**MODULE NAME: MACHINE LEARNING**

**MODULE CODE: 7072CEM**

**JAHNAVI PRIYA RATAKONDA**

**STUDENT ID: 14155136**

[**ratakondaj@uni.coventry.ac.uk**](mailto:ratakondaj@uni.coventry.ac.uk)

**TITLE: IDENTIFYING THE CENSUS INCOME OF ADULTS BY UTILIZING MACHINE LEARNING TECHNIQUES**

**1.ABSTRACT: *In the past few years, income inequality has been one of the major issues faced by different countries. Reducing economic inequality can help decrease in poverty, a critical concern in the United States and worldwide. Governments are taking proactive measures to address this issue The main purpose of this research is to use machine learning to identify the census income of adults which aims to provide insight into income inequity and help address this issue more efficiently. The task in progress is to predict whether an individual’s annual income in the United States falls into one of two categories: greater than $50K or less than or equal to $50K. This prediction completely depends on the specific attributes of the individual***

* ***KEYWORDS: Machine Learning, Classification, Gradient Boost classifier***

***GITHUBLINK:*** <https://github.com/ratakondaj/Machine-learning/blob/main/Machine%20learning%201.ipynb>

**2.INTRODUCTION:** In the present era, gross domestic product (GDP) is the main indicator used to measure a country's economic growth. The GDP offers a glance into a country's economic state for decision-making and regional development. In General, income approach is one of the methods that are used to measure a country’s economic output. In this paper a model is created to predict an individual’s GDP based on their adult income data. A model built to predict GDP depending on adult income data is important. Addressing income inequality is crucial for fair distribution of wealth. Detailed assessments can furnish valuable insights to enhance income. Our main objective is to research the most effective classification method for adult income data in the United States. I analyzed the most efficient classification method after using five classifiers. The Gradient booster model displayed the highest accuracy of 87%. This report details the problem and dataset, methodologies and future suggestions.

**3.PROBLEM AND DATASET:**

The Adult Income dataset is an important resource for experimenters who try to explore the connection between socioeconomic factors and income levels. Its main aim is to predict whether a person earns more than $50K per year, based on demographic and employment-related features. This dataset was particularly created for research purposes, and it involves a range of technical information and features that can be used to build and test predictive models. Researchers can explore the basic factors that drive income inequality.

This dataset has been collected from the UCI Machine Learning repository and contains 48842 instances each categorized by 14 attributes. The data in the dataset is provided in the CSV (Comma Separated Values) format and involves demographic and socio-economic features. This dataset can also be useful for studying benchmarking algorithms, exploring techniques in machine learning, the relationship between various demographic and socio-economic factors and an individual’s income level.

**This dataset contains:**

**A) INDIVIDUAL PERSONAL ATTRIBUTES:**

The **age** of the person is numeric (data type) **Description**: Age at the contact data

The **workclass** of the person is categorical (data type) **Description:** It refers to the type of work or occupation that an individual holds.

The **fnlwgt** of the person is numeric (data type) **Description:** In survey sampling, this technique is used to assign weights to the general population sample.

The **Education** of the person is categorical (data type) **Description:** This denotes the educational level attained by the person that is the highest.

The **Education-Num** of the person is numeric (data type) **Description:** It matches the level of education achieved by the person.

The **Marital Status** of the person is categorical (data type) **Description:** It shows the marital status of the person

The **Occupation** of the person is categorical (data type) **Description:** This attribute denotes the line of work that a person is involved in.

The **Relationship** of the person is categorical (data type) **Description:** This attribute pertains to the family role of a person's relationship with other members of their family.

The **Race** of the person is categorical (data type)

**Description:** This attribute denotes the ethnic group to which the person belongs.

The **sex** of the person is categorical (data type)

**Description**: This feature indicates the gender of the person.

The **capital gain** of the person is numeric (data type)

**Description:** This feature denotes the profits made by the person.

**Capital-Losses** of the person is numeric (data type)

**Description:** This feature denotes the losses made by the person.

**Hours-per-week** of the person is numeric (data type)

**Description:** This feature indicates the total hours an individual dedicates to work every week.

**Native country** of the person is categorical (data type) **Description:** This feature indicates the country of the origin of the individual

**B) TARGET VARIABLE:**

The target variable of this dataset is income, which is classified into two **classes**: >50K and <=50K specifying whether an individual’s annual income meets above $50K or not

**4. METHODOLOGY**

***a) Data Gathering:***

The Adult Income Dataset, collected from the U.S. Census Bureau database and available at the UCI Machine Learning Repository, is an in-depth collection of demographic information and documented annual income of individuals aged 16 and older from several regions in the United States.

***b) Creating an Environment:***

The Anaconda prompt is responsible for creating the environment and installing the packages, libraries, and modules in the system. In this environment, after installation, we use Jupiter notebooks to import libraries and modules. Jupiter Notebook has the greatest advantage in that is, it allows users to modify the code as frequently as possible. This will be beneficial for people who are working in data and software development and who can experiment with various applications and produce the best results.

***c) Data Pre-Processing:***

* I utilized the ‘read.csv ()’ function from pandas to load the dataset from the CSV file; after loading the dataset, I displayed the data frame to analyze its structure and contents.
* I utilized info () to get a brief overview of the data frame and obtain a summary.
* Missing values are checked using isnull(), and unique values are identified with nunique(). All the categorical columns are encoded with the label Encoder from ‘sci-kit-learn’.
* Split the data into training and testing sets using train\_test\_split() from scikit-learn.

***d) Classification:*** To achieve the highest accuracy, among all the classifier models these four models are used:

***1. Decision Tree Classifier:***

* The Decision Tree Algorithm is a fundamental technique widely used in machine learning in classification tasks.
* The decision tree algorithm is capable of handling both numerical and categorical data, and it can be easily interpretable and divides the data in the best way into different classes.

***2. Random Forest Classifier:***

* Random Forest classifier is a supervised learning algorithm
* This model is created from subsets of data and used to make predictions and categorize the information in the data frame.
* This model excels in recognizing and filling in any missing values within the data frame.

***3. Gradient Booster Classifier:***

* This algorithm has been formed to ensure maximum speed and performance
* It enables feature selection and comprehension of the dataset.
* This classifier trains new models in a sequential manner to rectify the mistakes present in the prior models, as each new

the model release focuses on its own mistakes

***4. Logistic Regression:***

* This algorithm is used to identify the relationship between independent variables and binary dependent variables.
* Logistic regression is commonly viewed as a classification method in machine learning instead of a regression technique.
* This algorithm is developed to estimate categorical results instead of continuous values specially in binary classification problems

***5. Naive Bayes:***

* It is a significant tool in machine learning for computational efficiency and interpretability.
* Naïve Bayes is highly effective in terms of computation and can handle large datasets with numerous features.
* This algorithm evaluates the likelihood of each possible result based on the input features and then chooses the outcome with the highest probability as the expected result.

***e) Evaluation Metrics:***

After executing five distinct classification models and choosing the main optimal one that produced the highest accuracy, we employed metrics comprising model accuracy, precision, recall, and f1-score to support our evaluation criteria. Through this process, we were able to calculate our model's performance efficiently.

***Accuracy:*** It refers to the number of forecasted instances.

|  |
| --- |
| Accuracy = (TP+TN)/(TP+FP+TN+FN) |

***Precision:*** This method helps to determine the number of data instances that hold positive predictive values.

|  |
| --- |
| Precision = TP/TP+FP |

***Recall:*** Recall is the percentage of true positive results accurately identified.

|  |
| --- |
| Recall = TP/TP+FN |

***F1 Score:*** The F1 score is a metric that takes into record both recall and precision.

|  |
| --- |
| F1 score = 2\*Precision\*Recall/Precision+Recall |

***Support:*** Support points to the number of instances estimated when computing a particular metric.

|  |  |  |
| --- | --- | --- |
|  | Y-new  (subscription) | Precision |
| Decision Tree | 0 | 0.89 |
| 1 | 0.60 |
| Random Forest | 0 | 0.89 |
| 1 | 0.74 |
| XGB Calssifier | 0 | 0.90 |
| 1 | 0.77 |
| Logistic Regression | 0 | 0.85 |
| 1 | 0.71 |
| Naïve Bayes | 0 | 0.82 |
| 1 | 0.68 |

Table 1: Classification comparison table.

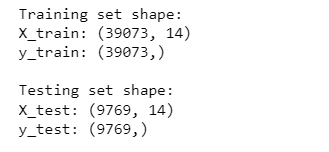
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Y-new  (subscription) | Recall | F1  Score | Sup-  port |
| Decision Tree | 0 | 0.87 | 0.88 | 7479 |
| 1 | 0.63 | 0.62 | 2290 |
| Random Forest | 0 | 0.93 | 0.91 | 7479 |
| 1 | 0.64 | 0.69 | 2290 |
| XGB Calssifier | 0 | 0.94 | 0.92 | 7479 |
| 1 | 0.68 | 0.72 | 2290 |
| Logistic Regression | 0 | 0.94 | 0.89 | 7479 |
| 1 | 0.45 | 0.55 | 2290 |
| NaïvBayes | 0 | 0.95 | 0.88 | 7479 |
| 1 | 0.33 | 0.45 | 2290 |

Table 2. Classification Comparision table2

**5) EXPERIMENTAL RESULTS AND ANALYSIS:**

**a) Experimental setup:**

This experiment is a study that utilized the adult income dataset. Python packages were integrated into one application using an anaconda prompt. Jupyter Notebook was used for developing and evaluating machine learning models by splitting the dataset into training and testing modules. By using a random split and determines its performance established on the test dataset. This platform helps to perform all the machine learning models in the given assignment.



**ROC CURVE:**

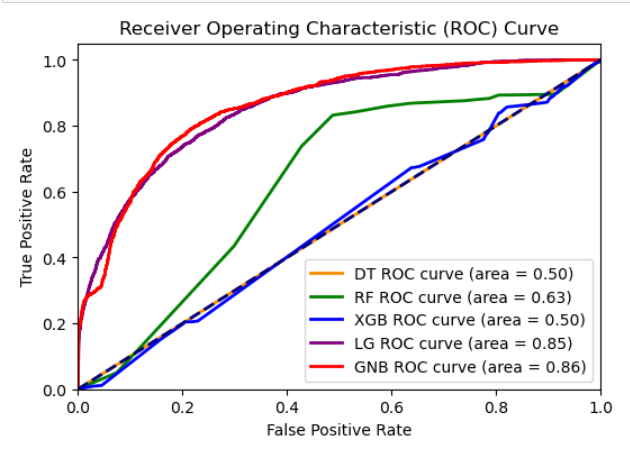


Fig 1. ROC curve of all the algorithms

The ROC curve assists the experts to identify the most efficient model for the selected dataset and it labels with the false positive rate on the x-axis and true positive rate on the y-axis. It gives the valuable perception into the performance of different classifiers in difference between positive and negative classes.

In the above curve states that the gaussian Naive Bayes classifier have the highest accuracy score which is closer to the top left corner indicates the highest performance among all the other classifiers. And the random classifier is indicated by a diagonal dashed line.

**b) Result and Discussion:**

|  |  |
| --- | --- |
| Classifier | Test  Accuracy |
| XGB classifier | 0.88 |
| Random Forest | 0.86 |
| Logistic Regression | 0.83 |
| Decision Tree | 0.82 |
| Naive Bayes | 0.81 |

Table 3: Model Accuracy

Among all the above classifiers that are used to calculate the performance of the dataset Gradient booster classifier got highest accuracy with 88% whereas random forest got 86% accuracy and decision tree got 82%, logistic regression got 83% and Naive Bayes got 81% of accuracy.

According to the adult income dataset after performing all the models we can clearly find that there are more adults who are earning income <=$50K when compared to adults who are earning income >$50K.

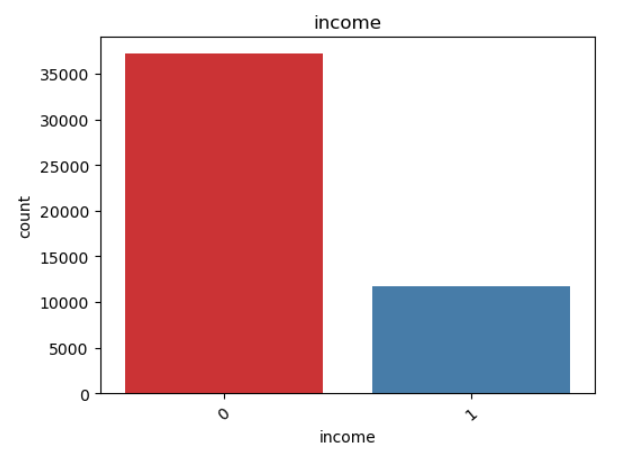


Fig 2. Comparision of income variable.

**6) SOCIAL, ETHICAL, LEGAL AND PROFFESSIONAL CONSIDERATION**

***Social Consideration:*** It means that this dataset is used to create a positive change in society by handling the individual data in a sensitive way.

***Legal Consideration:*** It means personal information like age, gender etc. should be treated as confidential data to avoid legal issues.

***Ethical Consideration:*** It means that informing prior to the individuals whose data is being used in this dataset is better to avoid these ethical issues***.***

***Professional Consideration:*** Finally, the data professionals can take advantage of the adult income dataset to create perception for social policy and economic growth.

**7) FUTURE WORKS AND CONCLUSION:**

This research paper applied five algorithms which have been used to predict people’s income levels. Among them the Gradient Booster model has achieved the highest accuracy, 88% ever recorded by any income prediction model. I believe this method could be further refined by integrating other advanced data processing methods, machine learning and deep learning methods thereby enhancing the model’s overall accuracy. This research could have significant consequences for a wide scope of industries, including finance, healthcare, and insurance where predicting income levels is vital for developing knowledgeable decisions.

**8.REFERENCES:**

***Reference 1:***

Chakrabarty, N., & Biswas, S. (2018, October). A statistical approach to adult census income level prediction. In *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)* (pp. 207-212). IEEE.

***Reference 2:***

Chockalingam, V., Shah, S., & Shaw, R. (2017). Income Classification using Adult Census Data.

***Reference 3:***

Moe, E. E., Win, S. S. M., & Khine, K. L. L. (2023, February). Adult Income Classification using Machine Learning Techniques. In *2023 IEEE Conference on Computer Applications (ICCA)* (pp. 91-96). IEEE.

***Reference 4:***

Deepa, N. (2022, July). Improving Performance Analysis in Classification with Accuracy of Adult Income Salary using Novel Gated Residual Neural Network in Comparison with Logistic Regression Algorithm. In *2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)* (pp. 1-5). IEEE.

***Reference 5:***

Chen, J. (2021). *Feature Significance Analysis of the US Adult Income Dataset*.

***Reference 6:***

Thapa, S. (2023). Adult Income Prediction Using various ML Algorithms. *Available at SSRN 4325813*.

**APPENDIX:**

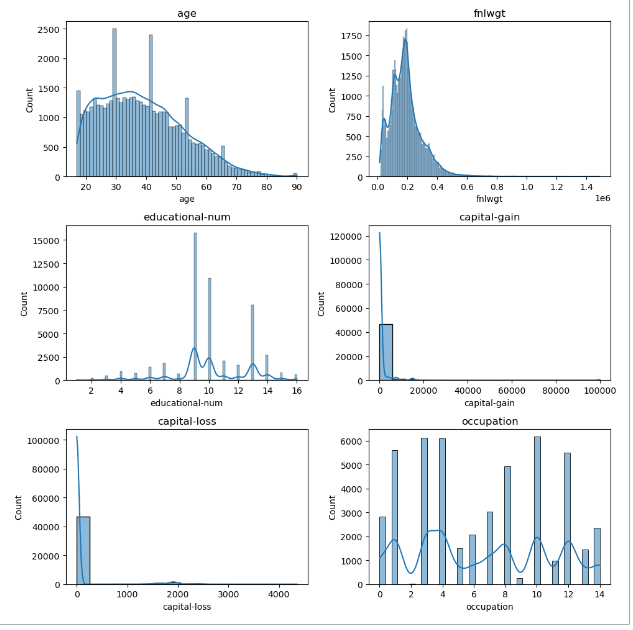


Fig 1. graphs of numerical columns

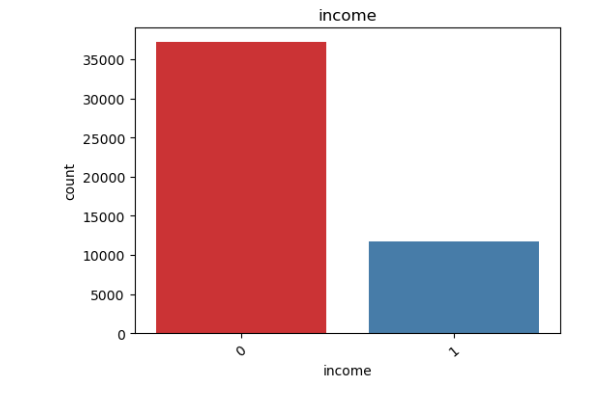


Fig 2. bar plot of income variable

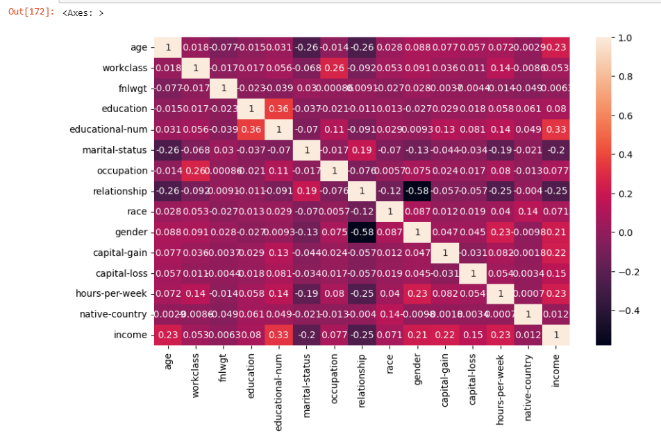


Fig 3. Heat map of all columns

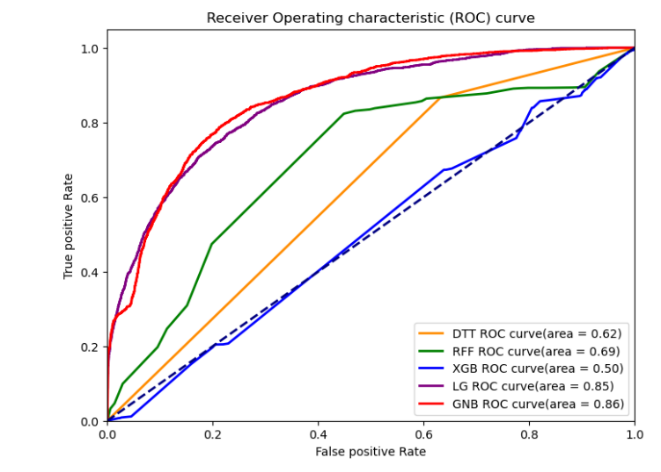


Fig 4.ROC curve of all the algorithms

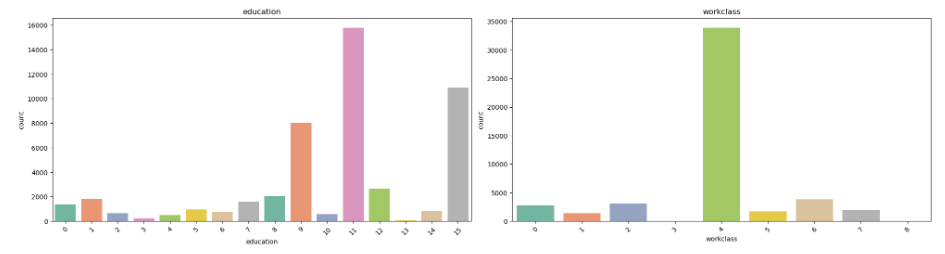


Fig 5. Bar plot of education and workclass

**#**