6/24/2022

Machine Learning Classification Project

R and Statistical Analysis

Dr.Shadman

Mina Kanaani
Ferdowsi University of Mashhad

List of Contents

List of Figures

List of Tables

Machine Learning : Classification Project

Mina Kanaani

2022-06-24

Abstract

Classification Project

This is the Report on the first R and Statistical Analysis's project on Classification and machine learning.

In this report I aim to predict the chance of occurring Stroke on hospital patients based on some important features such as their age, heart disease, hypertension(high blood pressure) ,glucose level they have, etc.

Based on their characteristics, I used two methods : None Parametric methods such as KNN and Regression Tree and Parametric methods such as Logistic Regression and Discrimination models.

At the end of each method, I tested some important criteria based on problem's characteristic to examine the prediction to determine whether they showed us good results or not.

Why Stroke?

This report is an analysis of a data set of hospital patients for the purpose of assigning risk factors to strokes and making recommendations to help prevent this event. Strokes are the second leading cause of death in the world according to the World Health Organization and are responsible for 11% of total deaths. Every year 15 million people suffer a stroke, 5 million of which pass away and another 5 million of which are left with a permanent disability (WHO, 2021). Many of these strokes are preventable through healthy habit forming and monitoring those at the highest risk can have a significant improvement in outcomes. For these reasons, it is an area that requires further study to prevent this event from impacting more lives than necessary. The following report focuses on determining risk factors for strokes and makes recommendations on how to prevent them.

Flowchart

Using R Codes with the help of "DiagrammeR" library, I aimed to create a Flowchart to illustrate the procedure of this Project in a better way. The figure 1 below displays the project flowchart.

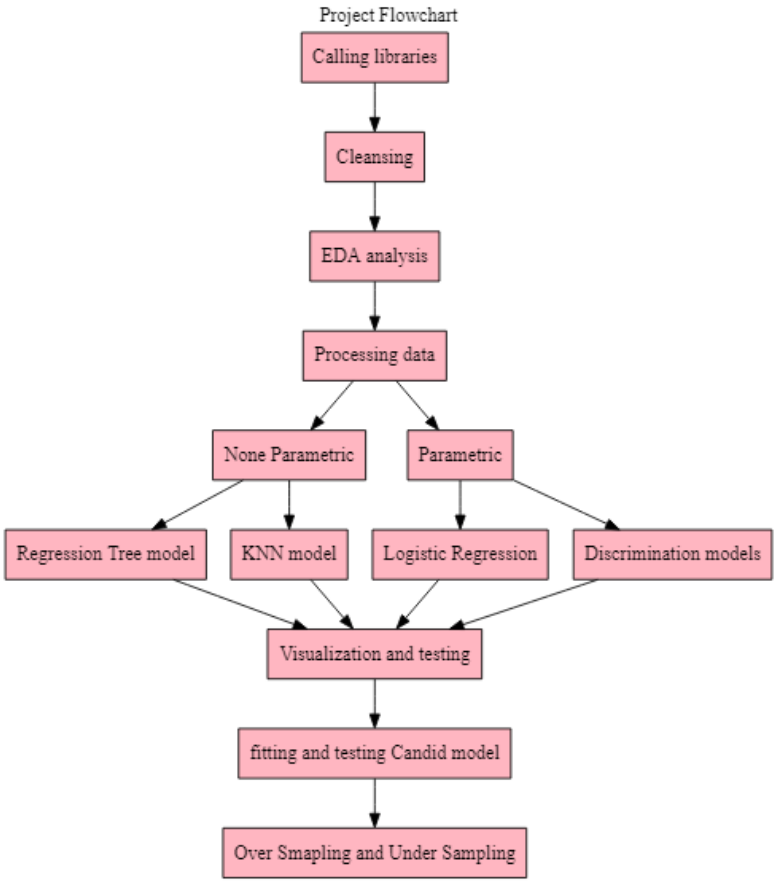


Figure 1-Flowchart

Recalling Libraries and Data

In this Part, we can take a short look at the data we gained from Kaggle.com, doing so, we read the excel file attached to this file to be able to access the data. Table 1 below shows the first five record of data.

Id	Gender	Age	Hypertension	Heart_Disease	Ever_Married	Work_Type	Residence_Type	Avg_Glucose_Level	Bmi	Smoking_Status	Stroke
9046	Male	67	0	1	Yes	Private	Urban	228.69	36.6	Formerly Smoked	1
51676	Female	61	0	0	Yes	Self-Employed	Rural	202.21	N/A	Never Smoked	1
31112	Male	80	0	1	Yes	Private	Rural	105.92	32.5	Never Smoked	1
60182	Female	49	0	0	Yes	Private	Urban	171.23	34.4	Smokes	1
1665	Female	79	1	0	Yes	Self-Employed	Rural	174.12	24	Never Smoked	1

Table 1-First look at data

Cleansing

In this part I aimed to clean the data from discrepancies and nulls. I divided the cleaning part into 3 parts.

First part is cleaning the BMI features:

BMI

Since there are 201 NA records in BMI features, I replace them with mean of BMI column in order to access to a more integrated dataset. Table 2 displays the BMI summary before and after the cleansing.

Before:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
10.30	23.50	28.10	28.89	33.10	97.60	201

After:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
10.30	23.80	28.40	28.89	32.80	97.60

Table 2-Summary before/after the cleansing

part 2 of cleaning is for Gender feature:

Gender

In this part, we have one record that is neither Female or male, since the majority of Data is Female, we add that one to Female category. Table 3 shows the Frequency of each before and after the cleansing.

Female	Male	Other
2994	2115	1

Var1	Freq
Female	2995
Male	2115

Table 3-Frequency of Female/Male

part 3 and last part of cleaning is for Smoking status.

Smoking Status

Since there is an "unknown" category, we put the data in that column in other categories based on their probability. Calculate the probability of formerly smoker, current smokers and non-smokers given that there's only this three categories in the smoking_status column. (Further calculation in attached R code), Table 4 illustrates the new data after cleansing.

id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	stroke	smoking_status
9046	Male	67	0	1	Yes	Private	Urban	228.69	36.60000	1	formerly smoked
51676	Female	61	0	0	Yes	Self-employed	Rural	202.21	28.89324	1	never smoked
31112	Male	80	0	1	Yes	Private	Rural	105.92	32.50000	1	never smoked

Table 4-New data after cleansing

EDA Visualization

In this part, we use EDA analysis tools to visualize our features and their binary correlation with charges to determine which features have significant effect and which are not of much importance.

Bar Chart

The first tool we use is Bar charts to illustrates the relationship between Quality features such as stroke, heart disease, sex, residence, smoke, work type and hypertension.

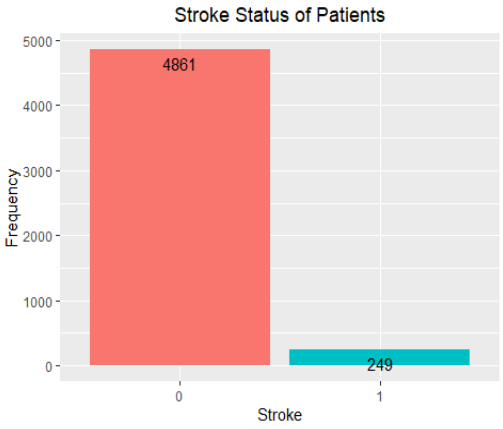


Figure 3-Stroke Status of Patients

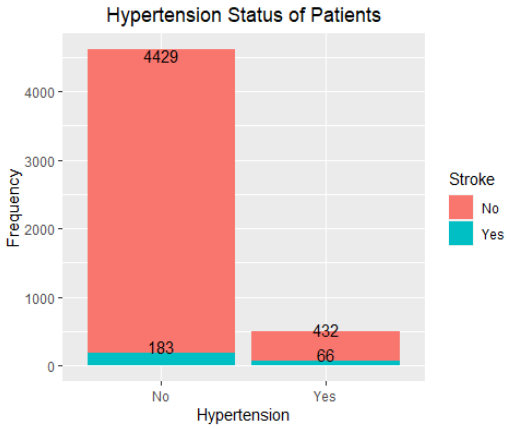


Figure 2-Hypertentsion Status of Patients

As it is illustrated in **Figure 3**, the majority of record have not experienced stroke , only 249 people did so. In conclusion we can see that we have an **imbalance** data.

As displayed in **Figure 2** ,We can have conclusion that people with blood pressure less than normal form the majority of our data, however the gap between the people who has hypertension and those who have not is slightly less than the plot with stroke patients. we can see that 15% of people with hypertension have experienced stoke and only 4 % of people with no hypertension had an stoke before. we can conclude the "hypertension" is one of important features effecting the stroke chance.

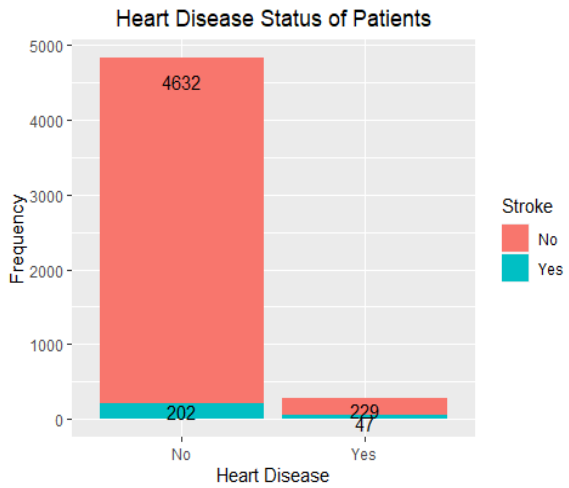


Figure 4-Heart Disease Status of Patients

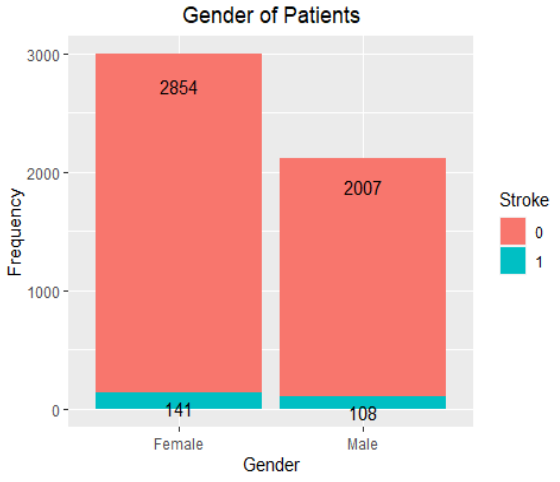


Figure 5-Gender of Patients

Again as it can be seen in **Figure 4** , the vast number of record do not have heart disease , 20 % of people who Have Heart disease have had stroke and only 4 % of people without heart disease experienced stroke, which illustrates that Heart disease have effect on number of stoke Heart disease have a relatively important effect on chance of stroke.

In **Figure 5** ,We can see the number of patients with stroke experience in male is slightly less than females which is logical due to more number of females in general, so we can conclude the gender has no significant effect on stroke.

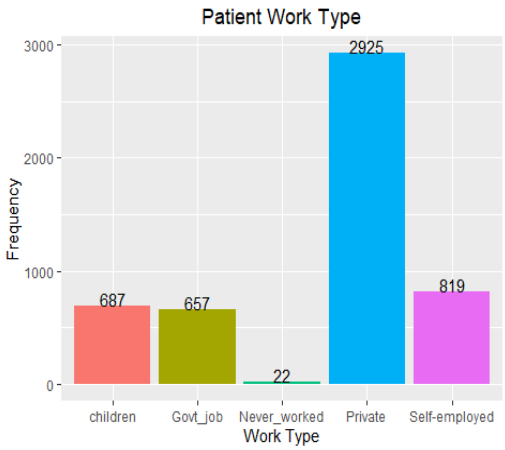


Figure 7-Patient Work Type

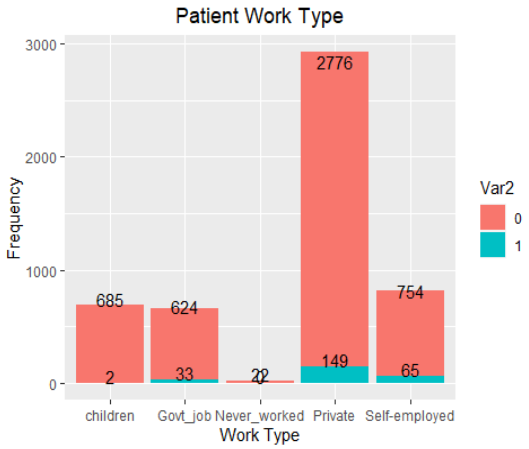


Figure 6-Patient with Stroke Experience Work Type

According to **Figure 7** ,The majority of data, had private work type, the least type is 22 people who have never worked. There are approximately even amounts of patients that are working government jobs, are self-employed, and are children.

As the **Figure 6** illustrates, the most patients with stoke experience, have the Private type of work, we can see 2 patients being under age had experience of stoke as well, not so much of a significant effect on stroke chance.

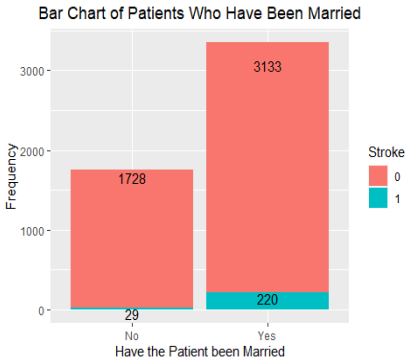


Figure 8 shows that the majority of record have been married before. Roughly double the amount of patients have been married before than those who have not.as shown, almost 7% of married people experienced stroke whereas only 1% of single people experienced the stroke, so we can conclude there is a slightly effect of married on stroke chance.

Figure 9-Bar Chart of Patients who have been married

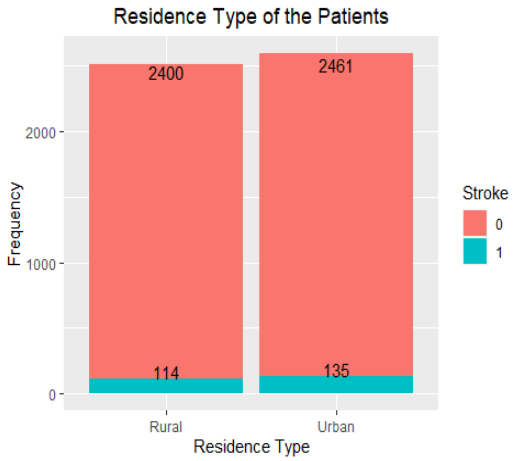


Figure 10-Residence Type of Patients

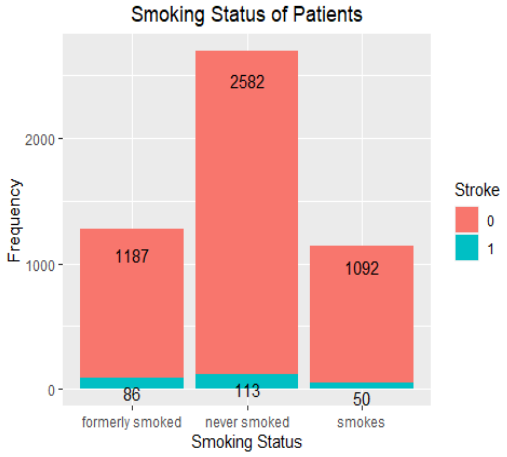


Figure 8-Smoking Status of Patients

According to **Figure 9**, the majority of data have never smoked. the data for formerly and currently smokers are similar. As illustrated, surprisingly there is no significant effect on people smoking status and stroke.

Based on **Figure 10**, there is virtually an even distribution between records in Rural residence and urban ones. Moreover, there is no significant effect of residency on stroke experienced number of patients.

Histograms

The second tool we use is Histogram charts to illustrates the relationship between Quantifiable features: bmi, age and average glucose level.

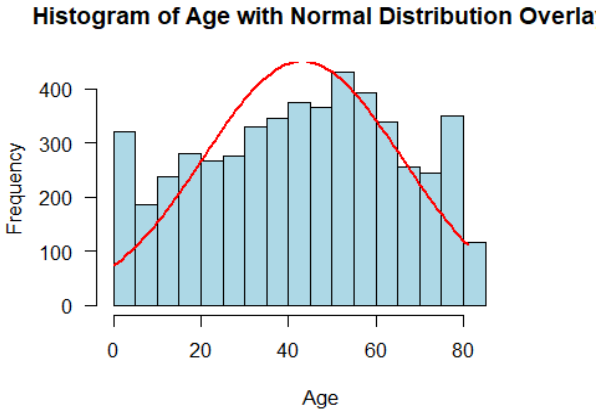


Figure 11-Histogram of Age

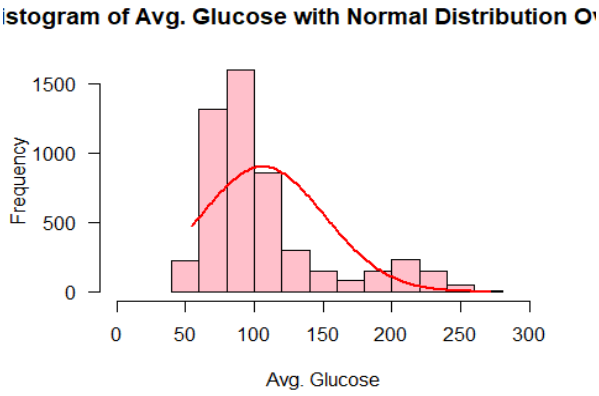


Figure 12-Histogram of Avg. Glucose

As illustrated in **Figure 11**, the age of patients in this study is close to normal distribution of age which got the mean of **43.22661** and based on this information we can conclude the majority of record are around **40** years old. Also, **Figure 12** shows that the average glucose levels of the patients in the study are **right skewed**, with mean of **106.1477** which is greater than median of **91.885**.

Box Plot

The third tool we use is Box plot for comparing one numeric and one quality features such as stroke and average glucose level and BMI.

The boxplot in **Figure 13** shows a relatively similar mean average glucose level in patients who suffered strokes and patients who have not, with lots of high outliers among non-stroke victims. we can conclude the average glucose level is not an effective feature on stroke.

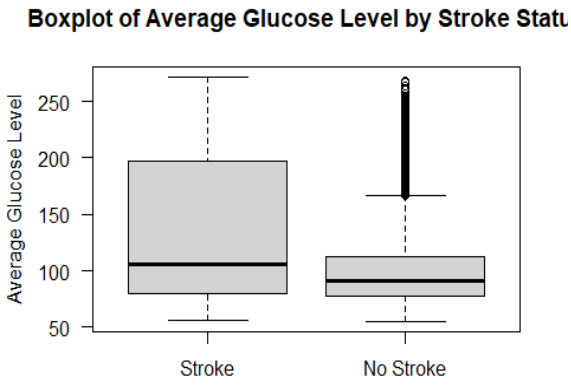


Figure 13-Boxplot of Average Glucose Level by Stroke Status

The mean of both patient with stroke and without it is relatively similar according to **Figure 14**, we can see the BMI is not the most effective feature on stroke.

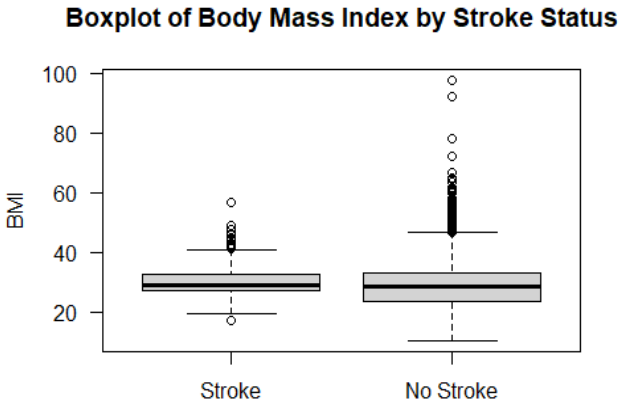


Figure 14-Boxplot of BMI by Stroke Status

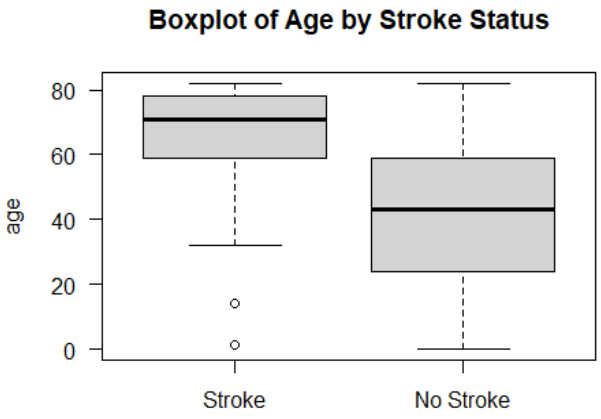


Figure 15-Boxplot of Age by Stroke Status

As it is shown in **Figure 15**, the older the patient is, the mean of stoke is higher than no stoke, so we can say age is a significant factor on stroke.

Point Plot

And as for the last tool, I used the ggplot2 library to draw a point chart for relationship between age and stroke chance. As the point charts illustrated, the majority of patients did not have stoke, but among those who had they are mostly from the age 40 to older. (mostly between 60 to 80.)

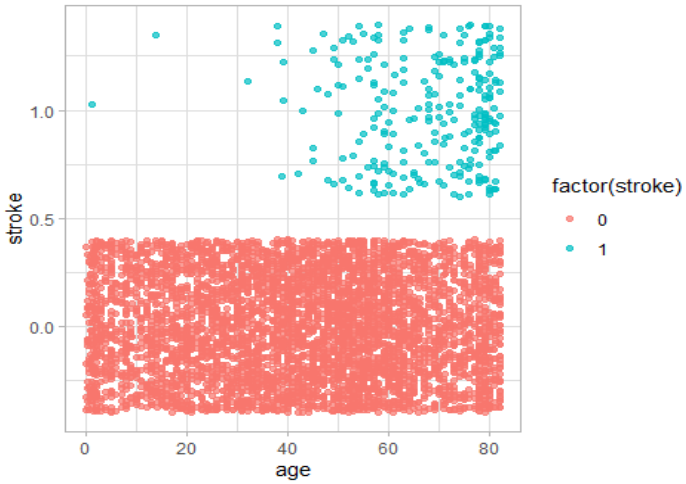


Figure 16-Relationship between Age and Stroke Chance

Processing Data

As we saw in above visualizations, we have several not significant variables: "gender", "residence type", "bmi", "work type", "id". So I decided to remove them from the main data. Also to prevent further confusion between the positive and vegetative class, I change stroke=1 to Y and stroke=0 to N. Also, after splitting the

Train data into Estimation and Validation ,in this part, I illustrated the Figure 17 with the help of GGallay library in order to illustrate all the important features of the main Data to have a sense of their relationship.

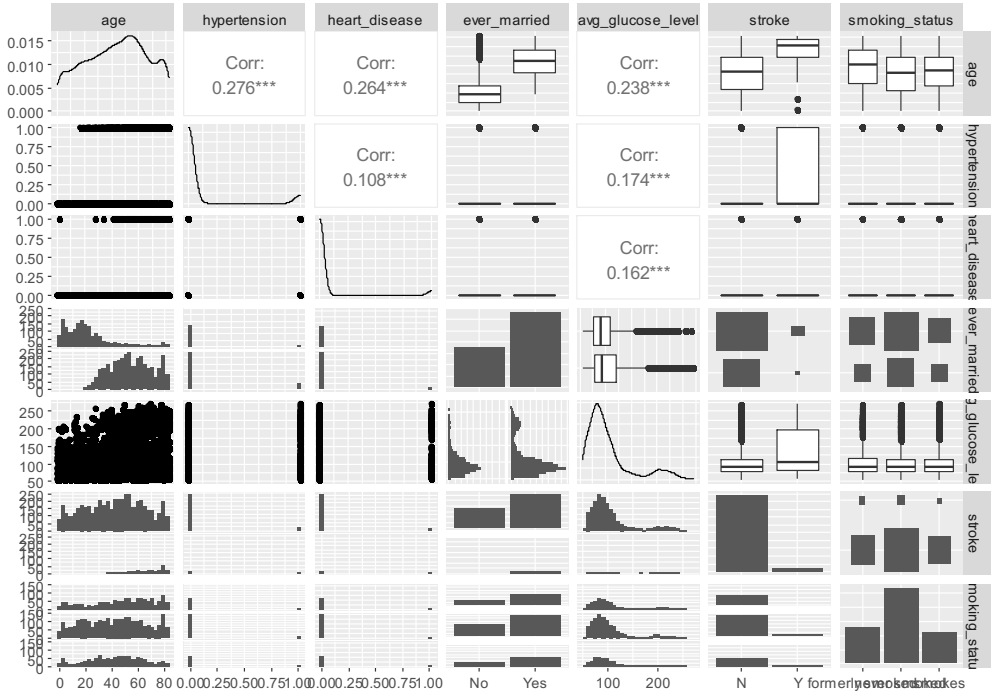


Figure 17-Important Features Chart

Criteria Function

In this question, it is important for us to predict the positive(Y) factors correctly, since it's important to know whether a person is likely to have **stroke** or not in order to be able to **prevent** it. So, the criteria we use is based on the model with **least False positive** because we do not want the patient **likely** to have stroke be left with **no alert** with wrong prediction. (The written function is available in R code attached to this file.)

None Parametric Methods

KNN

Steps that need to be done in this Section is as below:

1.normalization

Before we start with fitting candidate models of KNN we have to use Scale function to normalize the quantifiable features such as age and glucose level.

2.Tuning K

Now after we normalized our features, I created a function to generate K ranging from 1 to 100 and after comparing the Sensitivity and Accuracy we fit the final model. As the result shows in Figure 18, the KNN with k=1 with the 90% accuracy and 12% sensitivity is the candid model.

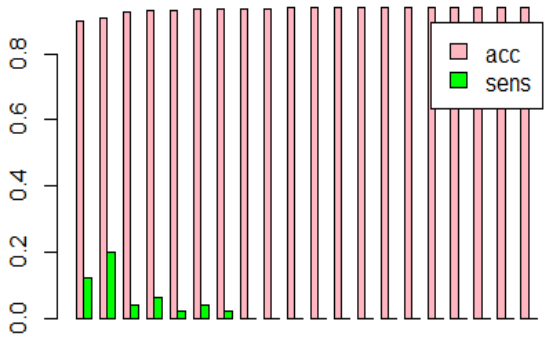


Figure 18-Tuning K

Regression Tree

In this method I try to fit suitable models for predicting stroke occurring, I use the Regression Tree to split the data based on important features and to achieve a final node. With regression tree, I create a list with 10 models, 5 consist of minsplit=5 and other 20 with cp ranging from 0 to 1. As illustration of Figure 19 shows, the model 2 with accuracy of 91% which is slighlty higher than model 1 and sensitivity of 8% is the candid one.

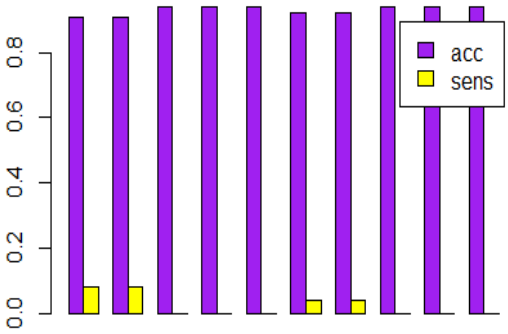


Figure 19-Regression Tree of Model 2

Parametric Methods

Logistic Regression

Before we start to fit the models and check different cutoffs, we should know that Stroke=Y is positive for us because it's the **important factor** in this problem. With each model, we include several different features and compare the result of their Cutoffs Sensitivity, Accuracy and Specificity. Tables 5 to 8 illustrate the comparative tables of each 4 model testes in this Section.

cutoffs	acc	sens	spec
0.1	0.8471883	0.52	0.8684896
0.3	0.9315403	0.06	0.9882812
0.5	0.9388753	0.00	1.0000000
0.7	0.9388753	0.00	1.0000000
0.9	0.9388753	0.00	1.0000000

Table 5

cutoffs	acc	sens	spec
0.1	0.8398533	0.54	0.8593750
0.3	0.9339853	0.06	0.9908854
0.5	0.9388753	0.00	1.0000000
0.7	0.9388753	0.00	1.0000000
0.9	0.9388753	0.00	1.0000000

Table 6

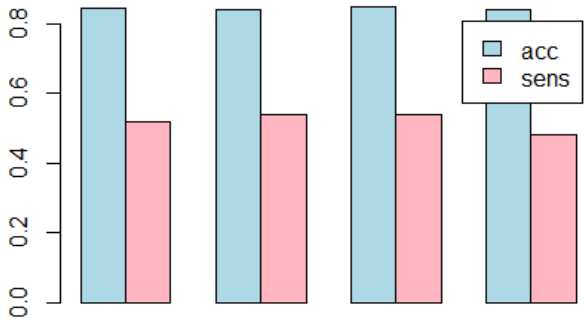
cutoffs	acc	sens	spec
0.1	0.8484108	0.54	0.8684896
0.3	0.9352078	0.10	0.9895833
0.5	0.9388753	0.00	1.0000000
0.7	0.9388753	0.00	1.0000000
0.9	0.9388753	0.00	1.0000000

Table 7

cutoffs	acc	sens	spec
0.1	0.8410758	0.48	0.8645833
0.3	0.9327628	0.08	0.9882812
0.5	0.9425428	0.06	1.0000000
0.7	0.9388753	0.00	1.0000000
0.9	0.9388753	0.00	1.0000000

Table 8

As a result of these 4 confusion matrix illustrates base on highest accuracy, the fourth model consist of all features with their binary interaction has the most Accuracy in cutoff 0.5. this cut off also has the sensitivity of 0.04 which means the number of True positives (stroke=1, pred=1) are not so much high. out of all the 50 positives(stroke=1), we only have predicted 2 of them right. (1-48/50)=0.04! on the other hand , the important factor, specificity is high. specificity indicates on (stroke=0 and pred=0) TN which is 768 out of 768, so the specificity is 1 and it means that we make no mistake in prediction of people unlikely to have stroke. accuracy is 94% which means that model is doing Alright specially predicting the healthy patients being healthy. What matters most in this modeling is that we should not have high FN. which means we do not say a patient who is likely to have stroke that they are healthy. in this model out of 50 people likely to have stoke, we wrongly predicted 49 of them that they are not going to have a stroke, so the model is not good at all. So, we should switch from the criteria of choosing the models based on the accuracy to highest **Sensitivity.**



As we can see in **Figure 20** the models with most sensitivity are model 2 and 3 with 0.54%. however our candid Logistic model is model 3, since the second important factor, accuracy is slightly higher.

Figure 20

Discrimination models

In this Part, I first organized a list of LDA, QDA and Naive Bayes models which I thought can work efficaciously, and then I put prediction of them in another list to be able to set the important criteria on them and compare them to determine the candid model of each method. The final result of Each method is illustrated in Charts below.

LDA

As we can see in **Figure 21**, the best model out of LDA model is the one with uniform prior probabilities , since our main criteria here is not misclassification or accuracy, we are looking for the model with highest sensitivity because it is important to predict right in order to raise survival rate of patients. But since the model is not perfect, we chose the third LDA model with uniform prior probabilities which had 77% accuracy and 68% sensitivity and doing better than other models in this criterion.

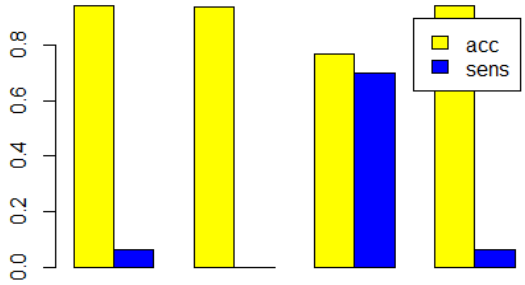


Figure 21-Comparing LDA models

QDA

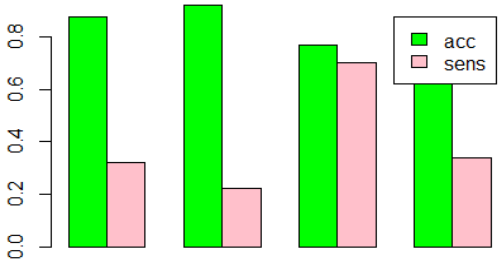


Figure 22-QDA

Naive Bayes

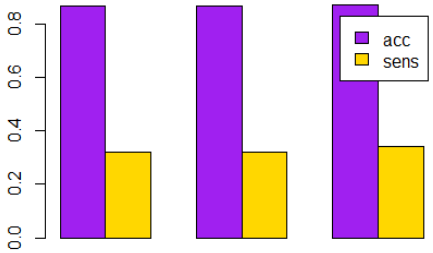


Figure 23-Naive Bayes

According to **Figure 22**, since our best model must have the highest sensitivity, The third model with accuracy of 76% and sensitivity of 70% is our candid model.

Based on **Figure 23**, The three models of Naive Bayes are all slightly different, the result shows that all of them have accuracy of nearly 87% and sensitivity around 35%. between three, the third model is with the most accuracy and sensitivity, so we pick the third model in this method.

Visualization and Testing

Comparing all

In this part, we use our important Criteria of Accuracy and Sensitivity to compare all the candid models from methods above, to pick a final candid model, to do

so, I used list of different models and using the criteria function available in R code attached, we concluded the Table 8-1 below.

	pred_knn	pred_tree	pred_glm	pred_lda	pred_qda	pred_nb
acc	0.900978	0.9095355	0.8484108	0.7689487	0.7689487	0.8716381
sens	0.120000	0.0800000	0.5400000	0.7000000	0.7000000	0.3400000

Table 9

Visualization

Now, using R tools, we illustrate the result of Table 9 in order to gain a better and more clear view of the models.

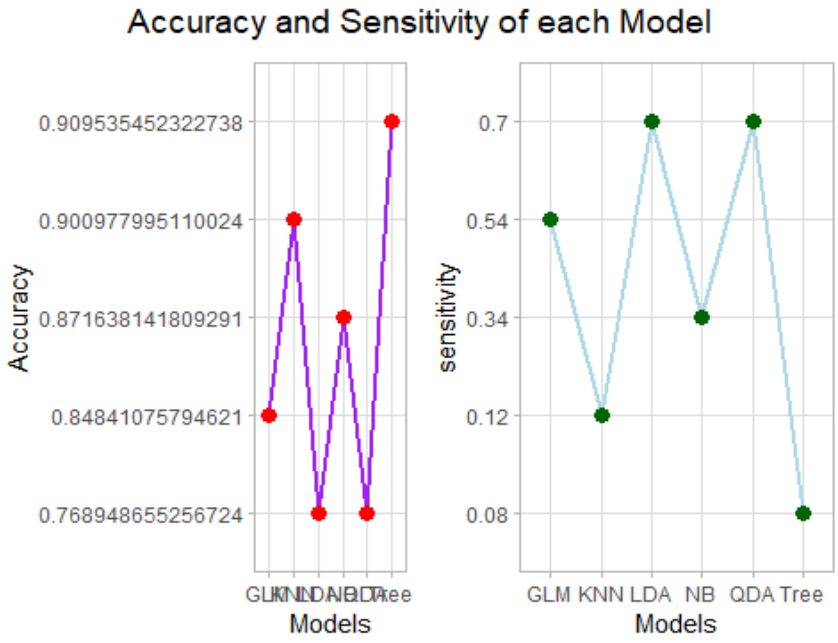


Figure 24-Accuracy and Sensitivity of Each Model

As both result of **Table 9** and **Figure 24** illustrate the highest sensitivity belongs to model LDA and QDA, since the accuracy of model QDA is 1% higher than LDA , we chose QDA model as our final Candid Model.

Testing

Now we fit the final Candid model, QDA, on train data and then test the prediction on test data. To do so, I used the confusion Matrix result to have a clear understanding of its accuracy and other important criteria. The result of the Confusion Matrix is shown in Table 10 , 11:

Confusion Matrix and Statistics

Prediction	N	Y
N	726	12
Y	252	32

Table 10

Accuracy : 0.7417
Sensitivity : 0.72727
Specificity : 0.74233

Table 11

We can see the final Model has the sensitivity of 70% and accuracy of 74%, in general the model is doing alright.

Over and Under Sampling

I guess the problem of not accurately predicting the patients with stroke is due to data being properly imbalanced, so we use Over and Under Sampling here to test the hypothesis.

Over

```
Accuracy    : 0.7378
Sensitivity : 0.72727
Specificity : 0.73824
```

Table 12

With this change shown in table 12, we can see that specificity and accuracy did not change significantly, sensitivity raised up about 2%.

Under

Accuracy :	0.726
Sensitivity :	0.72727
Specificity :	0.72597

Table 13

With this change, we can see that specificity and accuracy did not change significantly, but the sensitivity went up 2%.

Conclusion

This Report's content is created by R Studio and aims to illustrate the analysis and reasons behind the chunks of R codes of Machine learning Classification project on stroke occurring prediction and statistical calculations.

I hope you a have found this Report appropriate to your use.

Thank you for your attention.

Mina Kanaani

Resources

R code, Rmarkdown, And CSV files of data is attached to this file.

Other sources:

https://www.kaggle.com

https://www.kaggle.com/code/adrynh/stroke-prediction