**ADULT CENSUS INCOME PREDICTION USING RANDOM FOREST**

**Acknowledgement**

We hereby thank Dr. Mohan and Dr.Sirisha who has given us the great opportunity to work on this project. He has been a great source of inspiration and his timely support and guidance has helped in the successful completion, whose creative suggestions and timely interventions helped us a lot during the process.

Then we would like to thank Mr. Sai Radha Krishna and the mentors for their enthusiastic approach and dedication which has indeed been a great source of inspiration and support to us. We thank the department of Electrical and Communication, JNTU Hyderabad for providing us with necessary infrastructure to do our work.

**TABLE OF CONTENTS**

1. Abstract

2. Introduction

3. Methods

4. Results and discussions

5. Conclusion

6. References

**Abstract**

In this project report we have a summary of our analysis and exploration of the Adult Census Data to come up with meaningful, important and interesting attributes of the data. Further, after having sufficient knowledge about the attributes we have performed a predictive task of classification, whether income is greater than or less than 50K/yr.

We also aim to measure the accuracy of different models using Random Forest. While considering different factors such as age, work class, gender, marital status, education, race, Occupation etc. using exploratory analysis and classification algorithms.

**INTRODUCTION**

Income is instrumental in deciding a person’s standard of living and the financial status in the society. It plays a key role in determining growth of nation. Our aim is to  
identify meaningful insights which can be the basis for  
many cleverer decisions. Our dataset contains 32561 records with various attributes such as occupation, age, relationship, hoursperweek, education, income and soon. Exploratory analysis will be done between dependent and independent variables.

In our project, we worked on the Census data set. In our initial stages, we pre-processed the data and developed understanding of the data and its useful features that explain the variances by doing various types of exploratory analysis. Later, we moved on to a classification task of predicting whether the income is >=50k/year from a person’s attributes, by using important features. For the classification task, we implemented a machine learning model, that after the initial task would prove useful. In the whole process, we got help from previous works on the same dataset cited in various literature.

**OBJECTIVE OF RESEARCH:**

The scope of this analysis is to understand relationship of various parameters which impact the income.

Description of attributes:

● age: the age of an individual (Integer greater than 0)

● workclass: a general term to represent the employment status of an individual (Private, Self­emp­-not-­inc, Self­emp­inc, Federal­gov, Local­gov, State­gov, Without­pay, Never­worked.

● fnlwgt: final weight. In other words, this is the number of people the census believes the entry represents(Integer greater than 0)

● education: the highest level of education achieved by an individual. (Bachelors, Some­college, 11th , HS­grad, Prof­school, Assoc­acdm, Assoc­voc, 9th , 7th,­8th , 12th, Masters, 1st,­4th , 10th , Doctorate, 5th,­6th , Preschool).

● education­num: the highest level of education achieved in numerical form (Integer greater than 0).

● marital­status: marital status of an individual. Married­civ­spouse corresponds to a civilian spouse while Married­AF­spouse is a spouse in the Armed Forces. ( Married­civ­spouse, Divorced,Never­married, Separated, Widowed, Married ­spouse­ absent, Married­AF­spouse).

 ● occupation: the general type of occupation of an individual (Tech­ support, Craft ­repair, Other­ service, Sales, Exec­ managerial, Prof ­specialty, Handlers cleaners, Machine ­op­inspct , Adm­ clerical, Farming­fishing, Transport­ moving, Priv­ house­ serv, Protective­serv, Armed ­Forces).

 ● relationship: represents what this individual is relative to others. For example an individual could be a Husband. Each entry only has one relationship attribute and is somewhat redundant with marital status. We might not make use of this attribute at all. (Wife, Own ­child, Husband, Not­ in ­family, Other­relative, Unmarried).

 ● race: Descriptions of an individual’s race (White, Asian­Pac­Islander, Amer­Indian­Eskimo, Other, Black).

● sex: the biological sex of the individual (Male, Female)

 ● capital­gain: capital gains for an individual (Integer greater than or equal to 0)

● capital­loss: capital loss for an individual (Integer greater than or equal to 0)

● hours­per­week: the hours an individual has reported to work per week (continuous).

● native­country: country of origin for an individual (United­States, Cambodia, England, Puerto­Rico, Canada, Germany, Outlying­US(Guam­USVI­etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican­Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El­Salvador, Trinadad&Tobago, Peru, Hong, Holand­Netherlands).

● The label (income): whether or not an individual makes more than 50,000 annually. (<=50k, >50k The original dataset contains a distribution of 23.93% entries labelled with >50k and 76.07% entries)

**INDUSTRIAL PROFILE**

1. **Business understanding:**

This phase involves clearly defining the project objectives and goals, and translating these goals into a problem statement.

2. **Data understanding:**

This phase involves collection of data and performing a preliminary analysis on the data to evaluate the data quality. Data understanding phase may also contain making subsets of data that may have any actionable patterns.

3. **Data preparation:**

This phase is the most time taking one in the data mining process. It involves cleaning the data, performing certain transformations on the data to get the final dataset

4. **Modelling:**

This phase involves selecting the appropriate modelling technique

5. **Evaluation:**

The models generated during the modelling phase are evaluated for quality and also it’s determined whether the business objective is which means whether the problem statement is solved or not

6. **Deployment:**

In this phase the effective models are finally put to use. It may be making a simple report or using the insights in the daily functioning of a company

**DATA COLLECTION**

1. Name: Adult Census Income Prediction Using Random Forest.

2. Input variables:

* + age
  + workclass
  + fnlwgt
  + education
  + education\_num
  + marital status
  + occupation
  + relationship
  + race
  + sex
  + capital gain
  + capital loss
  + hoursperweek
  + native country

3. Output variable: Income

4. Number of Observations: 32561

5. Number of Attributes/Variables: 15

6. Missing Values: workclass---1836

Occupation---1843

Nativecountry---583

**METHODS**

Steps we followed to build our models for predicting the adult census income:

1. **Getting a feeling of the dataset**

First step is to see what kind of data you have at hand: How many attributes are available per instance, if the attributes are in numerical format, if pre-processing is required and more importantly if there are any missing attributes. The adult census income dataset has both numerical data and categorical data and has missing attributes which means that special pre-processing is required.

2. **Exploring our attributes**

By exploring the attributes we observed that few inputs are positively correlated and few are negatively correlated. From heat map these are made clear and observations are done. We have also used boxplots, histograms, bar graphs for data visualisation and for better understanding.

3. **Outlier Detection**

Even after our pre-processing there are some values which are highly unexpected in comparison to the rest of the values of the same attribute in the dataset. We call these outliers. We observe such extreme values mainly in hoursperweek, fnlwgt, and only very few on age and educationnum. The instances that have such extreme values in their attributes are <5% of the full dataset. We decided to remove completely these instances as they are expected to make learning more difficult and the scoring less accuracy.

4. **Exploring our target**

We have used many ways to represent the data like boxplots, histograms, heatmaps etc.

5. **Splitting for Testing**

Before applying any algorithm, we will split the dataset into training and testing data so that we can train the model on the training data. Based on the results obtained from the testing data, accuracy from various algorithms can be measured.

**6. Classification Task:**

Since our target variable is categorical in nature, so we will try using algorithms such as logistic regression, decision trees, Naive Bayes, Random Forest etc. But the given topic is, Adult census income prediction using Random Forest, so we have used Random Forest algorithm.

🡪***. Random Forest***

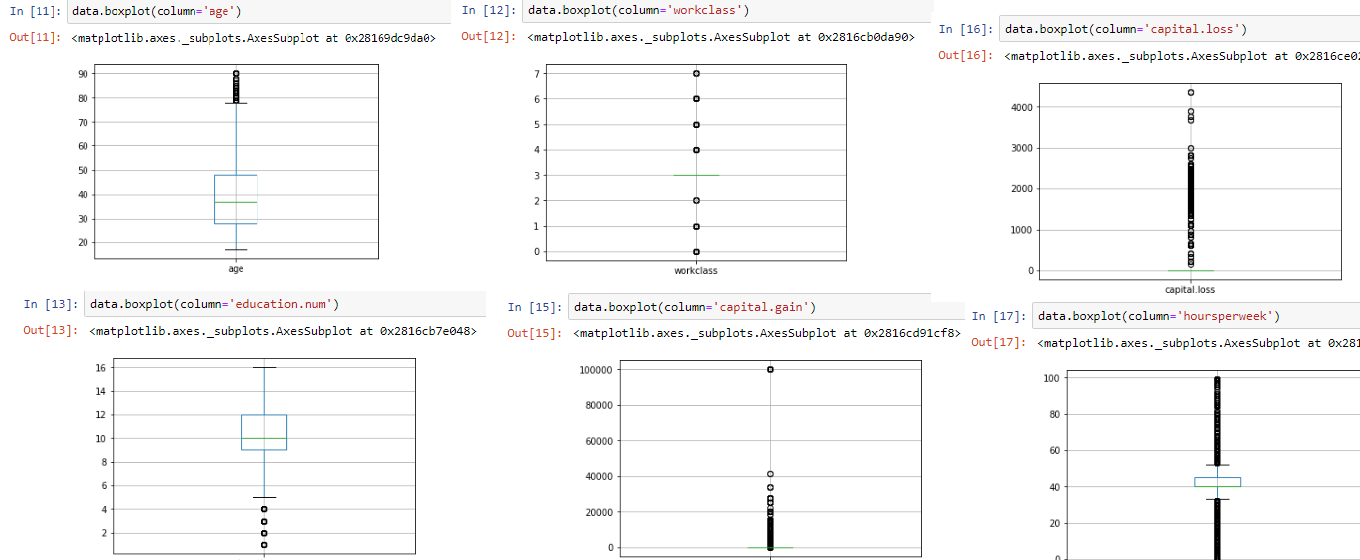
Random Forest is a combination of multiple trees. Each tree is provided with some random sample of data with replacement and gives different classifications. It takes votes from the result of all the trees and chooses the classification having maximum votes and when the dependent variable is continuous, it takes the mean from the outputs given by different trees. The number of attributes given to the trees for classification is considered by taking the square root of the total number of attributes. Although each tree works on different attributes. Thus each tree will have a different root node and split. So for the final result, the output of all the trees are considered. The number of tress to be taken can be tuned. One important feature of random forest is that it shows the most important as well as least important variables in the dataset The random forest algorithm can be used to solve both classification and regression problems. It can handle large dataset with n number of input variables. It automatically takes care of the missing values. Since it is combination of multiple trees, the accuracy is expected to be much higher than the single decision tree. Random forest is not as good for regression

problems as it is for the classification problems. With the help of our work, we will be predicting the income of the population and analyzing the factors which strongly affect the income.

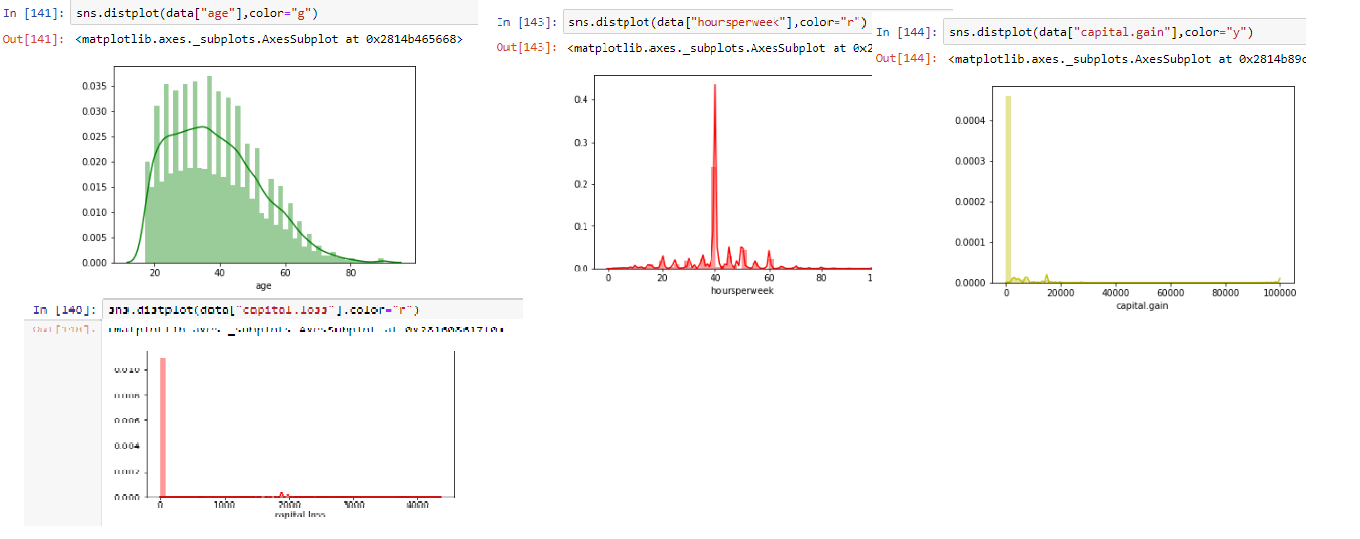
We will be giving suggestions based on the result obtained which level of qualification can lead to a higher income and people of which age group are earning more.

**4.1 Exploring Data Analysis**

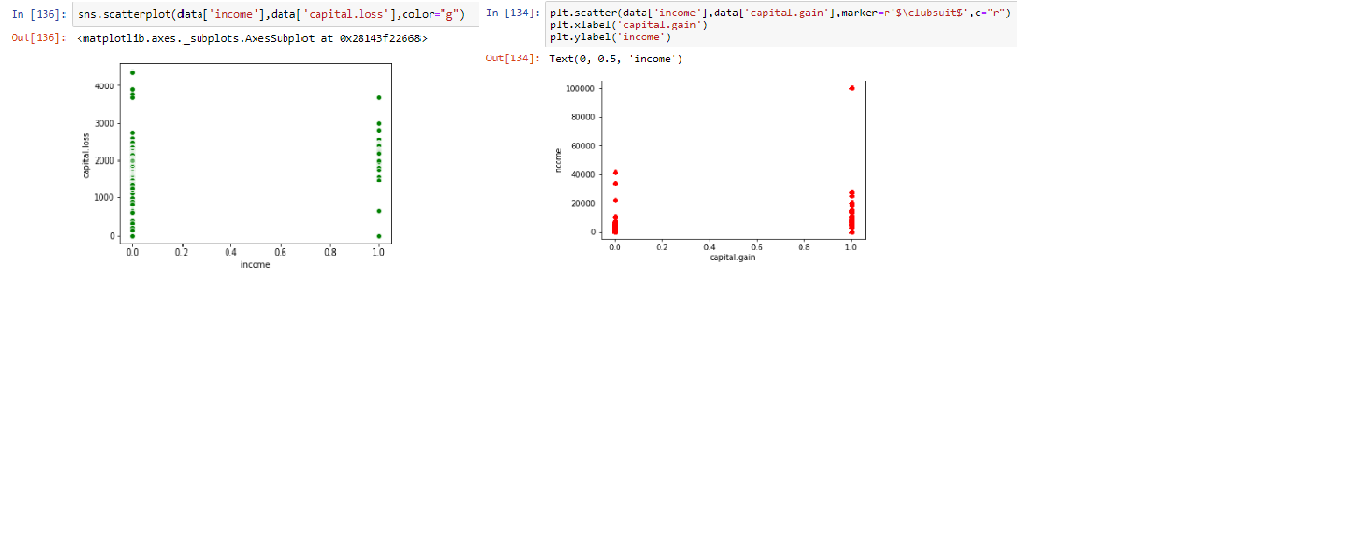
**>Outliers**

****

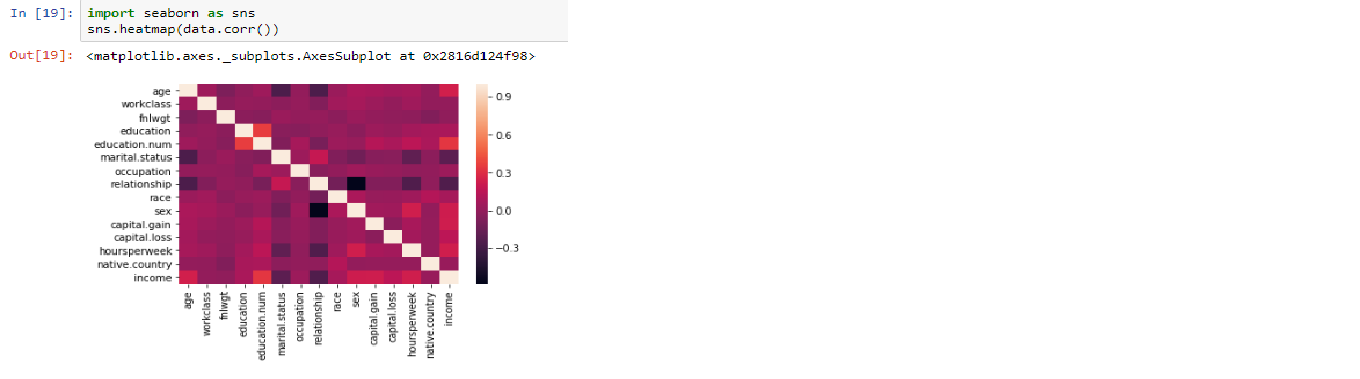
**>Distplots**

****

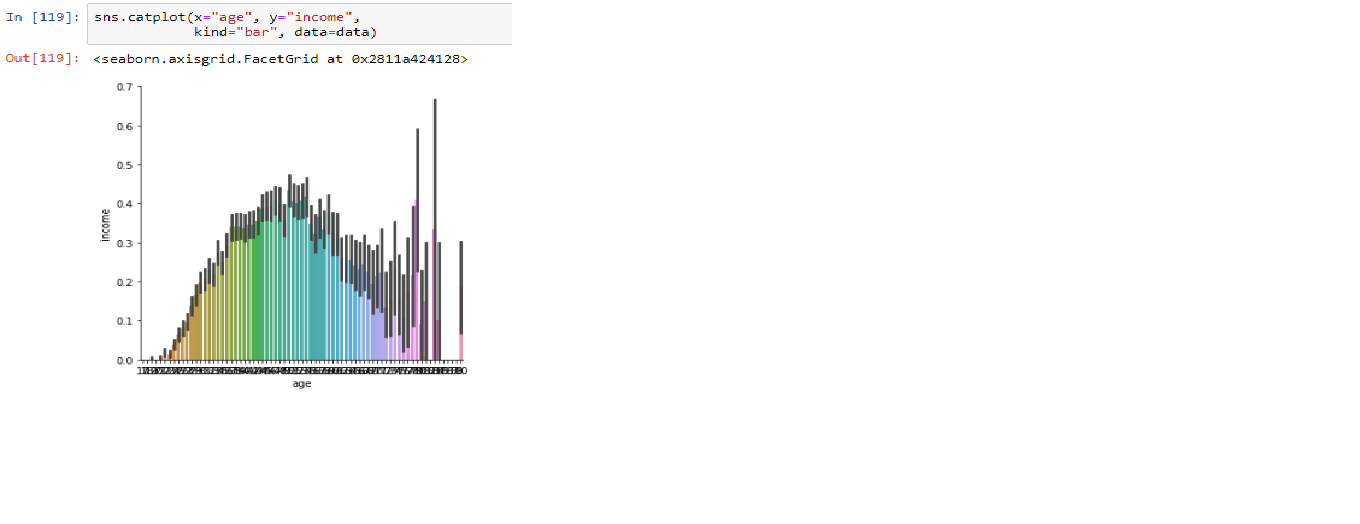
**>Scatter Plot**

****

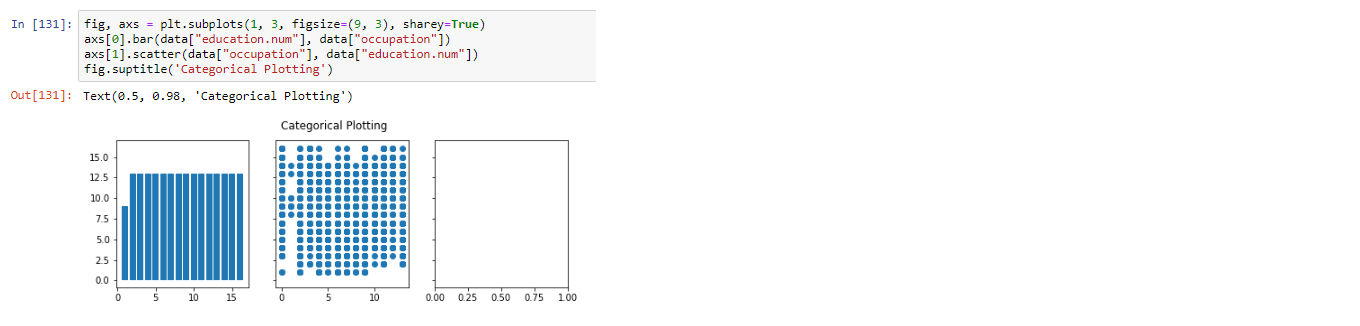
>Heatmap



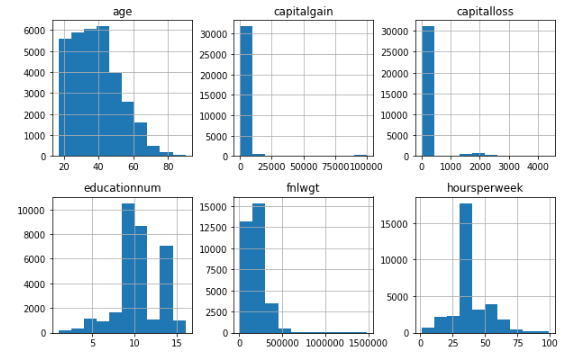
>Catplot

****

>Subplot

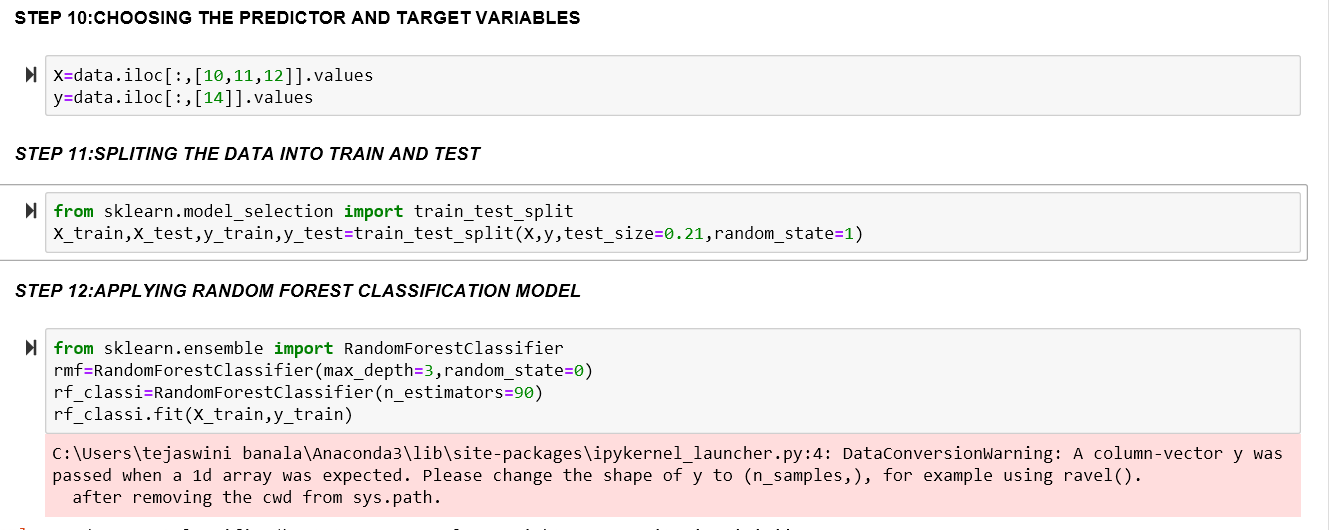


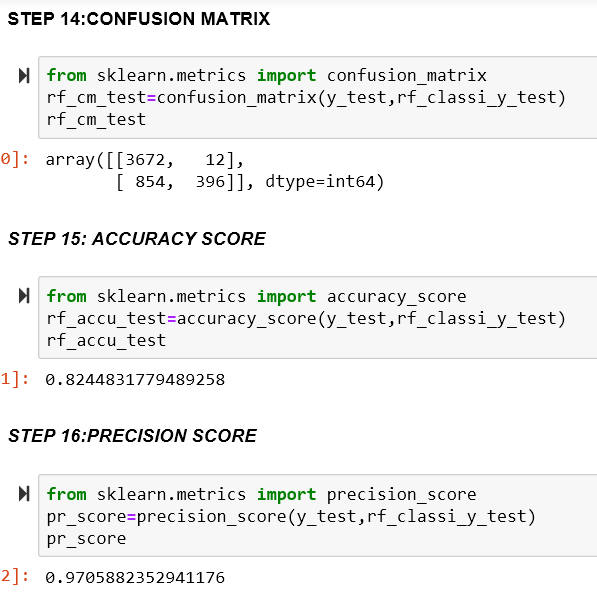
>Histograms



**Results and Discussions**

* With the Machine Learning Random Forest Model We are able to Predict the Income of An Adult with the Most Dependent Reasons that cause the Range of Income to Lie in a Particular Range
* Capital gain, Capital loss, Hours Per Week are good for predicting income which may lie >50K or <=50K
* Our Model is 83% Accurate in Calculating the Income with the Predictor Variables
* Random Forest is good enough in predicting the Income of an Adult by Providing the Most Dependent/Predictor Variables.



**CONCLUSION**

With the help of our work, we predicted the income of the population and analyzed the factors which strongly affect the income. We implemented model and predicted the accuracy using Random Forest. Also, we considered the output and determined the overall result.

**References**

**>** <https://www.kaggle.com/uciml/adult-census-income>

><https://arxiv.org/ftp/arxiv/papers/1810/1810.10076.pdf>

><http://cseweb.ucsd.edu/classes/sp15/cse190-c/reports/sp15/048.pdf>