallest HbA1c

	in the data set have heart disease.		level is 3.5, the highest is 9 while the mean is 5.528.
smoking history	6 categories: Around 35% of people in the data set have never smoked before while around 9.3% of them are smoking currently.	**blood glucose level**	Blood glucose level of people in the data set. The smallest blood glucose level is 80, the highest is 300 while the mean is 140.

Table 2. Summary of input variables

2. Association between the response variable and each input variable

2.1. Association between the response variable and categorical variables



```
          diabetes
gender            0          1
  Female 0.92381131 0.07618869
  Male   0.90251026 0.09748974
  Other  1.00000000 0.00000000
```

```
            diabetes
hypertension          0          1
           0 0.93069232 0.06930768
           1 0.72104208 0.27895792
```

```
             diabetes
heart_disease          0          1
            0 0.92470174 0.07529826
            1 0.67858955 0.32141045
```

```
               diabetes
smoking_history          0          1
    current     0.89791083 0.10208917
    ever        0.88211788 0.11788212
    former      0.82998289 0.17001711
    never       0.90465878 0.09534122
    No Info     0.95940362 0.04059638
    not current 0.89297348 0.10702652
```

Figure 3. Conditional probabilities of diabetes given each categorical variable

- For “gender”: There is an association between gender "Male" and "Female" with the response. Around 7.6% of females have diabetes while around 9.7% of males have diabetes. It seems that Male tends to have diabetes more than Female.
- For “hypertension”: There is quite a strong association between hypertension and the response. Among all people with hypertension, 27.9% of them have diabetes, while that of people without hypertension only stands at 6.9%. It seems that people with hypertension tend to have diabetes more than people without hypertension.
- For “heart_disease”: There is a strong association between heart disease and the response. Among all people with heart disease, around 32% of them have diabetes while that of people without diabetes is only 7.5%. It seems that people with heart disease tend to have diabetes more than people without heart disease.
- For “smoking_history”: There is a quite weak association between smoking history and the response. It seems that for almost all smoking history except "former" and "No Info", there are around 10% of people in each smoking history who have diabetes.

2.2. Association between the response variable and quantitative variables

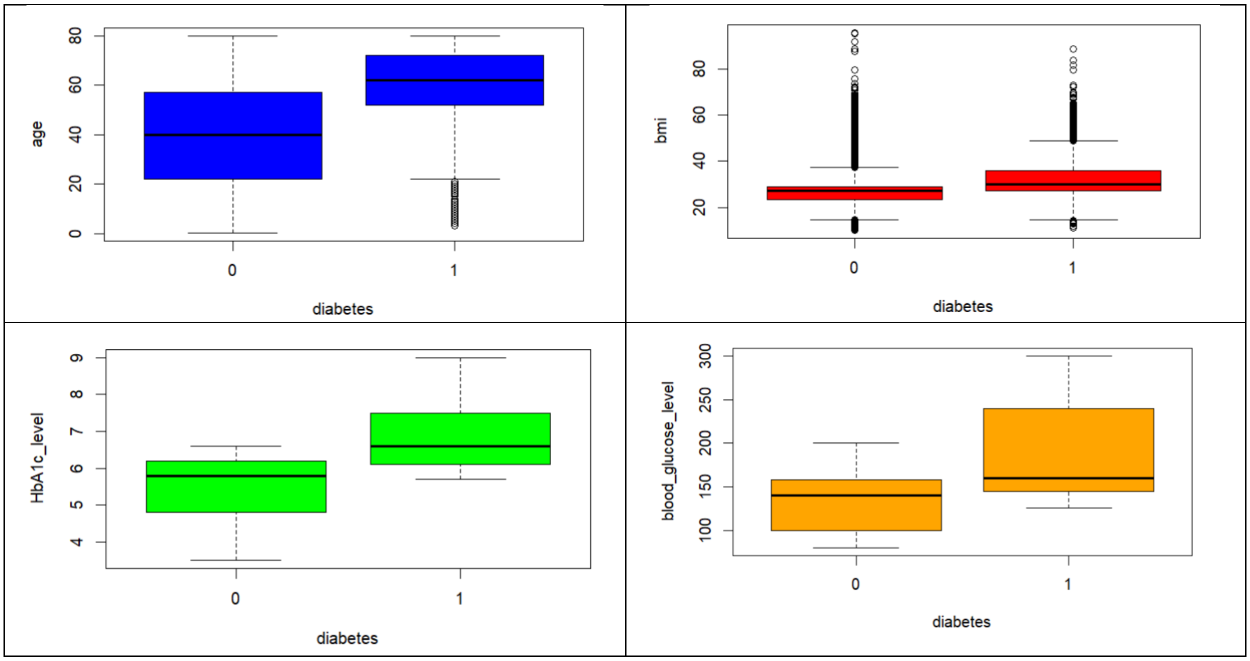


Figure 4: Box plots of each quantitative variable and the response variable

- For “age”: There is quite a strong association between age and the response. It seems that older people tend to have diabetes more than younger people. The median age of people without diabetes is around 40 years old, much younger than that of people with diabetes which is around 60 years old.
- For “bmi”: There is quite a strong association between BMI and the response. It seems that people with higher BMI tend to have diabetes more than people with lower BMI. While the median is approximately the same (at around 27), the range of the box plot of people having diabetes is larger and higher than that of people without diabetes.
- For “HbA1c_level”: There is a strong association between HbA1c level and the response. It seems that people with higher HbA1c level tend to have diabetes more than people with lower HbA1c level. The median HbA1c level of people with diabetes (around 6.7) is higher than that of people without diabetes (around 5.8).
- For “blood_glucose_level”: There is a strong association between blood glucose level and the response. It seems that people with higher blood glucose level tend to have diabetes more than people with lower blood glucose level. The median blood glucose level of people with diabetes is around 160, higher than that of people with diabetes which stands at around 140.

2.3. Summary

Overall, except for the input variable "smoking_history", all other input variables, including "gender," "age," "hypertension," heart_disease," "bmi," "HbA1c_level," "blood_glucose_level" have a quite strong association with the response variable "diabetes." Therefore, when forming classifier models, I will exclude the input variable "smoking_history" and include all other 7 input variables.

3. Building Models/Classifiers

3.1. Introduction

- Firstly, I divided the data set into training set and testing set with ratio 8:2, respectively. All the proposed models will have the same training set and testing set as other models. The training set is to build the model and determine the parameter for each model. Then, I will use the models built to test on the testing set and examine their goodness-of-fit.
- The metrics for goodness-of-fit I used are ROC and AUC, type 1 error, and type 2 error.
- We can see that the cost of predicting a person who has diabetes to be diagnosed as diabetes-free is extremely high because it will have a huge impact on the patient's health. Moreover, false negative can lead to a late diagnosis of diabetes, which can lead to catastrophic consequences.
- Therefore, when predicting whether a person has diabetes or not, type 2 error is very important. Type 2 error should be kept low while type 1 error can be tolerated.

3.2. Model 1: K-Nearest Neighbors (KNN)

For this model, I will only use 4 numeric input variables which are: “age,” “bmi,” “HbA1c_level,” “blood_glucose_level.”

Firstly, I have to standardize all the input features to make all the inputs contribute equally to the model. Then, I use 3-fold cross validation for the training data to find the best k. For k, I choose 9 values of k, ranging from k = 2 to k = 300.

After forming the model, based on the result, I would choose k = 2 which produces the lowest type 2 error and the highest AUC value.

Then, I form a new model with k = 2 to test on the testing set. At the end, I obtained the following metrics: Type 1 error: 0.02255229 = 2.26%, Type 2 error: 0.294013 = 29.4%, AUC value: 0.8417173.

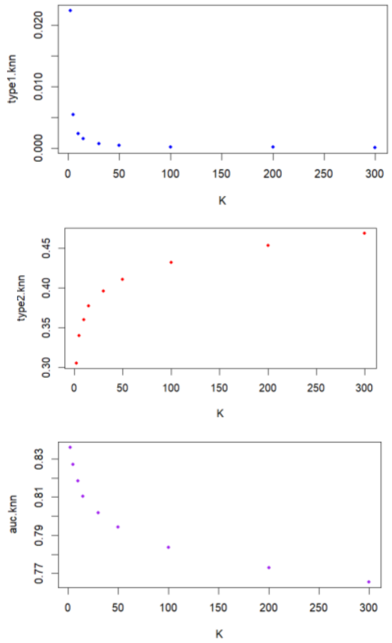


Figure 5. How K changes affect different metrics

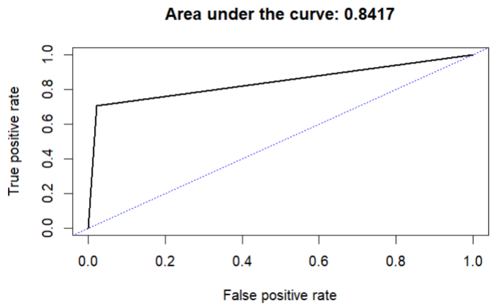


Figure 6. ROC of KNN model

3.3. Model 2: Decision Tree

For this model, I use 7 input variables (exclude “smoking_history”), including: “gender,” “age,” “hypertension,” heart_disease,” “bmi,” “HbA1c_level,” “blood_glucose_level.”

When forming the model, I choose parameter minsplit = 6. I chose this value because the data set is very large, hence choosing a small minsplit can over-capture the noisy data, hence it is not recommended to choose a small minsplit.

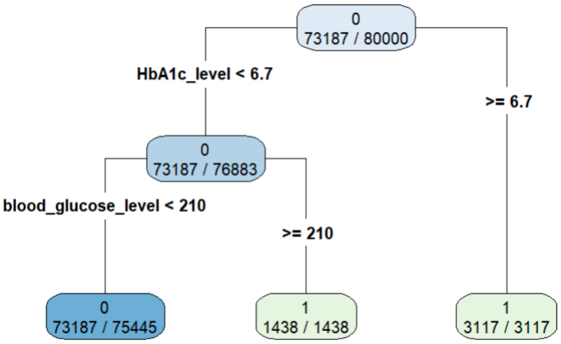


Figure 7. Decision Tree

After testing the model for the testing set, I obtained the following information: Type 1 error: 0, Type 2 error: 0.3289864 = 32.9%, AUC value: 0.8355068. As we can see, the type 1 error is very decent: 0%. The type 2 error is a bit high while the AUC value is not too high and not too low.

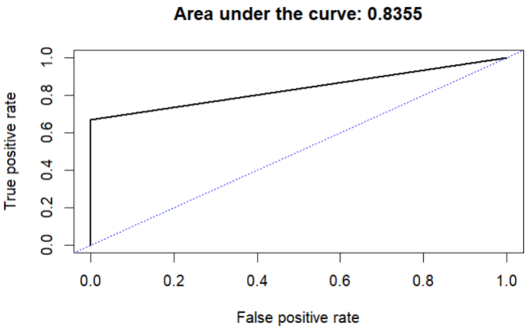


Figure 8. ROC of Decision Tree model

3.4. Model 3: Naïve Bayes

Using the same data set and methods as the Decision Tree model, I obtained these results for the Naïve Bayes model: Type 1 error: 0.01501665 = 1.5%, Type 2 error: 0.3615886 = 36.16%, AUC value: 0.9515408. As we can see, the type 2 error is high whereas the AUC value is high.

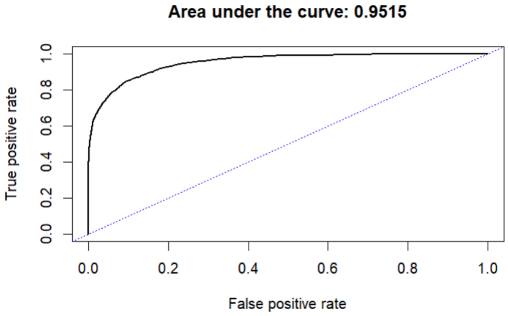


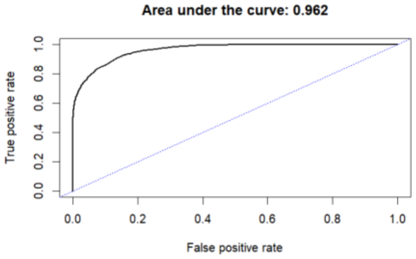
Figure 9. ROC of Naïve Bayes model

3.5. Model 4: Logistic Regression

For this model, I chose threshold = 0.1 after plotting how threshold changes will affect the TPR and FPR. This threshold balances both the TPR and the FPR.

After testing with the testing set, I obtained this information: Type 1 error: 0.102004 = 10.2%, Type 2 error: 0.1410788 = 14.11%, AUC value: 0.9619871. The type 2 error is the lowest among all 4 models while the type 1 error is quite high. The AUC value is also the highest among all models.

Figure 10. ROC of Logistic Regression model



3.6. Summary of goodness-of-fit of each model

	Type 1 error	Type 2 error	AUC value	Complexity
KNN	2.26%	29.4%	0.8417	Only use 4 input features. However, the code takes long time to run
Decision Tree	**0%**	32.9%	0.8355	Use 7 input features
Naïve Bayes	1.5%	36.16%	0.9515	Use 7 input features
Logistic Regression	10.2%	**14.11%**	**0.962**	Use 7 input features

Table 11. Goodness-of-fit of each model

Based on the goodness-of-fit, I choose the **Logistic Regression** model as the best classifier since it produces the lowest type 2 error and the highest AUC value among all models.

III. Summary

- Overall, the KNN model is the least complex among all models because it only uses 4 input variables. However, KNN code takes a long time to run, which is not recommended.
- The Decision Tree model has produced a decent type 1 error of 0%. However, its type 2 error, which is prioritized, is quite high.
- While the type 1 error and AUC of the Naïve Bayes model are quite good, it produced the highest type 2 error among all models.
- Finally, with the Logistic Regression model, the type 2 error is the lowest among all models and the AUC is also the highest among all models. Therefore, I choose **Logistic Regression** as my final model for predicting diabetes status, given its outstanding goodness-of-fit among all models.