**Project Report**

**On**

**Census Income**

**by**

**VAISHNAVI BALAJI**

**PG Program**

**In**

**Data Science, Machine Learning and Neural Networks**

**Batch-1844**

**TABLE OF CONTENTS**

CHAPTER 1  **INTRODUCTION**

* 1. INTRODUCTION
  2. PROBLEM STATEMENT
  3. AIM
  4. OBJECTIVES

CHAPTER 2 **EXPLORATORY DATA ANALYSIS**

CHAPTER 3 **DATA PRE-PROCESSING**

CHAPTER 4 **TRAINING AND TESTING MODEL**

CHAPTER 5 **RESULTS & CONCLUSION**

**CHAPTER 1**

**INTRODUCTION**

* 1. **Introduction**

Humans have become increasingly dependent on data and information in society over the past two decades, and as a result, technologies have developed for massive data storage, analysis, and processing. Data mining and machine learning have not only used them to get knowledge and make discoveries, but also to investigate certain hidden patterns and ideas that have helped make difficult-to-predict future events possible. In recent years, the issue of income disparity has received a lot of attention. The goal of eliminating this problem does not appear to be limited to improving the lot of the poor. Americans feel that the emergence of economic disparity is unacceptable and demands.

In this database, Ronny Kohavi and Barry Becker from the Census Government collected statistics from the 1994 Census Bureau Database (Data Mining and Visualization, Silicon Graphics). The following criteria were used to obtain a group of largely clean records-Average Age greater than sixteen, AAGI greater than hundred, average final weight greater than 1 and hours per week greater than zero. The weights on the CPS files, which measure the civilian non-institutional population of the US, are controlled to independent estimates. The Census Bureau's Population Division creates these each month. Three different sets of controls are employed which are: an estimation of the 16+ population in a single cell for each state, controls by age and sex for Hispanic origin, controls based on age, sex, and race.

In the weighting program, all three sets of controls are used and "rake" through them six times, returning to each set of controls at the end. Estimates are population totals obtained from CPS by constructing "weighted tallies" of any specified socioeconomic attributes of the population. People with comparable demographic traits need to weigh similarly. It's vital to keep in mind one important qualification to this assertion. This means that the assertion only holds true within states because the CPS sample is essentially a collection of 51 state samples, each with its own probability of selection.

* 1. **Problem Statement**

The stark wealth and income disparity is a significant worry, particularly in the US. The possibility, one good motivation to lower the global poverty rate is to economic inequality is increasing. The idea of universal moral equality guarantees long-term progress and enhances the strength of a nation's economy. different governments various nations have been making every effort to address this issue and give the best answer you can. This research tries to demonstrate the using data mining and machine learning methods to provide the answer to the issue of income equality.

* 1. **Aim**

The ultimate goal is to create a classification model that can determine, given certain characteristics like age, education, occupation, gender, race, etc., whether the annual income of a random adult American citizen is less than or greater than fifty thousand dollars.

* 1. **Objective**

(i)Perform Exploratory Analysis on the data set.

(ii) Analyze the relationship between the features and the income.

(iii) Perform data pre-processing algorithms on the data.

(iv) Develop models using Logistic Regression, Decision Tree Classifier, Random Forest

Classifier, K Neighbors Classifier, Support Vector Classifier, Gaussian NB Classifier, XGB Classifier.

(v) Testing the model and evaluating using the metrics and perform Hyperparameter tuning on the selected model.

**CHAPTER 2**

**EXPLORATORY DATA ANALYSIS**

Summary of the given dataset is as follows-

* There are 32560 rows and 15 columns.
* Out of the 15 columns- 6 columns are of integer data type and 9 columns are of object data type.
* Each entry in the dataset contains the following information-

1. Age- Age of the individual.
2. Workclass- a broad phrase used to describe someone's job position. For example, Private, Self­emp­not­inc, Self­emp­inc, Federal­gov, Local­gov, State­gov, Without­pay, Never­worked.
3. Fnlwgt- The weights on the Current Population Survey files, which measure the civilian non-institutional population of the US, are controlled to independent estimates. Here at the Census Bureau, the Population Division creates them on a monthly basis.
4. Education- the highest level of education pursued by the individual. For example- Bachelors, Some­college, 11th, HS­grad, Prof­school, Assoc­acdm, Assoc­voc, 9th, 7th­8th, 12th, Masters, 1st­4th, 10th, Doctorate, 5th­6th, Preschool.
5. Education-num- best academic achievement expressed as a number.
6. Marital-status- a person's state of marriage, while Married­AF­spouse refers to a spouse in the Armed Forces, Married­civ­spouse refers to a spouse in the civilian world, Married­civ­spouse, Divorced, Never­married, Separated, Widowed, Married­spouse­absent, Married­AF­spouse
7. Occupation-the type of occupation of an individual. Tech­support, Craft­repair, Other­service, Sales, Exec­managerial, Prof­specialty, Handlers­cleaners, Machine­op­inspct, Adm­clerical, Farming­fishing, Transport­moving, Priv­house­serv, Protective­serv, Armed­Forces
8. Relationship- depicts the person in relation to others. For instance, a person might be a husband. Each record only contains one relationship trait, which makes marital status fairly unnecessary.
9. Race-provides information about an individual’s race.
10. Sex-depicts biological gender of the individual.
11. Capital-gain-profit earned by an individual on the sale of an asset.
12. Capital-loss-loss incurred by an individual on selling an asset below the price it was purchased.
13. Hours-per-week- the number of hours a person has reported to work each week.
14. Native country-denotes a person’s country of origin.
15. Income- earnings of an individual on delivering services to an organization.

* In features-Work class and Occupation there are "?" values which are replaced by the mode of their respective columns.

**Univariate Analysis**

* Age of the members in the census are from 16-90 though most of the members are in the age group- 16-50.

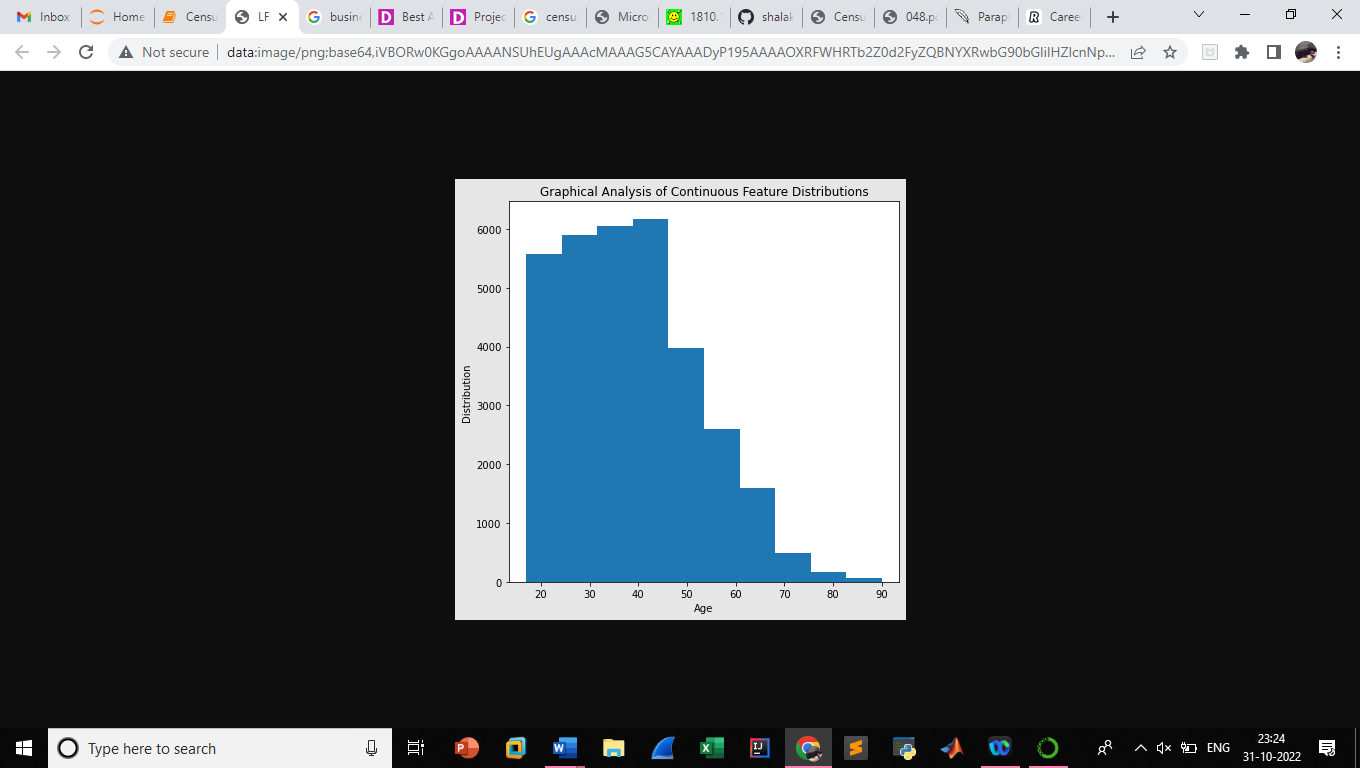
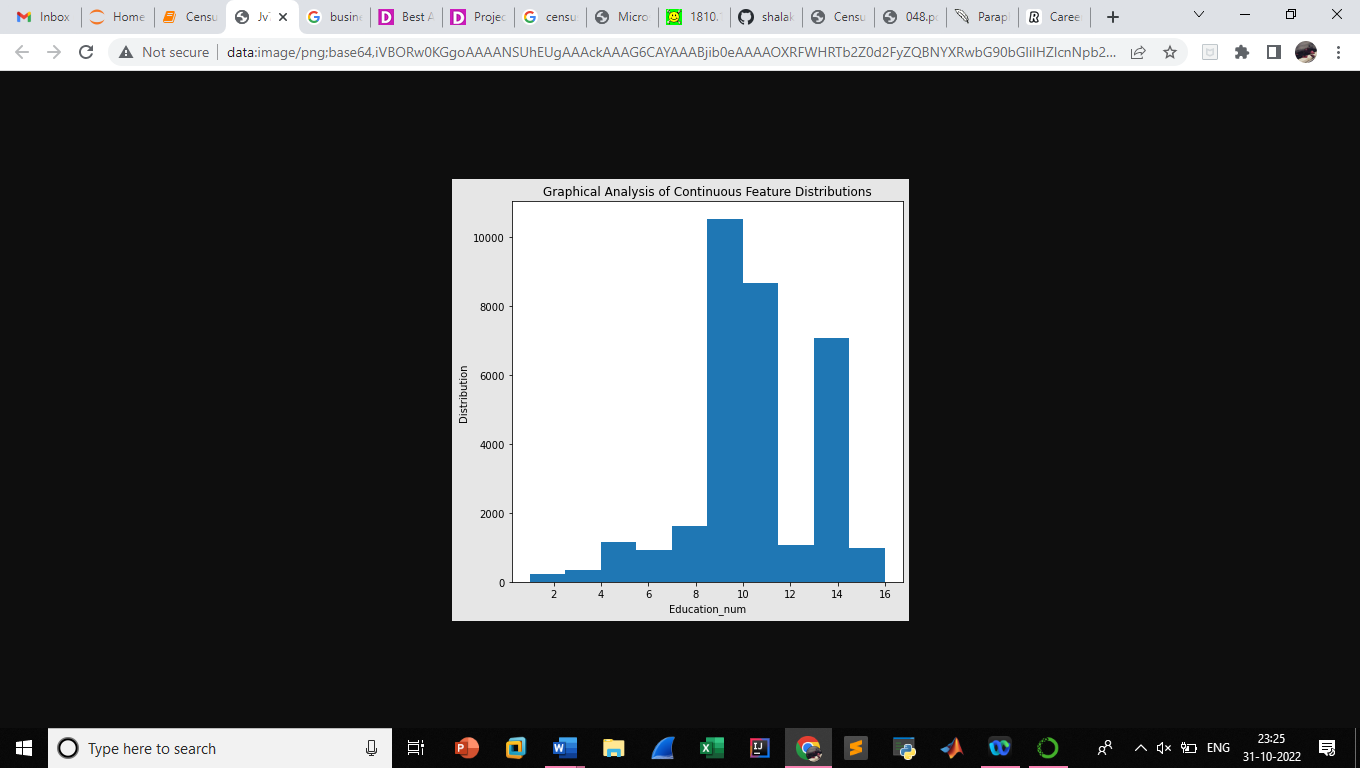
 

Figure 2- Distribution of Feature- Education\_num

Figure 1- Distribution of Feature- Age

* Fnlwgt is ranging from 0.01 to 0.3 for the majority, there are few members with fnlwgt from 0.3 to 0.8.
* Many members are having number of education years between 8.5 to 11 and 13.5 to 14.5.
* Capital gain is from 0- 10000 and few members have capital gain from 10000 to 30000.
* Capital loss has a peak from 0-400 and a small distribution from 1300 to 2600.
* Most of the members in the census work from 32 to 40 hours per week followed by 50-60 hours per week.
* Majority of the members are working in private sector followed by self-employed, local-government.
* Most of the members have an educational qualification of High school grad followed by members graduating from some college and bachelors and masters.

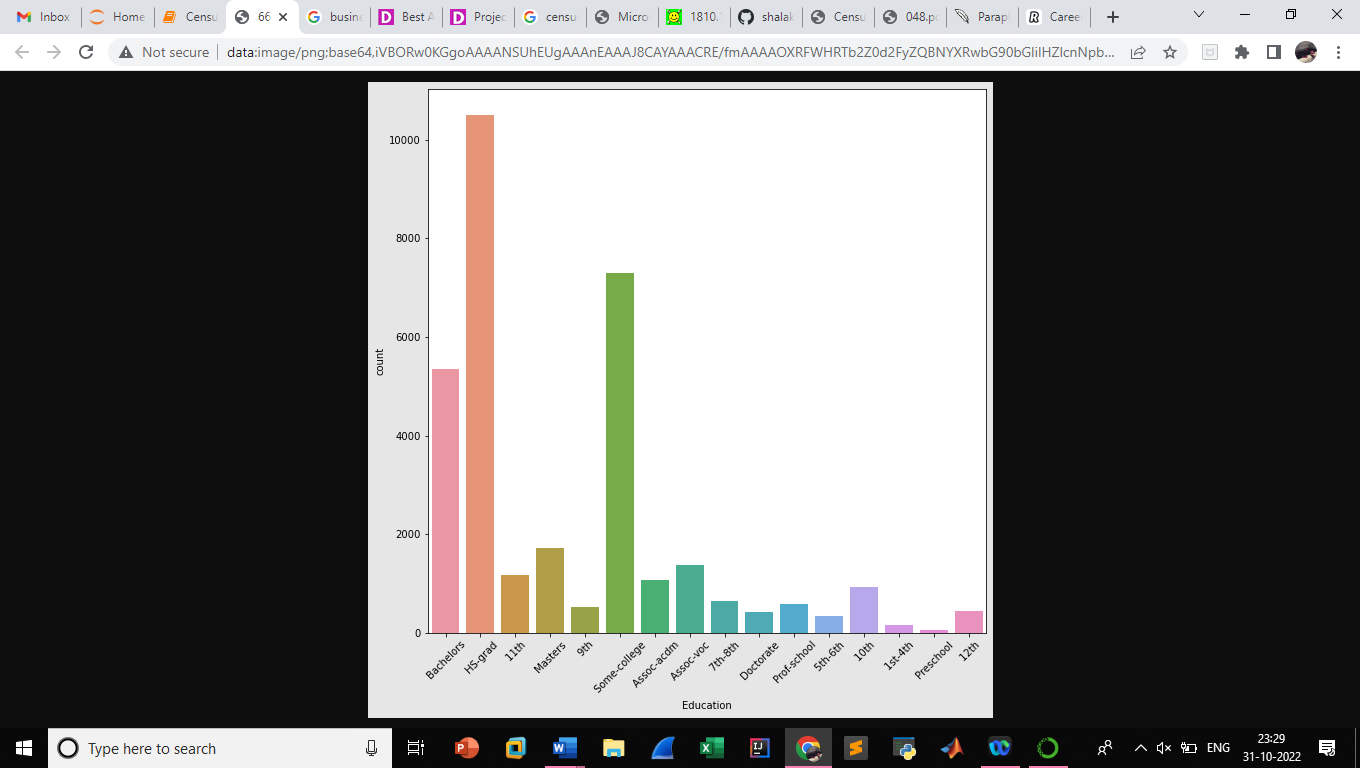


Figure 3- Distribution of Feature- Education

* Majority of the members are married to civilian spouse, followed by people who are not married, divorced. Approximately 80 people are separated and widowed.
* When it comes to occupation, members are majorly working in their profession-specialty followed by executive-managerial, craft-repair.
* After craft repair, members are in Adm-clerical occupation followed by sales and other services.
* Most of the members are husband in relation, almost 8200 members are not-in-family, 5000 members are own-children in their family.
* 3500 members are not married, 1600 are wife in relation and 700 members are other-relatives.
* 28000 members belong to white race followed by black race, Asian-Pac-Islander, Amer-Indian-Eskimo and others.
* Around 20500 are Male and 1300 are females.
* 25900 members belong to United States and small group of members belong to Mexico.
* Around 25800 members have an income less than or equal to 50k and around 7500 have income more than 50k

**Bivariate Analysis**

* In bivariate analysis, the effect of different features on the target which is the income is studied and the following results are obtained.
* Large number of members of age-36-48 have income greater than fifty thousand dollars.
* For lower age group and higher age group few members have income more than fifty thousand dollars.
* Bachelor is the largest group who earn more than 50k followed by High School graduates, college graduates and masters. High School graduates are the largest group who earn less than 50k followed by college-graduates and bachelors.
* People with 13 education\_num earn above 50k followed by 9,10 and 14. People with 9 education\_num are the highest number of ppl who earn less than 50k followed by 10,13 and 7.
* Never-married people are the highest number of people who earn less than 50k followed by married with civilian spouse and divorced. People who are married with civilian spouse are the highest number of people who earn greater than 50k.
* White people are the highest number of people for both less than 50k and above 50k category followed by black, asian-pac islander.
* Male members are high in both less than 50k and above 50k category.

**CHAPTER 3**

**DATA PRE-PROCESSING**

* Workclass, Occupation and Native\_country had “?” entries in their columns which were replaced with the respective column mode value.
* Native-country had 41 unique values which are difficult to encode hence eliminating the column.
* The categorical columns- Workclass, Education, Marital Status, Occupation, Relationship, Race, Sex, Income are encoded using the Label Encoder.
* When box plot was plotted for the data, the features- Age, Workclass, Fnlwgt, Education, Education\_num, Race, Capita gain, Capital loss, Hours per week have many data points beyond the range which were treated based on the z-score.

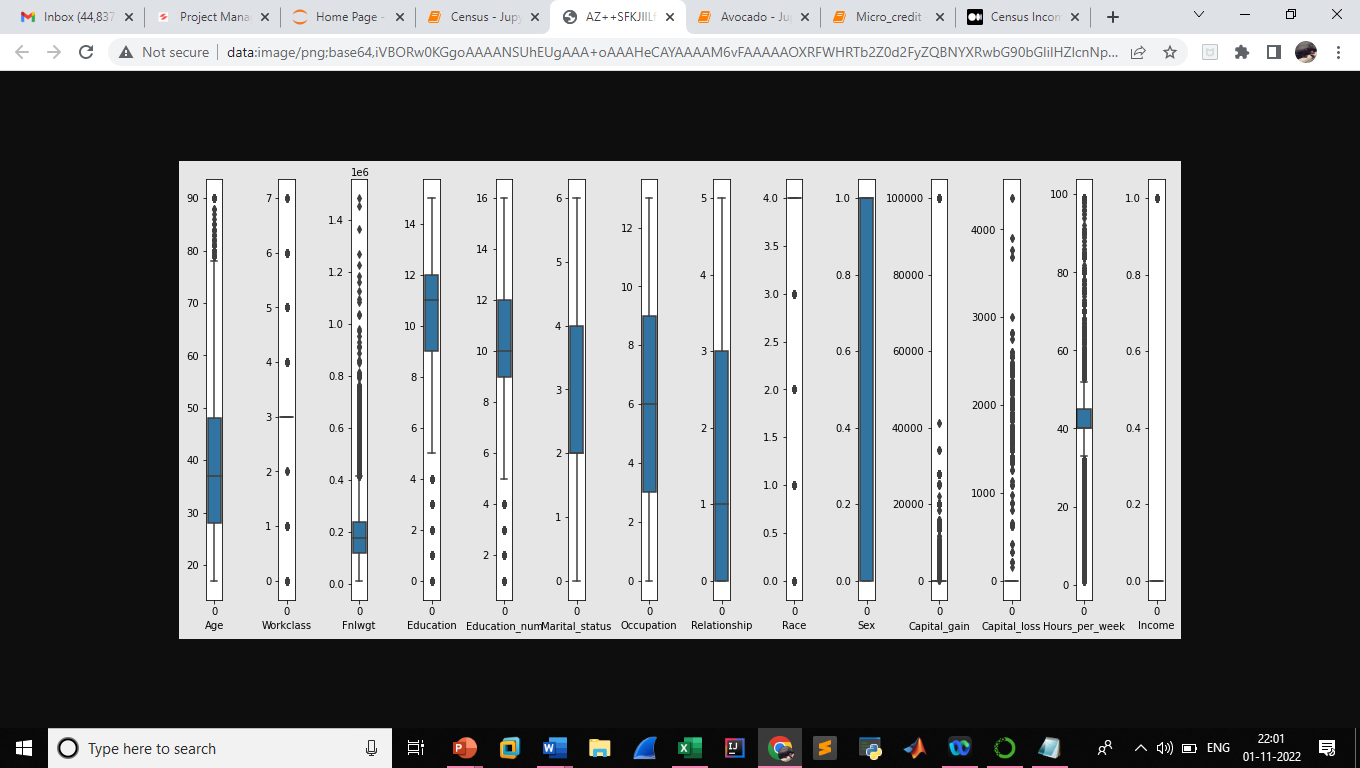
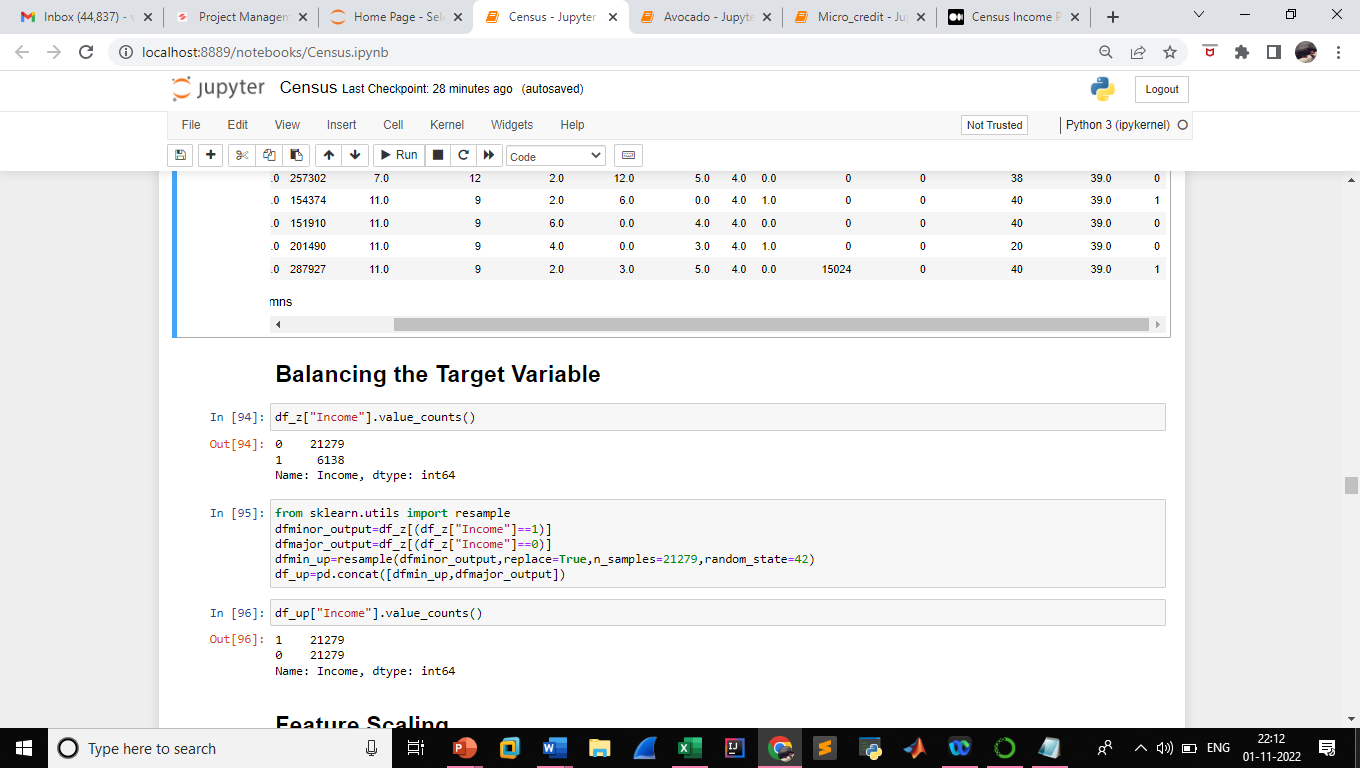
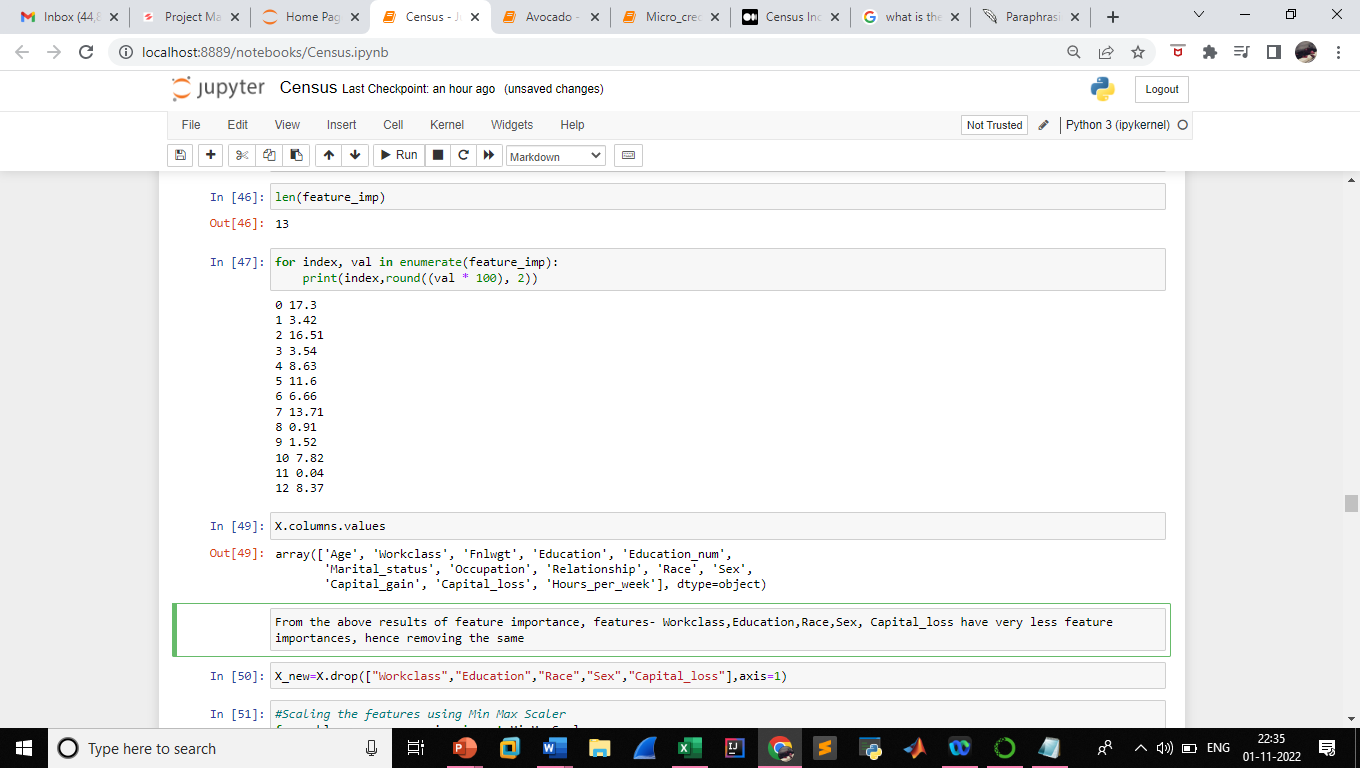


Figure 4-Box Plot of the dataset to detect outliers

* After removing the outliers, the percentage of data loss was 12.2%.
* The target in this project which is the income of an individual which has to be classified above fifty thousand dollars or not. The value for above fifty thousand dollars was encoded as one (“1”) and below fifty thousand as zero (“0”). The distribution of one and zero is very imbalanced which might be cause of inaccuracy during model development. Hence, the target was balanced using the resampling algorithm.



* After removing the outliers, the complete dataset was split into features and target, features represented by “X” and target represented by “Y”.
* There are thirteen features and one target, feature importance was analysed and the result obtained was the features Workclass,Education, Race, Sex and Capital\_loss have very less feature importance hence were eliminated from the features list.



* The features were scaled using the Min Max Scaler algorithm, this estimator scales and translates each feature separately so that it falls inside the training set's predefined range, such as between zero and one.
* Further, the features were subjected to power transformation so that the mean and standard deviation was valued at zero and one respectively. A class of parametric, monotonic adjustments called power transforms is used to give features a more Gaussian appearance. This is helpful when modelling heteroscedasticity (non-constant variance) problems or other circumstances when normalcy is preferred.

**CHAPTER 4**

**TRAINING AND TESTING MODEL**

* The income classification task was developed using following models- Logistic Regression Model, Random Forest Classifier, Decision Tree Classifier, K Neighbors Classifier, Support Vector Classifier, Gaussian NB Classifier, XG Boost Classifier. A loop was modelled to analyze the performance metrics for different random state values.
* Random Forest Classifier for random state-3 and test-size-0.25 gave accuracy score on training data – and testing data-. Compared to other models, Random Forest model had efficient accuracy score and less difference in scores of training and test data.
* With respect to training capabilities the efficient models are Random Forest, Decision Tree, K Neighbors and XG Boost, and with respect to less difference in accuracy of training and test data, the efficient models are Random Forest, XG Boost, Decision Tree and K Neighbors.
* The cross- validation scores of all the models were analyzed and the Random Forest Model achieved highest score of 94%, roc\_auc\_score of 93%, True Positive of 4854, True Negative- 5435, False Positive-157 and False Negative-647.

**CHAPTER 5**

**RESULTS AND CONCLUSION**

* Hyperparameter tuning was performed on the developed Random Forest Model using the Randomized Search CV algorithm for 7 samples.
* The best parameters for the model were the following: n\_jobs-2, max\_leaf\_nodes-None, max\_depth-70, criterion-gini, ccp\_alpha-0.0.
* After tuning the parameters, the False Negative has reduced from 647 to 644, F1 score on training data-100% and F1 score on testing data- 93% and the ROC Curve is shown below-

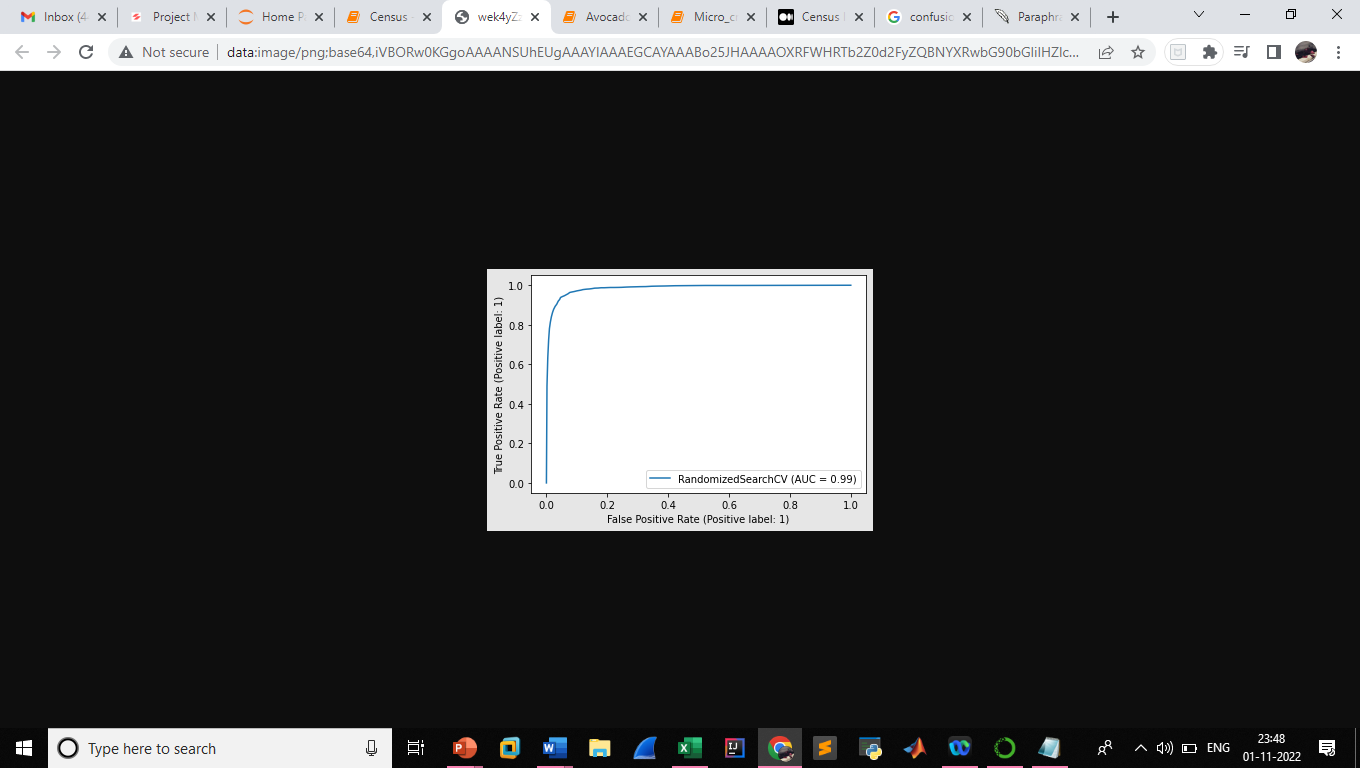


Figure 5-ROC Curve of the developed model