Assignment-3: A Mathematical Essay on Naive Bayes Classifier

Shania Mitra
Roll Number: CH18B067
Chemical Engineering
IIT Madras, Chennai
ch18b067@smail.iitm.ac.in

***Abstract*—In this study, we examine whether a person makes over $50K a year based on their age, educational qualification, marital status, occupation, race and other factors. A naive bayes model is used to model the importance of these factors and predict the income group of individuals.**

***Index Terms*—Naive Bayes, Visualization, 1994 Census**

I. Introduction

This study empirically analyzes the factors affecting personal income using survey data of the 1994 Census database. We consider education level as an important indicator, and perform classification using the Naive Bayes model. We find a number of factors, such as sex, age, education, and marriage that significantly affect personal income. In addition, differences between different occupations are also investigated. Naive Bayes is a classification technique based on Bayes' Theorem with an assumption of conditional independence among predictors. It assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

In this study, we use Naive Bayes to model the income category of individuals based on education, age, socioeconomic factors, marital status etc.

We begin by gathering, cleaning and preparing the data, following which we perform exploratory analysis. Finally, we build statistical models and perform visualizations to provide quantitative and visual evidence of the relationships observed. In the next section, we highlight the key principles underlying Naive Bayes. In section 3, we discuss the insight and observations drawn from the data and the models. Finally, in section 4 we outline the salient features of the study and present further avenues of possible investigation.

II. Naive Bayes

Naive Bayes is a mathematical technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. All naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.

Naive Bayes is a conditional probability model - given a problem instance to be classified, represented by a vector $\mathbf{x} = (x_1, \ldots, x_n)$ representing some n features (independent variables), it assigns to this instance probabilities

$$p\left(C_k \mid x_1, \ldots, x_n\right)$$

for each of K possible outcomes or classes C_k The problem with the above formulation is that if the number of features n is large or if a feature can take on a large number of values, then basing such a model on probability tables is infeasible. The model must therefore be reformulated to make it more tractable. Using Bayes' theorem, the conditional probability can be decomposed as

$$p\left(C_k \mid \mathbf{x}\right) = \frac{p\left(C_k\right) p\left(\mathbf{x} \mid C_k\right)}{p(\mathbf{x})}$$

Using Bayesian probability terminology, the above equation can be written as

$$\text{posterior } = \frac{\text{prior } \times \text{ likelihood}}{\text{evidence}}$$

In practice, there is interest only in the numerator of that fraction, because the denominator does not depend on C and the values of the features x_i are given, so that the denominator is effectively constant. The numerator is equivalent to the joint probability model:

$$p\left(C_k, x_1, \ldots, x_n\right)$$

Using the chain rule for repeated applications of the definition of conditional probability:

$$\begin{aligned} p\left(C_k, x_1, \ldots, x_n\right) &= p\left(x_1, \ldots, x_n, C_k\right) \\ &= p\left(x_1 \mid x_2, \ldots, x_n, C_k\right) p\left(x_2, \ldots, x_n, C_k\right) \\ &= p\left(x_1 \mid x_2, \ldots, x_n, C_k\right) p\left(x_2 \mid x_3, \ldots, x_n, C_k\right) p\left(x_3,\right. \\ &= p\left(x_1 \mid x_2, \ldots, x_n, C_k\right) \cdots p\left(x_{n-1} \mid x_n, C_k\right) p\left(x_n \mid\right. \end{aligned}$$

The "naive" conditional independence condition assumes that all features in $\mathbf{x}$ are mutually independent, conditional on the category C_k. Under this assumption,

$$p\left(x_i \mid x_{i+1}, \ldots, x_n, C_k\right) = p\left(x_i \mid C_k\right)$$

Thus, the joint model can be expressed as

$$\begin{aligned} p(C_k \mid x_1, \ldots, x_n) &\propto p(C_k, x_1, \ldots, x_n) \\ &\propto p(C_k)\, p(x_1 \mid C_k)\, p(x_2 \mid C_k)\, p(x_3 \mid C_k) \cdots \\ &\propto p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k) \end{aligned}$$

where $\propto$ denotes proportionality. This means that under the above independence assumptions, the conditional distribution over the class variable C is:

$$p(C_k \mid x_1, \ldots, x_n) = \frac{1}{Z} p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k)$$

where the evidence $Z = p(\mathbf{x}) = \sum_k p(C_k)\, p(\mathbf{x} \mid C_k)$ is a scaling factor dependent only on $x_1, \ldots, x_n$, that is, a constant if the values of the feature variables are known. Constructing a classifier from the probability model [edit] The discussion so far has derived the independent feature model, that is, the Naive Bayes probability model. The Naive Bayes classifier combines this model with a decision rule. One common rule is to pick the hypothesis that is most probable; this is known as the maximum a posteriori or MAP decision rule. The corresponding classifier, a Bayes classifier, is the function that assigns a class label $\hat{y} = C_k$ for some k as follows:

$$\hat{y} = \underset{k \in \{1, \ldots, K\}}{\operatorname{argmax}}\, p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k)$$

-

III. The problem

In this study, we examine whether a person makes over $50K a year based on their age, educational qualification, marital status, occupation, race and other factors.

A. Data Preparation

The training dataset used in this study consists of 32561 people and 14 features. Interpretation of the features is as follows:

- Age: Age of the person; ranges from 17 to 90
- Workclass: Category of employment of individual such as private, without-pay, state govt, etc.
- Fnlwgt
- Education: Level of education of the individual
- Education Years: Number of years of education of the individual
- Occupation: Occupation of the individual
- Relationship: Relationship of individual in the family
- Race: Race to which the individual belongs
- Sex: Gender of the individual
- Capital Gain, Loss
- Working hours: Hours per week the indivual works for on an average
- Native country: Background of the individual

Missing values are present in Workclass and occupation and are denoted by " ?".Upon dropping them, the performance of the model deteriorates and hence, they are not dropped but treated as a separate category. The aim is to predict the binary feature wage which is 1 if wage earned by the individual is greater than $50K/year and 0 otherwise.

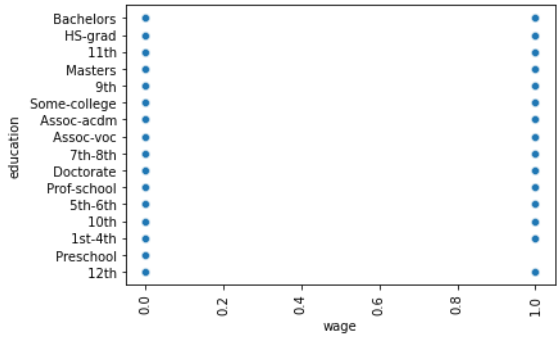

Fig. 1. Plot showing presence of individuals having various educational backgrounds in each wage class

B. Exploratory Analysis

In this section we look at the trends between various features.

A dot in the figures below (Fig-1 to Fig-3) indicate the presence of atleast 1 individual having the designated x and y coordinates.

In figure 1, which shows us the wages earned by individuals having various educational backgrounds. We can clearly see that no individuals having gone through only preschool receive wages above $50K. For all other degrees, individuals receiving both below and above $50K exist.

Similarly, in Figure 2, which plots workclass vs race, we observe that people from all races are employed in all jobs. Howver, among those who have never worked or are working without pay, only whites and blacks are found. Similarly, among those who work without pay, Whites, Blacks and Asians can be found. Further, Amer-Indian Eskimos and Blacks among others have the highest proportion of people with wages below $50K annually.

race	wage	wage
Amer-Indian-Eskimo	<=50K	0.884244
	>50K	0.115756
Asian-Pac-Islander	<=50K	0.734360
	>50K	0.265640
Black	<=50K	0.876120
	>50K	0.123880
Other	<=50K	0.907749
	>50K	0.092251
White	<=50K	0.744140
	>50K	0.255860

Figure 3 plots educational qualification vs workclass. We observe that for most government and private jobs all educational qualifications except preschooling suffice. However, as expected, we observe that those who have never worked belong to junior and middle school, while those who work without pay mostly belong to middle and high school.

In figure 4, we plot the distributions of working hours for each age. The green triangles indicate the mean of the working

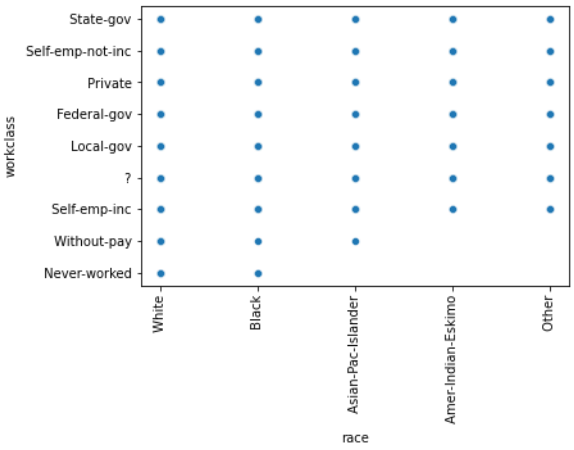

Fig. 2. Plot showing presence of individuals of different races in each work class

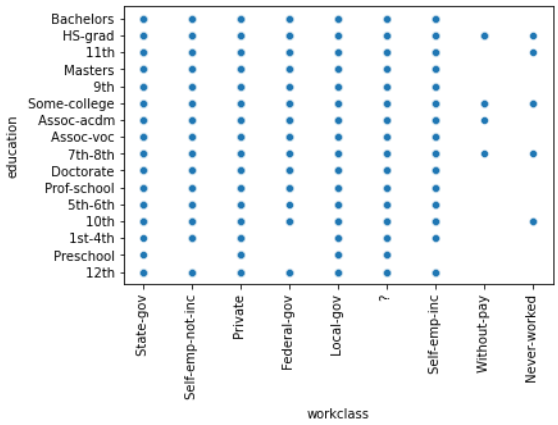

Fig. 3. Plot showing presence of individuals having various educational backgrounds in each work class

hours distribution for each age. Connecting the means we see that at lower ages (17-25) the average working hours are lower than 40hrs and increases with increasing age. From 25-61, the peak working age, working hours plateau at around 45 hours/week. Beyond 61, working hours decrease with increasing age, as individuals get older and their working capacity reduces.

In figure 5, we plot the distributions of working hours vs gender and observe that males have higher average working hours per week as compared females.

In figure 6, we plot the count of individuals of different workclasses having wages less than $50K in blue and those with greater than $50K in orange. From this we clearly see that maximum people in the population work in the private sector. Further, for all classes except those self employed, more people earn below $50K than above.

In figure 7 and 8, we plot the distributions of working hours and number of years of education for each of the two genders. We observe that the mean working hours for males is larger than that of females. However, the median number of years of education for both males and females is the same.

In figures 9 and 10, we plot the distributions of working hours and number of years of education for each of the two wage classes. We observe that the mean working hours as well as number of years of education for wage class >$50K is larger than that of wage class <$50K.

In figure 11, we plot the distributions of ages for people who earn below and above $50K/year. We can clearly see that the median age of those who earn above $50K is higher than those who earn below $50K.

In Figure 12 we plot the count of people of both genders for each of the two wage classes. It can be seen that the number of males in both wage classes is higher than that of females. Further, the ratio of people with income above $50K to those below $50K is larger for males as compared to females, as can be seen below as well as in the figure.

		wage
sex	wage	
Female	<=50K	0.890539
	>50K	0.109461
Male	<=50K	0.694263
	>50K	0.305737

In Figure 13, we plot count of people with different number of education years for each of the two wage classes. We observe that there are very few people with edu-years <=8 and the percentage of them earning more than $50K is extremely low. For edu-years > 9-13, the number of people increases but the number of people in wage class 1 is lower than that in wage class 0. For people with more than 14 years of education however, more people earn above $50K than below.

Finally, we observe that husbands and wives have the largest proportion of individuals earning more than $50K while those that live separately or are unmarried have the lowest proportion. This may be due to the fact that husbands and wives need to earn enough to make ends meet for themselves as well as their children, while those that are not in family need to earn enough to fend for themselves only.

		wage
relationship	wage	
Husband	<=50K	0.551429
	>50K	0.448571
Not-in-family	<=50K	0.896930
	>50K	0.103070
Other-relative	<=50K	0.962283
	>50K	0.037717
Own-child	<=50K	0.986780
	>50K	0.013220
Unmarried	<=50K	0.936738
	>50K	0.063262
Wife	<=50K	0.524872
	>50K	0.475128

C. Model

To train the model and gauge its performance, the dataset was split into train and test sets with an 80-20 split. Following are the metrics achieved by the Naive Bayes Model:

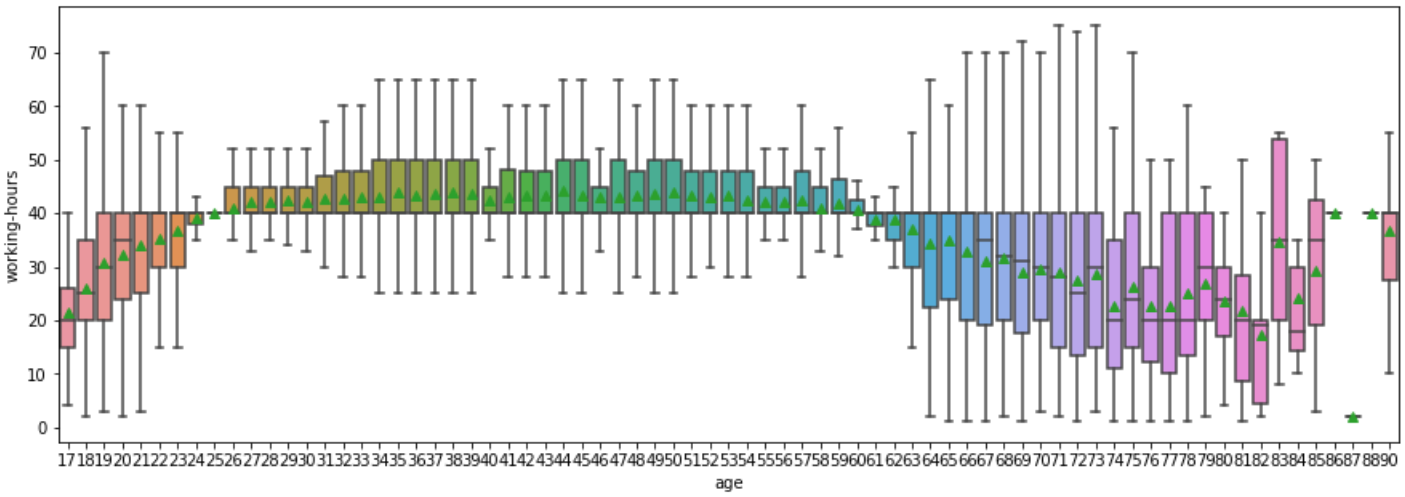

Fig. 4. Distribution of working hours for each age; green triangles signify the mean

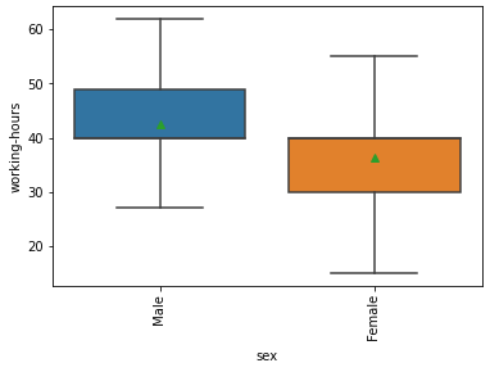

Fig. 5. Distribution of working hours vs gender

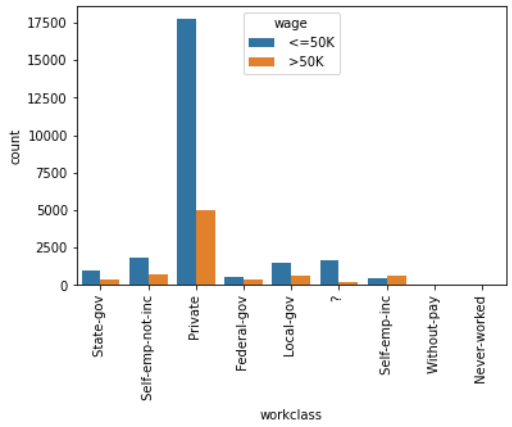

Fig. 6. Count of individuals of each wage class belonging to different workclasses

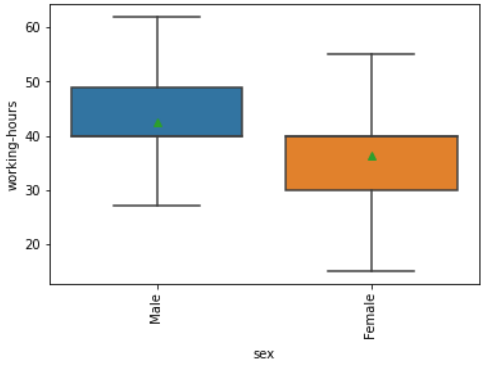

Fig. 7. Distribution of working hours for each gender

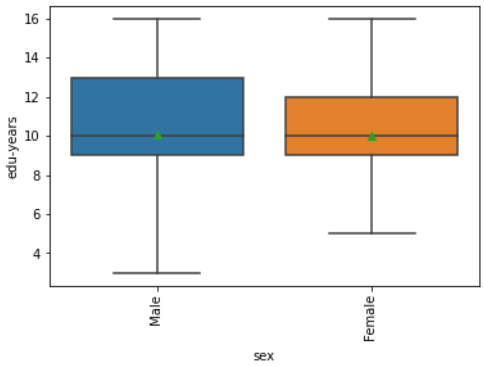

Fig. 8. Distribution of number of years of education for each gender

	precision	recall	f1-score	support
0	0.81	0.95	0.88	12323
1	0.68	0.32	0.43	3958
accuracy			0.80	16281
macro avg	0.74	0.63	0.65	16281
weighted avg	0.78	0.80	0.77	16281

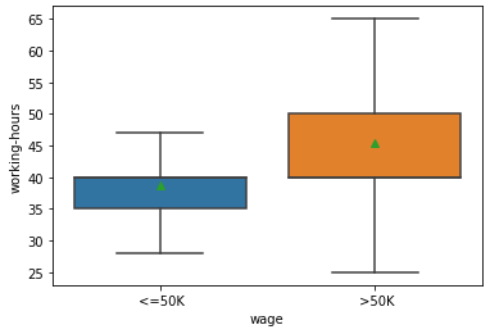

Fig. 9. Distribution of working hours for each wage class

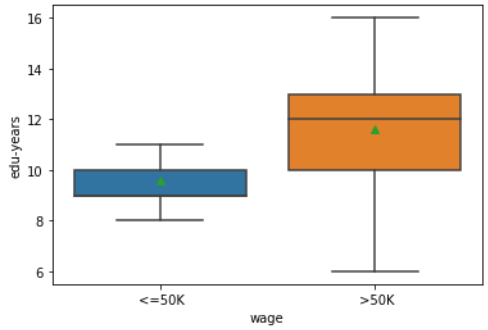

Fig. 10. Distribution of number of years of education for each wage class

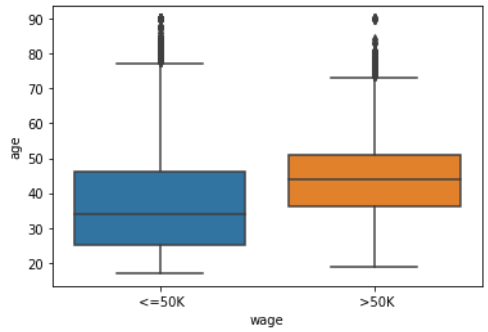

Fig. 11. Distribution of ages of individuals of different wage classes

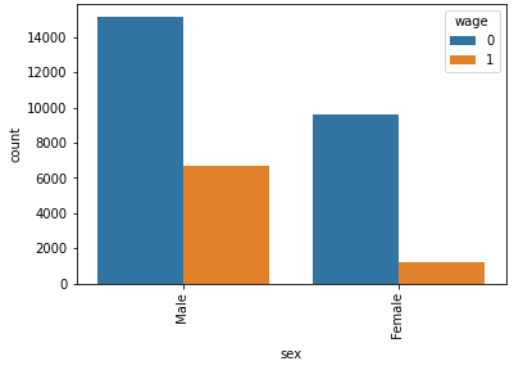

Fig. 12. Count plot for each gender split by wage class

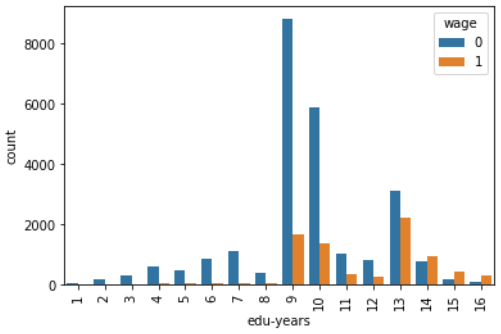

Fig. 13. Count plot of each education year split by wage class

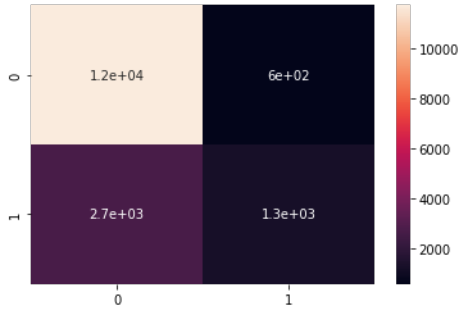

Fig. 14. Confusion Matrix for Validation Set

Thus, for the people who earn below $50K, we achieve an f1-score of 0.88 while for the people who earn above $50K, we achieve an f1-score of 0.43. Support refers to the number of samples in each class. For the 16281 validation samples, from the confusion matrix in Figure 14 we can see that many samples truly belonging to wage class 1 are classfied as wage class 0. This may be because the number of samples of wage class 1 are very few and the model is not able to learn properly.

IV. Conclusions

In this study, we observed that male, aged individuals (age>=45) with higher working hours are more likely to earn wages above $50K per annum. Further, most individuals study for around 9 years, but those having studied for more than 14 years and those that are self employed are most likely to earn wages above $50K per annum. Finally, we learnt that on an average, women and men are equally educated, however, women work for fewer hours and receive lower salaries. To improve the performance of the model, in future, non-linear models can be used.

References

[1] Naive Bayes classifier. (2021, October 21). In Wikipedia. https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[2] Zhang, Z. (2020, February 6). Naive Bayes Explained - Towards Data Science. Medium. Retrieved October 21, 2021, from https://towardsdatascience.com/naive-bayes-explained-9d2b96f4a9c0

[3] Ray, S. (2021, August 26). Learn Naive Bayes Algorithm | Naive Bayes Classifier Examples. Analytics Vidhya. Retrieved October 21, 2021, from https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/