# Assignment 3: Naive Bayes Classifier

Rohith D
*Electrical Engg.*
*IIT Madras*
Chennai, India
ee18b148@smail.iitm.ac.in

***Abstract*—This is a study on Data from the 1994 Census Bureau Database. The given data contains information about age, sex, race, occupation, etc. of many individuals. Using Naive Bayes Classifier, we try to determine whether a certain person makes over $50,000 a year.**

***Index Terms*—naive bayes, classification, accuracy, income prediction**

## I. Introduction

The Adult Dataset contains data from the 1994 US Census. It contains 32561 samples of data about individuals as well as ”fnlwgt” which details how much of the population that individual represents. We use this data to predict whether a person earns over $50K.

We do an exploratory analysis of the given data using some charts to get a preliminary idea of the distribution of the data. Using the charts and cleanliness of data, we determine which features to use. The model establishes a Naive Bayes Classifier between the input and output variables. Naive Bayes is a family of probabilistic classifiers based on applying Bayes' theorem with strong independence assumptions between the features. It is highly scalable, and can work with a large number of features.

The key task is to determine whether a person makes over $50K a year. We get an understanding on which socioeconomic features affect income the most.

We start with an overview of Naive Bayes Classifier. Then, we do an exploratory analysis based on some plots. We present our observations based on the data visually. We then analyze the fitted Naive Bayes classfier and run it on a test set. From these plots we arrive at a conclusion.

## II. Naive Bayes

Naive Bayes is a supervised machine learning algorithm. Naive refers to the strong assumption that the input features are independent, and Bayes refer to the Bayes' theorem which is what the algorithm is based on. We use the model to find the posterior probability of output classes, and thus make a prediction based on class with highest posterior probability. The inputs can be either categorical, continuous, or both (mixed). The output is categorical.

Bayes theorem is given as:

$$p(y|X) = \frac{p(X|y).p(y)}{p(X)}$$

where $p(X|y)$ is called the likelihood, $p(y)$ is the prior probability of the output variable (based on the train dataset), $p(X)$ is the prior probability of input.

Independence assumption:

$$p(X|y) = p(x_1|y).p(x_2|y)...p(x_n|y)$$

The Bayes Theorem becomes:

$$p(y|X) = \frac{p(x_1|y).p(x_2|y)...p(x_n|y).p(y)}{p(X)}$$

$$p(y|X) \propto p(x_1|y).p(x_2|y)...p(x_n|y).p(y)$$

For the above posterior, the prediction is done as:

$$\hat{y} = argmax_y(p(y|X))$$

We compare different Naive Bayes Classifiers using accuracy=$\frac{Predictions_{correct}}{Predictions_{total}}$ computed on test datasets.

## III. The Problem

There is a variety of information (features) present in the census data. We look at their distribution and see how they affect income.

We first clean the data before working with it:

- Column containing fnlwgt (final weight) has been removed from the data frame as that data is not used.
- Since the capital gain and capital loss have mostly zeroes, we drop these columns.
- We convert categorical income ($\leq 50K$ - 0, $> 50K$ - 1) into numbers.

To start, we try to draw some preliminary relations from the data using graphs. We check for the distribution of age in a scatter plot, and income according to age.

We see a much higher amount of individuals of age 20-40, and decreasing numbers as age continues increasing. We see that the percentage of high income population is maximum in the 40-50 age range, and lesser as age increases.

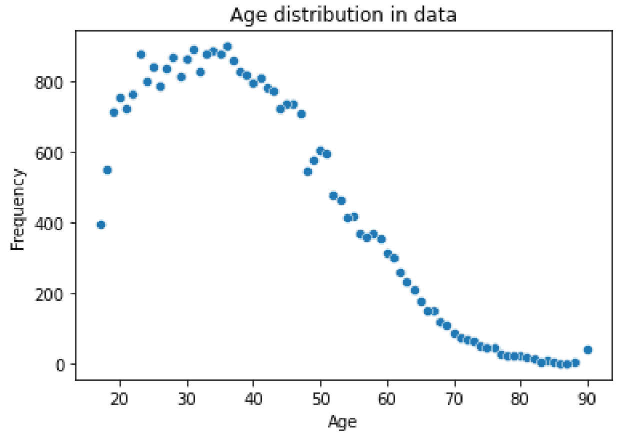

Fig. 1. Age distribution

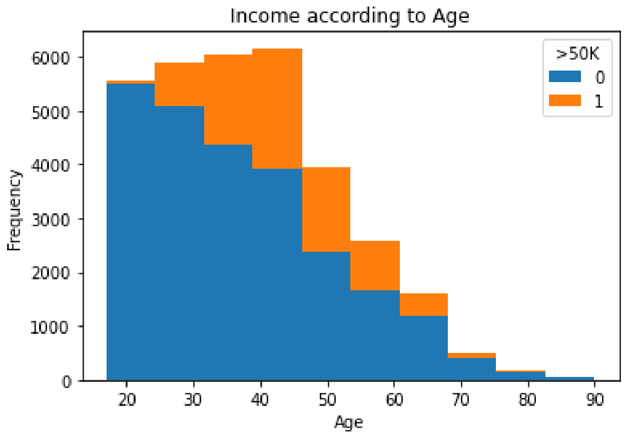

Fig. 2. Income according to Age

Next we analyze the education's effect on income.

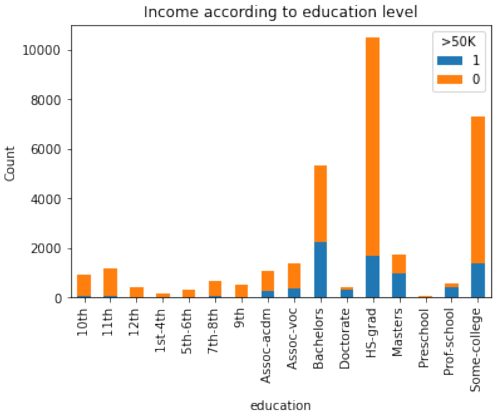

Fig. 3. Income according to Education

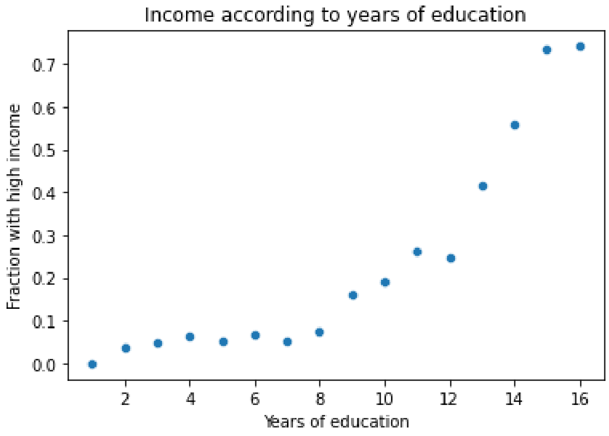

Fig. 4. Income according to years of education

We see that the data present on individuals with education less than 12th is very small. Also, as years of education go up, the average income also seems to go up. However, for those with 8 years of education (12th) or below, the income is fairly equal. The percentage of people with income greater than $50K peaks out at 74% for those with 16 years of education.

We look at distribution of work-class.

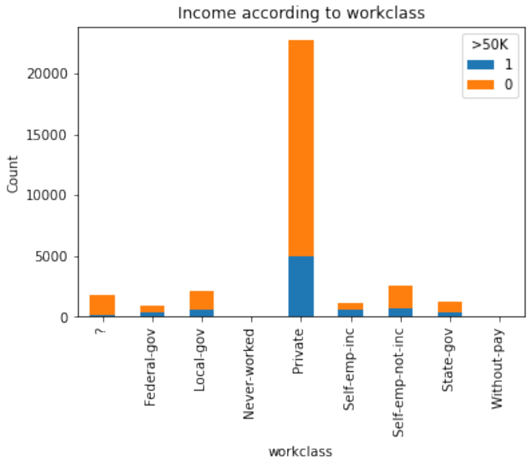

Fig. 5. Work-class distribution

We see that most of the data comes from the private sector, but not much information can be gained from this since the data is so concentrated under one label.

We look at distribution of occupation.

We see that the income levels vary highly among different occupations, so this is an important feature. Executive management and Professional specialities have a particularly high income.

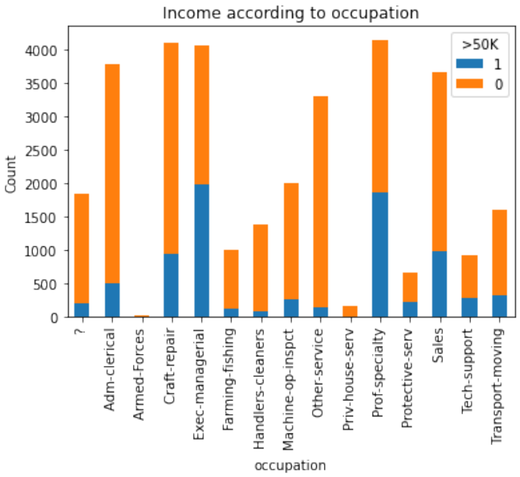

Fig. 6. occupation distribution

We look at distribution of working hours per week for both income classes.

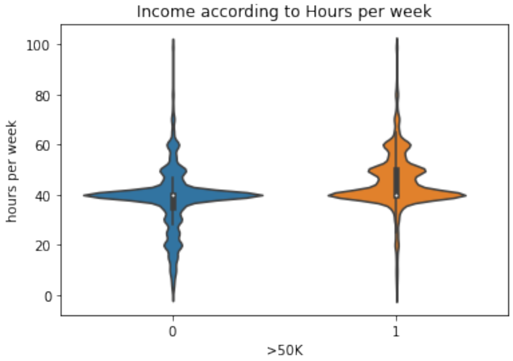

Fig. 7. distribution of hours per week

We see that 40 hours is the most common value. The higher income population has more people who work more than 40 hours per week, whereas the lower income population has more people who work less than 40 hours per week.

We look at distribution according to sex.

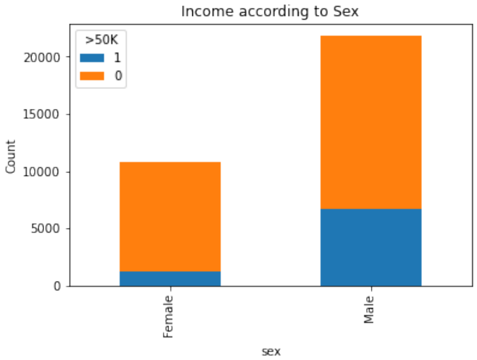

Fig. 8. Sex distribution

A higher percentage of males have income greater than $50K compared to females.

We look at distribution according to race.

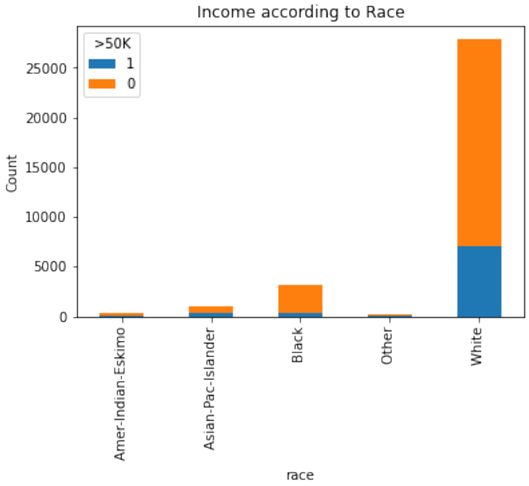

Fig. 9. Racial distribution

We see that most of the data comes from the white population, but not much information can be gained from this since the data is so concentrated under one label.

Lastly, we look at distribution according to Marital Status.

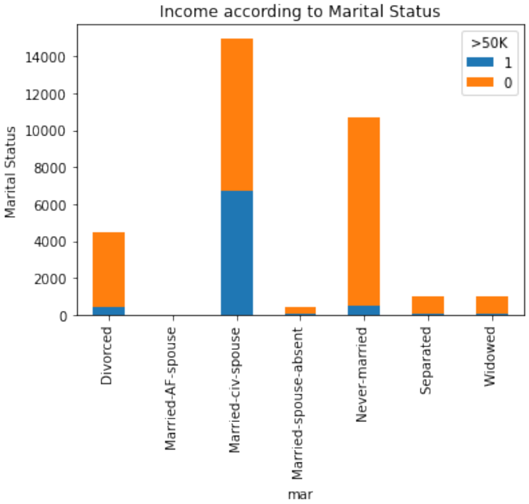

Fig. 10. Distribution according to Marital Status

We see that highest percentage of high income status is in married individuals.

We first separate the dataset into train and test in a 90:10 split to train our Naive Bayes Classifier model and run predictions respectively. We run a mixed Naive Bayes classifier where the continuous features are used in a Gaussian Naive Bayes, and the categorical features are used in a categorical Naive Bayes. The likelihoods of both are multiplied (or log likelihoods are added) to get likelihood of mixed Naive Bayes.

We start models for different combinations of input variables and measure accuracy. We first construct a model with the following inputs: Age, Education, Hours worked per week, Occupation, Sex, Marital Status and Race. We get the accuracy on test set at **78%**.

We now remove the Race input, since it is concentrated on a single label (White) and check accuracy, it goes up to **83.2%**. Since we get a high accuracy on the test data, we use this classifier.

We plot its confusion matrix.

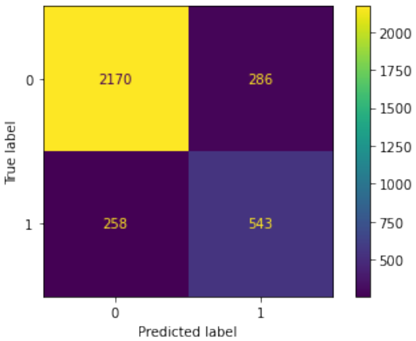

Fig. 11. Confusion Matrix

Now we compute the F1 score on the test data. We get an **F1** score of **0.66**.

## IV. Conclusions

We have seen which features affected the income of individuals in the US and how it is affected. We see that the individuals like to have income over $50,000 are between 40 and 50 years old, with higher years in education, and married. Those with higher hours working per week also more likely to earn more. The female population on average earns less. In Naive Bayes, we see that it is better to use features which have a more even spread, compared to using input features, like race which are concentrated around one label. Further study is possible on increasing sampling of the underrepresented labels in race and work class to get a better representation in all features.

## References

[1] https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[2] https://towardsdatascience.com/why-how-to-use-the-naive-bayes-algorithms-in-a-regulated-industry-with-sklearn-python-code-dbd8304ab2cf

[3] https://www.geeksforgeeks.org/naive-bayes-classifiers/