Abstract

Prediction of Salaries and their Classification in Different aspects

Data Science

Project Report

Contents

[Introduction: 1](#_Toc106359227)

[Requirements: 1](#_Toc106359228)

[Dataset Variables explanations: 1](#_Toc106359229)

[Data After loading: 2](#_Toc106359230)

[Data Cleaning: 2](#_Toc106359231)

[Exploratory Data Analysis: 3](#_Toc106359232)

[Does more education mean more money? 5](#_Toc106359233)

[Determining who earned more money Male/Female. 6](#_Toc106359234)

[Determining which workplace pays more money to their employees. 7](#_Toc106359235)

[Determining which occupations have dominancy. 7](#_Toc106359236)

[Data Preprocessing 9](#_Toc106359237)

[Native Countries and their entires. 10](#_Toc106359238)

[Splitting the Dataset for Training and Testing: 10](#_Toc106359239)

[Modeling: 11](#_Toc106359240)

[The result that we got after tuning the top classifiers. 11](#_Toc106359241)

[Comparison 11](#_Toc106359242)

[Second: 12](#_Toc106359243)

[Third: 12](#_Toc106359244)

[Graphical Representation: 13](#_Toc106359245)

[Logistic Regression: 13](#_Toc106359246)

[Random Forest Classifier: 14](#_Toc106359247)

[Conclusion: 14](#_Toc106359248)

Prediction of Salaries and their Classification

## Introduction:

The purpose of using this dataset is the classification of salaries in different aspects like education, marital status, occupation, ethnicity, and more.

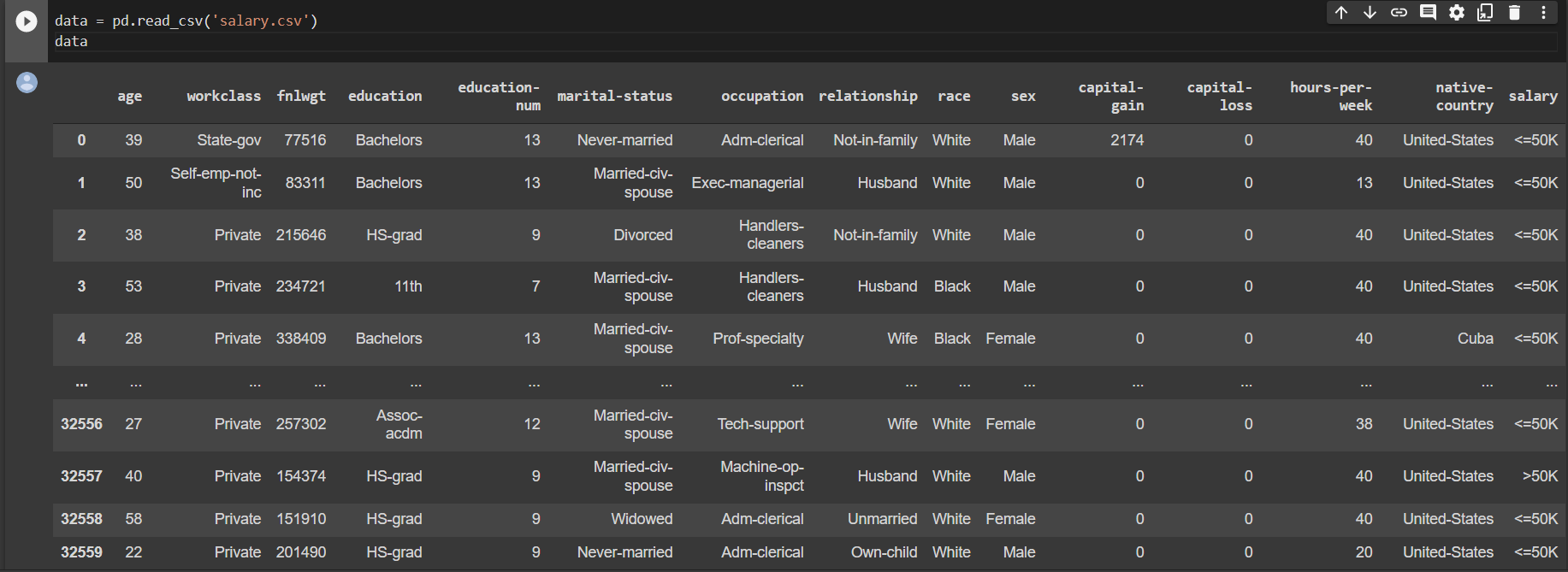
# Requirements:

The major goal of this dataset is to understand the data as much as possible, by comparing models of different farms and asking some questions that will allow us to understand the data better and predict if the salary is less than or greater than any particular amount.

# Dataset Variables explanations:

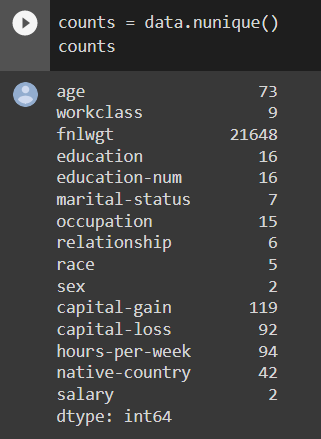
1. Age: continuous numbers
2. Work class: represents the employment status of individuals.
3. Private
4. Self-emp-not-inc/ Self-emp-inc
5. Federal-gov/Local-gov/State-gov
6. Without Pay
7. Never worked.
8. Fnlwgt: it is a continuous variable that represents the number of people that the census believes that entry represents.
9. Education:
10. Preschool, 1st to 12th
11. HS-grad
12. Prof-school
13. Assoc-acdm
14. Assoc-voc
15. Some-college
16. Bachelors
17. Masters
18. Doctorate
19. Education-num: it describes the education status in numbers from preschool to doctorate.
20. Marital status: marital status of individuals.
21. Married-civ-spouse: Status whether they are divorced, separated, widowed, or Never married.
22. Occupation: Defining their occupation and what they are doing including their designations.
23. Relationship: describe the relationship between each other.
24. Race: Explaining from which community he/she belongs.
25. Sex: Male, Female
26. Capital Gain: describing the capital income in numbers.
27. Capital loss: Describing how much they spend in numbers.
28. Hours-per-week: Describing how many hours they work in a week in numbers
29. Native-country: Representing the country from where they belongs.
30. Salary: Any particular amount is fixed for analysis.

# Data After loading:



# Data Cleaning:

The values that we get before cleaning the data are these.



It is clear from the image that fnlwgt has several unique values because it represents the number of people that have the same entry. That we have to remove them during the data cleaning process, and also have to remove those that we think that they are outliers for the dataset.

Result After removing the outliers.

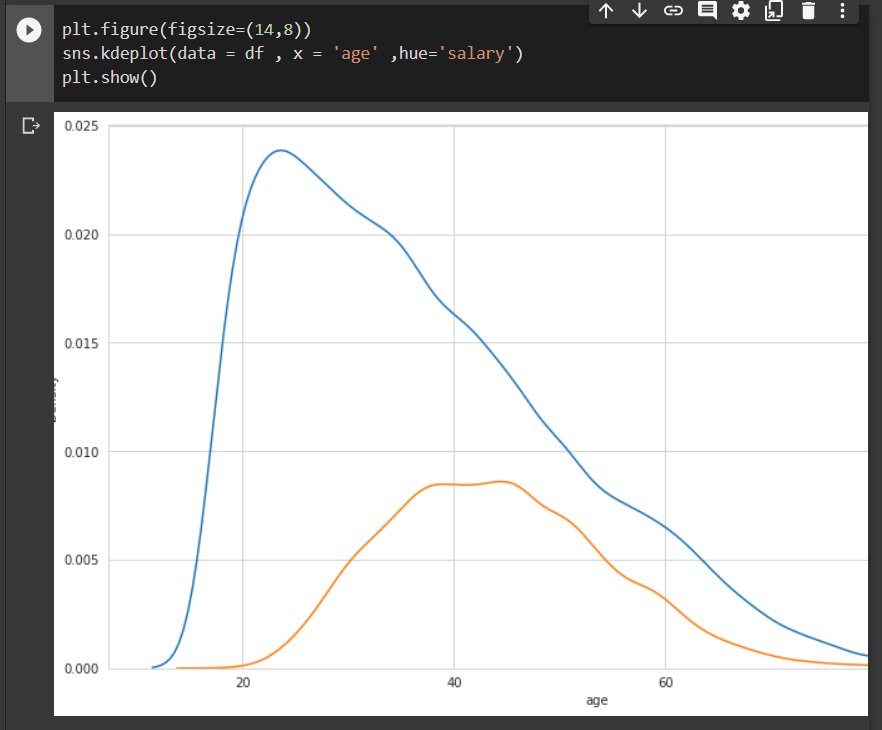
After that, we marked the cells that have null values and then removed them from the dataset for achieving better results in output forms.



# Exploratory Data Analysis:

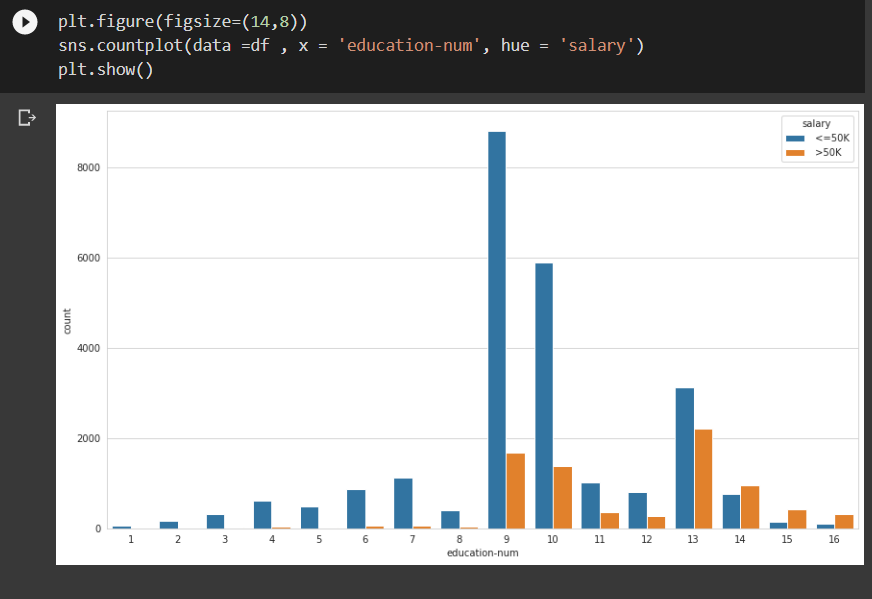
This approach is used to analyze the data by using visualization techniques and it discovers the trends, patterns, and also assumptions with graphical representation.

Now we have to analyze who get more money old or young peoples and visualization.



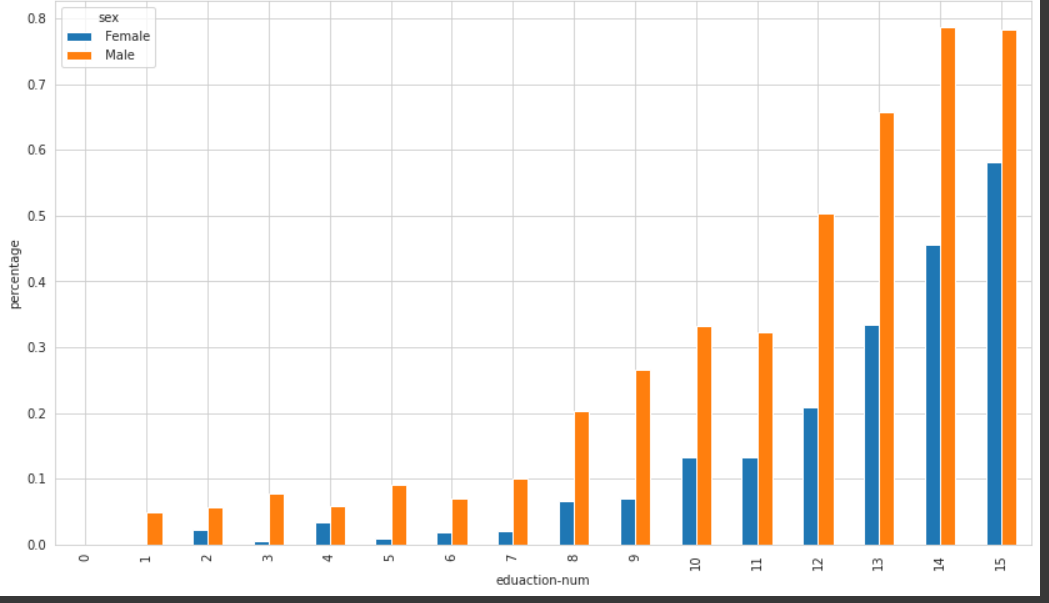
From the curves of this graph, we analyze that people in their early twenties start making more money till the peak in their forties, then start gaining less money.

# Does more education mean more money?



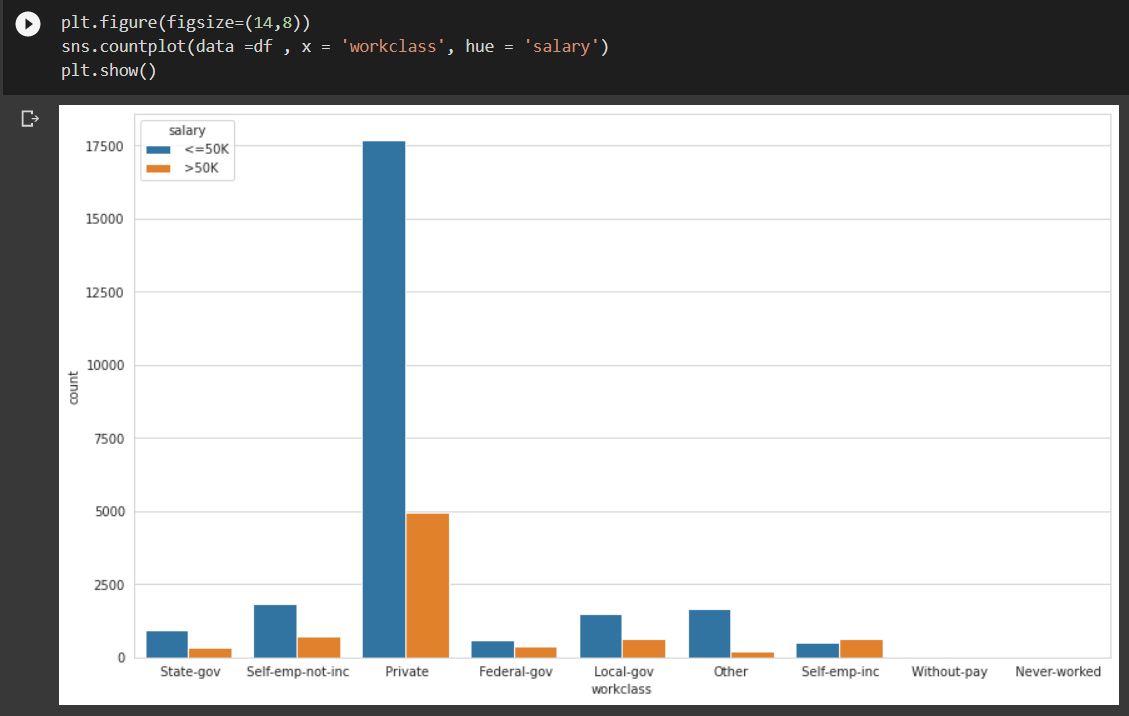
The bins of this chart show that people who are almost at 9 levels of education are earned more money as compared to others and then comes the second number is bachelor levels peoples earning.

# Determining who earned more money Male/Female.



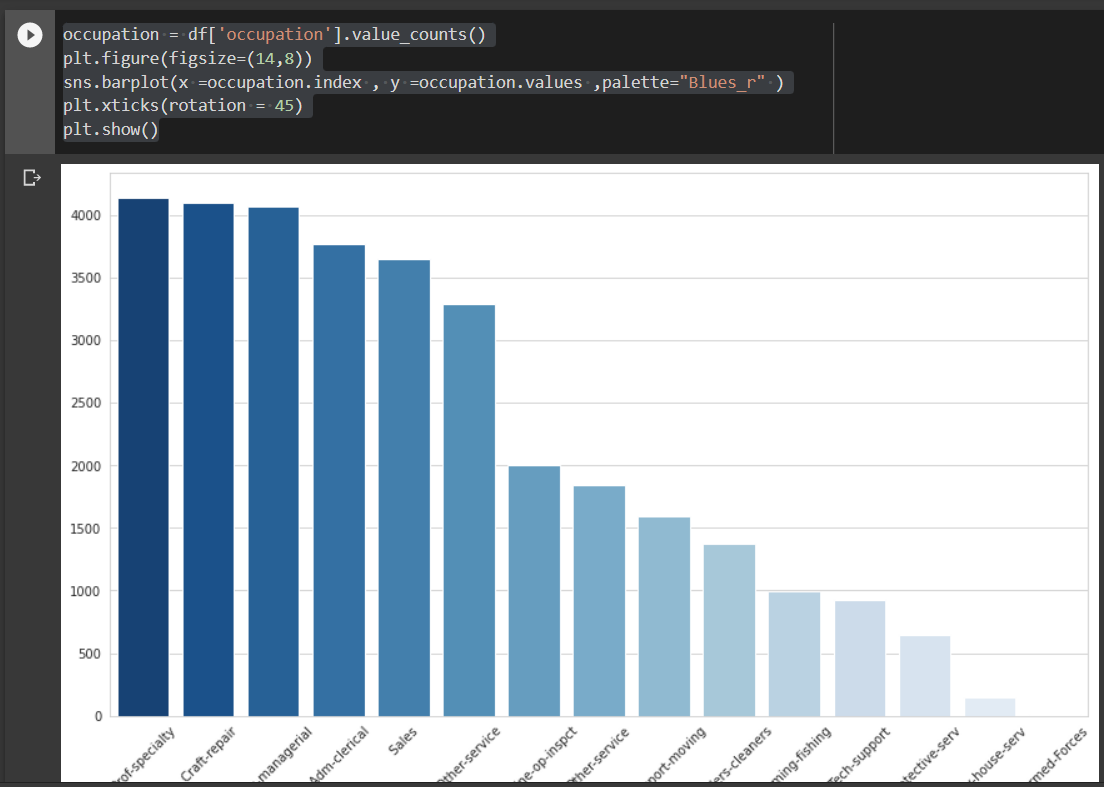
Bins of the chart are showing that some man education level is high and because of that, they also get more money as compared to females.

# Determining which workplace pays more money to their employees.



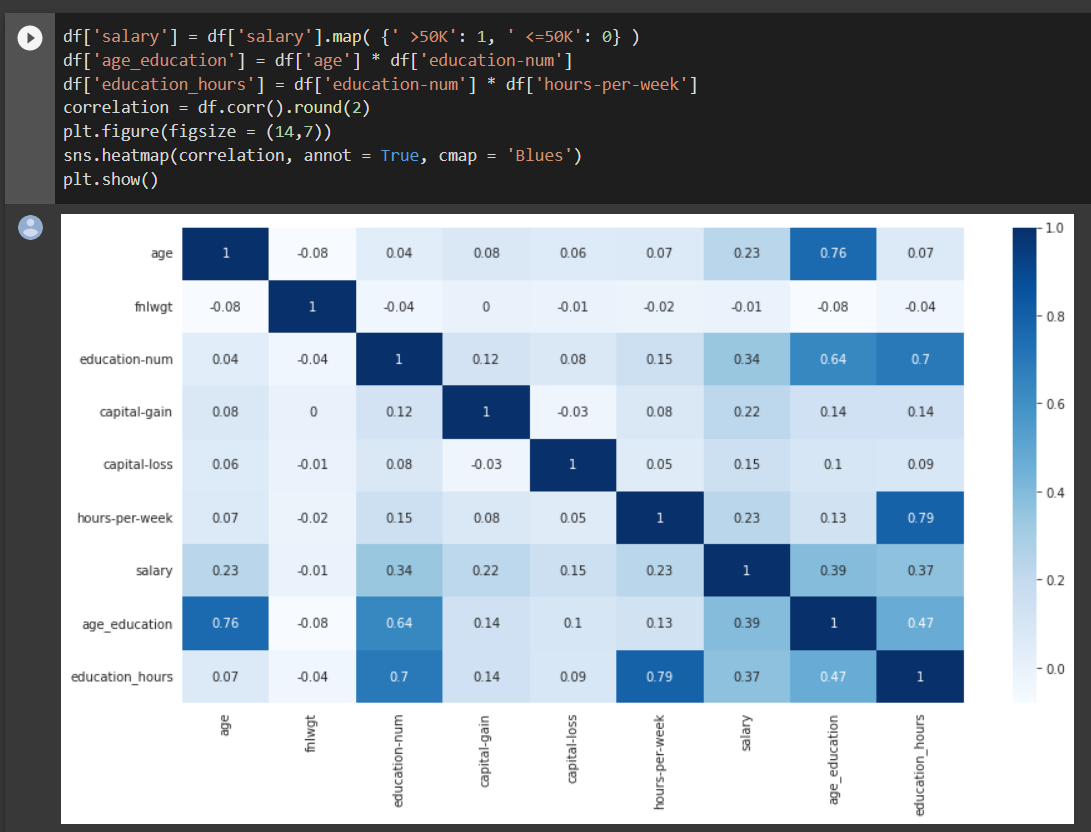
From given chart bins show that the private sector pays money less than 50k as compared to all others.

# Determining which occupations have dominancy.



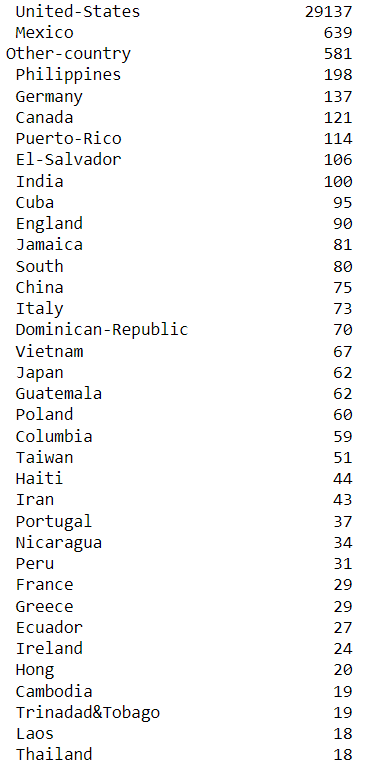
The bins of this graph show that Prof-specialty occupation is domination to all others, and armed forces values are almost equal to zero.

# Data Preprocessing

In this step first, we have to create the features for modeling data. By using the heat map.

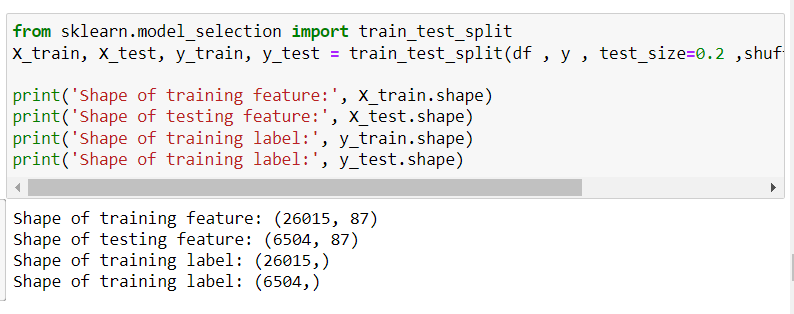
After creating features then handling Numerical data by normalizing it and categorical data by encoding it. Because the variable the education-num has categorical data of education for that it’s necessary to remove categories.

# Native Countries and their entires.



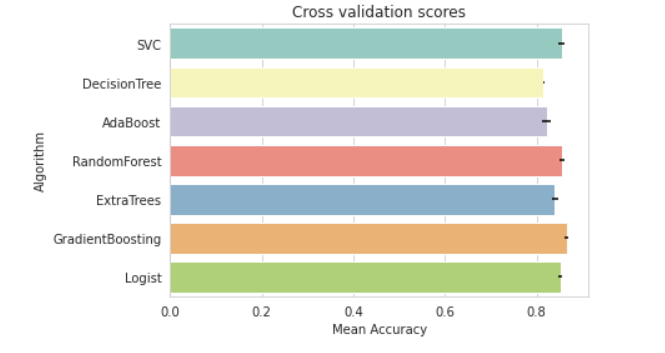
# Splitting the Dataset for Training and Testing:

Now I split the dataset into two parts one for testing and the other one for training and it depends on us for setting the ratio of training and testing. My training and testing values are these.



# Modeling:

In this step I searched values that fit the best for model by classifying the data. From that we get the cross validation scores. By setting Mean Accuracy score at x axis and Algorithms on Y axis.their results are that.



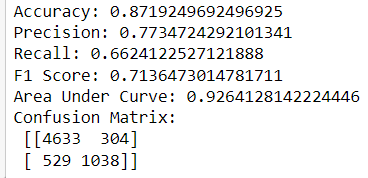
# The result that we got after tuning the top classifiers.

First fitted 5 fold for each of 72 candidates, the total that fits is 360. And their result.

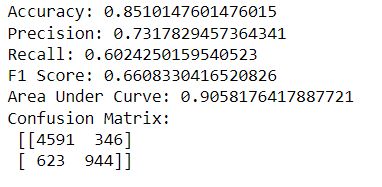
Second Fitted 5 folds for each of 54 candidates, the total that fit is 270, and their result. 0.8628099173553719

# Comparison

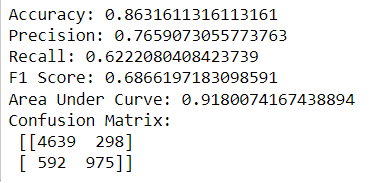
Now in the final, we compared their accuracy, precision, recall, scores, the area under curve, and confusion matrix values by changing their training and testing values and fitted values.



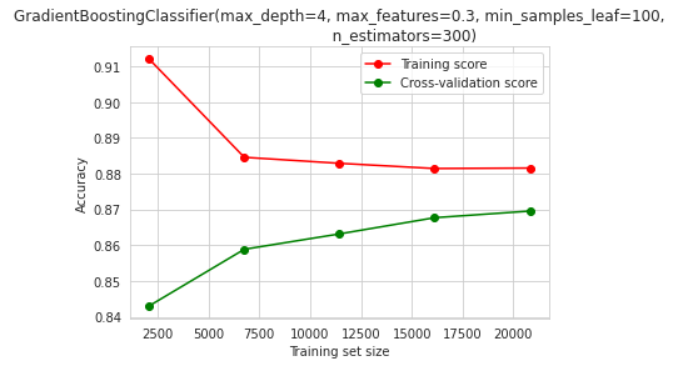
# Second:



# Third:

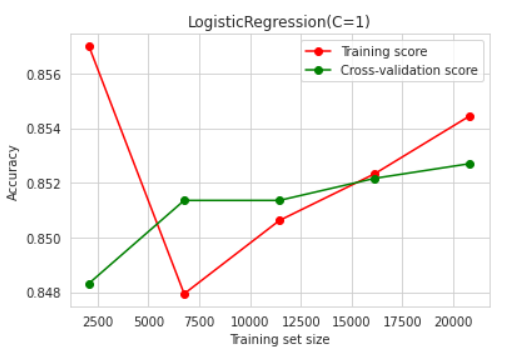


# Graphical Representation:

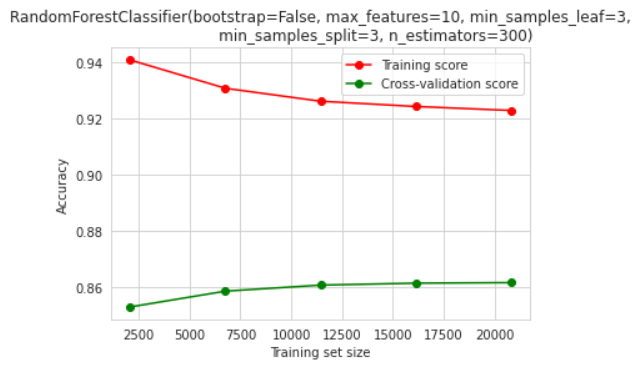


In this model, I set the value 4 for max\_depth 0.3 for features and minimum\_sample\_leaf value is 100 and their number of estimators value is 300.

# Logistic Regression:



# Random Forest Classifier:



# Conclusion:

The results we got from this dataset are that the ratio of the majority belongs to twenties that their income is less than 50k and also the private who paid less than 50k. And in the final, we trained the model in different scores and also get accuracy values.

The End.