Assignment 3: Naive Bayes Classifier

Prajwal Dinesh Sahu
ME18B114
Dept. of Mechanical Engineering
Indian Institute of Technology, Madras
me18b114@smail.iitm.ac.in

***Abstract*—This paper gives a brief introduction of naive bayes classifier and the basic mathematical theory behind it. We also look at some metrics to evaluate the performance of our model. Further, we apply the classifier on a real world data-set containing the socio-economic details of the people of USA, collected by US Census Bureau in 1996, to predict whether a person has greater or lesser than 50k income. Through exploratory data analysis and predictive naive bayes classifier model, we investigate trends in data to predict the persons who have higher income. Then we evaluate our model using confusion matrix and F1 score. Our findings reveal some interesting trends.**

I. Introduction

There are multiple types of algorithm methods used in machine learning. One such popular and commonly used machine learning method is Naive Bayes Classifier. It is a supervised learning algorithm used to train a model to predict the behaviour of our data based on given parameters.

Naive Bayes Classifier is a classification technique based on Bayes' Theorem which classifies the data into various classes based on prior probabilities and the likelihood, with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Mathematically,

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\ p(\mathbf{x} \mid C_k)}{p(\mathbf{x})} \tag{1}$$

where,
C is the class.
x represents the features of the model.
K is the number of classes.

In this paper, we apply Naive Bayes Classifier to data which was extracted from the 1994 Census bureau database by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics). The key task is to determine whether a person makes over 50K a year, adult.csv contains the dataset required to solve the task..

The paper will first give a brief explanation of logistic regression. Then, we will try to solve the Income prediction and model our data-set by
1. Cleaning the data.
2. Exploratory and visual analysis.
3. One-Hot Encoding.
4. Checking for Correlation.
5. Scaling the Data.
6. Training the model.
7. Evaluating the model.
8. Interpreting the results.

II. Naive Bayes Classifier Theory

A. *Classification*

Classification is a process of categorizing a given set of data into classes. Predicting a qualitative response for an observation can be referred to as classifying that observation, since it involves assigning the observation to a category, label or class.

Many real world problems require classification rather than regression. In the classification setting we have a set of training observations (x1, y1), . . . , (xn, yn) that we can use to build a classifier.

Regression is not useful in problems like if we are trying to predict the medical condition of a patient in the hospital and we have 3 possible diagnoses- stroke, drug overdose, and epileptic seizure based on the symptoms of the patient. We can label these diseases at 1,2 and 3. In this case, even if we create a metric to accurately convert the symptoms to numeric data, regression won't be able to classify the data as it cannot give the output as 1,2 or 3. Alternatively, we can create a threshold that if predicted value is between 0.5 to 1.5, we will assign it to label 1. This brings us to the basic idea behind classification algorithms which can give qualitative response.

B. *Naive Bayes Classifier*

*1) **The Bayes' Theorem**:* Bayes' theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event. It is named after Thomas Bayes, an English statistician and philosopher. It is mathematically stated as -

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{2}$$

where A and B are events and
• P(A/B) is a conditional probability: the probability of event

A occurring given that B is true. It is also called the posterior probability of A given B.
• P(B/A) is the probability of event B occurring given that A is true.
• P(A) and P(B) are probabilities of observing A and B respectively without any given conditions.

*2) **Using the Naive Bayes Classifier**:* Naive Bayes is a classification algorithm based on Bayes theorem. Given k possible outcomes or classes, and vector x representing n features, the model calculates the probability of the entry being is class k. Mathematically, it calculates -

$$p(C_k \mid x_1, \ldots, x_n) \tag{3}$$

It is called Naive classifer because it makes a naive assumption to the Bayes' theorem, which is, independence among the features, i.e $x_1, ..., x_n$ are all independent.

In other words, the probability is calculated as -

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\ p(\mathbf{x} \mid C_k)}{p(\mathbf{x})} \tag{4}$$

which translates to -

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}} \tag{5}$$

The denominator is constant and the numerator is equivalent to joint probability of C and x. After applying the naive independence condition, the joint probability boils down to -

$$p(C_k \mid x_1, .., x_n) \propto p(C_k, x_1, \ldots, x_n) \tag{6}$$

$$\propto p(C_k)\ p(x_1 \mid C_k)\ p(x_2 \mid C_k)\ p(x_3 \mid C_k) \tag{7}$$

$$\propto p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k) \tag{8}$$

The constant of proportionality, the scaling factor is P(X)

$$Z = p(\mathbf{x}) = \sum_k p(C_k)\ p(\mathbf{x} \mid C_k) \tag{9}$$

Therefore, the final distribution is -

$$p(C_k \mid x_1, \ldots, x_n) = \frac{1}{Z} p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k) \tag{10}$$

*3) **Working of the classifier**:* Once we have the distribution for the joint probability for C and x, the classifier used a decision rule called Maximum a Posteriori (MAP). It classifies the entry to the k_th class which has the maximum probability.

$$\hat{y} = \underset{k \in \{1,\ldots,K\}}{\operatorname{argmax}}\ p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k) \tag{11}$$

C. *Evaluating the model*

In model evaluation, we measure the extent to which the model fits the data.

*1) **Confusion Matrix**:* A confusion matrix is a table that shows the performance of a machine learning model. It displays how many of the model's predictions were right and how many were incorrect and which classes did the model succeed in and which did it fail in. It offers us an understanding of how the model works. If a model fails for a certain class, we may investigate it, figure out why, and try to improve the model.

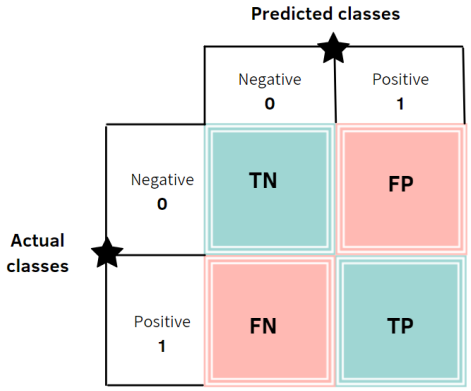

Fig. 1. Sample Confusion Matrix

There are various metrics based on confusion matrix. Some of them are listed below

*2) **Precision and Recall**:* Precision is the ratio between the True Positives and all the Positives. For our model, it will be the measure of persons who we correctly identified having income greater than 50k

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive + False Positive}} \tag{12}$$

Recall is the measure of our model correctly identifying True Positives. For all the people who have higher than 50k income, recall tells us how many we correctly predicted.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive + False Negative}} \tag{13}$$

*3) **F1 Score**:* F1 score is the measure which combines both precision and recall. It is calculated as the harmonic mean of precision and recall -

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{14}$$

III. THE PROBLEM

Given the data extracted by 1994 Census database by (Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics). We have to determine whether a person makes over dollar 50K a year or not based on the below given variables-

A. *The Variables*

Age - Age in years.
Work class - The working class out of Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
fnlwgt - This is the number of people the census believes the entry represents.
Education - Level of education.
Education-num - No. of years of education
Marital-status - Marital status of the person.
Occupation - Occupation of the person.
Race - The race of the person.
Sex - Male or Female.
capital-gain - Profit resulting from an investment or sale of a property.
capital-loss - Loss resulting from an investment or buying of a property.
Native Country - The native country of the person.

B. *Cleaning the Data*

We check for null values in all columns. No data is missing in the dataset.

C. *The target variable - Income*

We have 2 type of entries in the income column - greater and less than 50K, which labels the number of people having annual income in the above ranges. We have to build a Naive Bayesian Classifier to classify the test set. The target variable doesn't have a high degree of imbalance.

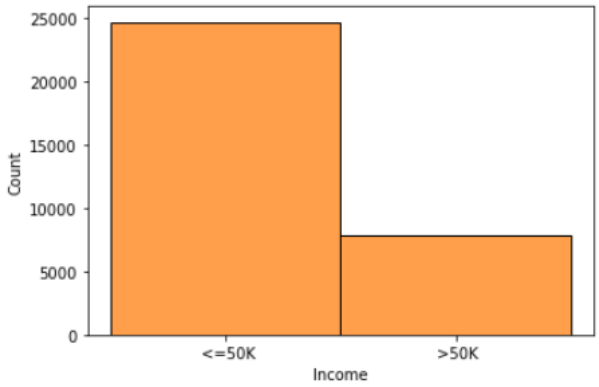

Fig. 2. Income countplot

D. *Cleaning and Visualizing the Numeric Variables*

1) ***Age****:* The age is well distributed with a mean of 38.58 years.
We visualize it using a boxplot.

The mean age is higher for the people having greater than 50K income, which is quite intuitive as professionals gain experience and reach higher posts with higher age.

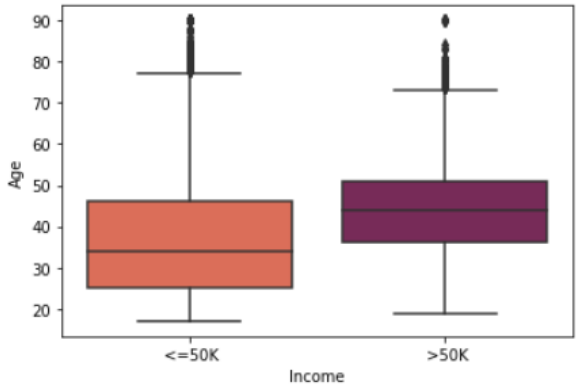

Fig. 3. Age vs Income

2) ***Education num****:* Most of the population have studied for 10 or more years. People who have studied more and gotten higher education tend to have higher income.

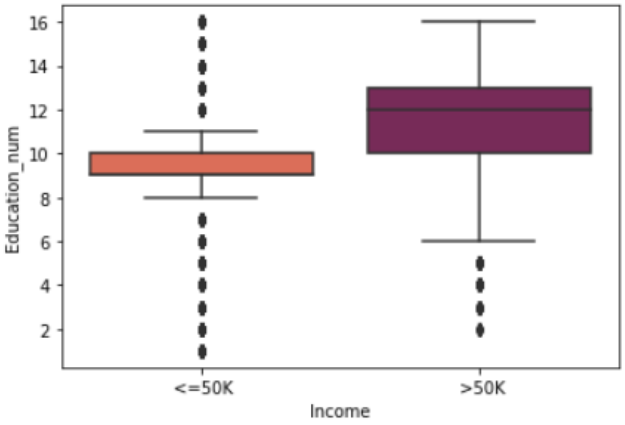

Fig. 4. Age vs Income

3) ***fnlwgt****:* This is a variable included by census committee to indicate the number of people the census believes the entry represents.

This variable does not seem to have any effect on our classifier.

4) ***Hours per week****:* Through the boxplot, we can see that increased number of hours directly leads to higher income.

5) ***Capital gain and Capital loss****:* Some people get profits/losses from transactions of property or investments. This supplements their annual income. In our original dataset, the number of people having greater than 50K income is only 24.08 %.

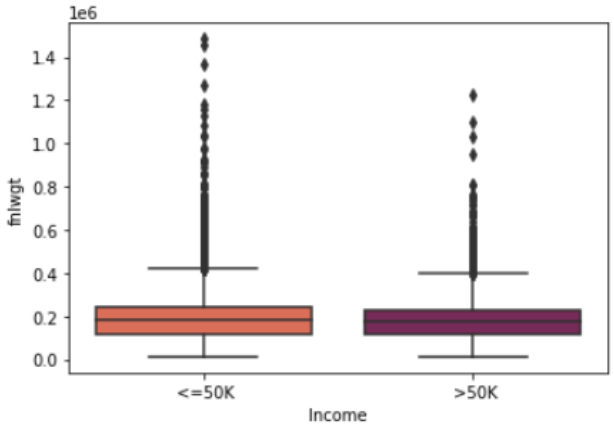

Fig. 5. fnlwgt vs Income

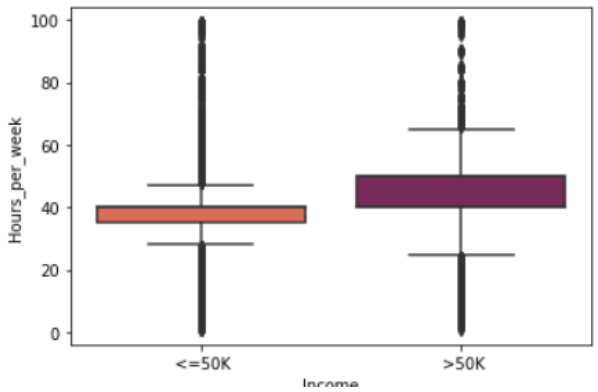

Fig. 6. Hours per week v Income

But if we conditionally select the people who have made gains more than 0, the percent changes to 61.85 %.

E. Cleaning and Visualizing the Categorical Variables

In this section, we will visualize the trends for the categorical variables.

1) ***Work Class*** *:* This variable represents the type of job the person has.
We find that some entries are '?'. We replace them with the mode value, which is 'Private'.

To visualize how Work Class effects the income, we plot a bar plot between the work class and the fraction of people having income above 50K.

We see that self employed and federal government employees have higher fractions.

2) ***Education****:* This represents what kind of education the persons have received from Bachelors, HS-grad, Masters etc.
To visualize how Education effects the income, we plot a bar plot between the Education and the fraction of people having income above 50K.

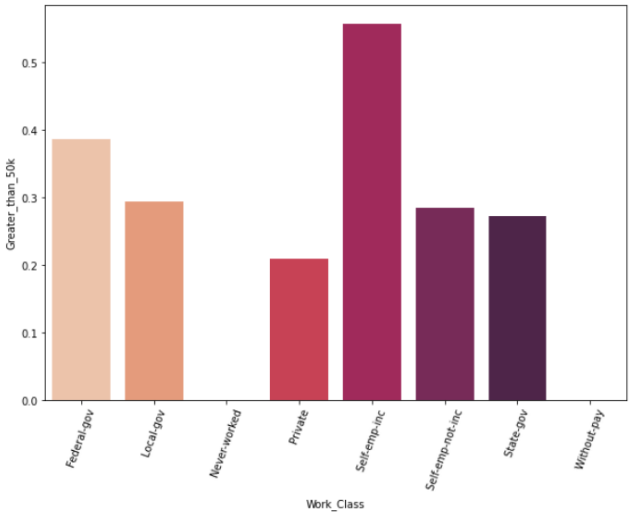

Fig. 7. Fraction of people having income greater than 50K vs Work Class

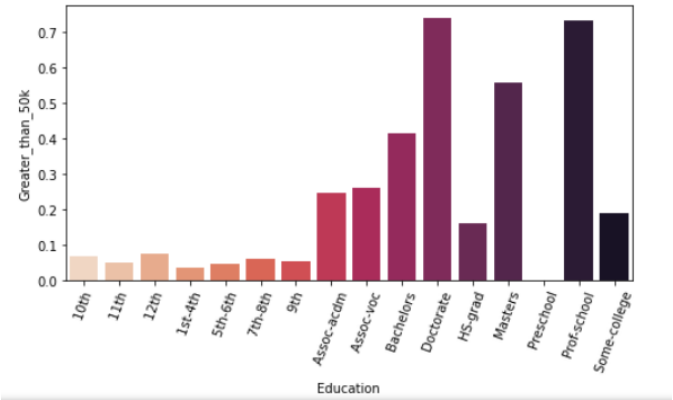

Fig. 8. Fraction of people having income greater than 50K vs Education

We can see that Doctorates and Prof-School graduates are among the top earners while the high school studied people have the lowest income.

3) ***Marital Status****:* This represents the marital status (divorced, married etc.)
To visualize how marital status effects the income, we plot a bar plot between the Marital Status and the fraction of people having income above 50K.

It is visible that married people constitute a larger fraction of people having higher income.

4) ***Occupation****:* This represents the type of job like sales, IT, clerical etc.
Some columns are missing some data and contain '?'. We replace it with the mode value.
To visualize how the job type effects the income, we plot a bar plot between the Occupation and the fraction of people having income above 50K.

It is visible that Executive managers tend to have higher

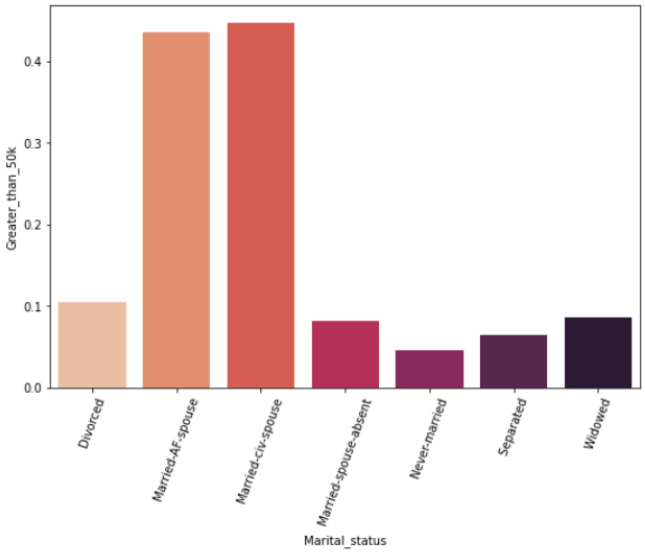

Fig. 9. Fraction of people having income greater than 50K vs Marital status

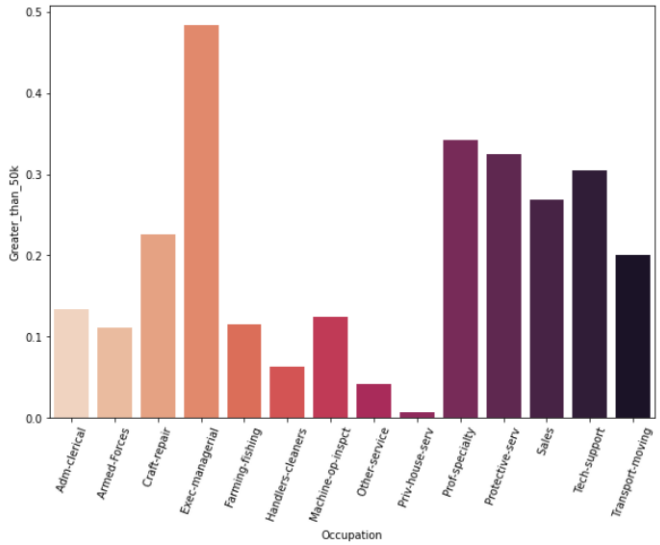

Fig. 10. Fraction of people having income greater than 50K vs Occupation

income.

5) ***Race:*** This variable represents the Race of the person. To visualize how race effects the income, we plot a bar plot between the Race and the fraction of people having income above 50K.

It is visible that Asian and White people have higher number of people above 50K income.

6) ***Sex:*** This represents the sex of the person
To visualize how marital status effects the income, we plot a bar plot between the Sex and the fraction of people having income above 50K.

Gender inequality is clearly evident from the graph. Out

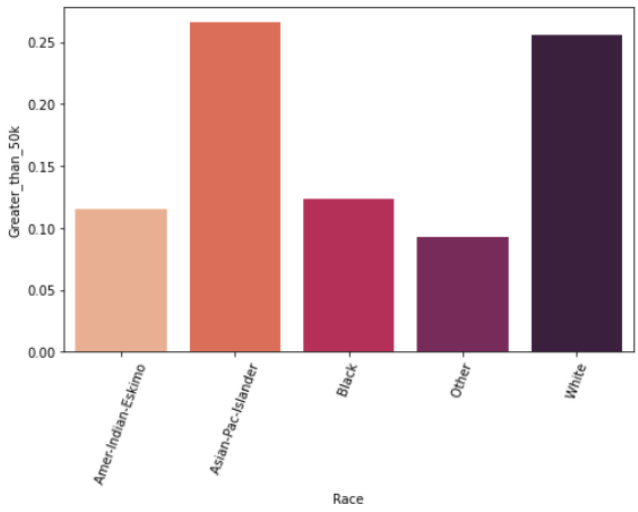

Fig. 11. Fraction of people having income greater than 50K vs Race

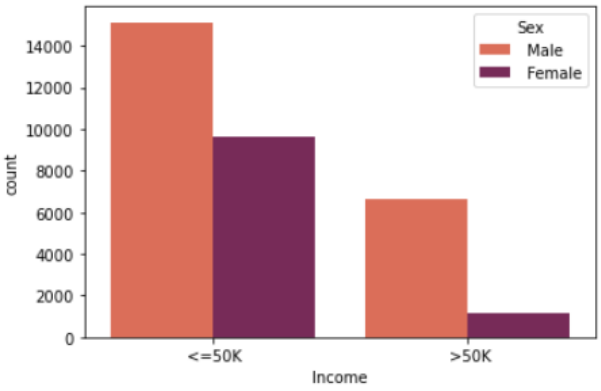

Fig. 12. Fraction of people having income greater than 50K vs Sex

of all the entries, female employees account for only 10.9 % above 50K while men acocunt for 30.57 %.

F. One Hot Encoding

To account for categorical variables in our model, we use One-Hot Encoder. With this, we convert each categorical value into a new categorical column and assign a binary value of 1 or 0 to those columns. Each integer value is represented as a binary vector.

G. Scaling the Data

Scaling refers to the process of putting values in the same range or scale such that no one variable dominates the others. Many Machine Learning algorithms use distance between two data points and if the features have varying magnitudes.

We have used scikit-learn's RobustScaler method to scale our data. RobustScaler transforms the feature vector by subtracting the median and then dividing by the inter-quartile

range.

H. *Checking Correlation*

Through the below heatmap, we can see no high values of correlation among our numeric variables-

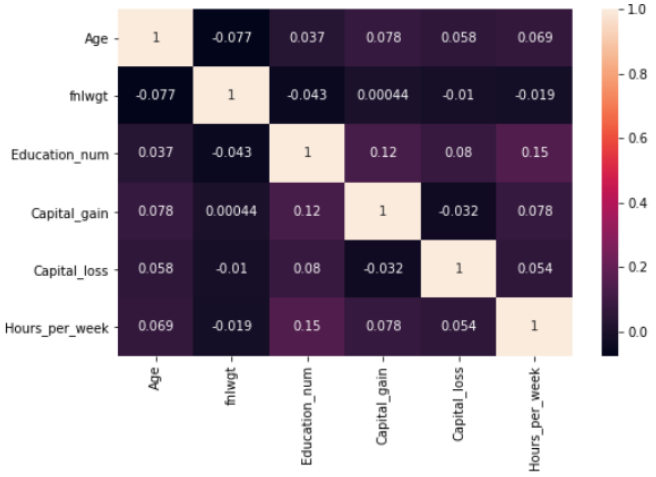

Fig. 13. Correlation Heatmap

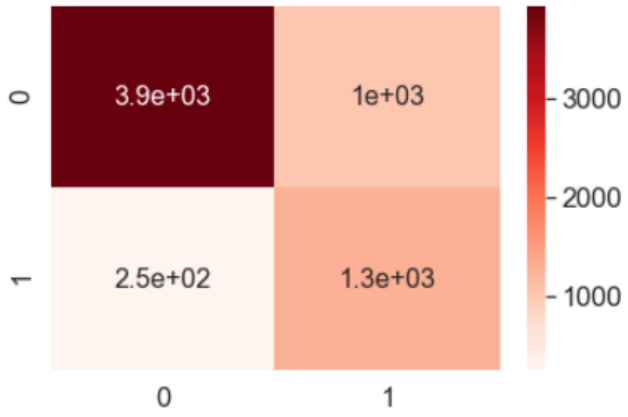

Fig. 14. Confusion Matrix for testing data

	precision	recall	f1-score	support
<=50K	0.94	0.79	0.86	4959
>50K	0.56	0.84	0.67	1553
accuracy			0.80	6512
macro avg	0.75	0.81	0.76	6512
weighted avg	0.85	0.80	0.81	6512

Fig. 15. Scores for testing data

IV. The Model

A. *Train Test Split*

We have the following X and y for our final model, which we divide our data into training and testing data..
X= 'Age', 'Work Class', 'fnlwgt', 'Education', 'Education num', 'Marital status', 'Occupation', 'Relationship', 'Race', 'Sex', 'Capital gain', 'Capital loss', 'Hours per week', 'Native country', 'Income' .
Y = Income

B. *Training the Model*

Using scikit-learn's GaussianNB() function on our X and y variables, we develop a Gaussian Naive Bayes classifier model.

C. *Predictions and Metrics*

The confusion matrix for our testing set in shown below
Based on the matrix, we can calculate the following -

The f1 score is 0.8028.

V. Conclusions

From our observations, we can conclude the following about the significant features-

- **Age** -With increasing age, people tend to have higher income.
- **Sex** - There is a clear gender disparity in income, with males having higher income.
- **Hours per week** - More hours per week on the job leads to more income.
- **Education** - People with higher education and with doctorates and professional degrees have higher pay.
- **Occupation** - From the data, we observed that self employed people have higher income and among the professional workers, executive managers have the highest fraction of people above 50K income.
- **Race and Native Country** - The American people, especially the whites, have higher income, followed by people from Asia-Pacific countries.
- **Capital gain or loss** -Making capital gains via profits from sale of properties of investments directly supplements the income, making it higher.

Thus, as shown by both the qualitative and the quantitative analysis, the income of the people of the United Stated depends upon many social, cultural and educational factors.

REFERENCES

[1] G. James, D. Witten, T. Hastie, and R. Tibshirani, An introduction to statistical learning: With applications in r. New York, NY: Springer, 2021.

[2] "Naive Bayes classifier," Wikipedia, 18-Aug-2021. [Online]. Available: https://en.wikipedia.org/wiki/NaiveBayesclassifier. [Accessed: 12-Oct-2021].

[3] R. Gandhi, "Naive Bayes classifier," Medium, 17-May-2018. [Online]. Available: https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c. [Accessed: 12-Oct-2021].

[4] J. Brownlee, "What is a confusion matrix in machine learning," Machine Learning Mastery, 14-Aug-2020. [Online]. Available: https://machinelearningmastery.com/confusion-matrix-machine-learning/: :text=A