Questions 7 and 8

Mohammad Javad Pesarakloo
810100103

March 18, 2024

Question 7

This section is implementation of naive bayes classifier which will be trained on dataset of lung cancer.

Part A

Naive bayes classifier is a classifier which assumes that the value of a particular feature is independent of the value of any other feature.Although this assumption might seem weird but statistics have shown that this classifier has decent accuracy.As we learned through the course, Bayes classification is an ideal idea that makes perfect sense when we have prior distribution of data;but when distribution is not available, Naive bayes classifier considers three probabalistic models:

- Gaussian naive Bayes
- Multinomial naive Bayes
- Bernoulli naive Bayes

In this handson, we use the **Guassian naive Bayes** and fit normal distribution to data to calculate posterior distribution. The reason for which we might use naive bayes is that considering dependency between features might add complexity to the model and also distribution is not always available.But drawback to this method is that the distribution might not even be close to Gaussian or Multinomial or Bernoulli and consequently we are loosing some detail of our train data. To develop a model on a dataset, a preprocessing phase should be done on that.There are multiple steps in this phase:

- Scaling and normalizing : To adjust the range of features and to standardize the data
- Encoding : Converting categorical data to numerical format
- Handling missing values : using some statistical techniques to handle some missing values in the dataset, such as replacing the missing item with mean of other available
- Feature selection : To select features which best distinguish the class with minimum overlap

In this dataset, we have some missing values, and consequently need to handle them;As the number of missing items is not great, we can easily drop the rows containing missing values.The next thing we have to do is to encode the last column of the dataframe which is the label of the data.We encode *YES* as **2** and *NO* as **1**. Next we have to select features.To do so we visulaize data using pairplot which gives us the following figure:

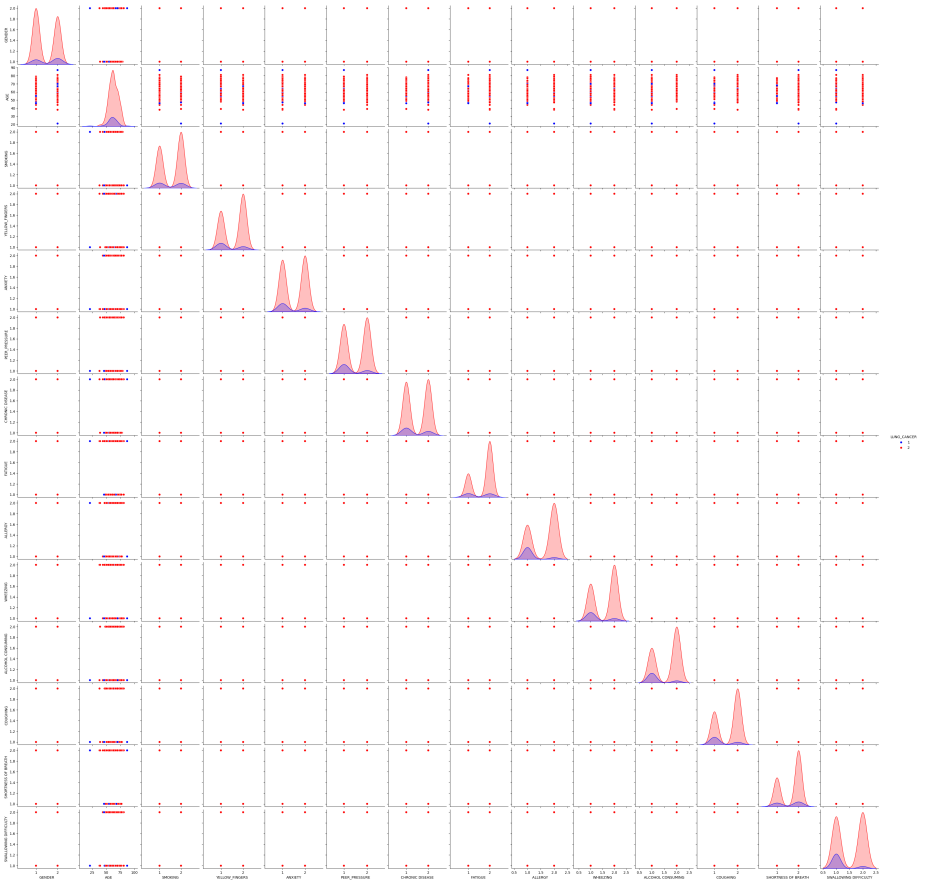

Figure 1: pairplot

part B

To develop a Naive Bayes model from scratch, we used it's guassian alternation;which means we fit a guassian distribution for perior probabalities.As all the probabalities are between 0 and 1 and we multiply multiple of them, to prevent underflow of numbers, we get logarithm from them and convert multiplication to summation:

$$p(y_i|X) = \frac{p(X|y_i)p(y_i)}{p(X)}$$

$$= \prod_{j=1}^{n} \frac{p(x_j|y_i)p(y_i)}{p(x_j)}$$

To find the label of qeury we have to apply argmax operation on labels:

$$LabelOfQuery = argmax(\prod_{j=1}^{n} \frac{p(x_j|y_i)p(y_i)}{p(x_j)})$$

$$= argmax(\prod_{j=1}^{n} p(x_j|y_i)p(y_i))$$

$$= argmax(\log(\prod_{j=1}^{n} p(x_j|y_i)p(y_i)))$$

$$= argmax(\sum_{j=1}^{n} \log(p(x_j|y_i)) + log(p(y_i)))$$

The evaluation of model is as follows:

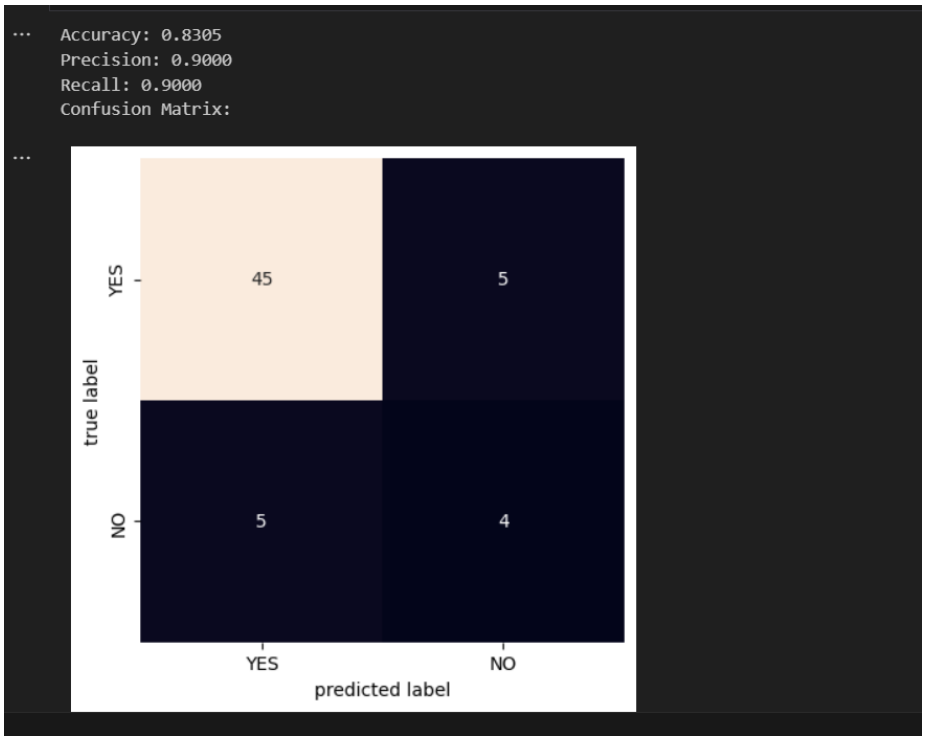

Figure 2: evaluation of model

From the above confusion matrix, we conclude that our model is more bias to positive classification.

part C

I got the exact same results(See LungCancer.ipynb file) Naive Bayes is giving about 83% accuracy which is not much good for a classifier.Also its precision and recall is 90% which is good in this case because it decreases the likelihood of a person having lung cancer and not being diagnosed.But the reason for which it's accuracy is low can be the assumptions about i.i.d data and fitting normal distribution to data.

part D

On this dataset, results of my own implementation and sklearn Naive Bayes classifier had very slight differences.Evaluation of my own implementation is:

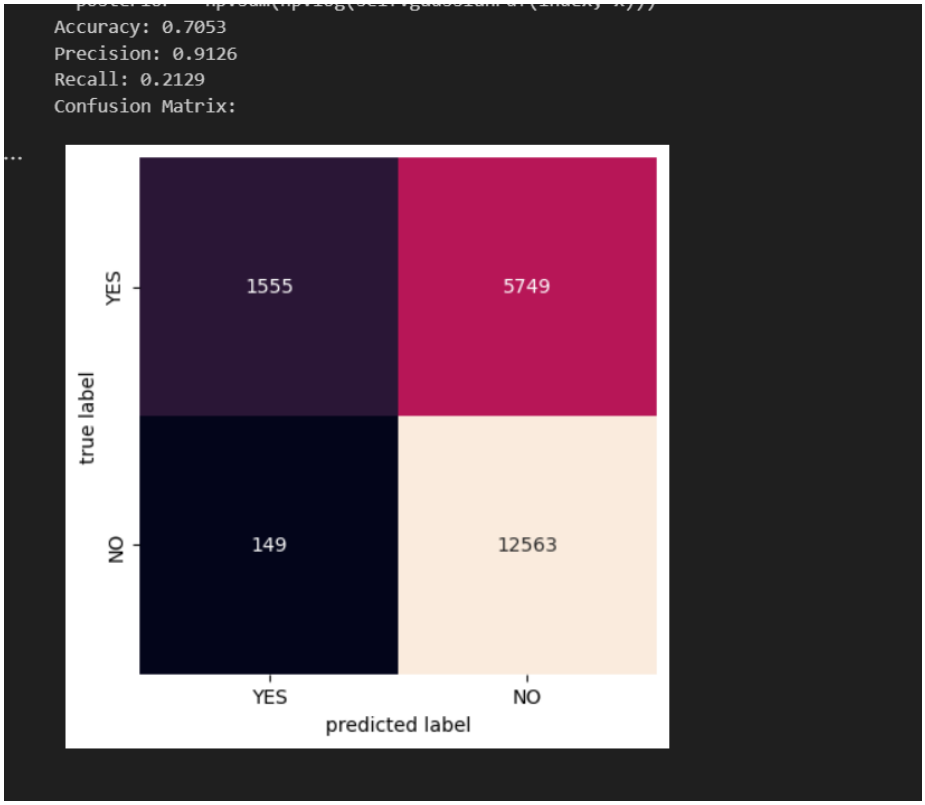

Figure 3: my own implementation

and sklearn evaluation is:

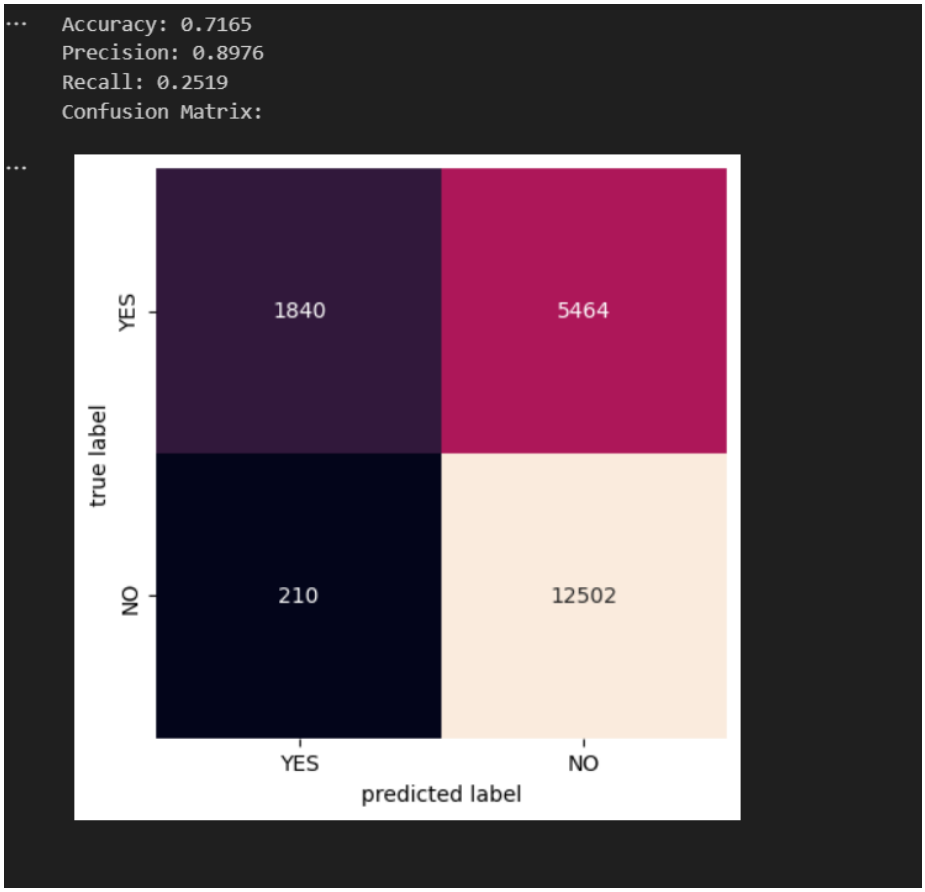

Figure 4: sklearn evaluation

As shown above, we are getting about 70% accuracy which is less than the previous dataset despite being larger.The reason is that every model has a bias.In naive bayes for example the assumption of independancy of features and gussian distribution, imposes a bias to the model.When new data is far from our assumption, increasing size of train data with these new data will increase this bias and decrease accuracy of model.

Question 8

In this part I implemented an algorithm which classify images analysing their color.The more they look yellow, the more probable to be classified as **s**.There is a parameter named thereshold which specifies amount of this similarity.By

changing this parameter we can get better accuracy.The accuracy I got was 67.5%.Datas which are classified wrongly are the pictures with yellow clouds.To make the model work better, we can add another condidition to distinguish clouds.Such as consequtive changes of colors that occur on a cloud.