# Naive Bayes: Using Bayesian Statistics for Classification

Vaibhav Mahapatra
*Data Science Department*
*Indian Institute of Technology, Madras*
Chennai, India
me19b197@smail.iitm.ac.in

***Abstract*—Income classification is an important step in public policy processes , wherein we collect and analyze the data collected via census surveys to determine different classes of beneficiaries. Most of the classification approaches focus on feature selection or reduction. Some features are irrelevant and redundant, resulting in a lengthy detection process and degrades the classifier's performance. The purpose of this study is to identify essential input features in building an effective income classifier. In this article, we aim to discover the relationship between different features and an individual's income. We will build a Naive Bayes Classifier on some given data to determine whether an individual earns more than $50k annually or not.**

**Improvements: Added mathematical theory on Linear Discriminant Analysis to the Naive Bayes Classifier Section, added more conceptual visuals.**

***Index Terms*—Binary Classification, Bayes theorem, conditional probability, Gaussian distribution, features, predictors.**

## I. Introduction

Classification of objects with many features and classes becomes difficult because estimating the probabilities would necessitate an enormous number of observations. The effect of a variable value on a given class is assumed to be independent of the values of other variables by Naive Bayes classifiers. This is known as class conditional independence. It is designed to simplify computation and is therefore considered naive. The assumption is relatively broad and is not always applicable. However, various studies have benchmarked the Naive Bayes classifier's performance as equivalent to those of Classification trees and Neural Networks. When applied to large datasets, they have also demonstrated high accuracy and speed.

Naive Bayes Classifier is a classification technique where the target variable is categorical in nature. Abstractly, it is a *conditional probability model* on given data. Given a data instance to be classified, represented by a vector $X_i = (x_{i1},\ldots,x_{in})$ representing $n$ independent features, it assigns to this instance a probability $p(C_k|X_i)$ for each of $K$ possible classes $C_k$. The probability $p(C_k|X_i)$ is calculated using the Bayes' Theorem which is stated mathematically as:

$$p(C_k|X_i) = \frac{p(C_k)p(X_i|C_k)}{p(X_i)}$$

where,
$p(C_k|X_i)$: Posterior probability
$p(C_k)$: Prior probability
$p(X_i|C_k)$: Likelihood
$p(X_i)$: Evidence
The data instance is then assigned with a class where the probability $p(C_k|X_i)$ is highest.

This document looks at a conditional probability model that works on data extracted from the 1994 Census bureau database by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics). The key focus here is to analyze the impact of various factors on an individuals income, identifying the most relevant features and using the data to create a naive Bayes classifier which will be able to bucket individuals into 2 buckets, those with more than $50k annual income and those with less. Ultimately, we would like to understand the data characteristics which affect the performance of naive Bayes classifier, and hence understand how each feature is related to each class.

In the next section, we will establish the necessary background and definitions. Section II focuses on the data description, analysis, handling of missing values and visualization to gain more insights into the problem, while Section III demonstrates that the use of *Naive Bayes Classifier*. A summary and conclusions are given in Section V.

## II. Naive Bayes Classifier

This section focuses on the mathematical aspect of how to classify vectors of discrete-valued features, $x \in 1,\ldots,K^D$, where $K$ is the number of values for each feature, and $D$ is the number of features. Naive Bayes is a generative approach as we aren't drawing any sort of a decision boundary. It requires us to specify the conditional class distribution, $p(x|y = c)$. The most straightforward approach is to assume the features are conditionally independent given the class label. Therefore the conditional class density as a product of one-dimensional densities can be written as:

$$p(x|y = c, \theta) = \prod_{j=1}^{D} p(x_j|y = c, \theta_{jc}) \tag{1}$$

The resulting model is called a **naive Bayes classifier (NBC)**. The class-conditional density depends on the type of each feature. Some probability distributions are described below:

- For real valued features Gaussian Distribution is used i.e. $p(x|y = c, \theta) = \prod_{j=1}^{D} \mathcal{N}(x_j|\mu_{jc}, \sigma_{jc}^2)$ where $\mu_{jc}$ is the

mean of feature $j$ in objects of class $c$, and $\sigma_{jc}^2$ is its variance.

- A Bernoulli distribution is employed in case of binary categorical features i.e. $p(x|y=c,\theta) = \prod_{j=1}^{D} \mathrm{Ber}(x_j|\mu_{jc})$ where $\mu_{jc}$ is the probability that feature $j$ occurs in class $c$. This is known as **multivariate Bernoulli naive Bayes** model.
- For categorical features, $x_j \in 1,\ldots,K$, we can model using the Multinoulli distribution: $p(x|y=c,\theta) = \prod_{j=1}^{D} \mathrm{Cat}(x_j|\mu_{jc})$, where $\mu_{jc}$ is a histogram over the $K$ possible values for $x_j$ in class $c$.

### A. Maximum Likelihood Estimate for NBC

Probability for a single data is written as

$$p(x_i, y_i|\theta) = p(y_i|\pi)\prod_j p(x_{ij}|\theta_j) = \prod_c \theta_c^{\mathbb{I}(y_i=c)}$$

$$\prod_c \theta_c^{\mathbb{I}(y_i=c)} = \prod_c \prod_j p(x_{ij}|\theta_{jc})^{\mathbb{I}(y_i=c)} \tag{2}$$

Therefore, log-likelihood can be written as

$$\log p(\mathcal{D}|\theta) = \sum_{c=1}^{C} N_c \log \pi_c + \sum_{j=1}^{D}\sum_{c=1}^{C}\sum_{i:y_i=c} \log p(x_{ij}|\theta_{jc}) \tag{3}$$

(3) shows MLE getting decomposed into two terms, one concerning $\pi$ and other having $D \times C$ terms containing $\theta_{jc}$'s. Hence both terms can be optimized separately. The MLE for the class prior is given by the

$$\hat{\pi}_c = \frac{N_c}{N}; N_c = \sum_i \mathbb{I}(y_i = c) \tag{4}$$

The MLE for the likelihood terms depends on the assumption of type of distribution of each feature. In case of binary assumption ($x_j|y=c \sim \mathrm{Ber}(\theta_{jc})$) the MLE can is obtained as

$$\hat{\theta}_{jc} = \frac{N_{jc}}{N_c} \tag{5}$$

Models of this form are much more manageable, since they use prior probabilities of classes and independent probability distributions. If there are k classes and if a model can be expressed in terms of r parameters, then the corresponding naive Bayes model has (k - 1) + nrk parameters. In practice, often $k = 2$ (binary classification) and $r = 1$ (Bernoulli variables as features) are common, and so the total number of parameters of the naive Bayes model is $2n + 1$, where n is the number of binary features used for prediction.

### B. Making a prediction with the Model

At test time, the goal is to compute

$$p(y = c|x, \mathcal{D}) \propto p(y = c|\mathcal{D}) \prod_{j=1}^{D} p(x_j|y = c, \mathcal{D}) \tag{6}$$

The correct Bayesian procedure is to integrate out the unknown parameters. Fortunately, this is easy to do, at least if the posterior is Dirichlet. In particular, the posterior predictive density can be obtained as $\frac{\alpha_j + N_j}{\alpha_0 + N}$. By simply plugging in the posterior mean parameters $\bar{\theta}$. Hence

$$p(y = c|x, \mathcal{D} \propto \bar{\pi}_c \prod_{j=1}^{D} (\bar{\theta}_{jc})^{\mathbb{I}(x_j=1)} (1 - \bar{\theta}_{jc})^{\mathbb{I}(x_j=0)} \tag{7}$$

$$\bar{\theta}_{jc} = \frac{N_{jc} + \beta_1}{N_c + \beta_0 + \beta_1} \tag{8}$$

$$\bar{\pi}_c = \frac{N_c + \alpha_c}{N + \alpha_0} \tag{9}$$

where $\alpha_0 = \sum_c \alpha_c$.

If tha psoterior is approximated by a single point, $p(\theta|\mathcal{D}) \approx \delta_{\hat{\theta}}(\theta)$, where $\hat{\theta}$ may be the ML or MAP estimate, then the posterior predictive density is obtained by simply plugging in the parameters, to yield a virtually identical rule:

$$p(y = c|x, \mathcal{D}) \propto \hat{\pi}_c \prod_{j=1}^{D} (\hat{\theta}_{jc})^{\mathbb{I}(x_j=1)} (1 - \hat{\theta}_{jc})^{\mathbb{I}(x_j=0)} \tag{10}$$

The only difference that the posterior mean $\theta$ is replaced with the posterior mode or MLE $\hat{\theta}$. However, this small difference can be important in practice, since the posterior mean will result in less over-fitting. The corresponding classifier from (6) can also be simply written as follows

$$\mathrm{classify}(f_1, \ldots, f_n) = arg\max_c p(C = c) \prod_{i=1}^{D} (F_i = f_i|C = c) \tag{11}$$

Although far-reaching independence assumptions are frequently incorrect, the decoupling of the class conditional feature distributions means that each distribution can be estimated independently as a one-dimensional distribution. This, in turn, aids in alleviating problems caused by the curse of dimensionality, such as the requirement for data sets that scale exponentially with the number of features. Like all probabilistic classifiers under the MAP decision rule, it arrives at the correct classification as long as the correct class is more probable than any other class; thus, class probabilities do not have to be estimated very well. [2]

### C. Linear Discriminant Analysis

A special case of the Naive Bayes model has a specialised name: Linear Discriminant Analysis. This is for the case where the prior class conditionals share the same variance but different means. By simplifying the math, we can correlate the score of x belonging to class k as follows:

$$\delta_k(x) = log(\pi_k) + x^T \Sigma^{-1} \mu_k - \frac{\mu_k^T \Sigma^{-1} \mu_k}{2} \tag{12}$$

As the above formula, a given x vector is classified into the class which gives the highest $\delta$ value. If we play closer attention, we notice that is it a linear function, implying linear decision boundaries. A sample case can be visualised in Fig. 1.

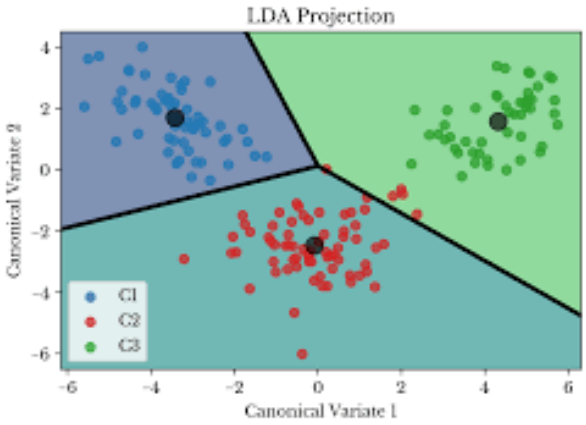


Fig. 1. Sample case for Linear Discriminant Analysis

## III. Applications in Real-Life Scenario

### A. Data Description and Missing Values Handling

The dataset comprises information about households divided into two categories based on their annual income, i.e. the households who earn more than 50K and those who are below the 50k mark. Information about the educational background, relationship status, sex, race, native country, change in capital, the working class is also given. The analysis aims to devise a generalised model that can determine whether a person earns more than 50K annually or not. Once the data is imported it is highlighted that there are missing values in the columns containing information about *workclass (5.6%)*, *occupation (5.7%)* and *native country (1.8%)*. From Fig. 2 it is observed that most people belong to the private working-class, works in an Exec-managerial role and belong to the United States. The number of missing values is significantly less. Therefore they had been imputed with the corresponding mode values.

### B. Exploratory Data Analysis

The following section will focus on the relationship between different features and identify the most suited ones for modelling the target variable. The number of people from each category who earn more than 50K is used a parameter to identify the relationship between income and given variables.

From Fig. 3 it can be conclude that the individuals with significant amount of education (like Bachelors, Masters, Doctorate, Prof-school) are more likely to fell in the category of earning more than 50k annually. On the other hand, from Fig. 4 it is observed both husband and wife are more likely to generate higher income in comparison to other groups.

It is observed that the number of high earning males is more than twice that of the number of females. In Fig. 5 the country-wise distribution of high income generating household is given. People from countries like India, Germany, France, England are more likely to earn more than$50K annually. In contrast, nearly all the nations have a higher number of males with a high income, validating our earlier observation.

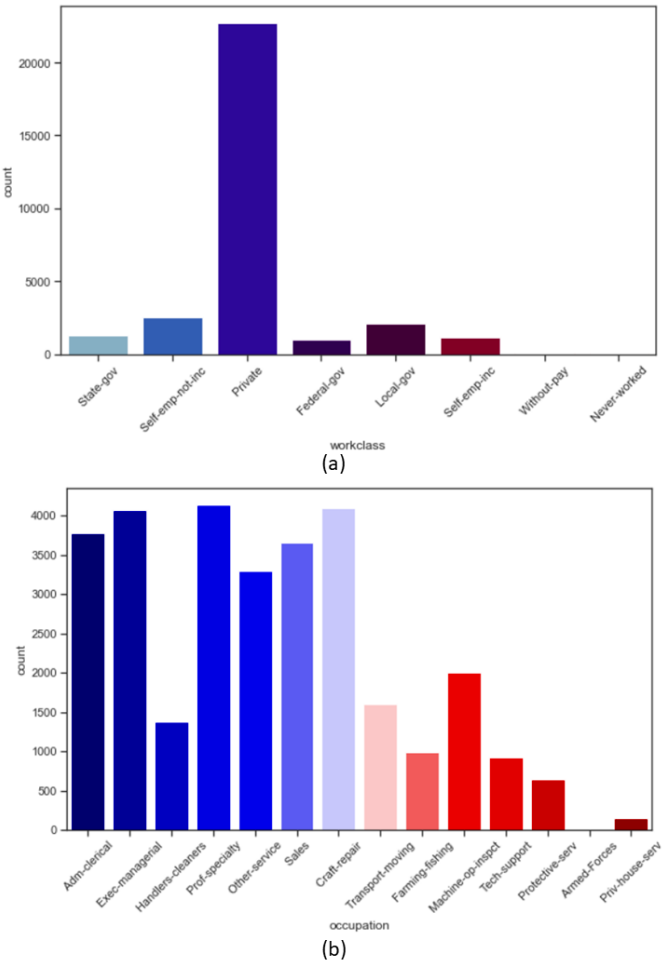


Fig. 2. (a) Frequency of individuals from different working class; (b) Frequency of individuals from different occupational backgrounds

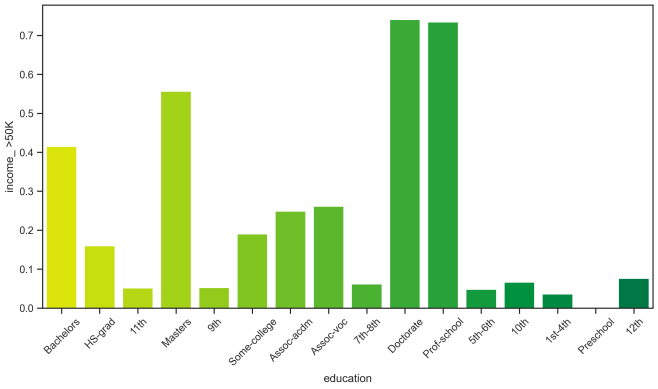


Fig. 3. People earning more than 50k based on education level

On analysing the distribution of income concerning the working class of people, it is observed that a higher percentage of self-employed people are earning more than $50 K per year compared to other groups. This evidence is against the fact that the most number of people works in the private sector. From Fig. 6 showcases that African-American people

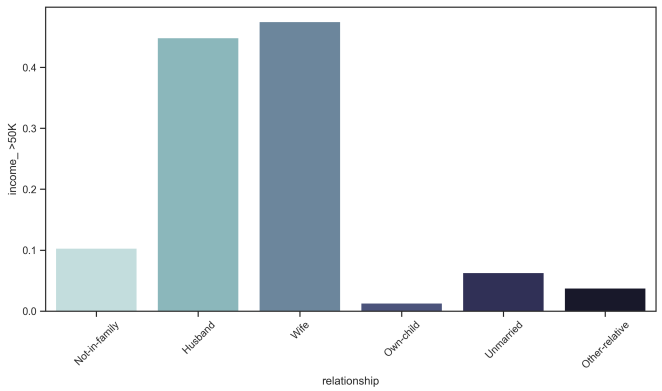


Fig. 4. Relationship vs high income distribution

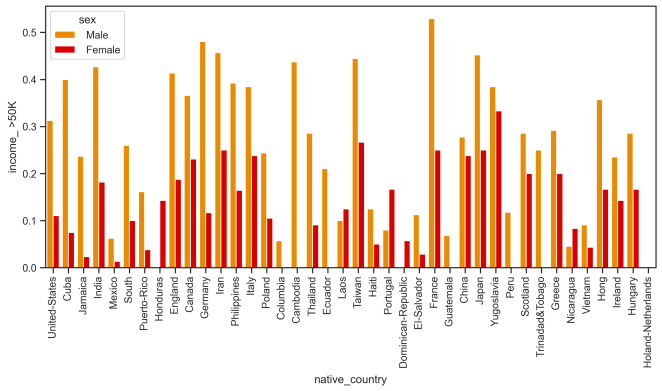


Fig. 5. Gender wise distribution of individuals earning more than $50K based on native countries

are less likely to earn a higher income than people from other races despite being in the same working-class and have a very high unemployment rate even after being a minor part of the population.

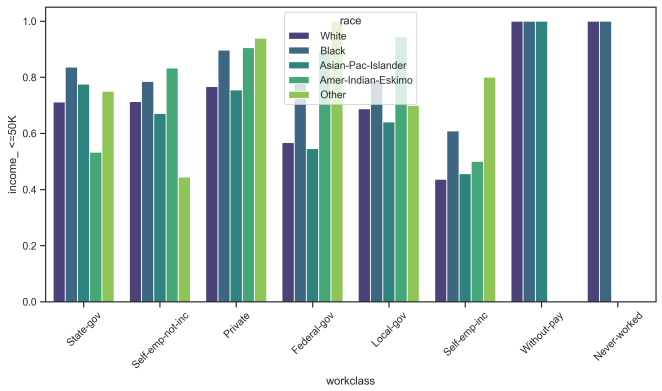


Fig. 6. Working class vs income less than $50k according to the race of individual

The number of people generating higher income is directly proportional to the number of years spent on education as evident in Fig. 7. This also validates the earlier observation that people who have achieved higher education are more likely to fall in the category of high-income earners. Less educated individuals are more likely to earn below $50K annual, which indicates that education is an essential factor in determining the yearly income of individuals.

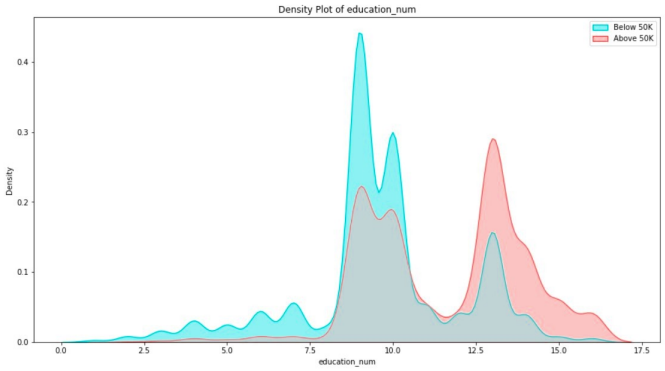


Fig. 7. Years spent on Education vs income

The variable *fnlwgt* doesn't have a significant impact on the income status of people, and it is distributed evenly across the two given income groups. The distribution of martial_status with income indicates that households with two earning individuals (i.e. husband and wife) are more likely to fall in the category of high-income earners. [3]

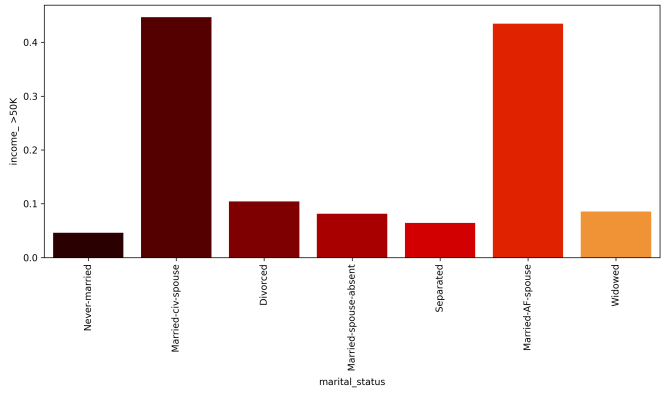


Fig. 8. Marial status vs income more than 50K

Similarly it is also observed that on an average people with high earning are putting more hours per week for work in comparison to their counter parts. See Fig. 9.

Based on the above observations, the set of features I will use to build the model are *age, workclass, fnwlgt, education, education_num, marital_status, occupation, relationship, race, sex, capital_gain, capital_loss* and *hours*.

The correlation heat-map in the Fig. 10 doesn't indicate any strong correlation between the given variables, strengthening the assumption that these features are independent of each other and the chances of a naive Bayes classifier yielding good results are high.

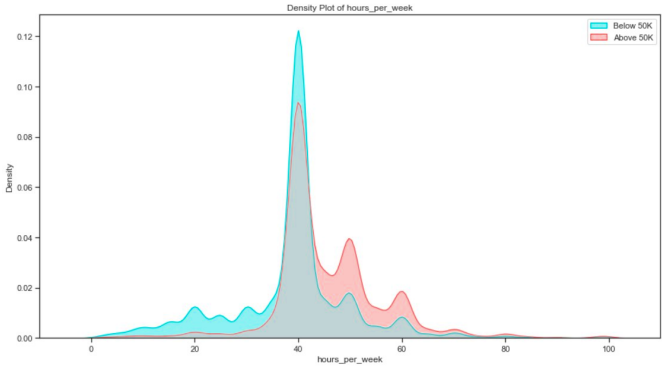


Fig. 9. Hours per week vs income

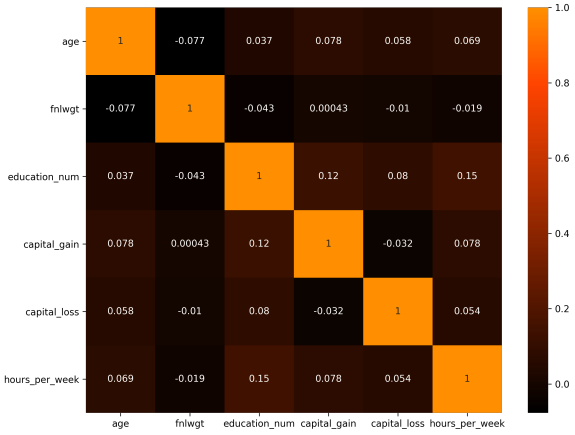


Fig. 10. Correlation heat-map

### C. Naive Bayes Classifier Model and Results

The following section will focus on the relationship between different features and identify the most suited for modelling purposes. The purpose of this paper is to identify the income group of a house-old and want to identify its relationship with the other features, the most crucial being age, education, race, native_country and sex. To do so, a generative model is fitted (which aims at learning the distribution of target classes rather than just identifying the decision boundary) is deployed. The mathematics behind the estimation of the line of best fit has already been described in detail in the previous sections. For our given problem our predictor is income and regressor consist of features describes in Sec. III-B.

The python package *sklearn* is used for performing the task of fitting a naive Bayes classifier. Scikit-Learn, which is primarily written in Python, is built upon NumPy, SciPy and Matplotlib. The sklearn package provides different classes for the classification problem, including **GaussianNB**. Before proceeding with the modelling, one-hot encoding is performed on the categorical features like work-class, education, marital_status, occupation, relationship, race, sex and native_country. MinMaxScaler is used to scale the data such that each value falls in the range [0,1]. Once the scaling is done the data is split into train and test data. Afterwards a Gaussian naive Bayes model is fitted on the training data and the labels for test data is predicted.

The mean accuracy score on the test data obtained by the naive Bayes classifier is $\mathbf{0.81}$, whereas that on the training dataset is $\mathbf{0.82}$. The majority of households lies in the category of annual income less than \$50K. The **null accuracy** obtained is $\mathbf{0.7582}$. Using a **random guess strategy** indicates that using a probabilistic outcome of $0.8$ will result in mean accuracy of $\mathbf{0.69}$, which further indicates that the classifier is doing a decent job. The confusion matrix looks as follows:

$$\begin{bmatrix} 5817 & 2007 \\ 1290 & 655 \end{bmatrix}$$

True Positive (TP): $\mathbf{8917}$, True Negative (TN): $\mathbf{2007}$, False Positive (FP): $\mathbf{1290}$, False Negative (FN): $\mathbf{655}$.

The classification report suggest the following

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| $<= 50K$ | 0.93 | 0.80 | 0.86 | 7407 |
| $> 50K$ | 0.56 | 0.81 | 0.66 | 2362 |
| accuracy | | | 0.80 | 9769 |

## IV. Improvements

From our above observations, it can be inferred that the naive Bayes classifier gives decent results for identifying households where the annual income is more than \$50K. Black people, people from 3rd world countries and females are less likely to be in the higher income region than others.

We used a Gaussian Naive Bayes classifier and obtained **80.67%** and **80.83%** accuracy on the train and test sets, respectively, without overfitting. In the test set, we found **7407** data points with income $\geqq$ 50K and **2362** with income ¡ 50K. We obtained a null accuracy score of **0.7582**. We achieved a classification accuracy of **0.0803**, an error of **0.1917**, a precision of **0.8099** and a sensitivity (or recall) of **0.9281**. As per our findings, privately employed/self-employed individuals, graduates, married individuals, exec-managerial positions or those with professional capacity, white people, husbands, males and US natives are more likely to earn higher salaries, and about the problem in hand, tend to make over \$50K a year.

This problem can be tackled using various other classifiers, like Logistic Regression, Decision Trees, Support Vector Classifier, etc. One can also use Ensemble methods like Random Forest Classifier and other bagging and boosting algorithms. In general, the ensemble methods seem to be more robust and capture inter-feature relationships better.

## V. Conclusion

The study of the question: *what sorts of households are more probable to earn more than \$50K annually?* is essential to understand the bias while providing different private and

public policy benefits to people. If the education status of people is known, then those with higher qualifications are preferred in a job over others. Similarly, men's are more likely to earn higher salaries. From our observations, it can be reasoned that there is some kind of bias for people belonging to different races and women, which highlights a social issue and would come in handy for future policy-making. The EDA and naive Bayes classifier approach give some decent idea about the relationship between predictors and regressors. It doesn't explain the underlying pattern thoroughly. The reason might be the presence of high cardinality in features like native_country and improper handling of parts. Therefore, different models have been used to find out the best performing one, and the *decision tree* turns out to be the best with a training accuracy of 0.97 and test accuracy of 0.84.

## REFERENCES


[1] Deisenroth, Marc Peter, Faisal, A. Aldo and Ong, Cheng Soon. Mathematics for Machine Learning. : Cambridge University Press, 2020, pp.348–350

[2] Murphy, Kevin P. Machine Learning: A Probabilistic Perspective. Cambridge, MA: MIT Press, 2012. pp.34–51, pp.82–87

[3] Boyd, Stephen, and Vandenberghe, Lieven, Introduction to Applied Linear Algebra, Cambridge University Press, 2018, pp.38

[4] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning : with Applications in R. New York :Springer, 2013, pp.138–149