DAL Assignment 3: A Mathematical Essay On Naive Bayes Classifier

C Vamshi Krishna Reddy
CH20B112
EE5708: Data Analytics Laboratory
IIT Madras, TamilNadu, India
ch20b112@smail.iitm.ac.in

***Abstract*—This document serves as a model for showcasing the application of Naive Bayes to build a model that helps us to predict whether a person makes over 50k dollars per year or not based on parameters like age, gender, work class, race, occupation, level of education, etc.**

***Index Terms*—Classification, Naive Bayes Classifier, statistical modeling**

I. Introduction

Machine Learning problems can be broadly classified as the collection of Classification and Regression. Classification constitutes a wide range of problems tackled with the help of Machine Learning apart from Regression. These models are basically supervised models where the algorithms use labeled datasets and create patterns between input and output data. These algorithms are used to predict the outcomes of new input data. In this paper, we are using a Supervised Learning model which is the Naive Bayes Classifier.

In a Naive Bayes Classifier, a probabilistic approach is followed for the classification tasks and is primarily based on the Bayes theorem. Let us consider two events A and B. When event B has already happened, we may use the Bayes theorem to calculate the likelihood that event A will also occur. Here, event A is known as the hypothesis, and event B is referred to as supporting evidence. When using this algorithm, we assume that the predictors and features are independent. That is, the presence of one feature does not change the behavior of another. So the classifier is called "naive". This classifier is a family of algorithms rather than a single method, and they are all based on the idea that every pair of features being classified is independent of the other. In reality, these assumptions of Naive Bayes, like the independence assumption are practically incorrect but still work pretty well in real-life scenarios.

This paper aims to use Naive Bayes Classifier in explaining the effect of age, gender, work class, education, occupation, and race to determine whether a person makes up 50k dollars per year or not.

The paper provides a detailed overview of Naive Bayes Classifier by solving a real-world problem and gives insights into the relation between variables like Age, gender, work class, level of education, occupation, race, and Income of the person.

II. Naive Bayes Classifer

Classification is a widely employed application in machine learning, especially in supervised settings. It essentially helps us understand how certain independent factors determine which class it belongs to. In simpler terms, it's like a tool used in machine learning to make classifications. The core idea is to teach algorithms how different independent variables are linked to an outcome or dependent variable. Once the model grasps this relationship, it becomes capable of predicting what might happen when faced with new, unexpected data or completing missing parts in existing data.

The naive Bayes model is very easy to develop and remains extremely useful when we encounter very large data sets. Despite its simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods. This classifier is also fast to predict the class of the test data set and it performs extremely well in the case of multi-class prediction. The naive Bayes classifier is also known to perform better compared to other classification models like logistic regression and also uses less training data when the assumption of independence holds true. They perform extremely well in the case of categorical input variables compared to numerical variables. In the case of numerical variables, Naive Bayes assumes normal distribution which may not fit some training data. Based on the type of distribution of the feature, the Naive Bayes classifier is classified into three subgroups: Gaussian, Multinomial, and Bernoulli.

- **Gaussian Naive Bayes:** In this Naive Bayes classifier, all continuous input variables associated with the output feature are assumed to be distributed according to Gaussian (or) Normal Distribution.
- **Multinomial Naive Bayes:** Multinomial Naive Bayes classifier is one of the most popular classifications in supervised learning that is used for the analysis of categorical text data, This classification enables us to classify data, that cannot be represented numerically. The main advantage of this classification is its ability to significantly reduce the complexity of the problem and perform classification using small training sets.
- **Bernoulli Naive Bayes:** As the name suggests this naive Bayes classifier employs the Bernoulli distribution and is used

for discrete data. The primary characteristic of Bernoulli Naive Bayes is that only binary values such as true or false, yes or no, success or failure, 0 or 1 are accepted for features. Therefore, we are aware that we must apply the Bernoulli Naive Bayes classifier when the feature values are binary.

A. Mathematical Equation of Naive Bayes Classifier

Bayes theorem is the foundation of the Naive Bayes classifier. Bayes theorem can be shown below

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

Here A and B refer to a set of events. We use the Bayes theorem to calculate the likelihood that A will also occur given that B has already occurred. Here, A is the hypothesis and B is the supporting evidence. Here, it is assumed that the predictors and features are independent. That is, the presence of one feature does not change the behavior of another so the classifier is termed "naive".

In this model, we have decided to use Gaussian or Normal Naive Bayes. Here each of the numeric variables is assumed to follow a normal distribution and the Bayes theorem can be defined as below:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp \frac{-(x_i - \mu_y)^2}{2\sigma_y^2} \tag{2}$$

B. Cost Function

The naive Bayes classifier model does not involve optimization of a cost function, instead, it finds the class y of the output based on probability as shown below.

$$y = argmax_y P(y) \prod_{i=1}^{n} P(x_i|y) \tag{3}$$

C. Evaluation Model

In classification models like Naive Bayes Classifier, confusion matrix, and f1 score are used to evaluate the performance of the model. Here since our problem is a Binary Logistic regression we have a confusion matrix of 2 X 2 dimension, And F1-Score is defined as the harmonic mean of recall and precision values of the classification problem. The formula for F1-Score is as follows:

$$F1 = \frac{2 * precision * recall}{precision + recall} \tag{4}$$

where Precision measures how many of the "positive" predictions made by the model were correct. Recall measures how many of the positive class samples present in the dataset were correctly identified by the model.

III. The Problem

The objective is to develop a model to predict whether a person makes up 50k dollars per year or not based on factors like age, gender, work class, race, occupation, and level of education. The process initiates with data cleaning and then exploratory data analysis where we find out factors that affect the income of a person. Subsequently, the correlation between these determinants and the response variable, Survival is established.

A. Data Cleaning

The given dataset contains around 33,000 rows and 15 columns. It contains around 5000 duplicates. We have to remove those duplicates as they affect the model. Talking about our data, we have a total of 15 variables of which 9 are categorical variables and the remaining 6 are numerical variables. work class, education, marital status, occupation, relationship, race, sex, native country, and income comes under categorical variables and age, fnlwgt, education num, capital gain, capital loss, and hours per week are numeric. Upon exploring categorical variables we can see that there are several variables like workclass, occupation and native country which contains the missing values which were represented by '?'were replaced with null values. They should be eliminated or modified using common methods such as mean/median/mode estimation-based filling, distribution-based filling, etc. The columns that contain null values are work class, occupation, and native country. These null values are filled with the most frequent values repeated in that column.

B. Exploratory Data Analysis and Visualisations

The provided data set includes person information like age, education, work class, occupation, etc. These details of the person are important and will be used to train a model that can predict whether he/she earns more than 50k dollars, based on the "ground truth" i.e. known data. So our main motive for this project is to develop a naive base classification model using a training data set and then predict the label of income on unseen data.

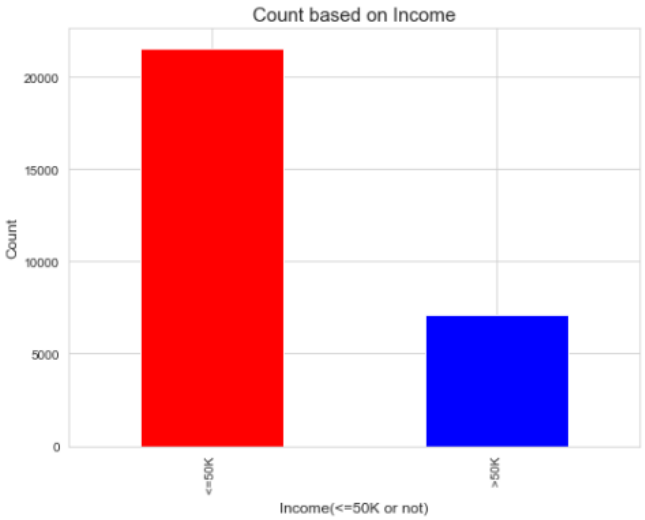

Fig. 1. Count based on Income.

Figure 1 shows a bar graph plot that tells the count of people who has income greater than 50k and less than 50k.

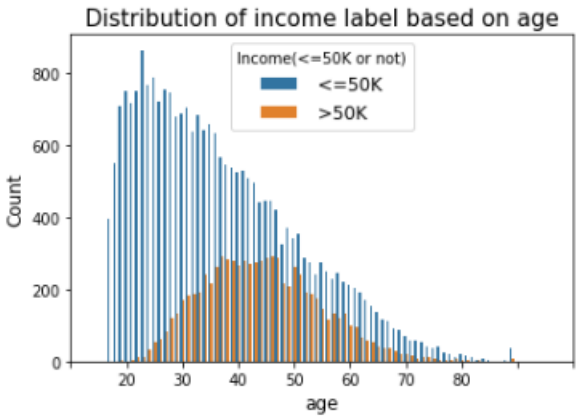

Fig. 2. Distribution of age based on Income.

Figure 2 shows that young people are dominant at having an income less than 50k dollars whereas middle-aged people are dominant at having an income greater than 50k dollars.

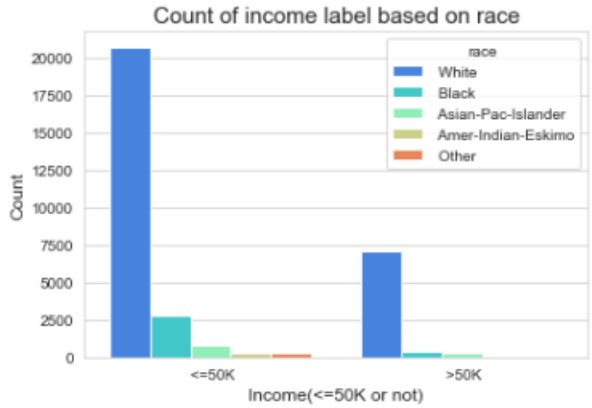

Fig. 3. Count of Income based on race.

Figure 3 shows that white people are dominant in both incomes greater than 50k dollars and less than 50k dollars.

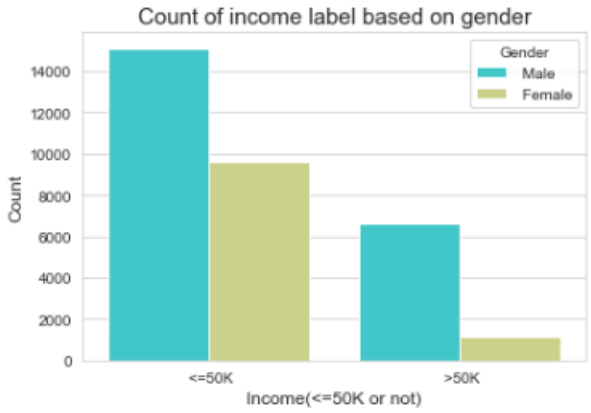

Fig. 4. Count of Income based on gender.

Figure 4 shows the Count of Income based on gender. We see that the proportion of men is significantly higher in the category of income greater than 50k dollars compared to the other category.

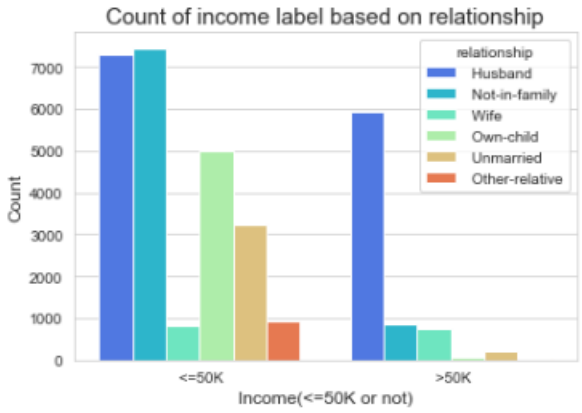

Fig. 5. Count of Income based on relationship.

Figure 5 shows the Count of Income based on the relationship. It helps us to visualize that most of the people who earn more than 50k dollars per year are husbands and they are about 3 times more probable than all the other relationships combined. And also almost all the people who are not in family earn income of less than 50k dollars per year.

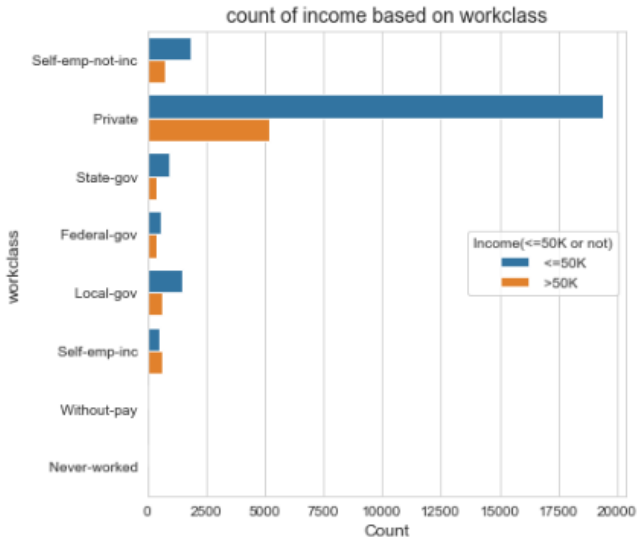

Fig. 6. Count of Income based on workclass.

Figure 6 shows the Count of Income based on the workclass. It shows that most people belong to the private sector and most of them in that private sector earn an income of less than 50k dollars per year.

From Figure 7, we can see near identical distributions for both less than and greater than 50k classes, the reason being that Fnlwgt (referring to final weight) is used to group people of similar characteristics, i.e. people having similar

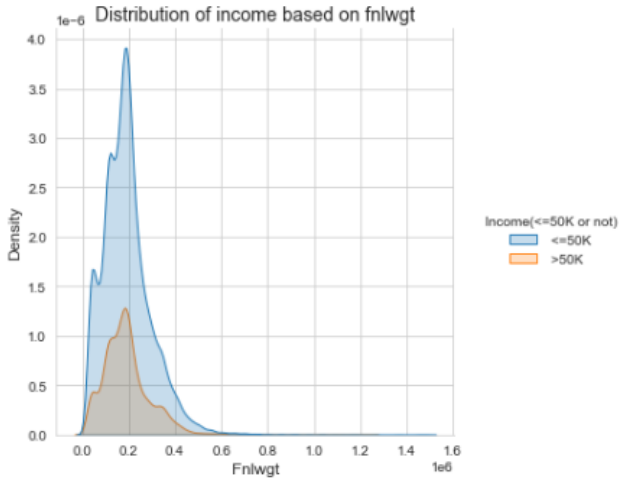

Fig. 7. Distribution of Income based on fnlwgt.

fnlwgt values are more likely to be of the same race, similar educational and social background, etc.. but they are not standardized across different states. So using it will not yield any insights about the income label of the person.

C. Model

Before developing a Naive Bayes classifier, the raw data is transformed into useful features for understanding a little more about our model and increasing the overall efficiency of the mode. This process is also referred to as Feature Engineering. Our first step is to label encode all the categorical variables in the data set. This process is called feature scaling. Now, we split the data set into a test set and train set using the train test split from sklearn model selection with a test size of 0.3. Now a Gaussian naive Bayes classifier is fit on the training data and is tested on the test data set to check its accuracy. Our model has obtained an accuracy of 0.81. The null accuracy score of the given data is 0.7582. As the predicted accuracy of our model is greater than 0.7582, we can conclude that our Gaussian Naive Bayes Classification model is doing a good job of predicting the class labels.

	precision	recall	f1-score	support
<=50K	0.81	0.97	0.88	6531
>50K	0.74	0.30	0.43	2077
accuracy			0.81	8608
macro avg	0.78	0.63	0.65	8608
weighted avg	0.79	0.81	0.77	8608

Fig. 8. Classification Report.

D. Insights and Observations

The heat map is plotted using a clean data set to identify the correlation between different features with the Income column. From this heat map of correlation, we can see that the Income column is primarily correlated with age, gender, marital status, and occupation.

Fig. 9. Heap map of correlation.

From this heat map of correlation, we can see that Income is highly correlated with age, gender, marital status, and occupation. We can see that Income is negatively correlated with marital-status which tells that the people who were married have Income greater than 50k dollars per year.

IV. Conclusion

Analysing the classification report of our project we can say that our Gaussian Naive Bayes Classifier model yields excellent performance in predicting whether a person makes over 50K a year or not. We can conclude that the income of a person will be affected based on race, gender, level of education, marital status, and occupation. The accuracy of our model is 0.8083 on the test data and it is about the same range as the accuracy on the training set, 0.8067 confirming that there are no signs of overfitting in our model. Comparing the model accuracy score,0.8083 with the null accuracy score which is 0.7582 we can confidently conclude that our classifier model is doing a very good job in predicting the class labels.

References

[1] Bishop, Christopher M., Pattern Recognition and Machine Learning, 2006.
[2] https://seaborn.pydata.org/tutorial/categorical.html.
[3] https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c.
[4] https://www.geeksforgeeks.org/naive-bayes-classifiers/.
[5] https://towardsdatascience.com/the-f1-score-bec2bbc38aa6.