Naïve Bayes Classifier

A Seminar Report

Submitted by

Mohammed Kashif
(21MA60R39)

MASTER OF TECHNOLOGY

COMPUTER SCIENCE AND DATA PROCESSING

at

DEPARTMENT OF MATHEMATICS

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

WEST BENGAL-721302, INDIA

Contents **Page No.**

ABSTRACT

The different classifier used for making decisions. Out of many machines learning algorithms, Naïve Bayes classifier uses knowledge from probabilistic approach. It uses making future prediction. This approach seen to be applicable for numeric as well as categorical data. The Naive Bayes classifier separates data into different classes according to the Bayes' Theorem, along with the assumption that all the predictors are independent of one another. It assumes that a particular feature in a class is not related to the presence of other features.

Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset. Naïve Bayes classifier is one of the simple and most effective classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object. Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.

1. Introduction

The Digital revolution just unlike any other revolution has changed everything around us. Nowadays, organizations of all types and across all industries utilize digital solutions to facilitate work operations, reduce costs, making error free decisions. People worldwide are using technology to work, study, socialize, entertainment, shopping, or online banking. On a common ground people making interactions with technology, this generates a large amount of digital data, combined with the data produced by organizations worldwide create a massive amount of digital data.

The total data created, captured, copied, and consumed worldwide is forecasted to increase rapidly, reaching 59 zettabytes in 2020, According to Statista. Most data are created digitally and never find their way into papers. Classification is a term refers for predicting that given data points are belong which class. The type of data analysis that extracts different models and describing important data classes is classification.

So, in this report we are discussing one of the Machine Learning techniques named as Naïve Bayes Classifier. Basically, there are following machine learning techniques,

1. Logistic Regression
2. Support Vector Machine
3. Decision Tree
4. Random Forest
5. Naïve Bayes
6. K-nearest Neighbor
7. Artificial Neural Network

Depending on performances of these technique in different application they are selected to use.

What is Data?

Data is information of individual person, facts, statistics that may be qualitative or quantitative. This information of person is useful or more specifically collectively for large group of people form dataset.

Let example of Employees in a company has their information related to their job such as employment Id, Age, Skills, joining date, work profile. These data for faster processing are being converted into different file format like CSV, Excel, Tabular, Text format etc.

What is Data classifier:

This representation can be done in terms of any data structure, pictorial representation that is in terms of graphs. For doing the same we have plenty of different techniques as discussed below,

There are many machine learning algorithms which employs various statistical, probabilistic and optimization methods to learn from past known experiences. These learning finds useful patterns from large, unstructured and complex datasets. These algorithms have wide range of applications, in marketing just like customer purchase behavior detection, in medical field for disease modelling, in information theory for junk e-mail detection and spam message detection. These applications are supervised learning variants rather than unsupervised learning. In supervised learning model, decisions are made by learning dataset. At learning stage labeling of data are known and this information used for making decisions of unlabeled examples.

Categorization of Data:

Data classification is a process of organizing and analyzing different data whether in structured form or unstructured form. This classification is based on file type, meta data and their contents.

In the field of data analysis, the different terms are being used while dealing with useful information.

1. Entity:
   In given information particular thing is called entity.
2. Attribute:
   An attribute is a measurable or observable property of entity.
3. Data:
   Different information parameters of attribute defines data. These parameters must be measurable.

Example:
In the example of online food ordering company. The order is entity, the customer is attribute and the details of order like order time, food type, cost etc are data information of customer attribute.

The data is classified in terms of measuring scales and its attribute type. In present we NOIR classification. Where N-Nominal data, O-Ordinal data, I-Interval data and R-Ratio data.

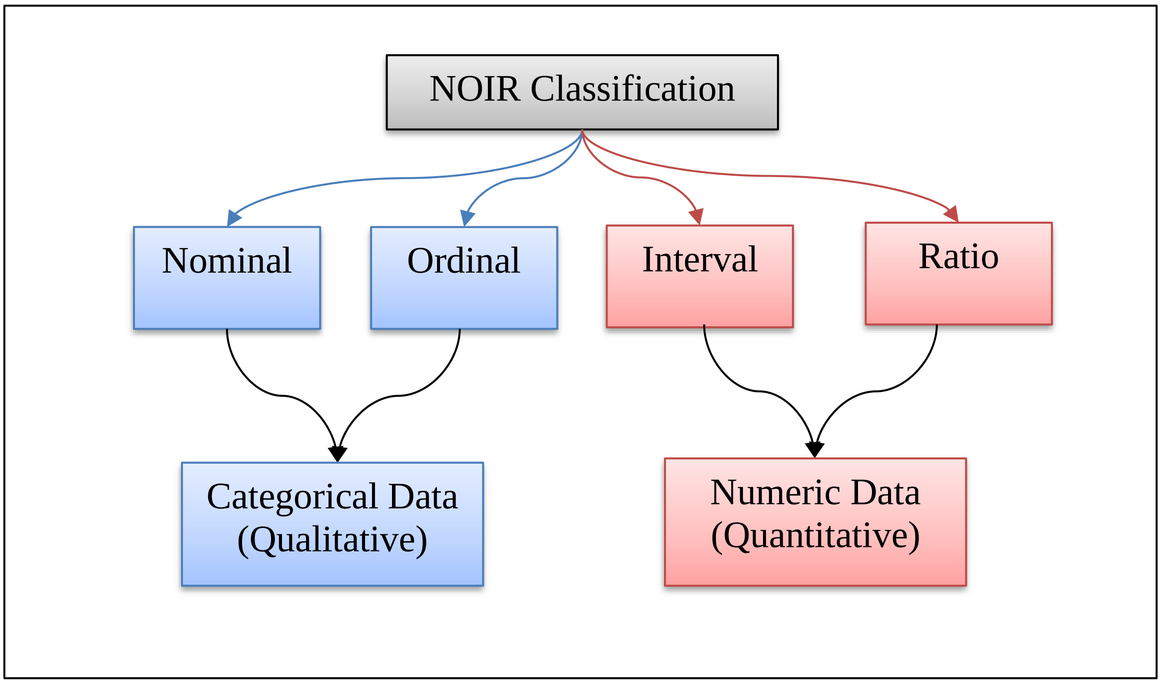


Fig. 1.1:Data Classification

Sr No.	Property	Operation	Type
1	Distinctiveness	= and !=	Categorical
2	Order	<,<=,>,>=	
3	Addition	(+) and (-)	Numerical
4	Multiplication	* and \	

Table 1.1. Data classification.

Bayesian Classifier:

In nutshell the bayes classifier uses the Bayes theorem to predict the class as the one that maximizes the posterior conditional probabilities.

- A statistical classifier,
  - Performs probabilistic prediction, i.e., predicts class membership probabilities
  - Foundation
  - Based on Bayes' Theorem.
  - Assumptions

The classes are mutually exclusive and exhaustive. The attributes are independent given the class. Hence called "Naïve" classifier because of these assumptions. Empirically proven to be useful.

In many applications, the relationship between the attributes set and the class variable is non-deterministic.

- In other words, a test cannot be classified to a class label with certainty.
- In such a situation, the classification can be achieved probabilistically.
- The Bayesian classifier is an approach for modelling probabilistic relationships between the attribute set and the class variable.
- More precisely, Bayesian classifier use Bayes' Theorem of Probability for classification.

Consider a training dataset D consists of n number of points x_j in a D dimensional space where j ϵ {1,2, 3,…..n}. Let y_i is the class for each point, with So here we considered there are k number of classes. The Bayes Classifier uses Bayes theorem to make decision about the class for a new test point. It estimates conditional probabilities P(C_i|x) for every class c_i, final class is selected which has highest probabilities. P(C_i|x) is also called posterior probabilities.

$$\hat{y} = \underset{i}{arg\,max}\{P(c_i|x)\}$$

The conditional probability that instance x_i has true for given class c_i is given as

$$P(c_i|x)=\frac{P(x|C_i)\times P(C_i)}{P(x)}$$

Where $P(x|c_i|)$ is called likelihood of instance x. $P(C_i)$ is the prior probabilities, and $P(x)$ is the probabilities of occurrence of x from any one of the classes.

$$P(x) = \sum_{j=1}^{k} P\left(x|c_j\right) \cdot P\left(c_j\right)$$

By using these relations, we can get,

$$\hat{y} = arg\max_{i} \left\{ \frac{P(x|c_i) \cdot p(c_i)}{p(x)} \right\}$$

The predicted class is essentially depends on likelihood and prior probabilities. For two attribute X and Y, It is been consider that the class conditional probability,

$$P(Y|X) \approx P(X|Y).P(Y)$$

Proof:

$$\begin{aligned} P(Y|X) &= \\ &= P(Y \wedge X)/P(X) \\ &= P(X \wedge Y)/P(X) \\ &= P(X|Y) * P(Y)/P(X) \\ &\approx P(X|Y).\ P(Y) \end{aligned}$$

P(X) is called the evidence (also the total probability) and it is constant.

Here for discussion of Naïve Bayes Classifier, we are going to see some important relations in joint probabilities.

1. Axioms used in probabilities:
   - All probabilities are between 0 and 1.
   - P(T)=1 and P(F)=0, where T-True and F-False.
   - P(AvB) = P(A) + P(B) - P(A^B).

Naïve Bayes Classifier:

Naïve Bayes classifier consider that attributes in given dataset are independent. Then also it is powerful tool found useful in many applications. This method is appropriate even when dimension n of dataset is large.

However, just known that there is another method commonly used of Naïve Bayes, called Gaussian Naïve Bayes Classification.

Supervised learning algorithm are being used in Naive bayes classifier, since NBC are trained using labelled data, that is labelled data has been pre-categorized into classes that are already available in classified form. In contrary the unsupervised learning is one where there is no prior data or information available. This type of learning gathers information naturally with no idea to how to classify the data.

Algorithm: (Naïve Bayesian Classification):

Input: Given a set of k mutually exclusive and exhaustive classes $C = \{c_1, c_2, \ldots .\ . , c_k\}$. Which have prior probabilities $P(C_1), P(C_2), \ldots.. P(C_k)$. There are n-attribute set $X = \{x_1, x_2, \ \ldots x_n\}$,which for a given instance have values $A_1 = a_1$, $A_2 = a_2, \ldots.., A_n = a_n.$

Step: For each $c_i \in$ C, calculate the class condition probabilities, i = 1,2,…..,k

$$p_i = P(c_i) \text{ x } \prod_{j=1}^{n} P(x_j|c_i)$$

$$p_x = max\{P_1, P_2, \ \ldots P_k\}$$

Output: C_x is the classification
Note: $\boldsymbol{\sum p_i \neq 1}$, because they are not probabilities rather proportion values (to posterior probabilities).

Example:

We receive messages from our friend and family and we also receive messages from unknown sources, spam. We exhausted by these messages, we wanted to filter out these messages. For those normal messages received from family or friends, make histogram for words commonly used. So histogram ease us to find probabilities of occurrences of these words.

Dear	Friend	Lunch	money
8	5	3	1

Table 1.2 Occurrences of word in Ham message

P(Dear | Normal)= $\frac{8}{17}$ = 0.47, P(Friend | Normal) = $\frac{5}{17}$ = 0.29

P(Lunch | Normal)= $\frac{3}{17}$ = 0.18, P(Money | Normal) = $\frac{1}{17}$ = 0.06

P(N) x P(Dear | N) x P(Friend | N) =0.67 x 0.47 x 0.29

= 0.09

P(N | Dear, Friend) ≈ 0.09

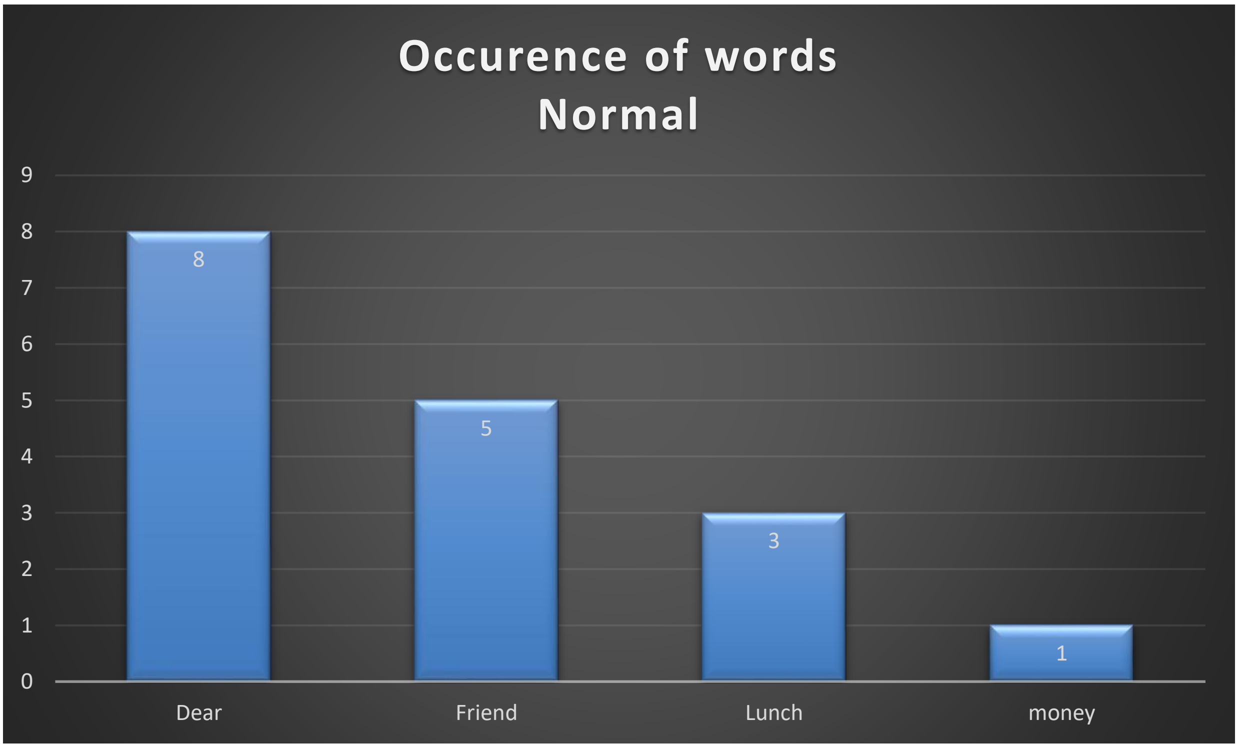


Fig 1.2 Occurrences of words

Similarly, for those spam messages receive, make histogram for words commonly used. So, histogram ease us to find probabilities of occurrences of these words.

Dear	Friend	Lunch	money
2	1	0	4

Table 1.3 Occurrences of word in spam message

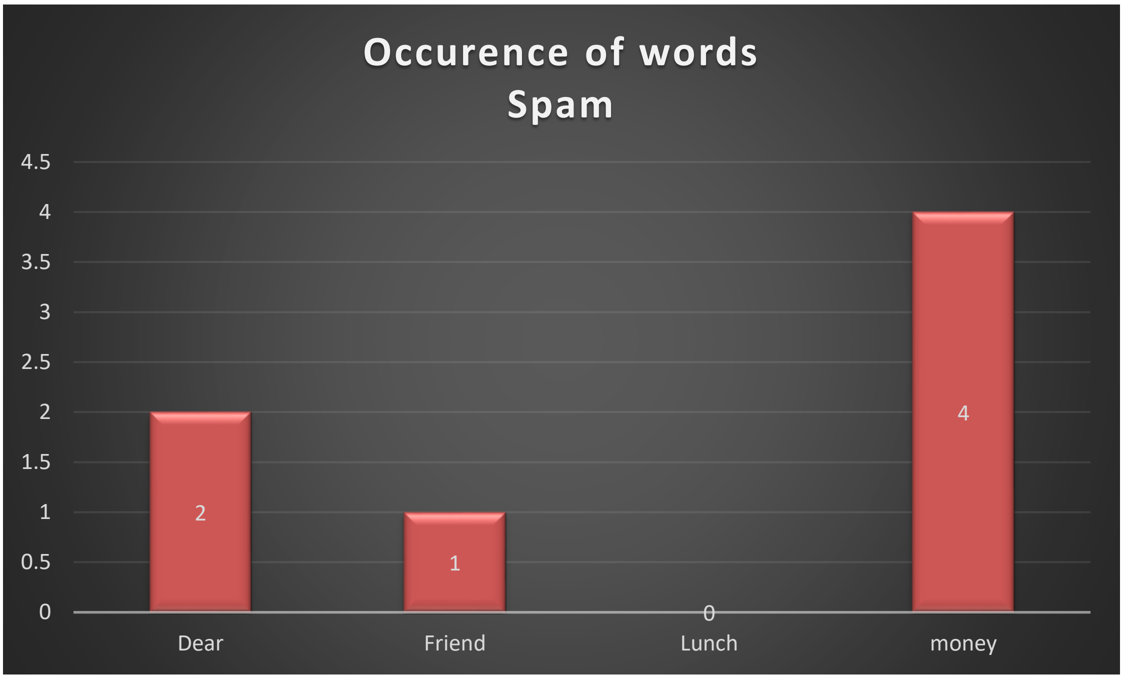


Fig 1.3 Occurrences of words

P(Dear | S) = $\frac{2}{7}$ $= 0.29$, P(Friend | S) = $\frac{2}{7}$ = 0.14

P(Lunch | S) = $\frac{0}{7}$ $= 0.0$, P(Money | S) = $\frac{4}{7}$ = 0.57

Now these probabilities as discussed earlier are called likelihoods. As we have calculated the probabilities of discrete, individual words, and not the probability of something continuous, like speed, weight or height.

P(S) x P(Dear | S) x P(Friend | S) = 0.57 x 0.14
= 0.01

P(S | Dear, Friend) ≈ 0.01

M-estimate of conditional probability:

The M-estimate is to deal with the potential problem of Naïve Bayes Classifier when data size is too poor. As seen in above example that class conditional probability depends on product of posterior probability. If one posterior probability of one attribute is vanishes then overall class conditional probability is zero. Even though some other attribute has significant occurrences, it might lead to inappropriate circumstances.

For the above example, Let consider a message “**Lunch Money Money Money**”. This message contains Money four times. As P(money | S)=0.57 is lot higher than P(money | N)=0.06, So it seen reasonable to predict this message as spam. So let try to verify:
For Normal class:

$$P(N) \text{ x } P(Lunch \mid N) \text{ x } \{[P(Money \mid N)]\text{^}4\}=0.000002$$

For Spam class:

$$P(S) \text{ x } P(Lunch \mid S) \text{ x } \{[P(Money \mid S)]\text{^}4\}=0.0$$

$$P(S) \text{ x } 0 \text{ x } \{[P(Money \mid S)]\text{^}4\}=0.0$$

$$\text{Since, } P(Lunch \mid S) = 0$$

This problem can be addressed by M-estimate approach.

To solve this problem, we add count 1 in each attribute. The number of count we add to each word is typically referred to with the m

$$P\left(x_j \middle| C_i\right) = \frac{n_{c_i}+mp}{n+m}$$

Where,

- n is number of training example for which Class = C_i,
- n_{c_i} number of examples for which Class = C_i and attribute x_j,
- m is a constant called equivalent constant size (equivalent sample size),
- p is prior estimate of probability.

Gaussian Naïve Bayesian classifier:

The Naïve Bayes approach is a very popular one, which often works well for categorical data but what if the data set contains numeric data. Certain advancements are made to make it eligible for numeric data.

Cons:

- It relies on all attributes being categorical.
- If data is less, then it estimates poorly.

Estimating the posterior probability (for continuous attributes):

In real life situation all the attribute need not necessarily be categorical. In fact, there is a mix of both categorical and continuous attributes. We can discretize each continuous attributes and then replace the continuous values with its corresponding discrete intervals.

Consider a training dataset D consists of n number of points x_j in a D dimensional space where j ϵ {1,2, 3,…..n}. Let y_i is the class for each point, with $y_i \epsilon \{C_1, C_{2\ldots 3} C_k\}$. So here we considered there are k number of classes. The Bayes Classifier uses Bayes theorem to make decision about the class for a new test point. It estimates conditional probabilities P(C_i|x) for every class c_i, final class is selected which has highest probabilities. P(C_i|x) is also called posterior probabilities. For each class C_i the posterior probability can be calculated using Gaussian Normal Distribution as follows:

$$P(x_j|C_i) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} \times e^{\left(-\frac{(x_j - u_{ij})^2}{2\sigma_{ij}^2}\right)}$$

Where,

- u_{ij} is mean of attribute x_j values belong to C_i,
- σ_{ij}^2 can be estimated from the calculation of variance of such training records.

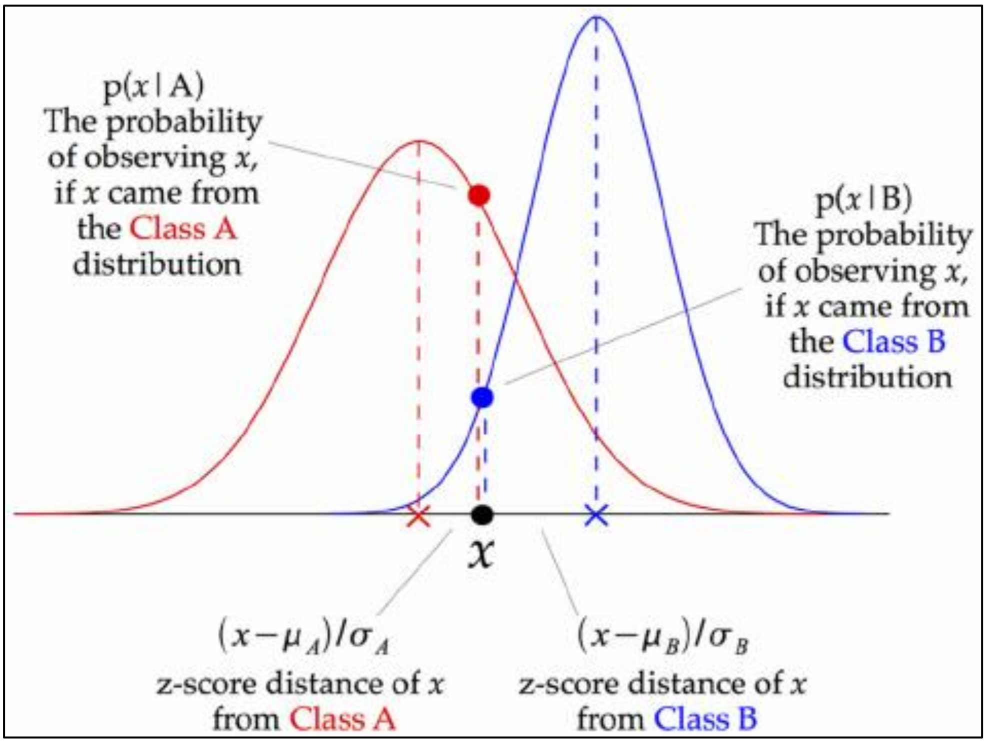


Fig 1.4 comparison of two numeric attribute

This model of Naïve bayes classifier can be fit simply by finding mean and standard deviation of the points within each attribute. So at every data point z-score is calculated namely distance of point from class C_i and divide by standard deviation of that class w.r.t attribute x_j.

Accuracy estimation:

We are to learn the metrics with which a classifier should be estimated. There are mainly to things to be measured for a given classifier.

i. Accuracy

ii. Performance

Accuracy:

If N is the number of instances with which a classifier is tested and p is the number of correctly classified instances, the accuracy can be denoted as,

$$\epsilon = \frac{P}{N}$$

Also, we can say the error rate (i.e., is misclassification rate) denoted by $\in^{\%}$ is denoted by

Error Rate = 1 – $\boldsymbol{\epsilon}$

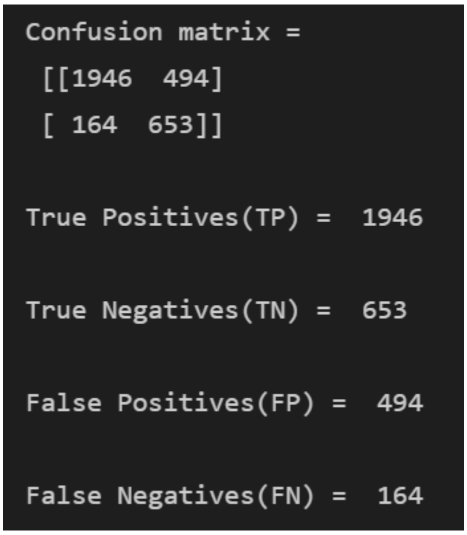
Performance:

There are four quadrants in the confusion matrix, which are symbolized as below,

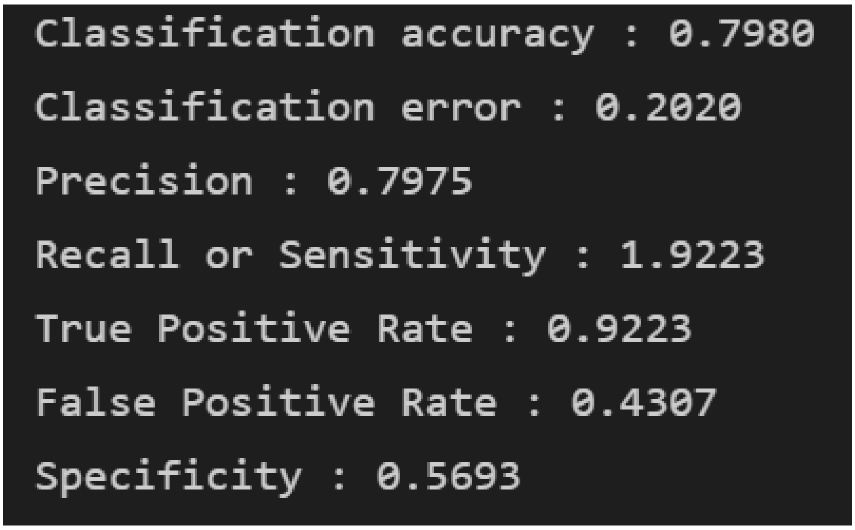
1) True Positive (TP: f++) : The number of instances that were positive (+) and correctly classified as positive (+v).
2) False Negative (FN: f+-): The number of instances that were positive (+) and incorrectly classified as negative (-). It is also known as Type 2 Error.
3) False Positive (FP: f-+): The number of instances that were negative (-) and incorrectly classified as (+). This also known as Type 1 Error.
4) True Negative (TN: f--): The number of instances that were negative (-) and correctly classified as (-).

Results

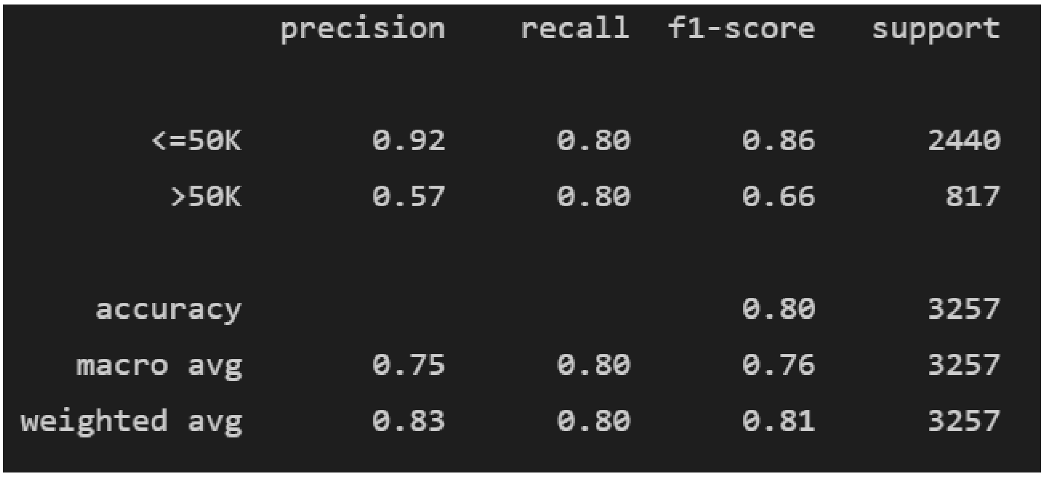
For 10% test data and 90% training data:



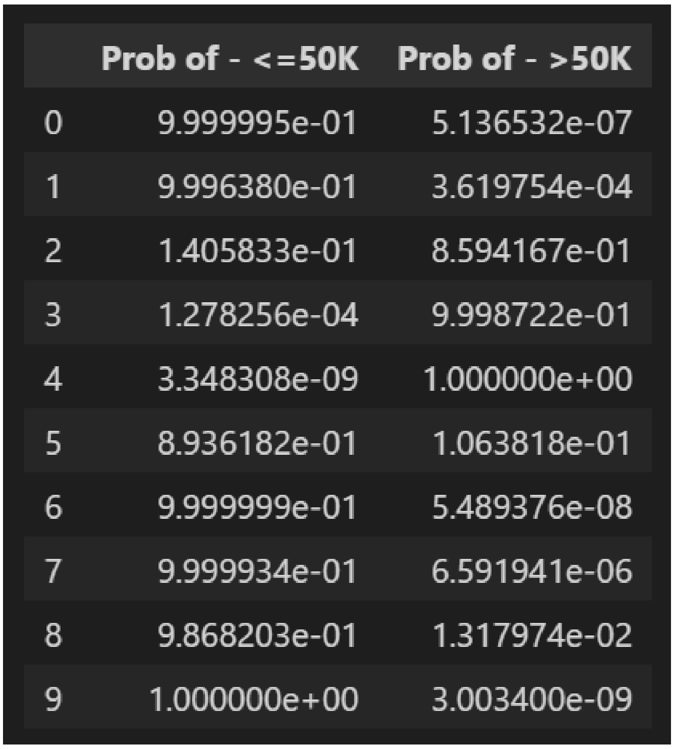
```
Confusion matrix =
 [[1946  494]
 [ 164  653]]

True Positives(TP) =  1946

True Negatives(TN) =  653

False Positives(FP) =  494

False Negatives(FN) =  164
```



```
Classification accuracy : 0.7980
Classification error : 0.2020
Precision : 0.7975
Recall or Sensitivity : 1.9223
True Positive Rate : 0.9223
False Positive Rate : 0.4307
Specificity : 0.5693
```

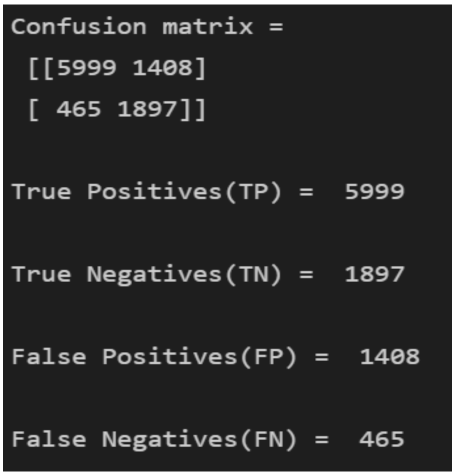


```
              precision    recall  f1-score   support

       <=50K       0.92      0.80      0.86      2440
        >50K       0.57      0.80      0.66       817

    accuracy                           0.80      3257
   macro avg       0.75      0.80      0.76      3257
weighted avg       0.83      0.80      0.81      3257
```

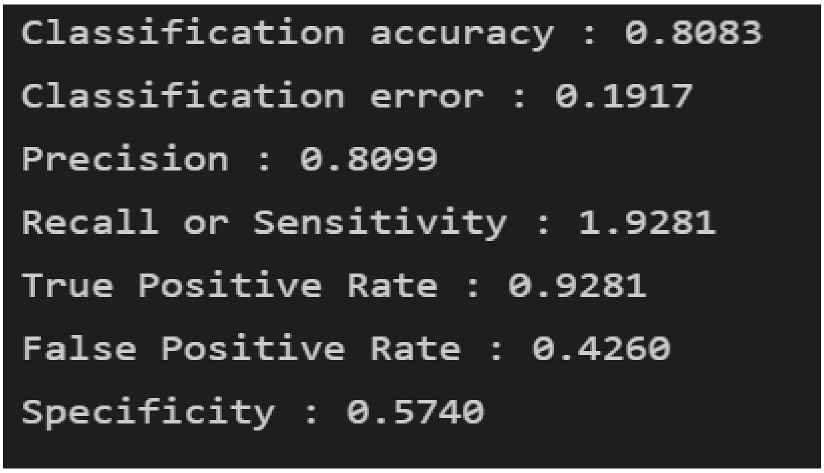


	Prob of - <=50K	Prob of - >50K
0	9.999995e-01	5.136532e-07
1	9.996380e-01	3.619754e-04
2	1.405833e-01	8.594167e-01
3	1.278256e-04	9.998722e-01
4	3.348308e-09	1.000000e+00
5	8.936182e-01	1.063818e-01
6	9.999999e-01	5.489376e-08
7	9.999934e-01	6.591941e-06
8	9.868203e-01	1.317974e-02
9	1.000000e+00	3.003400e-09

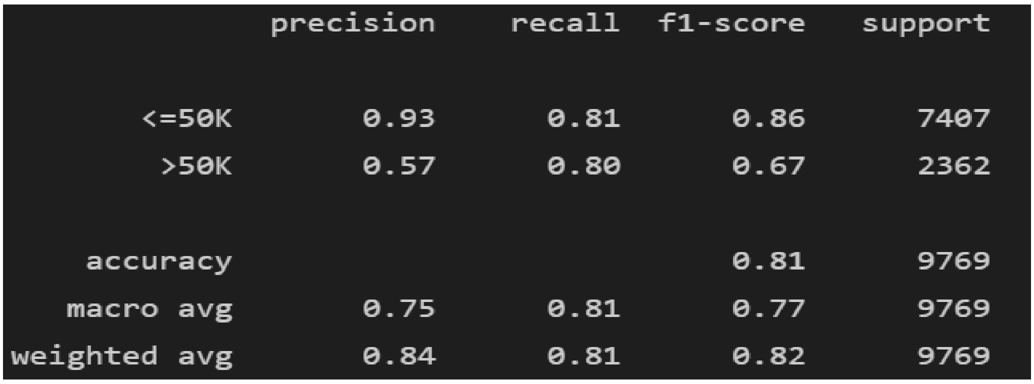
For 30% test data and 70% training data:



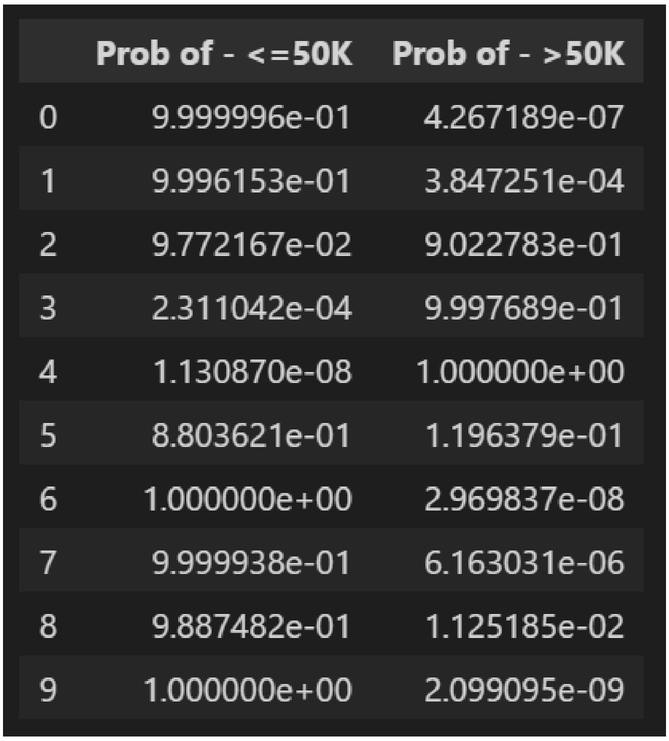
```
Confusion matrix =
 [[5999 1408]
 [ 465 1897]]

True Positives(TP) =  5999

True Negatives(TN) =  1897

False Positives(FP) =  1408

False Negatives(FN) =  465
```



```
Classification accuracy : 0.8083
Classification error : 0.1917
Precision : 0.8099
Recall or Sensitivity : 1.9281
True Positive Rate : 0.9281
False Positive Rate : 0.4260
Specificity : 0.5740
```



```
              precision    recall  f1-score   support

       <=50K       0.93      0.81      0.86      7407
        >50K       0.57      0.80      0.67      2362

    accuracy                           0.81      9769
   macro avg       0.75      0.81      0.77      9769
weighted avg       0.84      0.81      0.82      9769
```

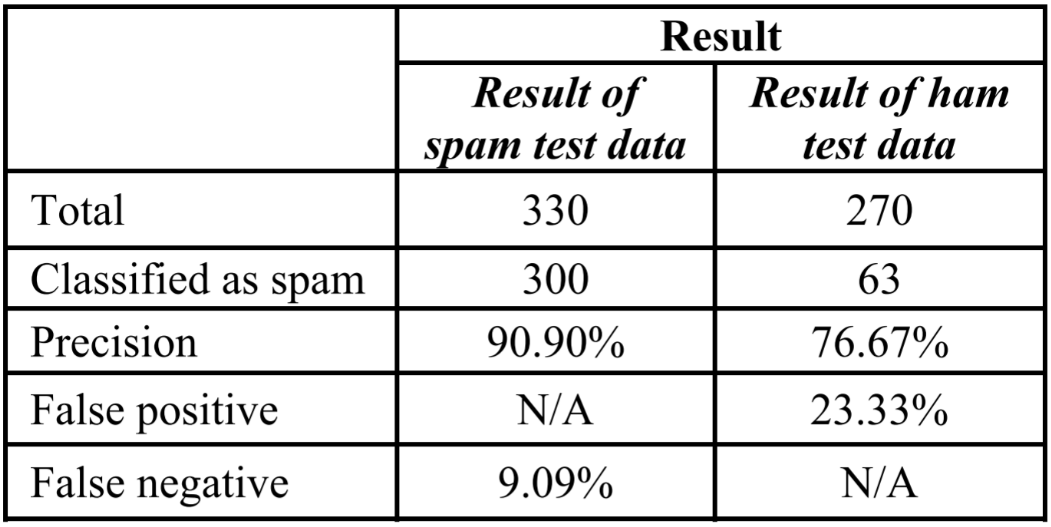


	Prob of - <=50K	Prob of - >50K
0	9.999996e-01	4.267189e-07
1	9.996153e-01	3.847251e-04
2	9.772167e-02	9.022783e-01
3	2.311042e-04	9.997689e-01
4	1.130870e-08	1.000000e+00
5	8.803621e-01	1.196379e-01
6	1.000000e+00	2.969837e-08
7	9.999938e-01	6.163031e-06
8	9.887482e-01	1.125185e-02
9	1.000000e+00	2.099095e-09

Performance of classifier changes as per selection of training and testing data.

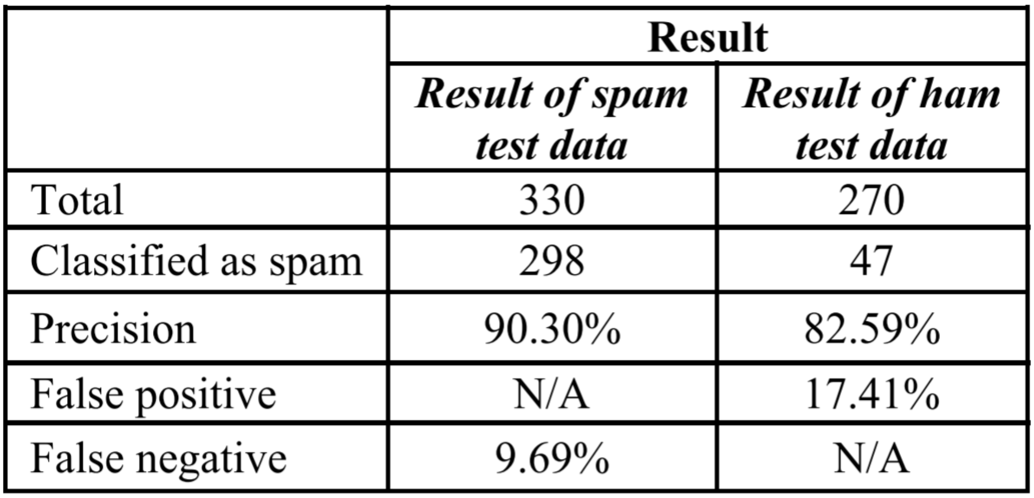
Importance of data cleaning:

The research paper referred in this report has signified that data preprocessing has positive impact on the results obtained by Naive Bayes Classifier. While comparing the results it is to be noted that the impact of False positive is more than that of False negative. Here without preprocessing False positive value for ham emails is 23.33% and after preprocessing False positive value for ham emails is reduced to 17.41%.



	Result	
	Result of spam test data	***Result of ham test data***
Total	330	270
Classified as spam	300	63
Precision	90.90%	76.67%
False positive	N/A	23.33%
False negative	9.09%	N/A

Without preprocessing



	Result	
	Result of spam test data	***Result of ham test data***
Total	330	270
Classified as spam	298	47
Precision	90.30%	82.59%
False positive	N/A	17.41%
False negative	9.69%	N/A

With preprocessing

After all discussion we found that Naïve Bayes Classifier have advantages and limitations:

Advantages:

1. Simple and very useful for large datasets.
2. It can be used for both binary and multi-class classification problems.
3. It can make probabilistic predictions and can handle both continuous and discrete data.
4. Performance of Naive Bayes Classifier is improved by data preprocessing

Limitations:

1. Classes must be mutually exclusive.
2. Presence of dependency between attributes negatively affects the classification performance.
3. It assumes the normal distribution of numeric attributes.
4. It requires more amount of training data.

Key Terms Used:

1. Meta-Data: According to oxford dictionary Meta-Data is "A set of data which describes and gives information about other data". Consider example that a document includes a different information grouped together just like a author of book, book size, date of publication, name of table and figure etc.
2. Information theory: It is the scientific study of the quantification, storage, and communication of digital information.

7. CONCLUSION

The availability of large amount of data provides us plenty of opportunities to utilize its insight for sole purpose of marketing strategies, business development and many more. The first and fore most step for drawing insight, we had to analyze it and represent in readable manner. The discussed Naïve Bayes Classifier is useful machine learning algorithm. It has found efficient in terms of time and space complexity. The attributes of given set if independent in nature, this classifier outperform other types of the classifier. Naive bayes classifier, together with its improved version that is Gaussian Naive bayes classifier, is works well with categorical as well as continuous data. As it is found that, it gives better results after preprocessing, its output performance is enhanced.

In certain cases, it has limitations. If there is introduction of new class then, that class given class probability is zero, it will unable to make classification. The other problem is the assumption of independence in the features which give the classifier its name. If the slightest of dependency exists the classifier will not make the best fit. This challenges the concept of naïve as well as becomes interpersonal to the architect of the classifier to be certain about independence. It works on the assumption of classes are mutually exclusive.

REFERENCES

1. Data Mining and Analysis: Fundamental Concepts and Algorithms:
   Mohammed J. Zaki Wagner Meira Jr.
2. Comparing different supervised machine learning algorithms for disease prediction: (*BMC Medical Informatics and Decision Making).*
   Shahadat Uddin, Arif Khan, Md Ekramul Hossain and Mohammad Ali Moni.
3. https://highdemandskills.com
4. https://data.world
5. Probability & Statistics for Engineers & Scientists:
   Ronald E. Walpole Roanoke College Raymond H. Myers Virginia Tech Sharon L. Myers Radford University Keying Ye University of Texas at San Antonio.
6. Classification of 20 News Group with Naïve Bayes Classifier:
   Abdulwahab O. Adi;Erbuğ Çelebi
7. An Implementation of Naive Bayes Classifier:
   Feng-Jen Yang
8. The Impact of Data Preprocessing on the Performance of a Naive Bayes Classifier:
   Priyanga Chandrasekar;Kai Qian