**Census Income**

* **Problem Statement**

As per the above statement Income Census is meant to awareness about how the income factor is impacting personal life and making future better. The inequality of wealth and income is a huge concern. Economic inequality leads to huge gap between the upper income class and lower income class. The principle of this census is to improve economic stability. Different Organisations and Governments are working on model to address this problem and to reach optimal solution. Our aim is to show the usage of machine learning to provide the solution for this economic inequality.

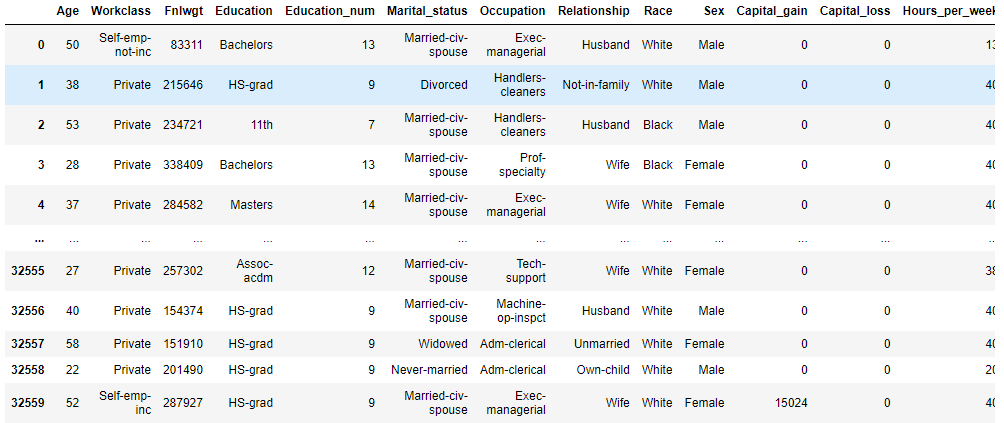
The provided data will surely help us to analyse the role of different features for individual income.

The dataset contains 32560 rows and 15 columns which mainly concentrating on factor of income whether it is above or below 50K. Few provided dataset features are as below.

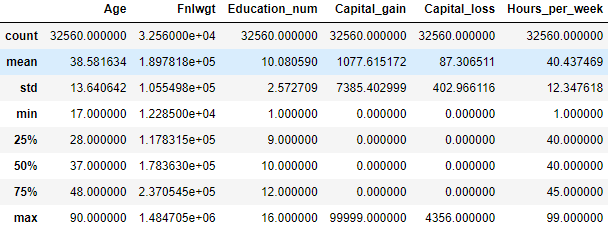
* **Data Analysis**

Data Analysis is the process of inspecting, cleansing, transforming, and modelling data with the goal of discovering useful information by informing conclusions and supporting decision making. In this analysis we were systematically apply statistical and/or logical techniques to describe and illustrate, condense and recap, and evaluate the dataset. Segregating and interpreting the data into a useful business insight for accurate decision-making.

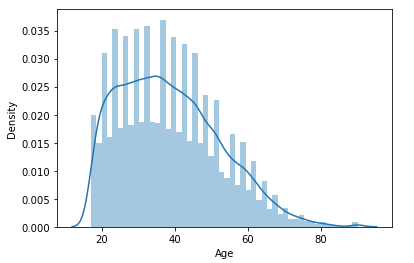
We see the results shown in the image as below when we check the information on our provided data.



As per the provided data we were having following ranges for different sections.

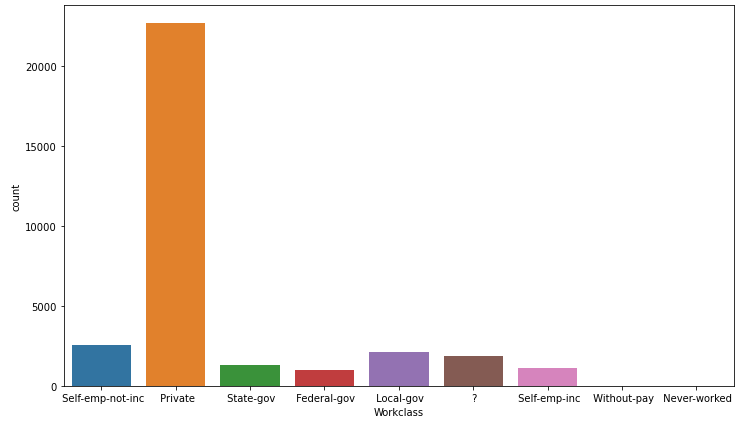


**Age** - Individuals age range is between 17 and 90. Our data is has right skewness, with majority of the ages falling in the 20–50. The count keeps on decreasing as the age increases.

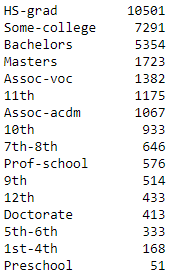


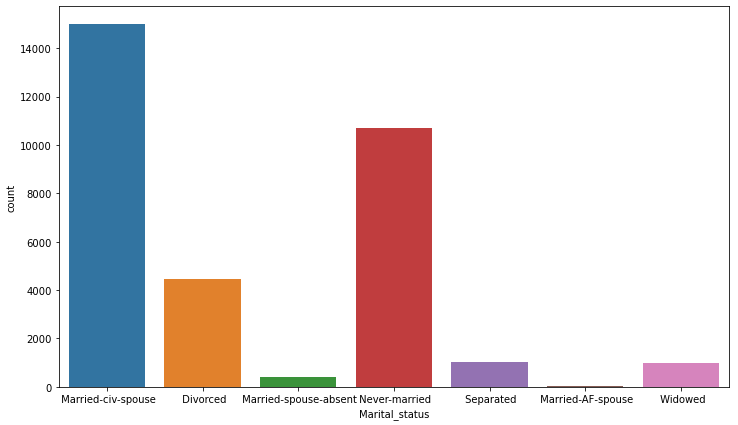
**Fnlwgt -** The age cFnlwgt has a range of 1.228500 to 1.484705

**Working Class** – Below graph shows that majority of the people were belonging to ‘Private’ sector.

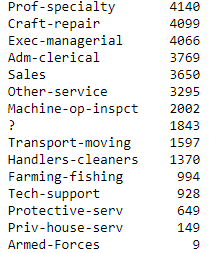


**Education & Education\_num** – Education and Education\_num both were similar. Education qualification as in below data showing us that HS-grad individual is having more number than any other qualification. The education number has a range of 1 to 16 who were having particular degree.



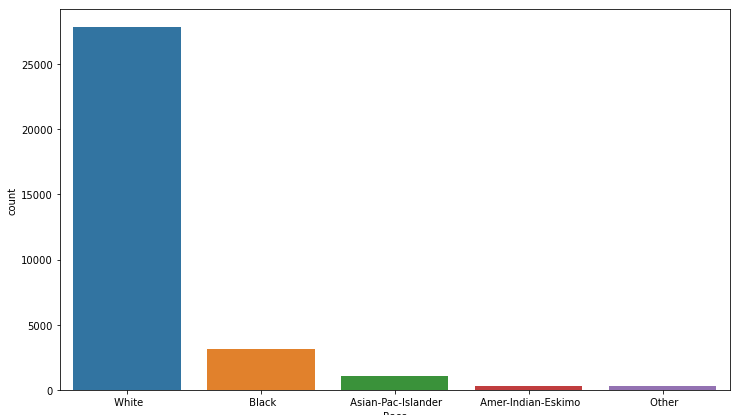
**Marital\_status -** As in below graph Majority of the people have "Married-civ-spouse", and lowest is "Married-AF-spouse". 

**Occupation**– Defines the profession of a person and their Specialty in performed job. As we seeing the below dataset there is a missing value for 1843 individuals which shows as ‘**?**”.

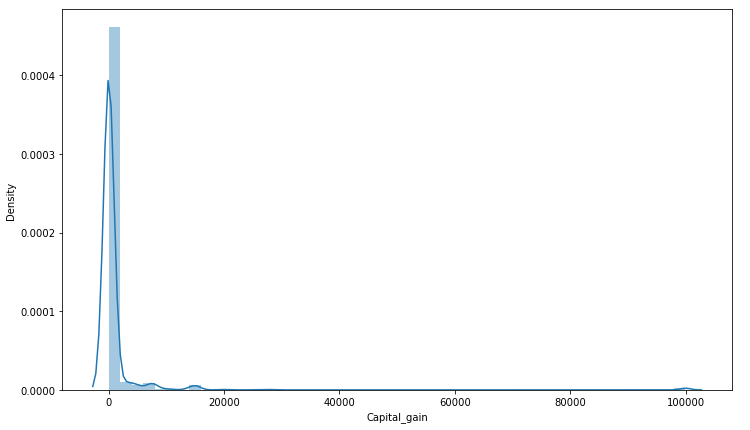


**Sex** – As per the available dataset there are 21789 male and 10771 Female.

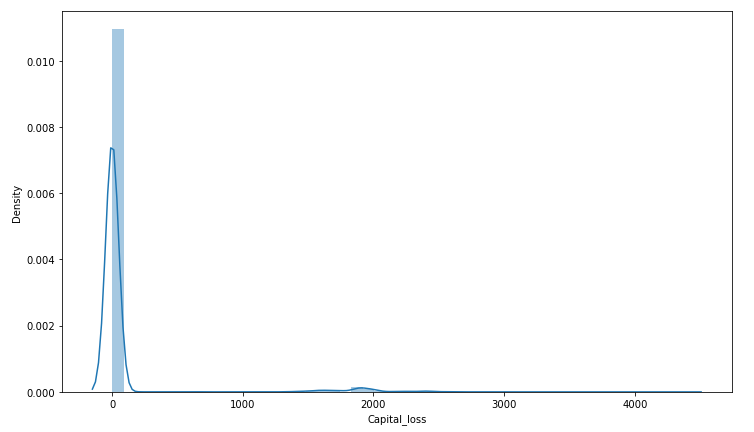
**Race** – Below data is showing the highest number of people have race as White.



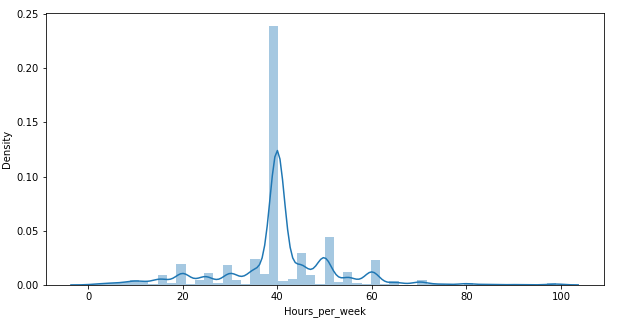
**Capital\_gain –** The capital gain starts from 0 and ends at 99999



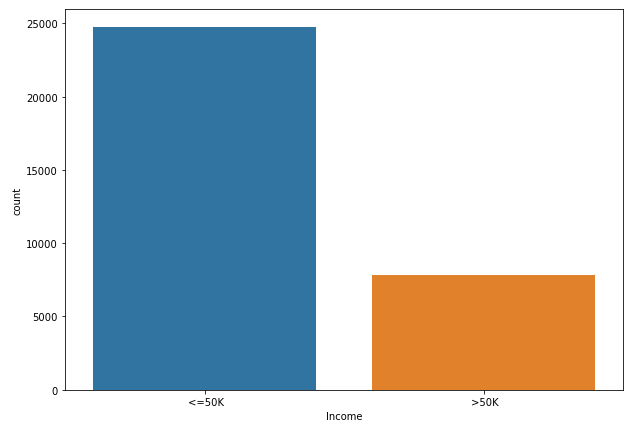
**Capital\_loss** - The capital loss starts at 0 and ends at 4356. The "Capital\_loss" column also has majority of the values set as 0, and the data is highly right skewed.



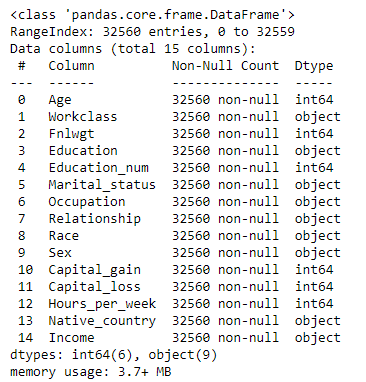
**Hours\_per\_week** - Hours per week range between 1–99.



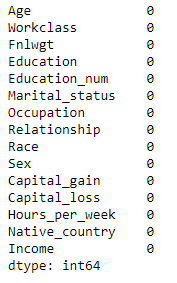
**Income –** As per the data we were having individuals whose income <=50K is 24719 and > is 7841.



**Basic summary of dataset**:



**Check the null values in our dataset**



**Check the columns name**

Index(['Age', 'Workclass', 'Fnlwgt', 'Education', 'Education\_num',

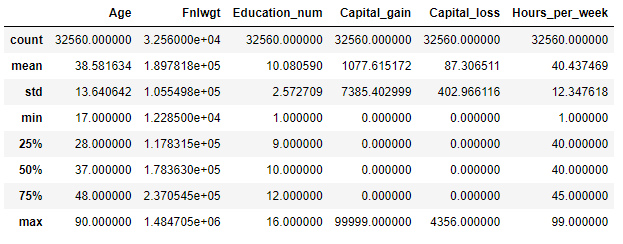
'Marital\_status', 'Occupation', 'Relationship', 'Race', 'Sex',

'Capital\_gain', 'Capital\_loss', 'Hours\_per\_week', 'Native\_county

'Income'],

dtype='object')

**Check statistical summary of numeric columns**



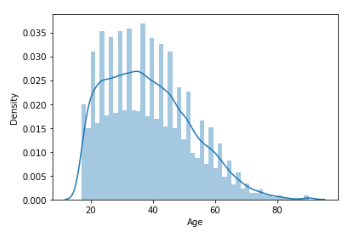
* **EDA Concludings**

In statistics, exploratory data analysis is an approach of analysing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods

Techniques and tools :--> Typical graphical techniques used in EDA are:-->

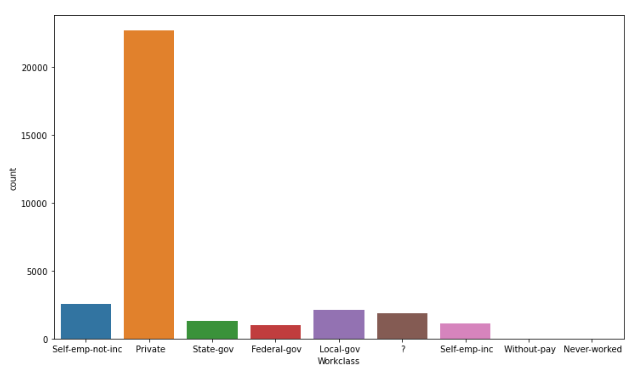
* Box plot
* Histogram
* Scatter plot
* Heat-map
* Bar chart

We can understand the future by using different graphs.



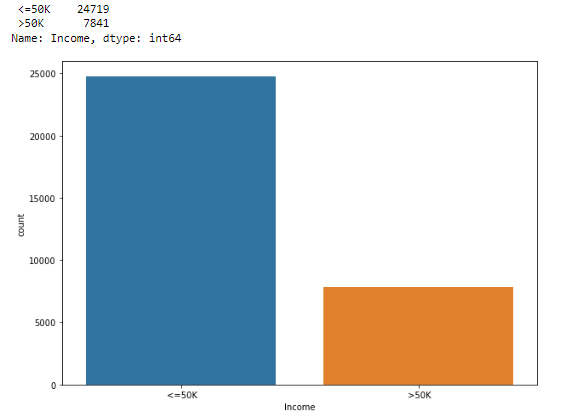
As we were seeing our data is having right skewness, with majority of the ages falling in the 20–50. The count keeps on decreasing as the age increases.

Majority of the people belong to ‘Private’ sector work class

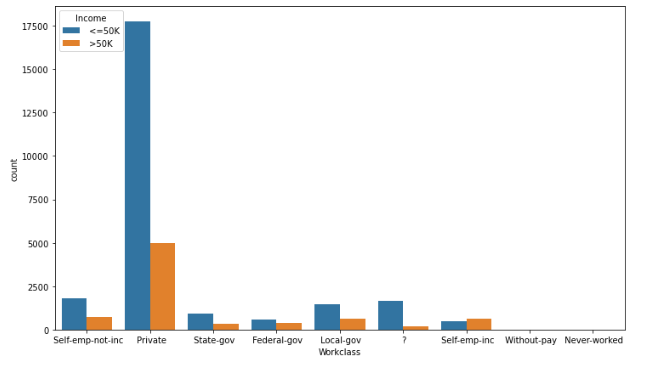


Sector wise data

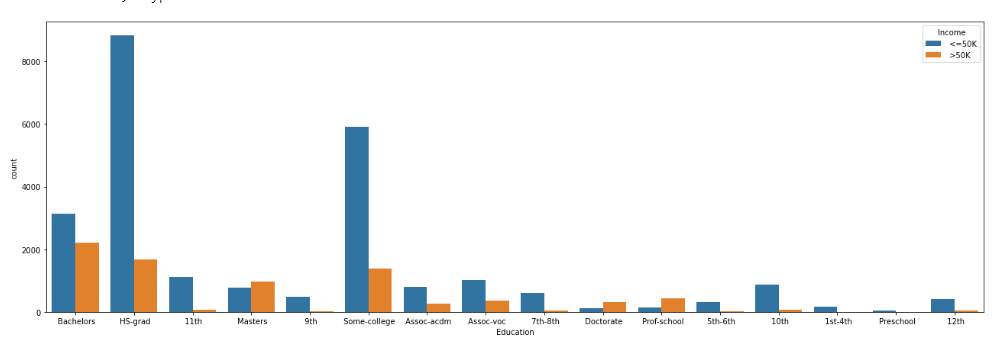
**Income plot**



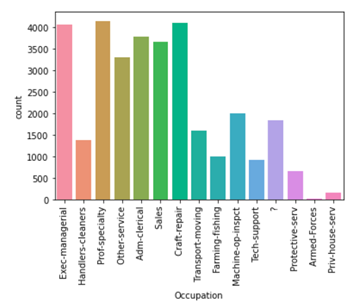
**Salary and work-class** comparison is showing number of people in Private sector is higher than the others.



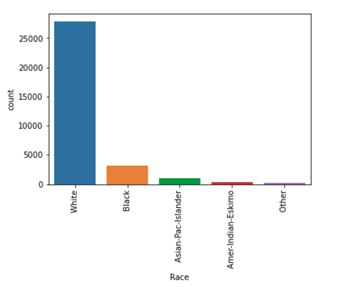
**Education vs income**



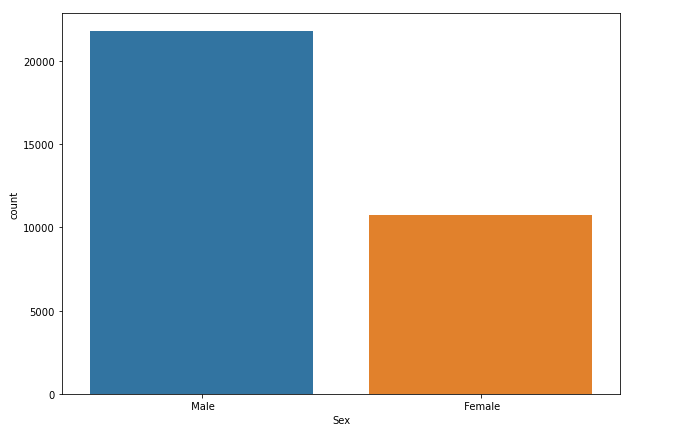
**Occupation**



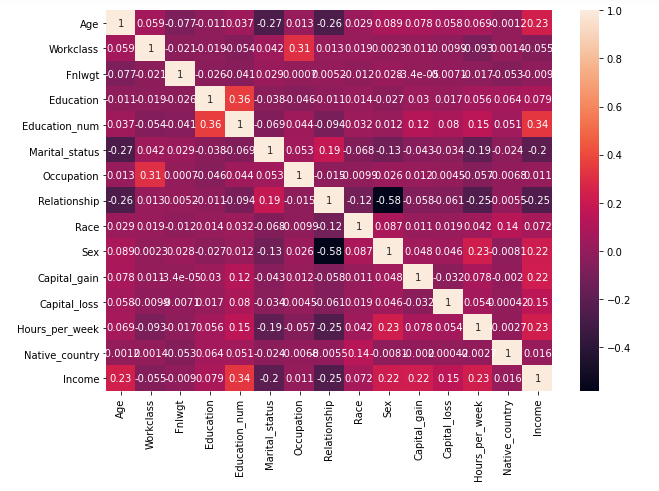
**Race : -** Highest number of people have race as ‘White’ (significantly high numbers).



**Male v/s female –** As in below graph the numbers of males are almost double to number of females.



**Correlation map: -**34% correlations with ‘Education\_num’, 23% correlation with ‘hours\_per\_week’



* **Pre-processing Pipeline**

Data pre-processing is a step in the data mining and data analysis process that takes raw data and transforms it into a format that can be understood and analysed by computers and machine learning.

Data pre-processing can refer to manipulation or dropping of data before it is used in order to ensure or enhance performance and is an important step in the data mining process.

Data cleansing

Data editing

Data reduction

Data wrangling

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm.

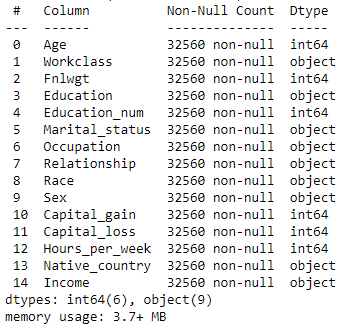
Data Pre-processing is a technique that is used to convert the raw data into a clean data set. In other words,

Whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis.

**Need of Data Pre-processing** - For achieving better results from the applied model in Machine Learning projects the format of the data has to be in a proper manner. Some specified Machine Learning model needs information in a specified format; for example, Random Forest algorithm does not support null values. Therefore to execute random forest algorithm null values have to be managed from the original raw data set.

**Data quality assessment: -** Take a good look at your data and get an idea of its overall quality, relevance to your project, and consistency. There are a number of data anomalies and inherent problems to look out for in almost any data set.

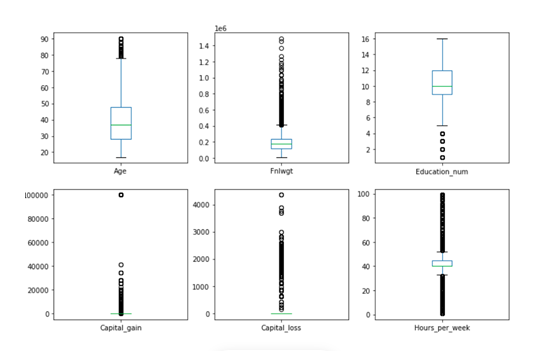
**Different/Mismatched data types: -** When we get or collect data from many different sources, it may come to us in different formats. While the ultimate goal of this entire process is to reformat our data for machines, we still need to begin with similarly formatted data.



Our data in 2d-type in integer and object.

**Data outliers:** Outliers can have a huge impact on data analysis results

**Find outliers and remove:-**



**Use Inter-Quartile Range (IQR) proximity rule:**

Q1 = data.quantile(0.25)

Q3 = data.quantile(0.75)

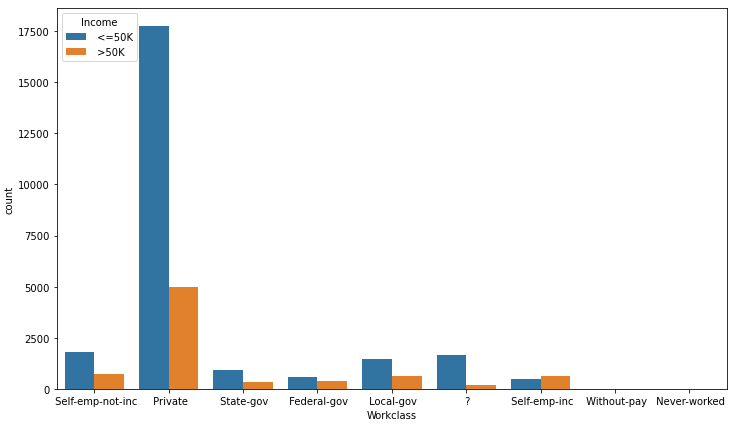
IQR = Q3 - Q1

print(IQR)

**z-score: -** The z-score is a score that measures how many standard deviations a data point is away from the mean. The z-score allows us to determine how usual or unusual a data point is in a distribution. The z-score allows us more easily compare data-points for a record across features, especially when the different features have significantly different ranges. Z-score calculations formula is as below.

**Z=x-mu/sigma**

**Missing data:** Take a look for missing data fields, blank spaces in text, or unanswered survey questions. This could be due to human error or incomplete data. To take care missing data, we have to perform data cleaning as our data set work-class, Occupation, native country, has “?” which is a missing values.



### 

### 

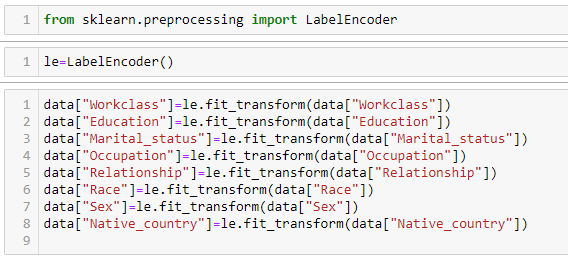
### Data cleaning: Data cleaning is the process of adding missing data and correcting, repairing, or removing incorrect or irrelevant data from a data set. Dating cleaning is the most important step of pre-processing because it will ensure that your data is ready to go for your downstream needs.

**Missing data***:*Manually fill in the missing data can be tedious, but is definitely necessary when working with smaller data sets.



**Drop:** This is only recommended for large data sets, when a few ignored tuples won’t harm further analysis

**Encoding Data** : - Since the majority of the classification models need input as int/float, and do not work on ‘string’ data, we have to encode our categorical columns using ‘Label Encoder’.



**Future scaling:**

**Normalization: -** Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as **Min-Max scaling.**

Here’s the formula for normalization:

[Normalization equation](https://cdn.analyticsvidhya.com/wp-content/uploads/2020/03/Norm_eq.gif)

Here, Xmax and Xmin are the maximum and the minimum values of the feature respectively.

When the value of X is the minimum value in the column, the numerator will be 0, and hence X’ is 0

On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator and thus the value of X’ is 1

If the value of X is between the minimum and the maximum value, then the value of X’ is between 0 and 1.

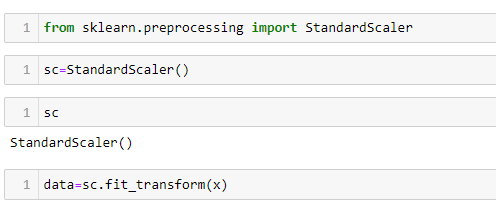
## Standardization: - Standardization is another scaling technique where the values were centred around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

**Here’s the formula for standardization**:

[Standardization equation](https://cdn.analyticsvidhya.com/wp-content/uploads/2020/03/Stand_eq.gif)

Feature scaling: Muis the mean of the feature values andFeature scaling: Sigmais the standard deviation of the feature values. Note that in this case, the values are not restricted to a particular range.

**I’m using standerdscaler in my dataset**



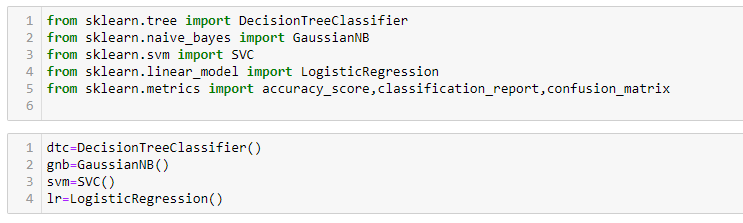
* **Building Machine Learning Models**

We now proceed to the main step of our machine learning, fitting the model and predicting the outputs. We fit the data into multiple classification models to compare the performance of all models and select the best model –

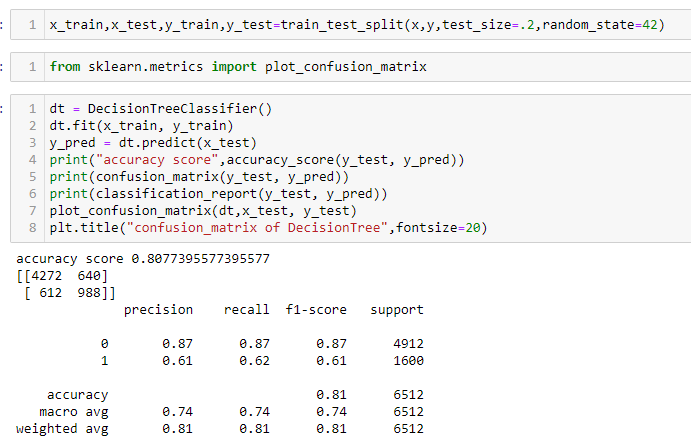
**Split the data in future variable and target variable**

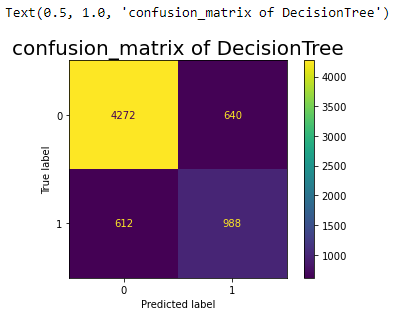
x=data.drop("Income",axis=1)

y=data['Income']



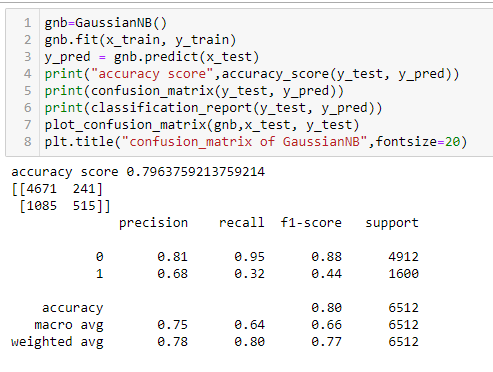
**DecisionTreeClassification**

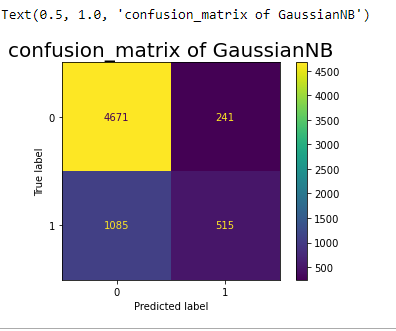




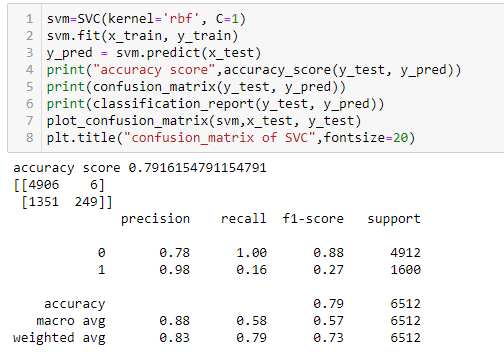
DecisionTree, which provides an accuracy of 80%

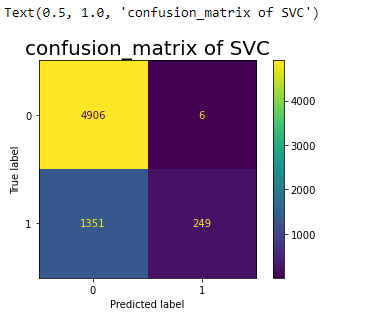
**GaussianNB**



  
GaussianNB, which provides an accuracy of 79.6%

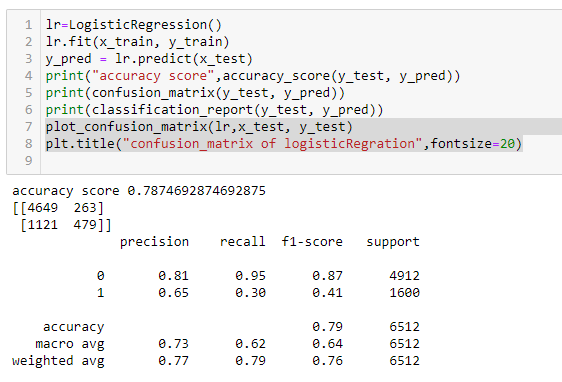
**Support vector classifier:**

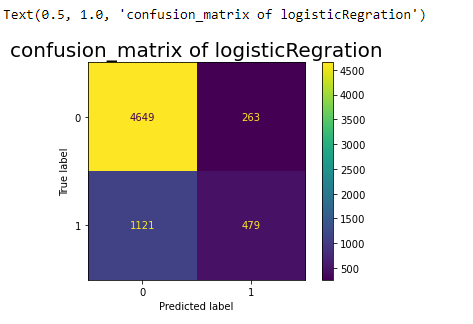




SVC, which provides an accuracy of 79.1%

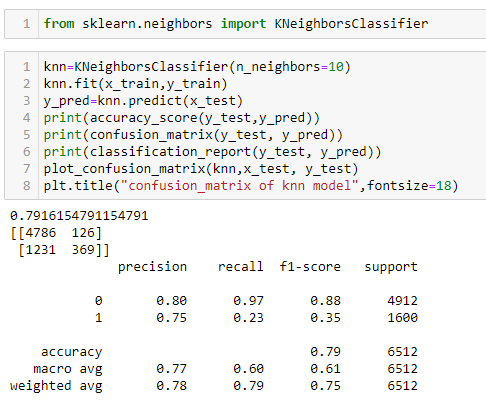
**LogisticRegration**

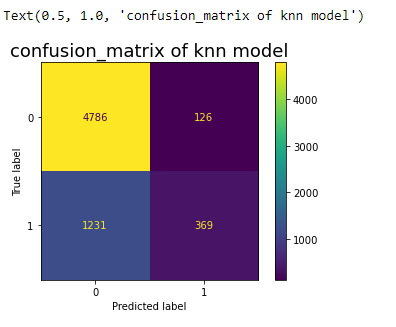




LOGISTICREGRESSION, which provides an accuracy of 78.74%

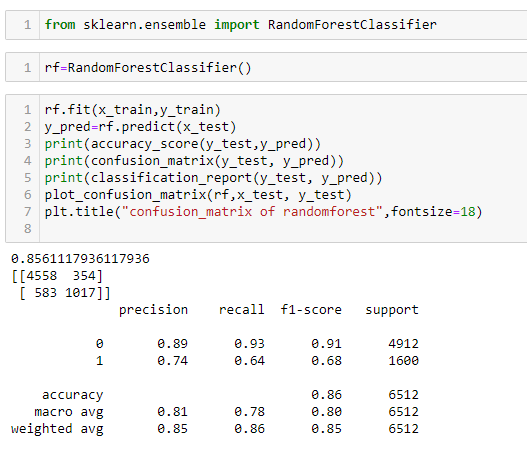
**KNeighbors Classifier**

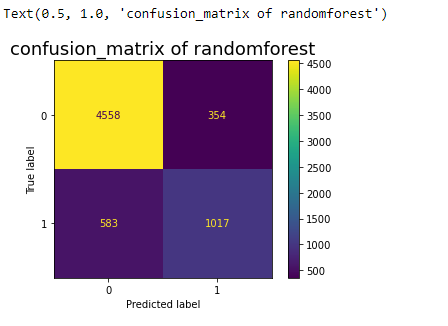




KNN, which provides an accuracy of 79%

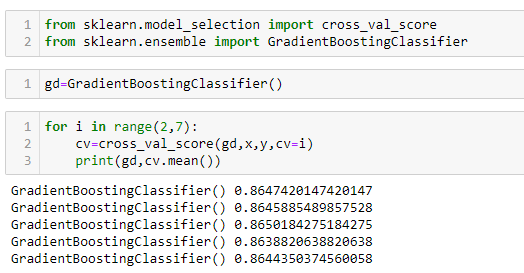
**RendomForestClassifier**





RANDOMFOREST, which provides HIGHEST accuracy that is 86%

**GradientBoostingClassifier**



All above model The best accuracy give RANDOMFOREST model that is 86%

**ROC curves** typically feature true positive rate on the Y axis, and false positive rate on the X axis. This means that the top left corner of the plot is the “ideal” point - a false positive rate of zero, and a true positive rate of one.

### 

* **Conclusion**

#### Finally, the Validation Accuracy, We have successfully trained our model to predict the income of a person, with an accuracy of 80% approx. which is, by the best of our knowledge. The final work involves achieving better set of results by using different models with inclusion of Machine Learning and by applying many other advanced pre-processing techniques without further depletion. The performances of the models are compared on the basis of classification report and ROC curve. We have done the step by step analysing, cleaning and modelling of the data to achieve the desired predictions. We also tuned the model to improve the accuracy, and were able to achieve a model with quite a good accuracy.

Thank you

Presented by

Ruchita Parmar