Assignment 3 Version 2: A mathematical essay on naive bayes classifier

S, Karthik
ME18B149 - IDDD Data Science
Indian Institute of Technology Madras
Chennai, India
me18b149@smail.iitm.ac.in

***Abstract*—This document is an overview of the concept of the naive Bayes classifier. It gives a general idea of the topic, along with the mathematical aspects, as well as an application involving naive Bayes classifier based on data from a real-world problem.**

***Index Terms*—Classification, Naive Bayes, Machine Learning, Data Science**

I. Introduction

Classification is the process of learning a predictive model that relates input features to discrete data classes or categories. This model can then be used to classify a new inputs to a particular class or category. If the outcomes are either positive or negative, then it is referred to as binary classification. If the outputs can take more than two classes, then it is referred to as multi-class classification.

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. 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. Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.

The specific problem discussed is a binary classification problem which determines whether the income of a person is above or below $50000. Naive Bayes classifier will be used to analyze the data and build a model which will then be able to predict the category for unknown data. Various insights and conclusions will be made based on the trends followed by the dataset.

This paper is a case study through which the principles of naive Bayes classifier are implemented. It aims to examine which parameters influence the income of the population obtained through the 1994 Census Data.

II. Naive Bayes Classifier

Naive Bayes is a probabilistic classifier, which means it predicts on the basis of the probability of an object. It is built using Bayes theorem. It works based on the assumption (naive assumption) that all the features are independent of each other. It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other feature

A. Bayes' Theorem

In probability theory and statistics, Bayes' theorem, named after Thomas Bayes, describes the probability of an event, based on prior knowledge of conditions that might be related to the event. Mathematically, it can be given as

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1}$$

where A and B are events and $P(B) \neq 0$

$P(A|B)$ is a conditional probability: the probability of event A occurring given that B is true. It is also called the posterior probability of A given B. $P(A)$ and $P(B)$ are the probabilities of observing A and B respectively without any given conditions; they are known as the marginal probability or prior probability.

B. Model Structure

Suppose y represents the output labels and $\mathbf{x}$ represents vector of input features, the posterior probability of a class given input features can be calculated using Bayes' theorem as follows

$$P(y|\mathbf{x}) = \frac{P(\mathbf{x}|y)P(y)}{P(\mathbf{x})} \tag{2}$$

Based on the probability $P(\mathbf{x}|y)$ and a chosen threshold, the class label of an input can be predicted

The posterior probabilities $P(\mathbf{x}|y), P(y), P(\mathbf{x})$ can be obtained from the data. The naive assumption that is made is that all these features are independent of each other. If $\mathbf{x} = (x_1, x_2, ..., x_d)$ are the d features, then

$$P(y|x_1, x_2, ..., x_d) = \frac{P(x_1, x_2, ..., x_d|y)P(y)}{P(x_1, x_2, ..., x_d)}$$

Since all the features are assumed to be independent of each other, the above equation simplifies to

$$P(y|x_1, x_2, ..., x_d) = \frac{P(x_1|y)P(x_2|y)...P(x_d|y)P(y)}{P(x_1)P(x_2)...P(x_d)} \tag{3}$$

The probabilities in the above formula can be calculated from the data. The conditional probabilities of all the output labels given the input data can be determined and the class corresponding to the maximum probability can be taken as the predicted output class.

C. Advantages

- They are extremely fast for both training and prediction.
- They provide straightforward probabilistic prediction.
- They are often very easily interpretative.
- They have very few (if any) tunable parameters.

D. Types

There are three types of Naive Bayes Model, which are given below:

1) Gaussian: The Gaussian model assumes that features follow a normal distribution. This means if predictors take continuous values instead of discrete, then the model assumes that these values are sampled from the Gaussian distribution.

2) Multinomial: The Multinomial Naïve Bayes classifier is used when the data is multinomial distributed. It is primarily used for document classification problems, it means a particular document belongs to which category such as Sports, Politics, education, etc. The classifier uses the frequency of words for the predictors.

3) Bernoulli: The Bernoulli classifier works similar to the Multinomial classifier, but the predictor variables are the independent Booleans variables. This model is also famous for document classification tasks.

III. The Problem

A. Overview and objectives

This data was extracted from the 1994 Census bureau database by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics). The key task is to determine whether a person makes over $50K a year, adult.csv contains the dataset required to solve the task

In this assignment, we attempt to build a predictive Naive Bayes classification model that answers the question: "what factors of people are more likely to influence the income?" using the above Census data (i.e age, gender, socio-economic class, etc).

B. Reading the data

We will be using python throughout this assignment. A single dataset titled "adult.csv" is provided which contains the details of some of the working professionals including whether they earn less or greater than $50K a year. Since we are creating a model, we would be splitting the dataset into train and validation sets with a fraction of 10% for validation.

IV. Data cleaning

After reading the data, the next task will be to filter out the attributes that are not required or redundant in this analysis. Let us first check the number of unfilled values in each of the parameters.

Table I gives the number of missing entries under every feature.

TABLE I
NUMBER OF MISSING ENTRIES UNDER EACH FEATURE

Feature	Number of missing entries
age	0
workclass	1836
fnlwgt	0
education	0
educational-num	0
marital-status	0
occupation	1843
relationship	0
race	0
gender	0
capital-gain	0
capital-loss	0
hours-per-week	0
native-country	583
income	0

1) Redundant data: Attributes like capital-gain and capital-loss have the value as 0 for most of the population. Hence, including that will be unnecessary in our analysis as it would not affect them.

education-num and education are related very much as the number of years of education is determined closely by the degree that is completed. Hence, education-num becomes redundant in our analysis and can be removed.

Under education, there are too many parameters which makes the model too complicated and would make it difficult to analyze trends. Hence, the entries HS-grad, 9th, 10th, 11th, 12th can be combined to a single parameter. Similarly, 1st-4th, 5th-6th and 7th-8th can be combined to a parameter named elementary.

Similar analysis can be done for marital-status where the number of categories is limited to just 4 - Married, Never-married, Separated, Widowed.

2) Unreported data: It is observed that workclass, occupation and native-country are unreported for some of the entries. All the missing values were filled with the mode of the respective parameter.

V. Exploratory Data Analysis

The datasets contains the population data for all the passengers onboard the Titanic. The complete details of all the attributes are shown in Fig 1.

Our first task is to analyze the data and form visible conclusions and trends from the same with the help of plots.

A. Gender and Age

From Fig 3, we observe that the proportion of females who earn more than 50K a year is much lesser than men. From Fig 4, we can see that most of the people who earn less than 50K a year are around the age of 25 and most of the people who earn more 50K a year are around the age of 40. This is reasonable since higher the age, more likely to be more experienced.

Variable	Definition	Key
age	Age	Continuous
workclass	Work class	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
fnlwgt		Continuous
education	Level of education	Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th,
		7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool
education-num	No. of years of education	Continuous
marital-status	Marital status	Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
Occupation	Occupation	Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces
relationship	Relationship	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
race	Race	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
sex	Gender	Female, Male
capital-gain	Capital gain	Continuous
capital-loss	Capital loss	Continuous
hours-per-week	Working hours / week	continuous
native-country	Native Country	United-States, Cambodia, England . . .

Fig. 1. Exhaustive list of all the attributes in the given dataset

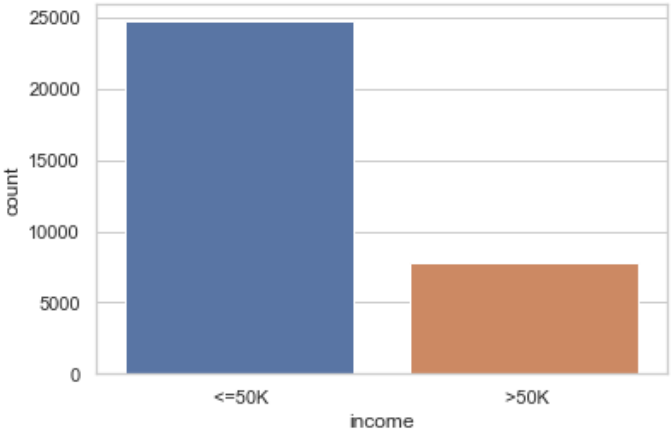

Fig. 2. Bar chart - Population division by Income

B. Occupation

From Fig 6, we can conclude that for the lower income category, the occupation Prof-speciality dominates over the others. Among the higher income population, along with prof-speciality, exec-manegerial also has a relatively high fraction compared to others. This is evident as managerial roles are offered for people with more experience which in turn leads to higher income.

C. Marital Status

From Fig 5, the fraction of people with less than 50K income is highest in non-married people while it is the married

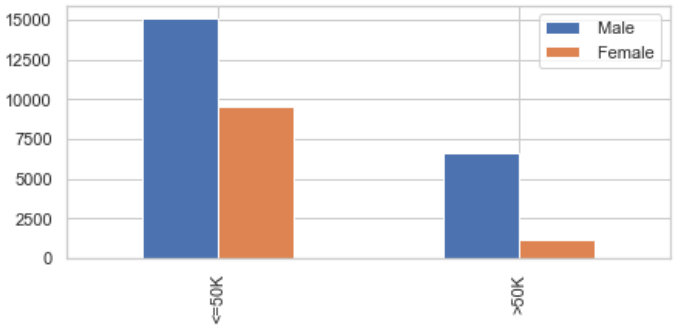

Fig. 3. Bar chart - Genderwise distribution

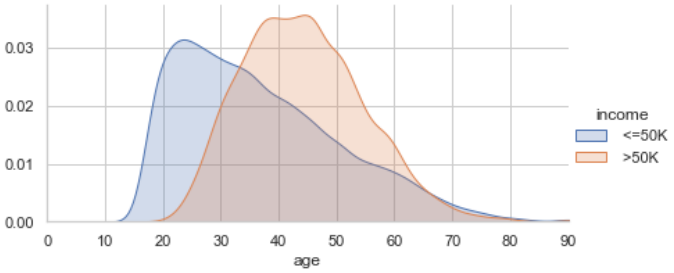

Fig. 4. KDE plot - Age distribution

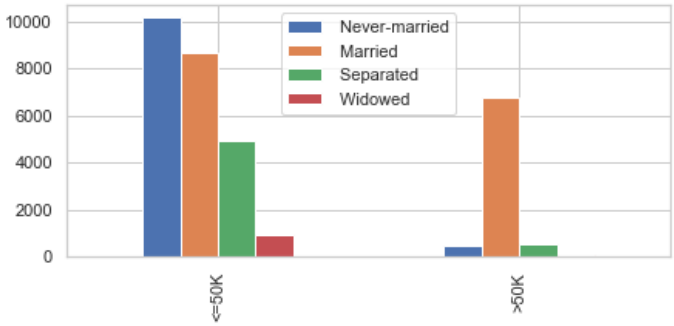

Fig. 5. Bar chart - Marital Status distribution

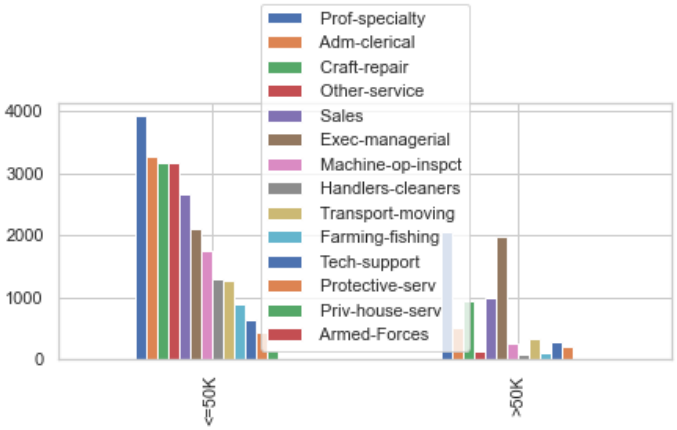

Fig. 6. KDE plot - Occupation distribution

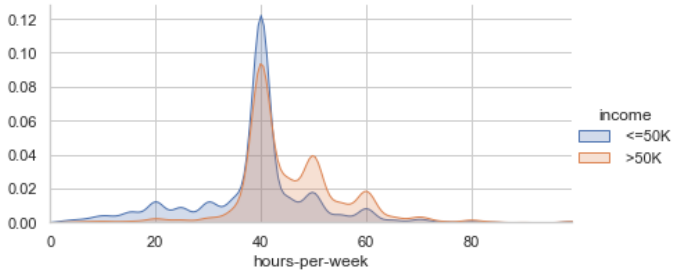

Fig. 7. Correlation matrix - training data

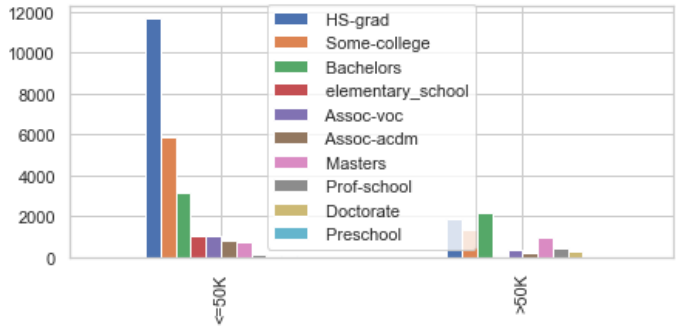

Fig. 8. Correlation matrix - training data

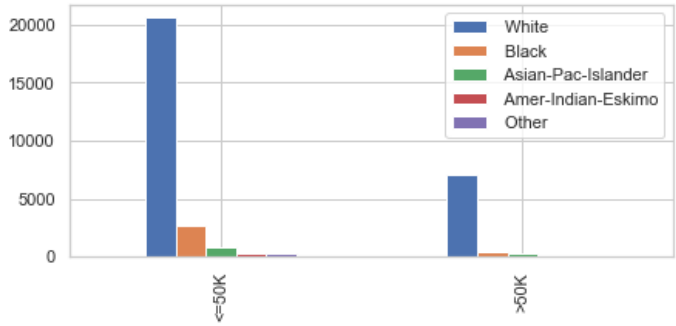

Fig. 9. Correlation matrix - training data

people who almost sweep the entire population with earnings more than 50K. This can be attributed to the fact that typically married people are older than unmarried ones and hence the same correlation as age comes into picture.

D. Other parameters

1) Hours per week: From Fig 7, we can see that for both the income sets, the peak is at the same point. Most of the population works around 40 hours a week irrespective of the income level.

2) Education: From Fig 8, it is clear that the low income population mostly are high school graduates wheras the high income population have mostly either high school graduates or have a Bachelor's degree.

VI. Data preparation

Since there are parameters which are continuous and almost unique if we directly consider the value(Age, hours-per-week) and also parameters which are strings (Embarked), we need to make all the parameters compatible for the logistic regression model.

TABLE II
CATEGORIZATION OF AGE

Greater than	Lesser than	Class number
0	30	0
30	40	1
40	60	2
60	-	3

TABLE III
CATEGORIZATION OF HOURS-PER-WEEK

Greater than	Lesser than	Class number
0	30	0
30	40	1
40	60	2
60	-	3

For Age and hours-per-week, it is necessary to group the values into classes. Based on the distribution (Fig 4 and Fig 7), the parameters were classified as shown in Table II and Table III

Since all other variables were discrete by themselves, it was decided to directly categorize them in terms of integers.

After data preparation, the datasets are compatible for the logistic regression model. Some sample entries of the datasets are shown in Table IV

VII. Model preparation, training and validation

Since the datasets are now model compatible, the next step is to create the model. The entire dataset was split into training and validation datasets with a fraction of 10% kept for validation. Validation accuracy, i.e, the fraction of entries for which the predicted and the actual value are the same will be used as the metric to compare between models. Different classes of Naive Bayes classifiers were implemented whose results are shown as follows Table V

Also, the confusion matrix which is the plot between the actual income and the predicted income is plotted for all the models and are shown in Fig 10, Fig 11, Fig 12.

From these models (Table V), the maximum validation accuracy was observed in the model with Categorical Naive Bayes Classifier with a value of 81.08%. This is expected as the dataset was made with all the variables being categorical in nature.

VIII. Attributes from the confusion matrix

We will try to find some useful parameters for the best model from the results.

TABLE IV
SAMPLE ENTRIES OF MODEL COMPATIBLE DATASET

S.No	age	workclass	fnlwgt	education	marital-status	occupation	relationship	race	gender	hours-per-week	native-country	income
0	1	0	2671	2	1	0	1	4	1	1	38	0
1	2	3	2926	2	0	3	0	4	1	0	38	0
2	1	2	14086	4	2	5	1	4	1	1	38	0
3	2	2	15336	4	0	5	0	2	1	1	38	1
4	0	2	19355	2	0	9	5	2	0	1	4	0

TABLE V
TRAINING AND VALIDATION RESULTS

Model No.	Type	Validation accuracy
1	Gaussian NB	75.78
2	Multinomial NB	75.59
3	Categorical NB	81.08

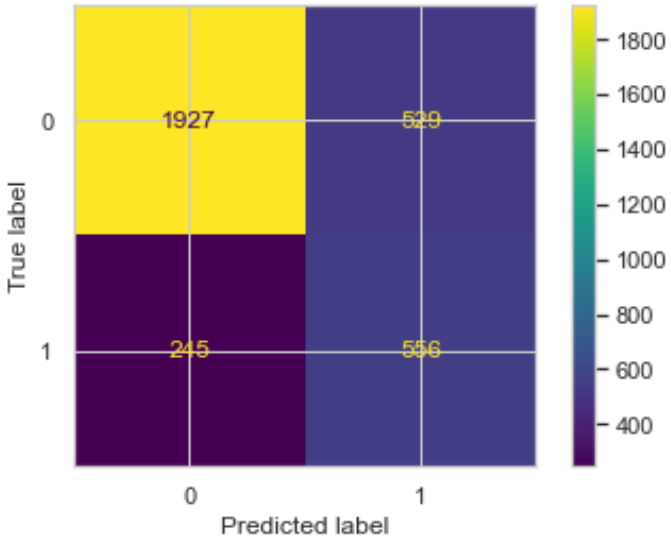

Fig. 10. Correlation matrix - training data

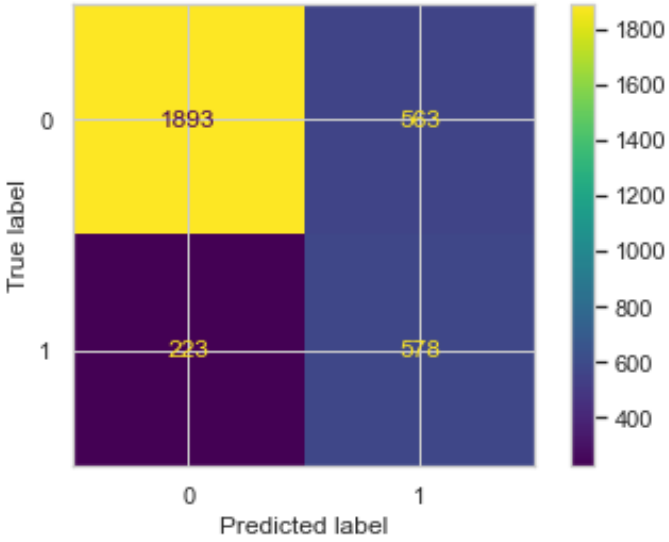

Fig. 11. Correlation matrix - training data

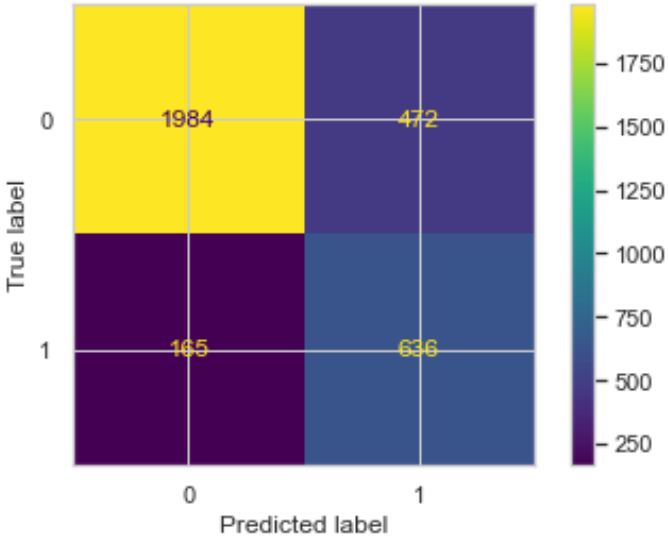

Fig. 12. Correlation matrix - training data

A. Precision

Precision can be defined as the percentage of correctly predicted positive outcomes out of all the predicted positive outcomes. It can be given as the ratio of true positives (TP) to the sum of true and false positives (TP + FP).

So, Precision identifies the proportion of correctly predicted positive outcome. It is more concerned with the positive class than the negative class.

Mathematically, precision can be defined as the ratio of TP to (TP + FP). In the best model, it comes out to be 82.45%.

B. Recall/Sensitivity

Recall can be defined as the percentage of correctly predicted positive outcomes out of all the actual positive outcomes. It can be given as the ratio of true positives (TP) to the sum of true positives and false negatives (TP + FN). Recall is also called Sensitivity.

Recall identifies the proportion of correctly predicted actual positives.

Mathematically, recall can be given as the ratio of TP to (TP + FN). In the best model, it comes out to be 91.63%.

C. True and False Positive Rate

True positive rate is same as recall (TP/(TP+FN)). False positive rate is given as (FP/(FP+TN)). In this model, True Positive rate is 91.63% and False Positive Rate is 41.17%.

TABLE VI
BEST MODEL ATTRIBUTES

S.No.	Attribute	Value
1	Classification accuracy	0.8109
2	Classification error	0.1891
3	Precision	0.8245
4	Recall or Sensitivity	0.9163
5	True Positive Rate	0.9163
6	False Positive Rate	0.4117
7	Specificity	0.5883

IX. ROC CURVES AND AUC

Another tool to measure the classification model performance visually is ROC Curve. ROC Curve stands for Receiver Operating Characteristic Curve. An ROC Curve is a plot which shows the performance of a classification model at various classification threshold levels.

The ROC Curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold levels. ROC AUC stands for Receiver Operating Characteristic - Area Under Curve. It is a technique to compare classifier performance. In this technique, we measure the area under the curve (AUC). A perfect classifier will have a ROC AUC equal to 1, whereas a purely random classifier will have a ROC AUC equal to 0.5.

So, ROC AUC is the percentage of the ROC plot that is underneath the curve. The ROC of all the models along with a logistic regression model is plotted in Fig 13. ROC AUC of our model approaches towards 1. So, we can conclude that our classifier does a good job.

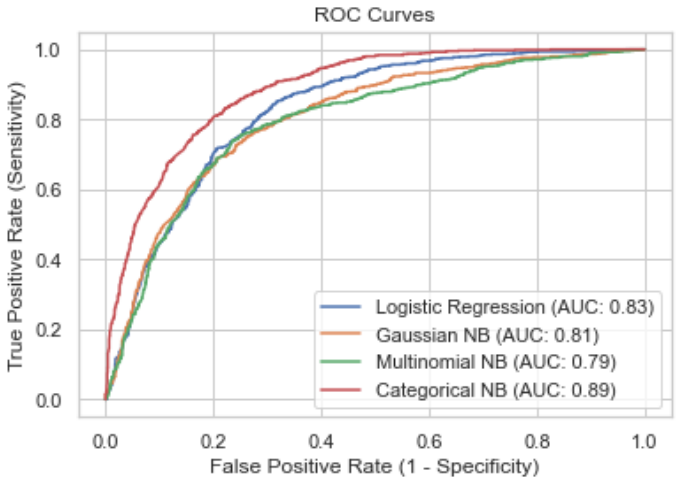

Fig. 13. ROC Curve with AUC values for all the models

X. CONCLUSION

The effects of various factors which could have affected the income of the population involved in the 1994 Census Bureau Database were analyzed in great detail. Various factors greatly affected the income like Gender, Age, education, marital status, occupation, hours per week and even the place from which they are from.

Few Naive Bayes Classification models were constructed on the basis of the known data which gave a validation accuracy of 81.08%. Using this model, the confusion matrix was plotted which gave the exact numbers predicted correctly and incorrectly.

Further avenues could be explored using this data where they can be classified using various other classifiers available to study how the parameters might possibly affect each other and also how they affect the income. Also, implementing techniques more advanced than Naive Bayes classifier might give a model with better accuracy and predictions which would improve on the current model.

REFERENCES

[1] Dr. Kaushik Mitra, Dr. Ramkrishna Pasumarthy, "Naive Bayes Classifier", EE4708: Data Analytics Laboratory - Week 5
[2] Naive Bayes Classifier — Detailed Overview, Towards Data Science.