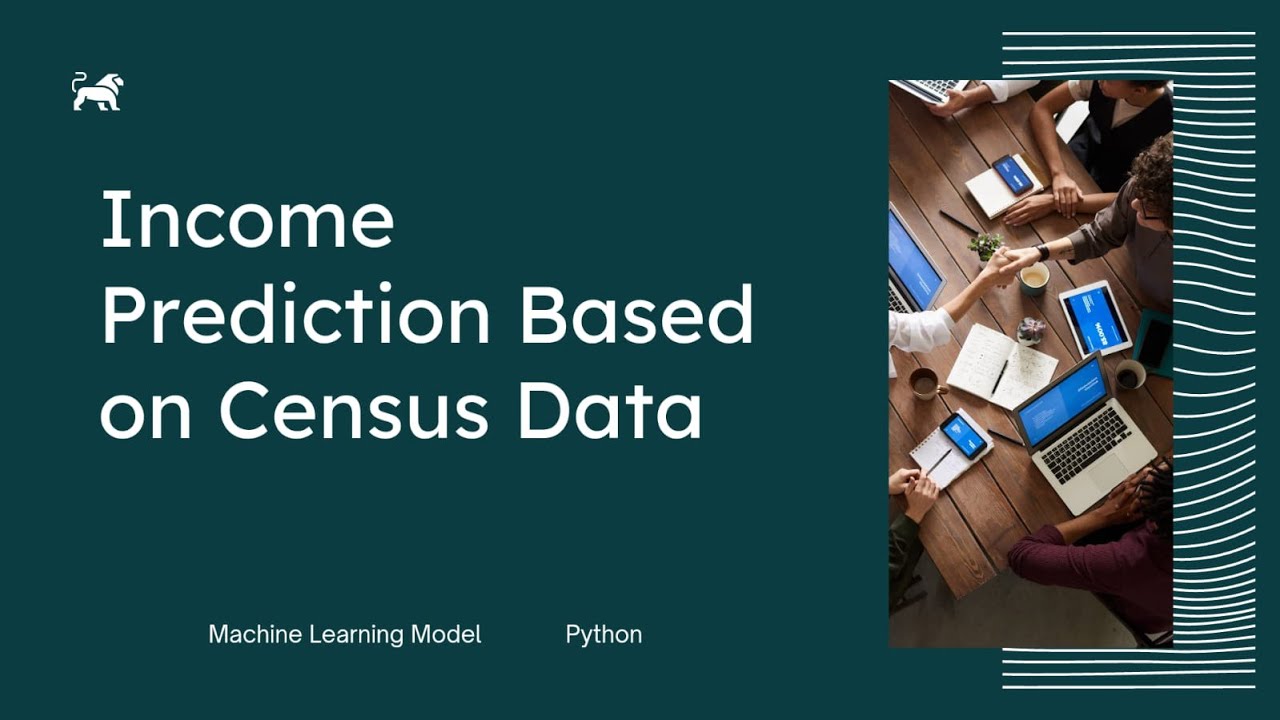
**Census Income Project**



The economic well-being of a Nation is highly driven by the income of the residents. Countless decisions in private and public sectors are based on Census data. Census data is the backbone of the democratic system of government, highly affecting the economic sectors. Census-related figures are used to distribute the federal funding by the government into different states and localities.

The Data Which I am using here can also used for post census population estimates and projections, economic and social science research, and many other such applications. Hence, the importance of this data and its correct predictions is very clear to us.

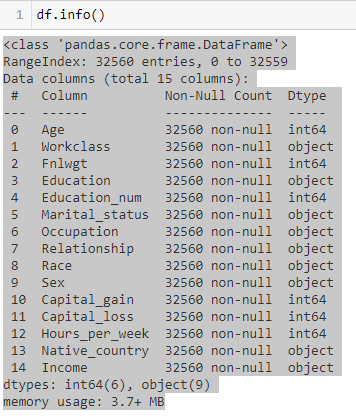
Data has always been the backbone of many important decisions. When an assumption is backed up by facts and numbers, the chances of incorrectness and bad decisions decrease.

**Problem Definition:**

The above introduction had an aim to increase the awareness about how the income factor actually has an impact not only on the personal lives of people, but also an impact on the nation and its betterment. We will today have a look on the data extracted from the 1994 Census bureau database, and try to find insights about how different features have an impact on the income of an individual. Even Though the data is quite old, and the insights drawn cannot be directly used for derivation in the modern world, but it would surely help us to analyze what role different features play in predicting the income of an individual.

**Data Analysis:**

The Dataset Contains Total of 32560 rows and 15 columns, In which some are numerical Data and some are Categorical Data

****

As per the Above Info Numerical Data are 6 columns and 9 are Cateogorical Data.

To know about the dataset we will analyze Each Column(**Both Independent and Dependent)**

**Independent Variables:**

1. **Age** — The age of an individual, this ranges from 17 to 90.

2. **Workclass** — The class of work to which an individual belongs.

3. **Fnlwgt** — The weight assigned to the combination of features (an estimate of how many people belong to this set of combination)

4. **Education** — Highest level of education

5. **Education num** — Number of years for which education was taken

6. **Marital Status** — Represents the category assigned on the basis of marriage status of a person

7. **Occupation** — Profession of a person

8. **Relationship** — Relation of the person in his family

9. **Race** — Origin background of a person

10. **Sex** — Gender of a person-(Male or Female)

11. **Capital gain** — Capital gained by a person

12. **Capital loss** — Loss of capital for a person

13. **Hours per week** — Number of hours for which an individual works per week

14. **Native Country** — Country to which a person belongs

**Dependent Variable:**

1. **Income** — The target variable, which predicts if the income is higher or lower than 50K$.

**Exploratory Data Analysis:**

In EDA we need to Pre-process the Data and Visualization:

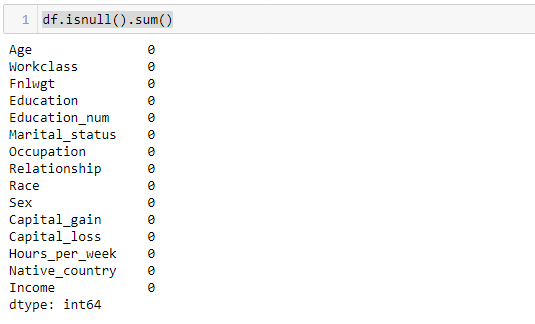
Steps include in Pre-Processing Data are

1)**Data Cleaning**: - Removing Outliers, Skewness and imputing Missing Values.

2) **Data Transformation**: - like Normalization by applying normalization we can improve the accuracy and efficiency of the models. And also reduce the errors.

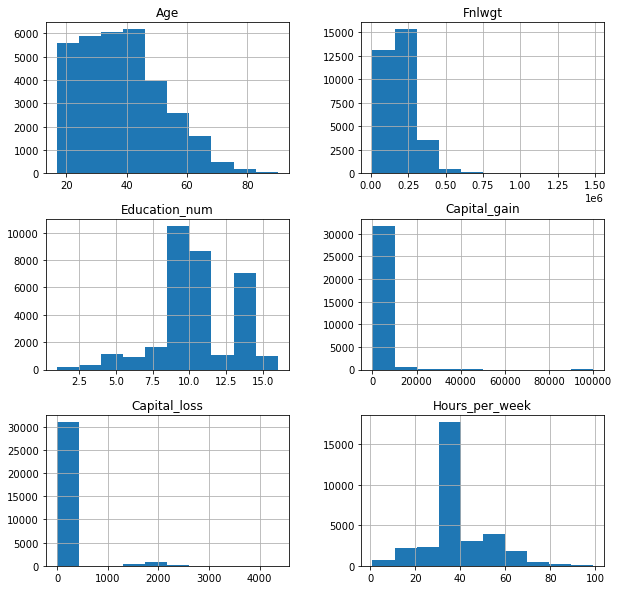
3)**Data Reduction**: By Reducing the no of features by Feature Selection Process, PCA And VIF

**1.Data Cleaning:** As a Part of EDA we need to do Data cleaning so firstly we need to check any null values in our data, From the below image shows we don’t have any null values, so no need to impute any data



For Doing Visualization Easy I Just Split the numerical and Categorical Data separately





As per Hist Graph our numeric Data is not Normalised we need to scale the Data

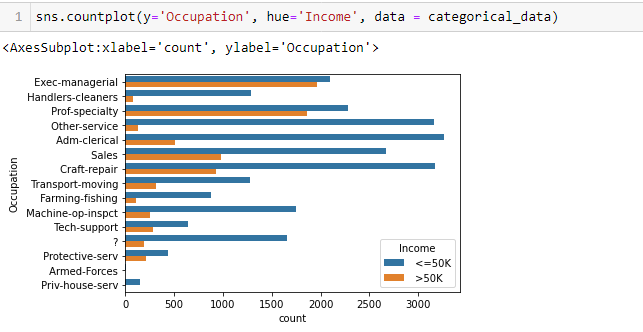
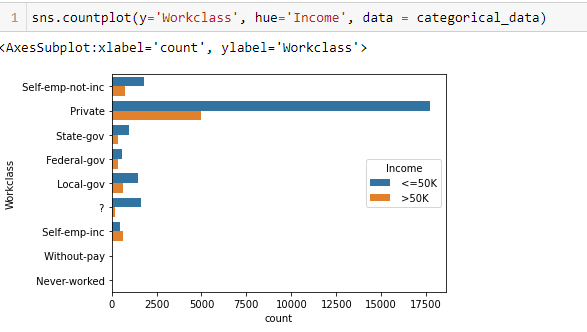
Observations made from the data

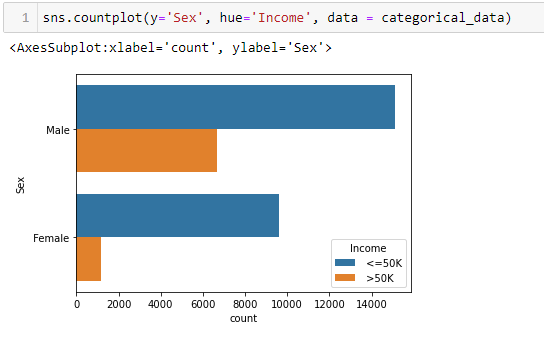
1: People age of 40-45 are more

2: fnlwght is more in between 0.20 to 0.30

3: Almost the education is more around 8-9 years

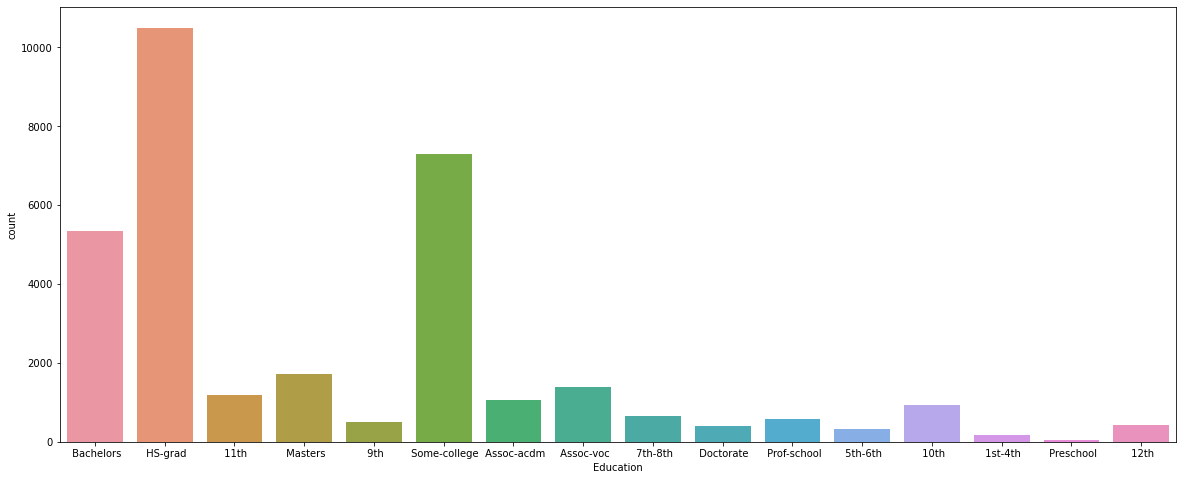
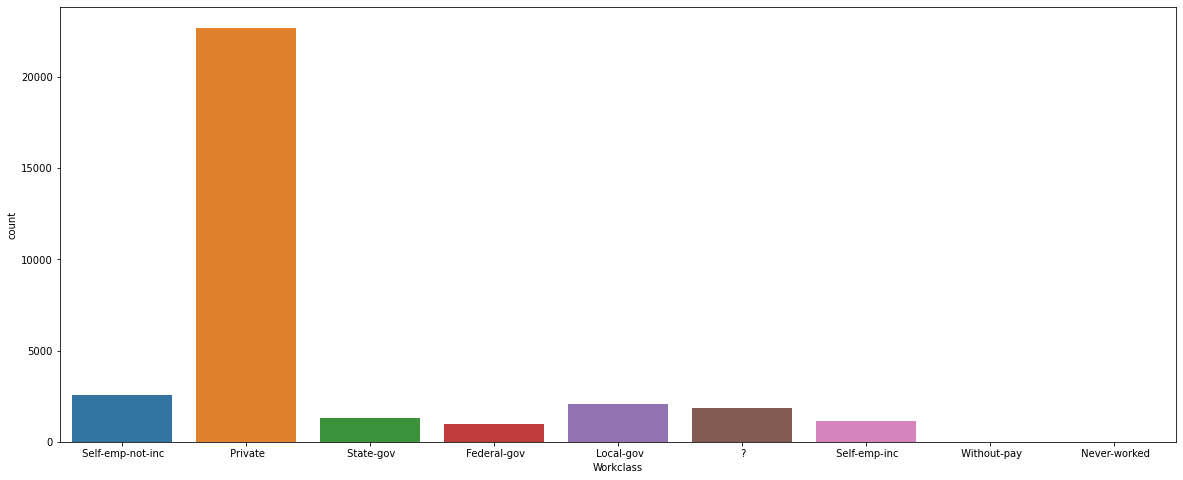
4: hours per week are more in 30hours

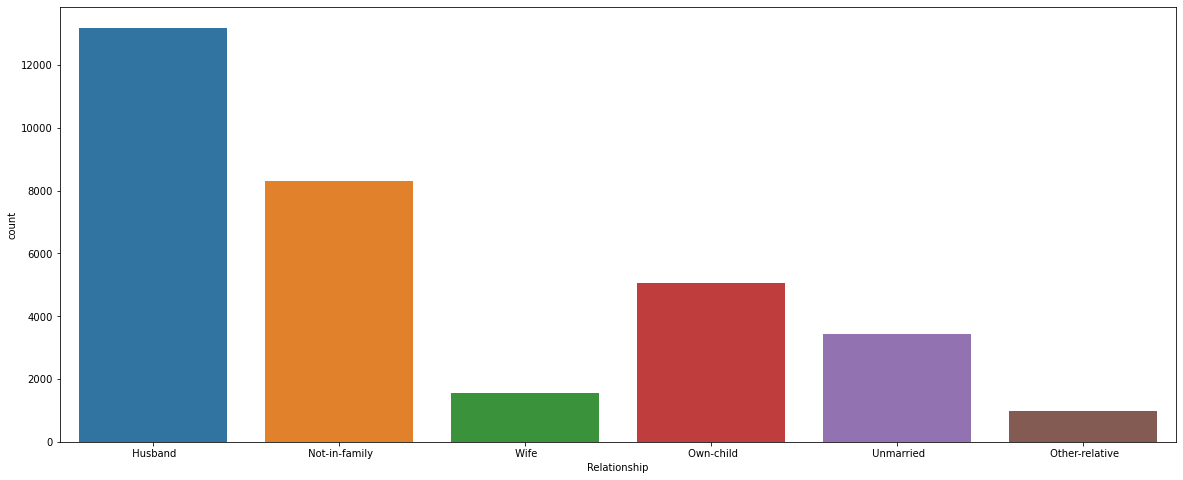




Observations:

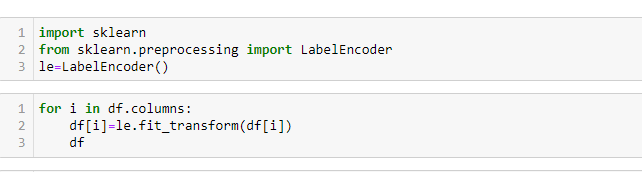
1. Male people are 14000 who are above 50k salary Female people are 9900 who are above 50k salary.
2. As per the graph private employees who earn more than 50k and less than 50K.





Label Encoding:

Label Encoding is necessary for the data to process to find any outliers are there as of our data consists of both numerical and categorical need to change categorical into numerical values using the encoding methods



After encoding Data need to check the any correlation of a Data





**Observations:**

1. Income has 34% correlation with ‘Education num’, 23% correlation with ‘hours per week’ and ‘age’, and 22% correlation with ‘Capital gain’. The correlations are moderate.

2. Education and final weight not that much affective to dependent variable so we can delete the variables Education and Education num are multi collinear so that is one of the reason to remove one variable

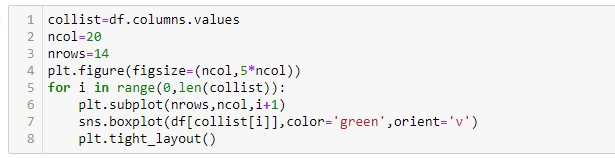
Next Step is dropping the Education and Fnlwgt variables so for that we are using the drop function

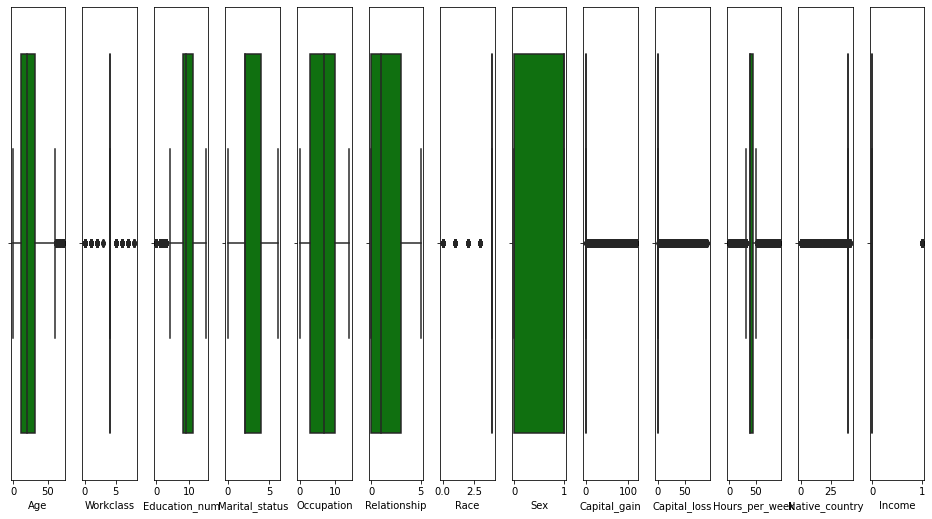


Checking Any Outliers in our Data and Remove it:

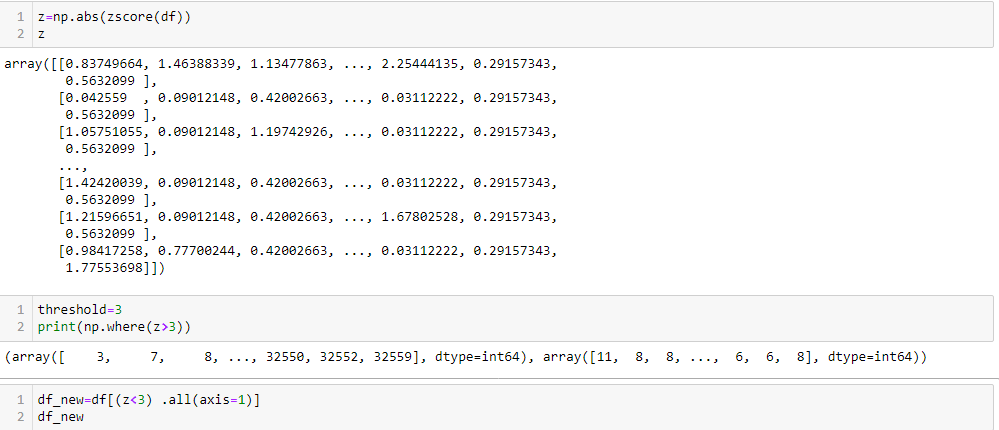
It is defined as the points that are far away from the same points.it can be happen because of the variability of the measurements and may be some error also. If possible, outliers should be removed from the datasets. There are servals methos to remove the outliers. 1)Z score 2) Quantile Method (Capping the data)

1)Z Score: it can call from the SciPy. Stats library. And for most of the case threshold values should be used 3.



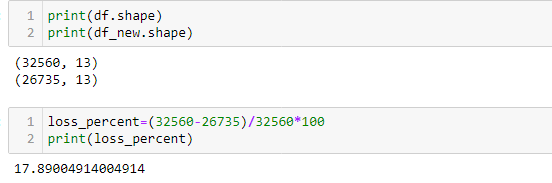


As above shows our data having the outliers so need to remove outlier by one of the prominent method called z-score method



From using above formula we are removing all the outliers

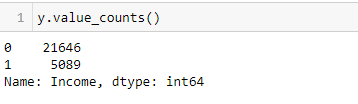
For checking how much data is clean we need to subtract the old data shape from new data shape and then divide with old data shape



Almost 17% of our data is removed

2)Quantile Methods: Inter Quantile Range is used to detect or cap the outliers. Calculate the IQR by scipy.stats.iqr Multiply Interquartile range by 1.5 Add 1.5 x interquartile range to the third quartile. Any number greater than this is a suspected outlier. Subtract 1.5 x interquartile range from the first quartile. Any number lesser than this is a suspected outlier.

Now our Data is ready for modelling as our data is clean so only thing need to do is normalizing the data before need to check our dependent variable



As our dependent classification is not equal so for that need to do sampling before training and testing the data

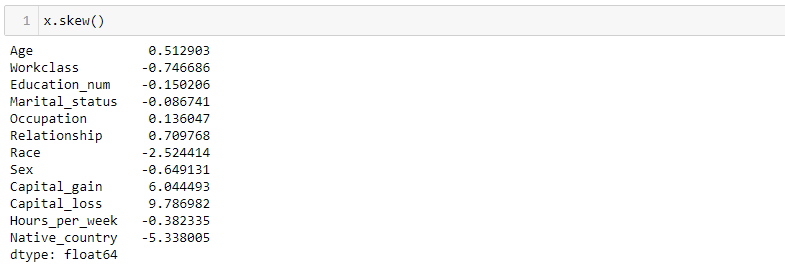
**Skewness of Data:**

As of our numeric data is skewed we need to do normalization before go for training and testing for that need to check the skewness of data if our data is Greater than 0.5% in both positive and negative sides ,then need to do power transformation and do scaling

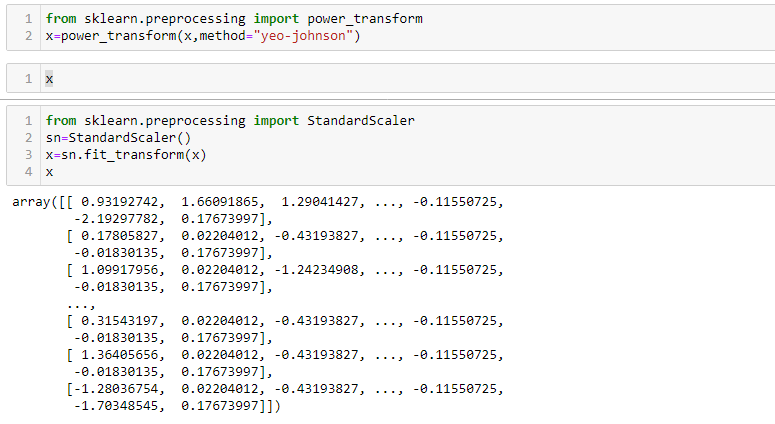
Scaling are of two types:

1.**Standard Scaler:** Standard scalar standardizes features of the data set by scaling to unit variance and removing the mean (optionally) using column summary statistics on the samples in the training set.

**MIN-MAX Scaler:** MinMaxScaler. For each value in a feature, MinMaxScaler subtracts the minimum value in the feature and then divides by the range. The range is the difference between the original maximum and original minimum. MinMaxScaler preserves the shape of the original distribution.



Workclass, Relationship, Race, Sex, Capital\_gain, Capital\_loss ,Native\_country these are positively and negitively skewed

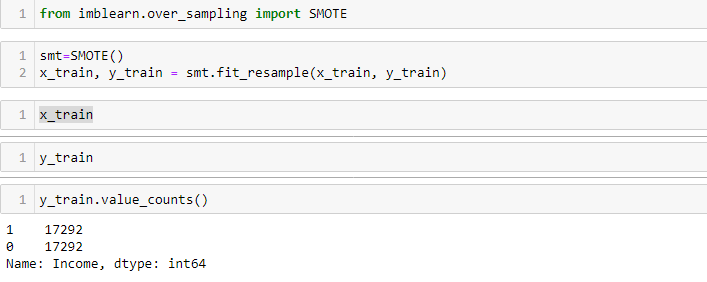


**Splitting Data Into train\_test\_split: -**

This function is in sklearn. Model selection splitting the data array into two arrays. Train and Test with this function we don’t need to splitting train and test manually.by default it make random partition and we can also set the random state.it gives four o/p like x\_train, x\_test, y\_train, y\_test. After Doing splitting we have to balanced our data.it can be by SMOTE or oversampling methods. Like Up Sampling, down sampling.

**Up sampling:** -This method used to modify the unequal data into the balanced data by increases the minor class or rare class. Advantage of this method is to no loss of information but from that model can be in overfitting.

**Down Sampling:** like the Up sampling its also balanced data but by reducing the size of the class which is high

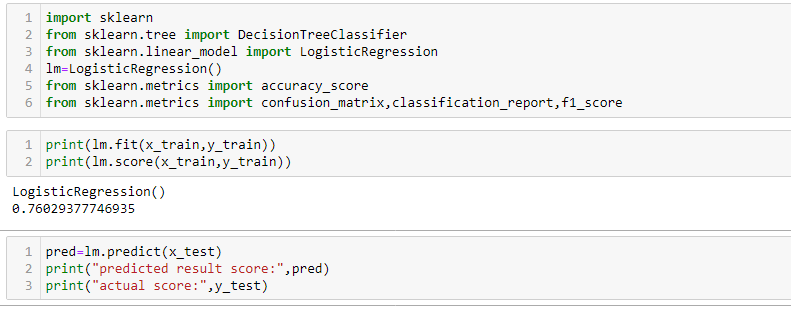


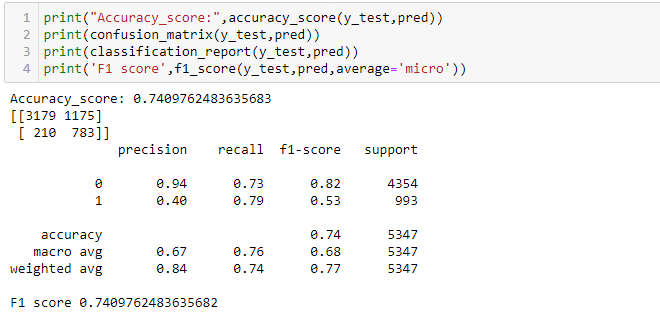
**But if we balanced our data before train test split means we balanced from our whole data set or form x. it means at that time our test data is leak. We have to isolate our test data. Here you expose it.so our f1 or recall or precision will be good. so, our model will already know which is positive or negative. And I can also say because of that there is bias or model Overfitting.to prevent this We balanced our data**

Now our data is ready to apply to the model.

Try Different Models….

1)**Logistic Regression:** -logistic regression is the supervised machine learning problem which is used for the classification problem and used to predict the probability of the classification.it is widely used for the binary classification problem. It is one od the simplest methos of ML

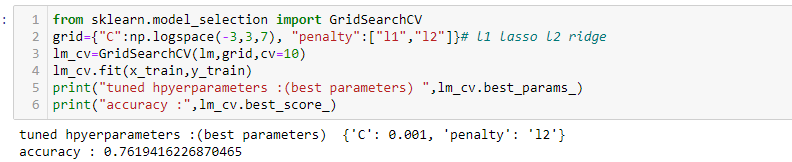




Here my Logistic Regression is giving an accuracy score of 74% and F1-Score :74%

**Hyper Tuning for Logistic Regression**:

**Definition of Hyper Parameter:** Hyper parameter optimization in machine learning is used to find parameters of given machine learning algorithm that perform best as measured on validation. I used GridSearchCV for hyper tunning.

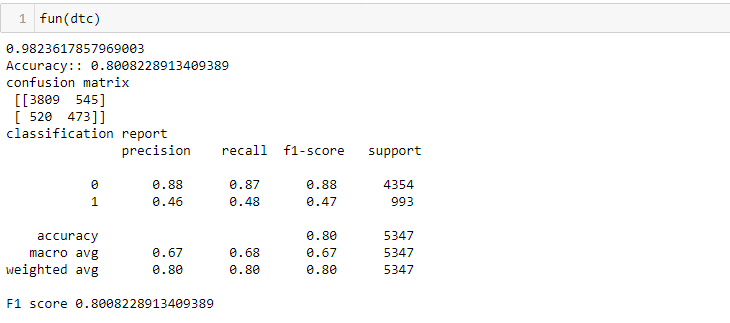


After tuning it is increasing an accuracy from 74% to 76%.

2) **Decision Tree Classifier:** DTC can be used by both classification and regression both. But mostly it’s used for the classification problem. Its structure is tree based. Where internal nodes represents the features of dataset and branches represents the decision rules and each leaf nodes represents the outcomes.



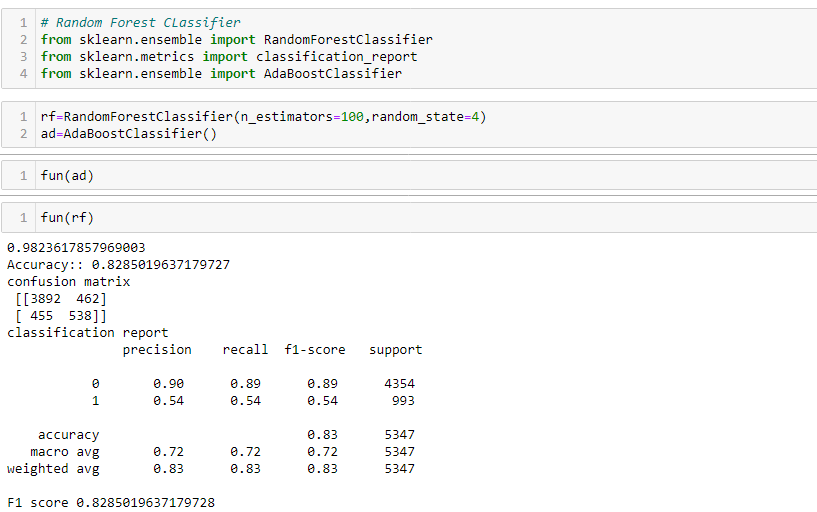
By using the function method we can call all the models one time no need to do coding separately



Decision Tree Giving the model training score:98%, Accuracy :80%, F1 Score :80%.



After Hyper tuning the Decision tree Classifier the Accuracy is same

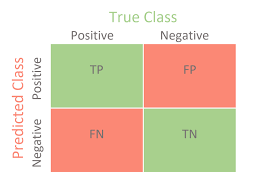


I tried svc, knn and Ada booster also same function used and now for random forest classifier my model is giving the model score:98.2% , Accuracy:82.8% and F1 Score is 82.85%

**Now lets see about each one in our output clearly:**

**Confusion Matrix:** It is the table that is used to describe the performance of classification model on set of tests data.by using different parameters.

We get the best score in Random Forest Classifier as Accuracy score is 82.85%,And model predicts 3892 as True Positive,462 False Positive , 455 False Negative and 538 True Negative

****

**Now lets understand the Recall Precision and f1-score**

**Accuracy:** it can be defined as the ratio of total number of correct classifications divided by total number of classifications.

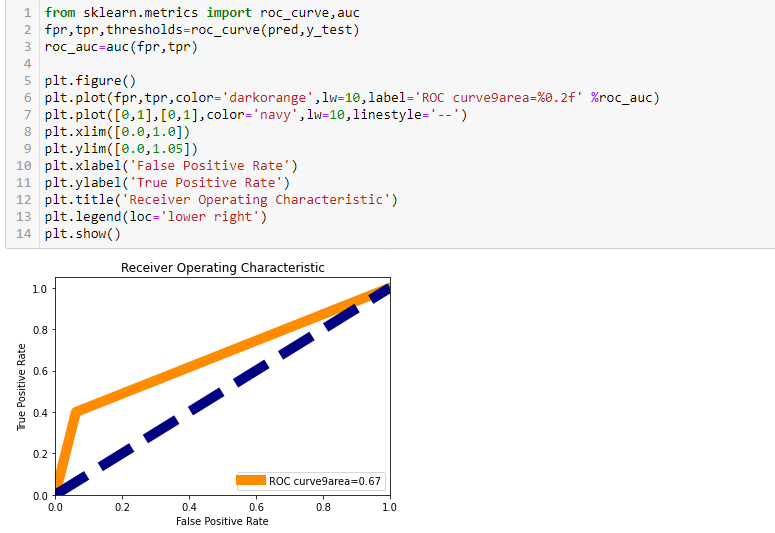
Accuracy=(TP+TN)/(TP+FP+TN+FN)

**Precision**: It is measure of all the positive predictions how many of them actually positive. Precision=TP/(TP+FP)

**F1-Score:** It give the combine result of Recall and Precision

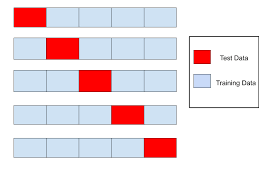
F1-score=2\*(Precision\*Recall)/ (Precision + Recall)

**ROC-AUC Curve:**- It is the performance measurement of the model at diff diff thresholds. ROC is the performance score and AUC is the separation score means how much mode classify 0 as 0 and 1 as 1.

****

**Cross Validation:** - This technique is used to check weather out data set is over fitting or under fitting. If model score is high and cv score is less it means model perform well in train dataset but did not perform well in unseen or test dataset. Feature selection is the best way to overcome the overfitting problem. There are 3 ways for the validation. KFold Cross validation score, Hold Out Methods and LOOCV.

**KFold:** - In this technique it will rotate the data into the k-fold times.

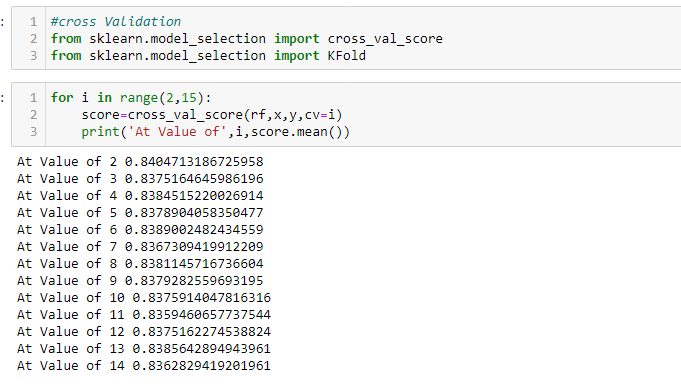


1 st Iteration: 1-3 as Test and 4-9 Train

2 nd Iteration:4-6 as Test and 1-3 & 7-9

Train 3 rd Iteration:7-9 as Test and 1-6 as Train

It means all the data (9 rows) go for training.

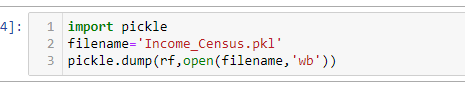
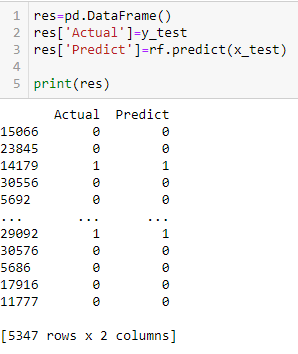
LOOCV: Leave one out cross validation It will take one row for test and remaining for training so each and every row go for test so its time-consuming processing

**Concluding Remarks:** - From this model we can predict the income of every person and predicts how it affects with each variable.

We used different classifiers like Logistic Regression, Decision tree Classifiers and Ada boosting classifiers,Random Forest Classifier. And also used the data Balanced process and also hyper parameter tunning for improving score.

We get good score in Random Classifier we got accuracy of 82.7% on training data,F1 score is 82.8% and cross validation value also is high in random forest classifier.the model performance is excellent.

We further proceed to test the object that we saved using pickle, and create a data frame of predicted values –

This marks the end of our process; we have successfully trained our model to predict the income of a person, with an accuracy of ~82.7%. We moved step by step, analyzing, cleaning and modeling the data, and applied various machine learning models 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.s