Data Analytics Lab: Assignment-3
A Mathematical Essay on Naive Bayes Classifier

Arjav Singh
Metallurgical and Materials Engineering
Indian Institute of Technology Madras
Chennai, India
mm20b007@smail.iitm.ac.in

***Abstract*—In this study, the correlation between whether an individual makes over $50,000 a year and various factors such as age, educational qualification, marital status, occupation, race, and other factors is examined. The importance of these factors is modeled, and the income group of individuals is predicted using a Naive Bayes model.**

***Index Terms*—Introduction, Naive Bayes, Data & Problem, Conclusion**

I. Introduction

This study uses survey data from the 1994 Census database to conduct an empirical analysis of the factors influencing personal income. Education level is a crucial indicator, and classification is performed using the Naive Bayes model. It is discovered that several factors, including gender, age, education, and marital status, have a significant impact on personal income. Additionally, variations among different occupations are also explored. Naive Bayes is a classification technique founded on Bayes' Theorem, assuming conditional independence among predictors and that one particular feature in a class is unrelated to the presence of any other feature.

In this study, the income category of individuals is modeled based on education, age, socioeconomic factors, marital status, etc., using Naive Bayes. The process begins with the gathering, cleaning, and preparing data, followed by exploratory analysis. Subsequently, statistical models are constructed, and visualizations are generated to provide quantitative and visual evidence of the observed relationships. In the next section, the key principles underlying Naive Bayes are highlighted. Section 3 delves into the insights and observations derived from the data and models. Finally, in section 4, the salient features of the study are outlined, and potential avenues for further investigation are presented.

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, with the class labels being 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. There are different types of Naive Bayes classifiers available, some of them are -

1) **Multinomial Naive Bayes**: In the case of a Multinomial Naive Bayes model, the samples (feature vectors) represent the frequencies at which certain events have been generated by a multinomial distribution with probabilities $(p_1, \ldots, p_n)$, where p_i is the probability that event i occurs. The Multinomial Naive Bayes algorithm is typically preferred for data that follows a multinomial distribution. It is one of the standard algorithms used in text categorization and classification.
2) **Bernoulli Naive Bayes**: In the multivariate Bernoulli event model, features are independent boolean variables (binary variables) describing inputs. Like the multinomial model, this model is also commonly employed in document classification tasks, where binary term occurrence features are used instead of term frequencies.
3) **Gaussian Naive Bayes**: When dealing with continuous attribute values, an assumption is made that the values associated with each class follow a Gaussian or Normal distribution. For instance, consider training data containing a continuous attribute x. First, the data is segmented by class, and then the mean (μ_i) and variance (σ_i^2) of x in each class are computed. Suppose there is an observation value x_i. Then, the probability distribution of x_i given a class can be computed using the following equation:

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

A. Model Structure

The Naïve Bayes Classifier utilizes Bayes' theorem to estimate the membership probabilities for each class, specifically the probability that a given record or data point belongs to a particular class. The class with the highest probability is considered the most likely class, often referred to as the Maximum A Posteriori (MAP) class.

The MAP for a hypothesis involving two events, A and B, is calculated as follows:

$$\text{MAP}(A) = \max(P(A|B)) \quad (1) \tag{1}$$

This can also be expressed as:

$$\mathrm{MAP}(A) = \max\left(\frac{P(B|A)\cdot P(A)}{P(B)}\right) \quad (2) \tag{2}$$

Simplifying further:

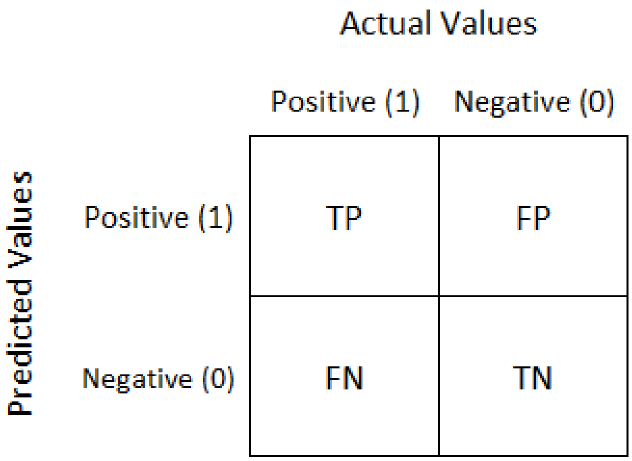
$$\mathrm{MAP}(A) = \max(P(B|A)\cdot P(A)) \quad (3) \tag{3}$$

Here, $P(B)$ represents the evidence probability, which is used for normalization purposes. It remains constant across calculations, so removing it does not affect the result.

The Naïve Bayes Classifier operates under the assumption that all features are mutually independent. In other words, the presence or absence of one feature does not influence the presence or absence of any other feature.

In real-world datasets, hypotheses are tested based on multiple items of evidence derived from various features. These calculations can become quite complex. To simplify this process, the feature independence assumption is applied to treat each item of evidence as independent, thereby simplifying the analysis.

B. Metrics for model evaluation



Predicted Values	Actual Values: Positive (1)	Actual Values: Negative (0)
Positive (1)	TP	FP
Negative (0)	FN	TN

Fig. 1. Confusion Matrix.

1) **Confusion Matrix**: It is used to summarize the performance of a classification algorithm on a set of test data for which the true values are previously known. Sometimes it is also called an error matrix. Terminologies of the Confusion matrix (Figure 1) are:
   - **True Positive**: TP means the model predicted yes, and the actual answer is also yes.
   - **True negative**: TN means the model predicted no, and the actual answer is also no.
   - **False positive**: FP means the model predicted yes, but the actual answer is no.
   - **False negative**: FN means the model predicted no, but the actual answer is yes.

   The rates calculated using the Confusion Matrix are:

   a) **Accuracy**: (TP+TN/Total) tells about overall how classifier Is correct.

   b) **True positive rate**: TP/(actual yes) it says about how much time yes is predicted correctly. It is also called "sensitivity" or "recall."

   c) **False positive rate**: FP/(actual number) says how much time yes is predicted when the actual answer is no.

   d) **True negative rate**: TN/(actual number) says how much time no is predicted correctly, and the actual answer is also no. It is also known as "specificity."

   e) **Misclassification rate**: (FP+FN)/(Total) It is also known as the error rate and tells about how often our model is wrong.

   f) **Precision**: (TP/ (predicted yes)) If it predicts yes, then how often is it correct.

   g) **Prevalence**: (actual yes /total) how often yes condition actually occurs.

   h) **F1-score**: f1 score is defined as the weighted harmonic mean of precision and recall. The best achievable F1 score is 1.0, while the worst is 0.0. The F1 score serves as the harmonic mean of precision and recall. Consequently, the F1-score consistently yields lower values than accuracy measures since it incorporates precision and recall in its computation. When evaluating classifier models, it is advisable to employ the weighted average of the F1 score instead of relying solely on global accuracy.

2) **ROC curve (Receiver Operating Characteristic)**: The Receiver Operating Characteristic (ROC) curve is a useful tool for assessing a model's performance by examining the trade-offs between its True Positive (TP) rate, also known as sensitivity, and its False Negative (FN) rate, which is the complement of specificity. This curve visually represents these two parameters.
   The Area Under the Curve (AUC) metric to summarize the ROC curve concisely. The AUC quantifies the area under the ROC curve. In simpler terms, it measures how well the model can distinguish between positive and negative cases. A higher AUC indicates better classifier performance.
   In essence, AUC categorizes model performance as follows:

   - If AUC = 1, the classifier correctly distinguishes between all the Positive and Negative class points.
   - If 0.5¡ AUC ¡ 1, the classifier will distinguish the positive class value from the negative one because it finds more TP and TN than FP and FN.
   - If AUC = 0.5, the classifier cannot distinguish between positive and negative values.
   - If AUC =0, the classifier predicts all positive as negative and negative as positive.

III. Problem

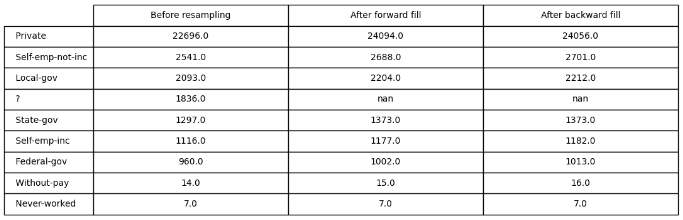
The problem at hand is centered around predicting whether the income of a person is more than $50,000 from the 1994 Census US income database. A naive Bayes classifier will be employed to predict the possibility for every person, incorporating various features such as age, relationship status, education, and other relevant factors for analysis.

A. Exploratory Data Analysis and Feature Generation

The training dataset employed in this study comprises 32,561 individuals and encompasses 14 distinct features. The interpretation of these features is as follows:

- *Age*: The individual's age, ranging from 17 to 90.
- *Workclass*: The employment category of the individual, which includes designations such as private, without-pay, state government, etc.
- *Fnlwgt*
- *Education*: The educational level of the individual.
- *Education Years*: The number of years of education completed by the individual.
- *Occupation*: The occupation of the individual.
- *Relationship*: The individual's role within the family.
- *Race*: The racial background to which the individual belongs.
- *Sex*: The gender of the individual.
- *Capital Gain, Loss*
- *Working Hours*: The average number of hours per week that the individual works.
- *Native Country*: The cultural or geographic background of the individual.

The dataset contains missing values denoted by "?" in the "Workclass" and "Occupation" features. Instead of discarding these entries, they are treated as a separate category due to the observation that removing them adversely impacts the model's performance. The primary objective is to predict the binary feature "Wage," which is equal to 1 if the individual earns an annual income greater than 50,000 dollars and 0 otherwise.

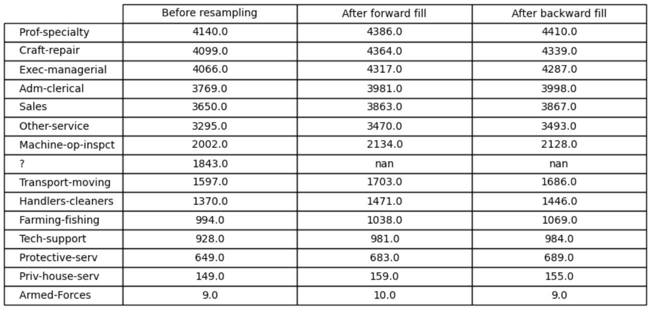


	Before resampling	After forward fill	After backward fill
Private	22696.0	24094.0	24056.0
Self-emp-not-inc	2541.0	2688.0	2701.0
Local-gov	2093.0	2204.0	2212.0
?	1836.0	nan	nan
State-gov	1297.0	1373.0	1373.0
Self-emp-inc	1116.0	1177.0	1182.0
Federal-gov	960.0	1002.0	1013.0
Without-pay	14.0	15.0	16.0
Never-worked	7.0	7.0	7.0

Fig. 2. Distribution of values before and after re-sampling of workclass feature.

The data is initially read into a pandas data frame, revealing a total of 32,561 data points and a total of 15 columns encompassing various person-related features. When the distributions of individuals who earn over 50K are visualized, it is observed that approximately 24.1% of the total population falls into this category (Figure 1). Among the 15 features, 9 are categorical, and 6 are numerical. Subsequently, an assessment is made to identify null values within the data, and it is determined that there are no NaN values. However, three columns, namely Workclass, Occupation, and Native Country, contain '?' marks in some data points, necessitating treatment. The initial step involves replacing these '?' with NaN values, then imputing them with the value preceded by them in each respective column, as the number of null values is very low since the effect of forward re-sampling and backward re-sampling had very low difference which can be observed in Figure 2, 3, and 4 hence 'forward fill' re-sampling method was opted.

The cardinality of each categorical feature is then examined, measuring the number of unique values each feature can assume. High cardinality can potentially lead to issues. It is observed that most features have no more than 7 attributes, with the Native Country feature having the highest number (Figure 2).



	Before resampling	After forward fill	After backward fill
Prof-specialty	4140.0	4386.0	4410.0
Craft-repair	4099.0	4364.0	4339.0
Exec-managerial	4066.0	4317.0	4287.0
Adm-clerical	3769.0	3981.0	3998.0
Sales	3650.0	3863.0	3867.0
Other-service	3295.0	3470.0	3493.0
Machine-op-inspct	2002.0	2134.0	2128.0
?	1843.0	nan	nan
Transport-moving	1597.0	1703.0	1686.0
Handlers-cleaners	1370.0	1471.0	1446.0
Farming-fishing	994.0	1038.0	1069.0
Tech-support	928.0	981.0	984.0
Protective-serv	649.0	683.0	689.0
Priv-house-serv	149.0	159.0	155.0
Armed-Forces	9.0	10.0	9.0

Fig. 3. Distribution of values before and after re-sampling of occupation feature.

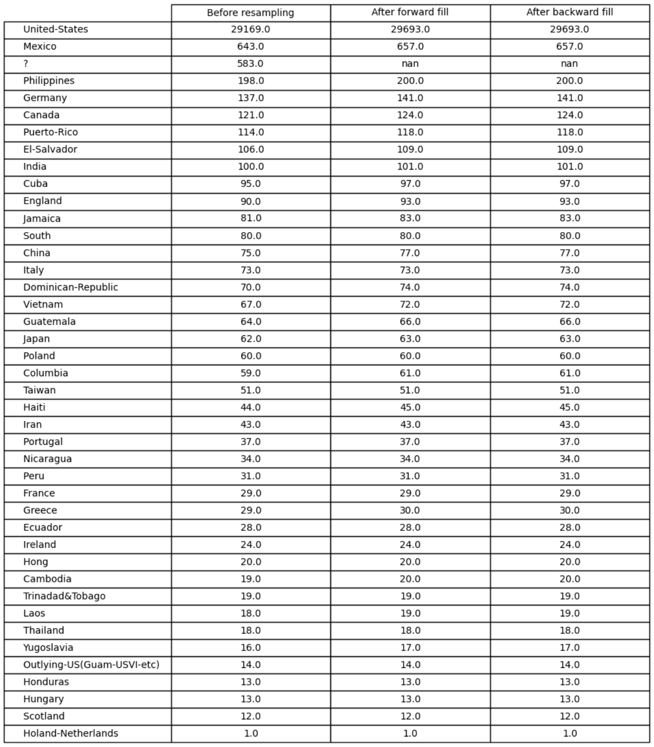
B. Visualization and Feature Generation

The features are examined one by one, beginning with "Workclass." It is observed that the primary categories are government employees, private sector workers, self-employed individuals, those without pay, and individuals who have never worked. Each category exhibits a different higher income rate, with the highest rate among self-employed individuals and the lowest among those in the private sector.

Moving on to "Education," the primary classes are university level, school level, and postgraduate level. A trend emerges where lower levels of education correspond to lower incomes, while individuals with doctorates and professional school degrees earn the highest salaries.

In the case of "Marital Status," individuals who are married and have a spouse tend to have the highest incomes. In contrast, all other marital statuses are associated with a lower likelihood of high income.

Analyzing "Occupation," it is evident that the number of classes increases significantly. Each occupation class enjoys



	Before resampling	After forward fill	After backward fill
United-States	29169.0	29693.0	29693.0
Mexico	643.0	657.0	657.0
?	583.0	nan	nan
Philippines	198.0	200.0	200.0
Germany	137.0	141.0	141.0
Canada	121.0	124.0	124.0
Puerto-Rico	114.0	118.0	118.0
El-Salvador	106.0	109.0	109.0
India	100.0	101.0	101.0
Cuba	95.0	97.0	97.0
England	90.0	93.0	93.0
Jamaica	81.0	83.0	83.0
South	80.0	80.0	80.0
China	75.0	77.0	77.0
Italy	73.0	73.0	73.0
Dominican-Republic	70.0	74.0	74.0
Vietnam	67.0	72.0	72.0
Guatemala	64.0	66.0	66.0
Japan	62.0	63.0	63.0
Poland	60.0	60.0	60.0
Columbia	59.0	61.0	61.0
Taiwan	51.0	51.0	51.0
Haiti	44.0	45.0	45.0
Iran	43.0	43.0	43.0
Portugal	37.0	37.0	37.0
Nicaragua	34.0	34.0	34.0
Peru	31.0	31.0	31.0
France	29.0	29.0	29.0
Greece	29.0	30.0	30.0
Ecuador	28.0	28.0	28.0
Ireland	24.0	24.0	24.0
Hong	20.0	20.0	20.0
Cambodia	19.0	20.0	20.0
Trinadad&Tobago	19.0	19.0	19.0
Laos	18.0	19.0	19.0
Thailand	18.0	18.0	18.0
Yugoslavia	16.0	17.0	17.0
Outlying-US(Guam-USVI-etc)	14.0	14.0	14.0
Honduras	13.0	13.0	13.0
Hungary	13.0	13.0	13.0
Scotland	12.0	12.0	12.0
Holand-Netherlands	1.0	1.0	1.0

Fig. 4. Distribution of values before and after re-sampling of native country feature.

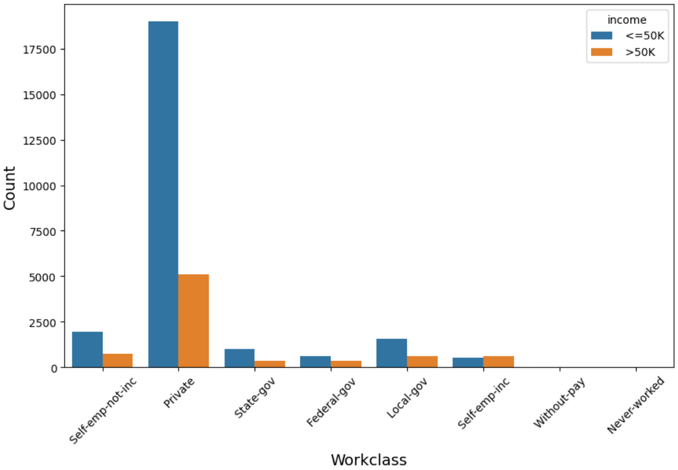


Fig. 5. Income variability over different work classes.

different income rates, with executive managers and professionals in specialized fields earning the highest incomes.

Exploring the ”Relationship” feature reveals that husbands and wives have the highest probability of having high incomes. In contrast, single individuals without families or those who are unmarried do not enjoy such high incomes.

In terms of ”Race,” there is a bias toward white individuals having higher salaries compared to others. A concerning observation is made when visualizing ”Sex,” with males having higher average incomes than females.

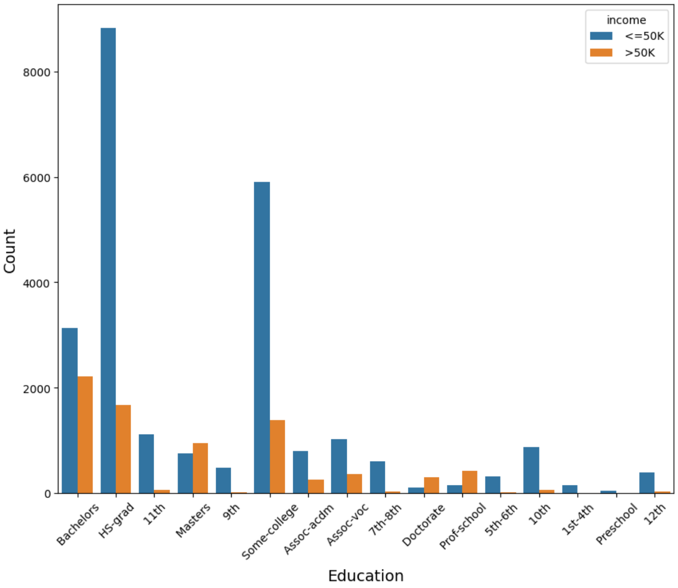


Fig. 6. Income variability over academic qualifications.

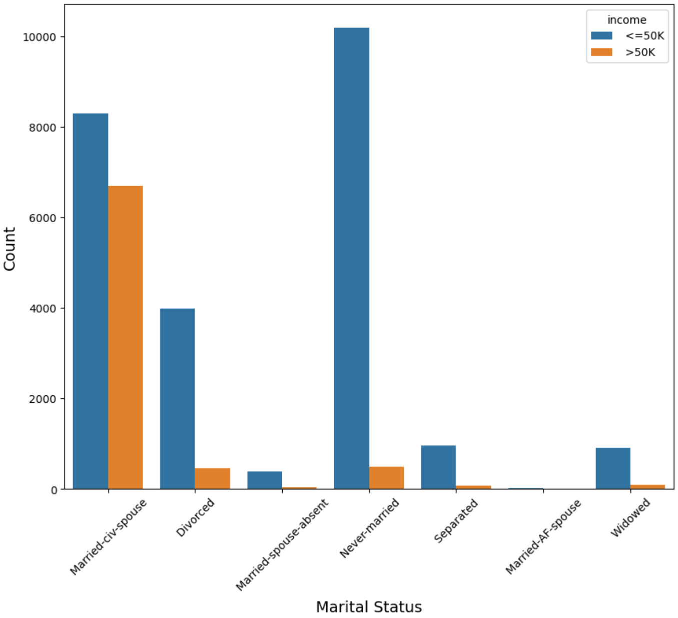


Fig. 7. Income distribution over marital status of an individual.

When exploring ”Age” using histograms and box plots, it becomes apparent that individuals in their 40s tend to have higher incomes. This can be attributed to the fact that most people do not earn well at the beginning of their careers, and as they grow old, in their 60s and 70s, they tend to lose more money than they earn.

Further investigation into ”Education-Num,” which represents the number of education levels, reveals a trend where individuals with higher education levels tend to have higher incomes. Similarly, ”Hours per Week (hourspw),” which represents the number of hours worked per week, indicates

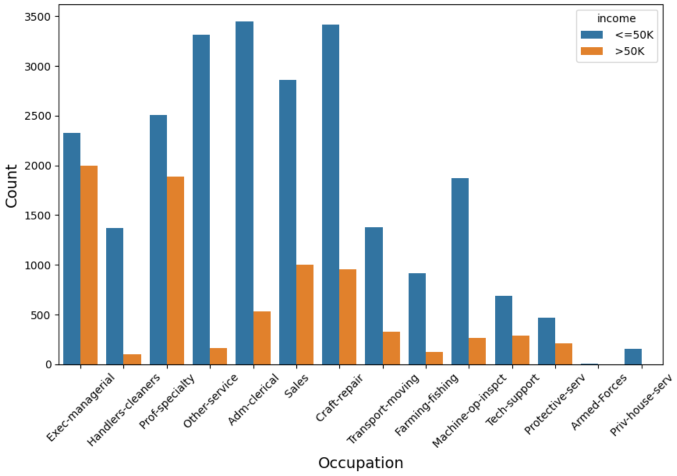


Fig. 8. Income distribution over different occupations.

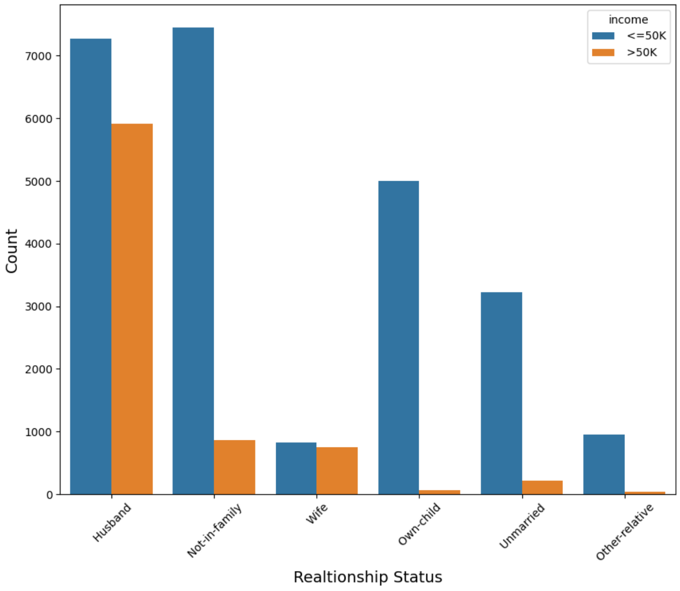


Fig. 9. Income over the relationship status of the individuals.

that individuals who work longer hours tend to have higher incomes.

Before moving to feature generation, Pearson correlation values were checked by converting all the categorical features into numerical ones by assigning each unique data point a number, and it was found that income is highly correlated to education-num, relationship, marital status, sex, age and number of hours per week.

With help of the above information new features were generated to improvise the modelling. The number of years of education is categorised into low, medium, and high based their values, similar categorization is implemented to hours per week. The occupation data was categorised into highskill if in managerial and specialty position and lowskill otherwise. Based on income distribution with respect to race it was found

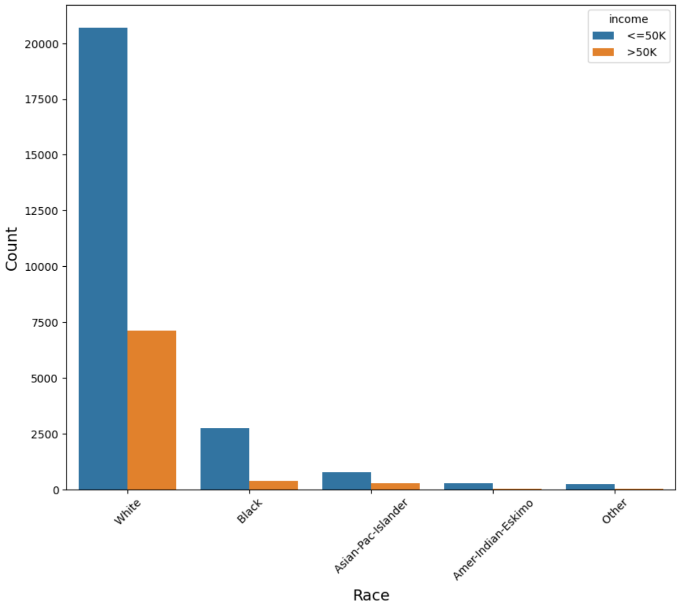


Fig. 10. Income variability over different races.

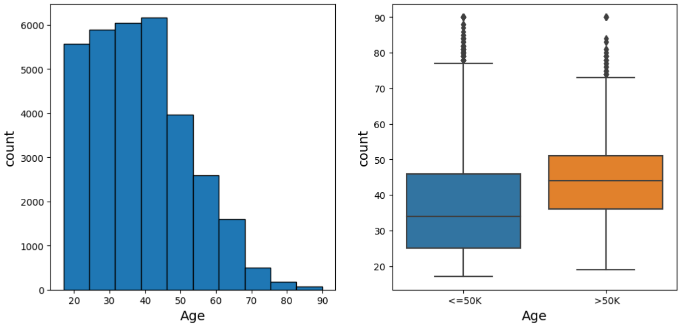


Fig. 11. Distribution of income over the age of the individuals.

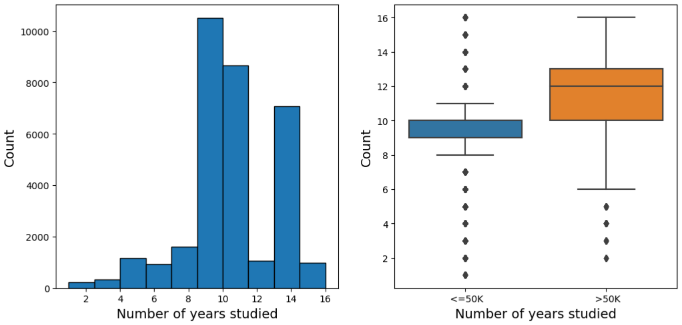


Fig. 12. Income vs number of years of education.

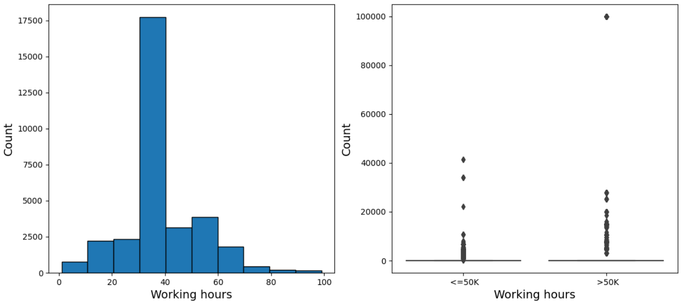


Fig. 13. Income distribution over the number of work hours.

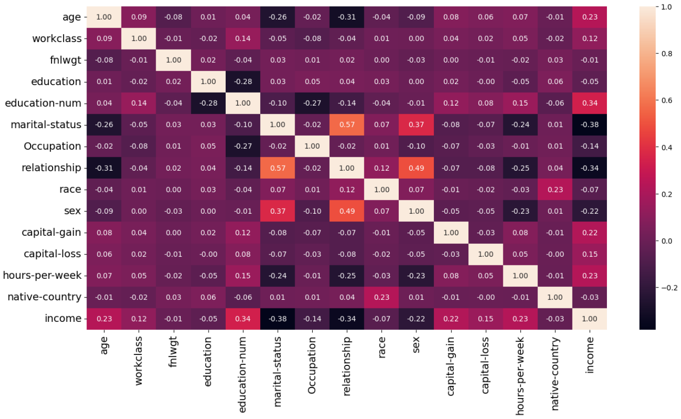


Fig. 14. Pearson Correlation heatmap.

that most of people who has more than $50000 income are white people hence race feature is categorised into white and other.

Finally, as part of postprocessing, we create dummies for the categorical variables to make them meaningful to the machine, also termed One Hot encoding.

C. Feature Selection

One hot encoding resulted in a very high dimensional data which is not suitable for model hence Variance threshold is utilized to reduce the dimension and choose the most relevant features. Variance Threshold is a univariate approach to feature selection. It removes all features whose variance doesn't meet some threshold. By default, it removes all zero-variance features, i.e. features that have the same value in all samples.

D. Guassian Naive Bayes Classifier Modelling

The modeling begins by utilizing the Gaussian Naive Bayes classifier from the sklearn library. Initially, the data is split into training and validation sets to assess model performance on unseen data. To ensure feature compatibility, sklearn's robust scaler is applied. The subsequent step involves fine-tuning the model's hyperparameters, specifically the var smoothing parameter, through randomized search cross-validation from sklearn's model selection toolkit. Following training and predictions using the tuned model, % accuracy of 83.27% is achieved on the training set and 83.12% on the test dataset. The similarity in accuracy values between the test and training sets suggests no signs of overfitting.

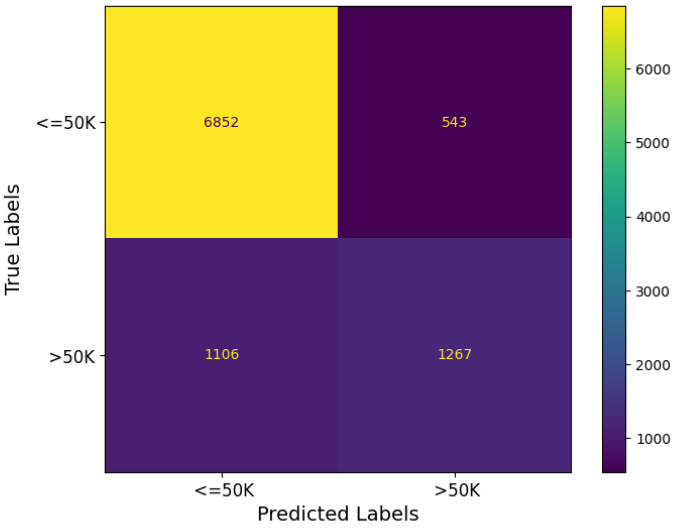


Fig. 15. Confusion Matrix.

Subsequently, evaluation metrics are examined, beginning with the Confusion Matrix. The analysis reveals 6,852 true positives, 1,267 true negatives, 543 false positives, and 1106 false negatives. Moving on to the classification report, an F1 score of 0.89 for incomes less than or equal to 50K and 0.61 for incomes greater than 50K is observed, with an accuracy of 0.83, a macro average of 0.76, and a weighted average of 0.83. These values indicate strong model performance.

TABLE I
CLASSIFICATION REPORT

Class	Precision	Recall	F1-Score	Support
<=50K	0.86	0.93	0.89	7395
>50K	0.70	0.53	0.61	2373
Accuracy			0.83	9768
Macro Avg	0.78	0.73	0.75	9768
Weighted Avg	0.82	0.83	0.82	9768

Next, the ROC curve is plotted, representing the false positive rate versus the true positive rate. The curve lies well above the y = x line, indicating good model discrimination, with an AUC value of 0.8843. Subsequently, variability in performance on testing and training datasets is assessed using 10-fold cross-validation. The mean accuracy is close to the original accuracy, with minimal deviation across folds, suggesting that the model's performance is not heavily reliant on the specific training data.

Finally, k-fold cross validation is done and it was found that the mean accuracy is close to the original one, and also there is not much deviation from the average for all the folds, thus it is clear that the model is not much reliant on the data on which it is being trained. Further more thresholds were tested to optimize accuracy on the test set. It is determined that a threshold of 0.8 yields the highest accuracy of 0.835.

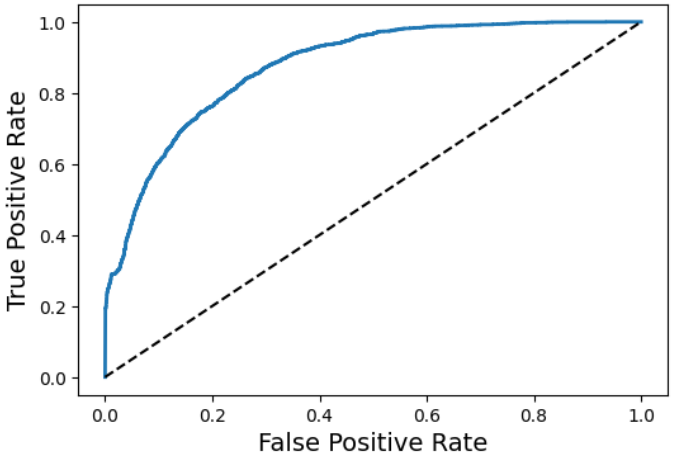


Fig. 16. ROC curve for Gaussian Naive Bayes Classifier for Predicting Salaries.

IV. Conclusion

In this study, it was observed that individuals who are male and aged (age ¿= 45) working longer hours are more inclined to earn annual wages exceeding $50K. Additionally, the data showed that most individuals typically undergo around 9 years of education. Still, those with more than 14 years of education and those who are self-employed are more likely to earn wages exceeding $50K annually. Furthermore, the analysis revealed that, on average, both women and men have similar educational levels, but women tend to work fewer hours and consequently receive lower salaries.

References

[1] "Naive Bayes Classifier," *Towards Data Science*, 2023. [Online]. Available: https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c.

[2] "Naive Bayes Tutorial," *DataCamp*, 2023. [Online]. Available: https://www.datacamp.com/tutorial/naive-bayes-scikit-learn.

[3] "Naive Bayes classifier," *Wikipedia*, 2023. [Online]. Available: https://en.wikipedia.org/wiki/Naive_Bayes_classifier.

[4] "AUC-ROC Curve & Confusion Matrix Explained in Detail." [Online]. Available: https://www.kaggle.com/code/vithal2311/auc-roc-curve-confusion-matrix-explained-in-detail.

[5] Analytics Vidhya. "K-Fold Cross-Validation Technique and Its Essentials." [Online]. Available: https://www.analyticsvidhya.com/blog/2022/02/k-fold-cross-validation-technique-and-its-essentials/.