***Census Income Project :***

1. **Problem definition:**

This data was extracted from the [1994 Census bureau database](http://www.census.gov/en.html) by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics). A set of reasonably clean records was extracted using the following conditions: ((AAGE>16) && (AGI>100) && (AFNLWGT>1) && (HRSWK>0)). **The prediction task is to determine whether a person makes over $50K a year.**

Description of fnlwgt (final weight)

The weights on the Current Population Survey (CPS) files are controlled to independent estimates of the civilian non-institutional population of the US. These are prepared monthly for us by Population Division here at the Census Bureau. We use 3 sets of controls. These are:

1. A single cell estimate of the population 16+ for each state.

2. Controls for Hispanic Origin by age and sex.

3. Controls by Race, age and sex.

We use all three sets of controls in our weighting program and "rake" through them 6 times so that by the end we come back to all the controls we used. The term estimate refers to population totals derived from CPS by creating "weighted tallies" of any specified socio-economic characteristics of the population. People with similar demographic characteristics should have similar weights. There is one important caveat to remember about this statement. That is that since the CPS sample is actually a collection of 51 state samples, each with its own probability of selection, the statement only applies within state.

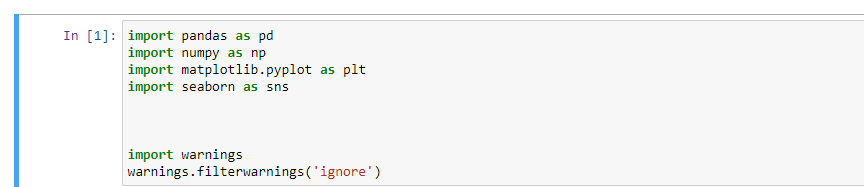
Downlaod Files:

* <https://raw.githubusercontent.com/dsrscientist/dataset1/master/census_income.csv>

Please find above is the raw dataset for which we have to do the analysis and give the best model for the prediction.

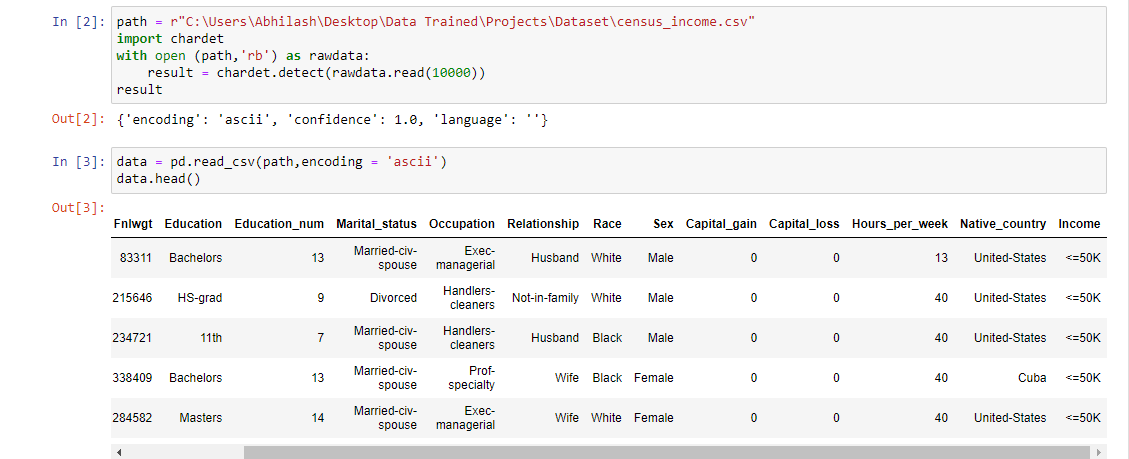
Hello all before we proceed anything further, we have been given a dataset with other features supporting such as age sex race etc, and our task is to predict whether a person makes over $50K a year or not.

Let me call upon or download all the important libraries before we proceed to touch any of the dataset and get an over view.



I have imported warnings in order to ignore any unnecessary warnings for the allotment we have redeemed for. It is not compulsory we have to call in warnings.

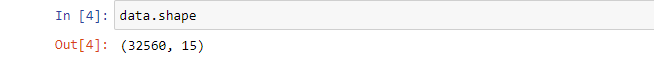
The data we have received is an excel which is in the csv format, so we have to import the data to our Jupiter notebook.



The file has been imported to the Jupiter book and data.head is indicating all the features and labels involved in that particular file.

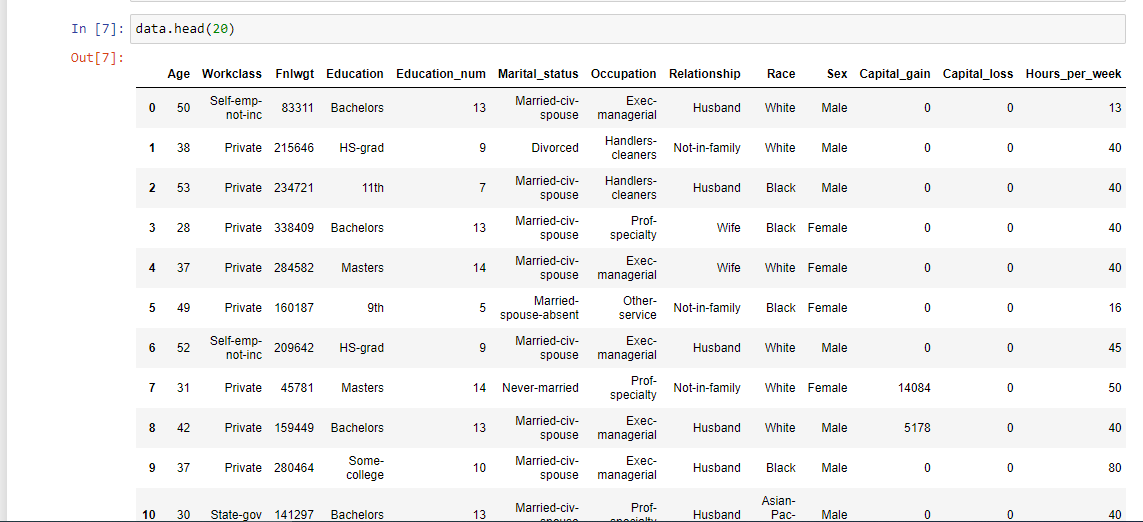
We are seeing there are many features and as declared in the problem statement **Income** is our label which is indicating earning greater than **$50K** or less than **$50K**.

I would like to check how big the data set is by simple to getting to know how many rows and columns



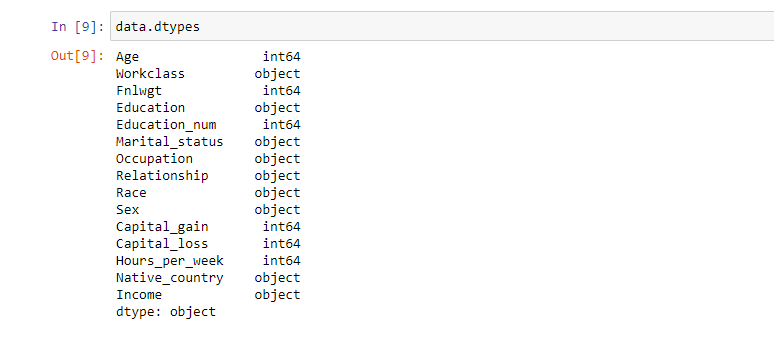
Here it means to say it has 32560 rows and 15 columns.

I would like to see more detail to just give an inspect to determine if I can get any information from the dataset I pull.



It is showing top 20 rows of all the attributes of the given data set, Let me see if I can find any similarities so that it would be useful for me.

Let me also see the type of data involved in the feature :



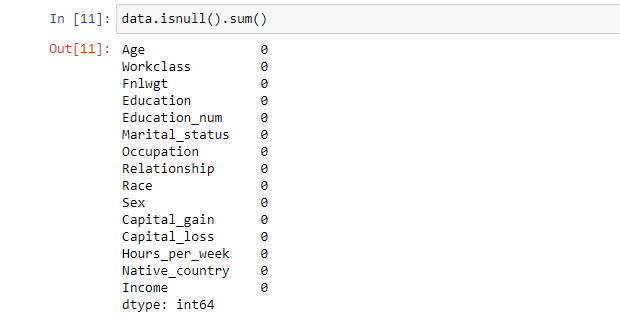
Here we see it is having :

1. Age,Flnwgt,Education\_num,Capital\_gain,Capital\_loss,Hours\_per\_Weer : Integers
2. Workclass,Education,Marital\_status,Occupation,Relationship,Race,Sex,Native\_country,Income : Object/String type

Also from the above dataset by overview we can say :

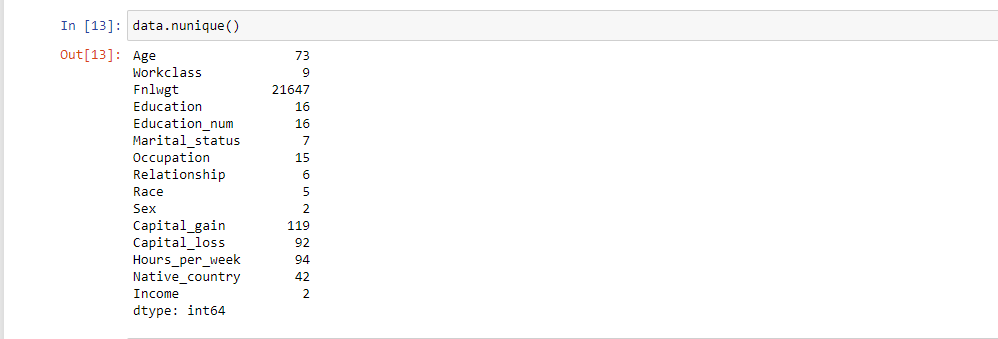
1. Some of them are Categorical data, for example **Workclass,Marital\_status,Relationship,Race,Sex** they all are in categorical format,I was able to observe that cause the data are repeating like for **Sex** its either **Male or Female** and its repeating.May be we can tune all those in the later stage.

Now let me look for if any of the data is missing .



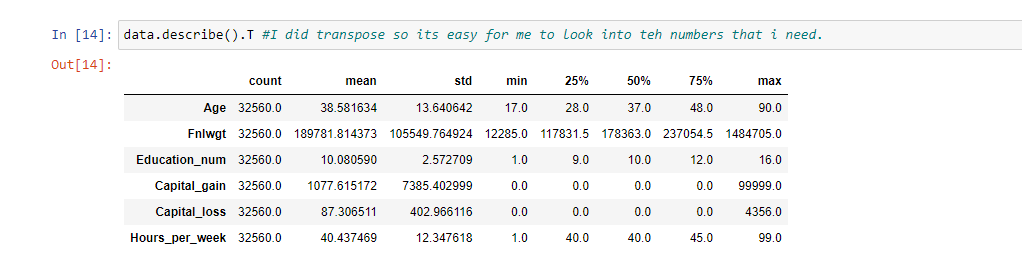
Wow great we have no data missing in our data set, if at all there was any missing data then we had to treat the data to stabilize the data.

However now since the data looks good we can proceed further.



Here I am trying to check how many unique contents are there in my dataset, we see a varying number Infront of my features and labels. **Sex -2** It means it say under **Sex** there are only two categories, since we have referred the dataset, I can tell its either **Male** or **Female**, so generally anything less than 10 or at max 20 we can easily categorise all that to make our dataset clean and easy to access and it would be uniform.

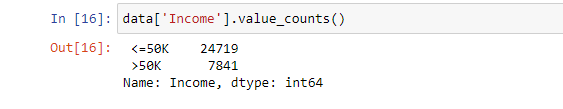
Now let me check the data, how well its been arranged and also try to inspect if there are any similarities amongst the data.

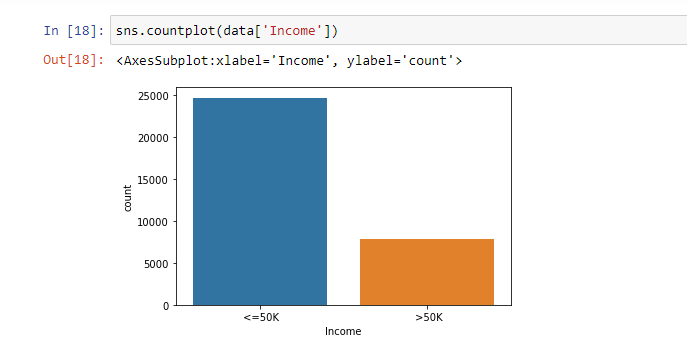


I have used **data.describe().T** it means to say transpose the data so that I can check column wise accordingly.

I see the data is good except for Capital gain and Capital loss, because when we saw the min 25% 50% and 75% it shows all as 0 however the max has some huge value, so I need to treat it on later wise to check for it.

Let me check the categorical value of the label **Income** how well it is distributed in the platform.

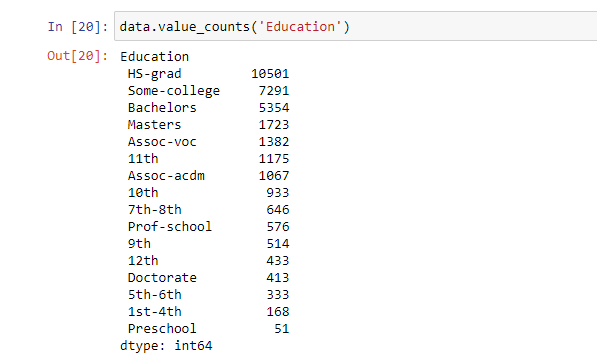




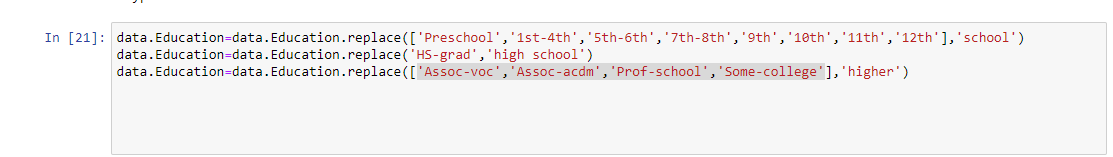
From the above we see the counts how many are earning <=$50K and how many >$50K

Now let me start manipulating my features in order to bring it to a better shape and simple to make my data look good.

I am starting with **Education** :

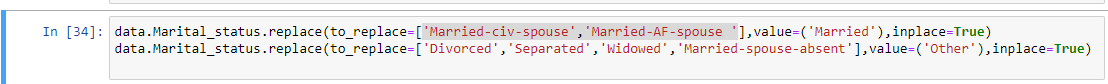


I see there are many sub columns under Education I would like to design to make it simple by categorising **1std to 12th std as School** , **'Assoc-voc','Assoc-acdm','Prof-school','Some-college'** all as higher so that we do not have all these so many category and it would tune in better.



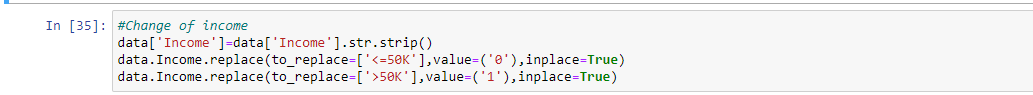
I would do the same **for Marital\_status like : 'Married-civ-spouse’, ‘Married-AF-spouse ' as “Married”**



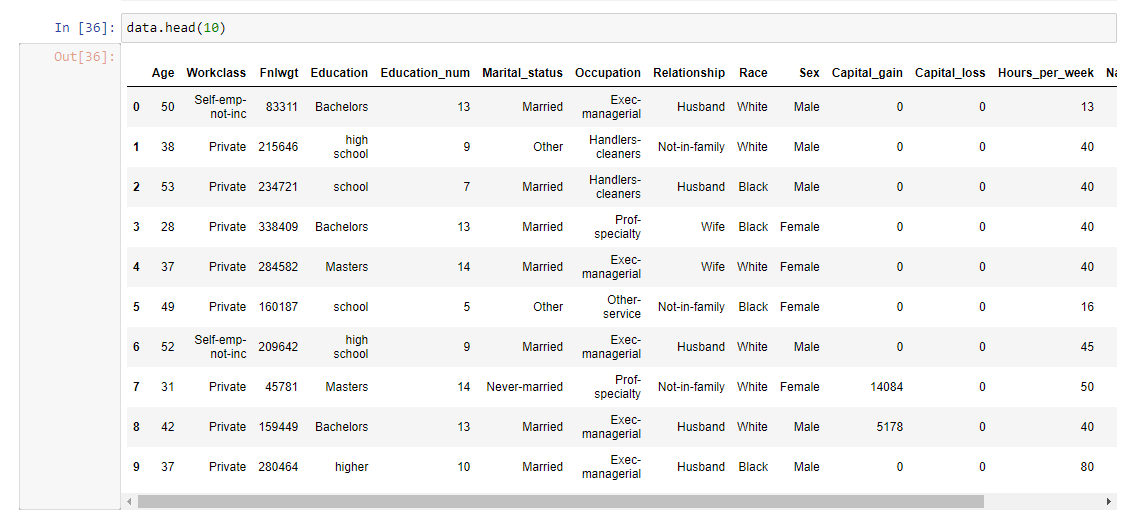


And

Income : <=$50K as **0** and >$50K as **1**



Now let me look how my data looks :

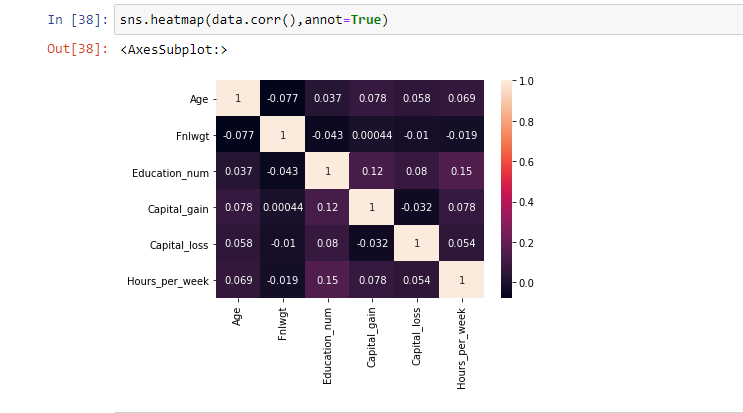




The data looks all good for the necessary design changes we have done .

I need to check for any correlation amongst the features included in the dataset, correlation here meaning to see if any column is related to the other column like similarities or dependent.

I would be using Heatmap technique to visualise if any correlation exist.



This is the output of the heatmap, there are colour range from the lightest colour to the dark and at the ide the numerical count is given per the colour vice.

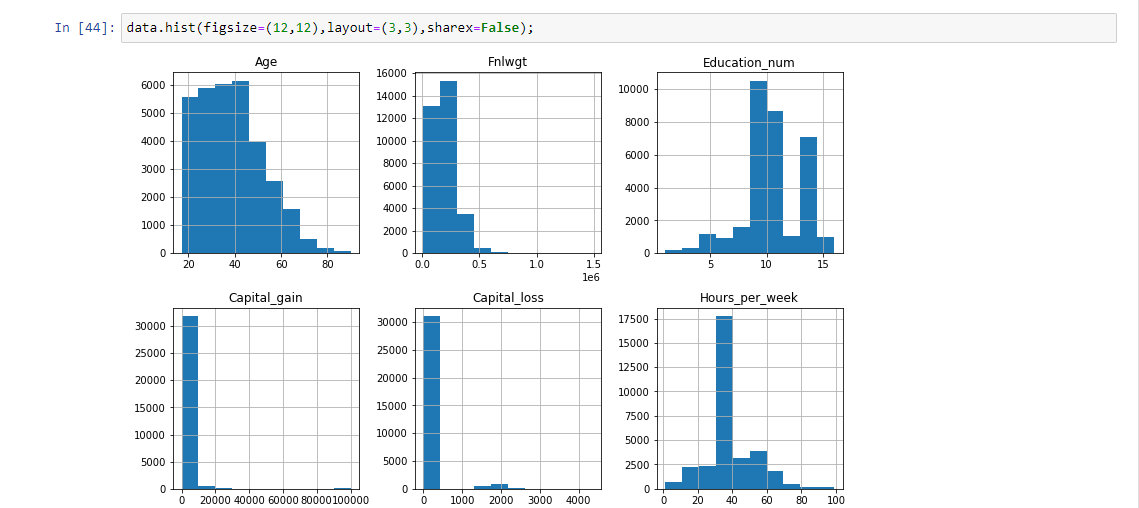
1. being the lightest which means to say Correlation exisit
2. being the darkest which has the least/no correlation

The graph in the above plotted are columns against column to check against each column wrt to other should be, here the diagonal all is the lightest with the count 1 since that columns is plotted against its own column which would be completely correlated.

Eg,Age vs Age it is 1 so lightest colour ,Flnwgt vs Flnwgt Age it is 1 so lightest colour , ……etc

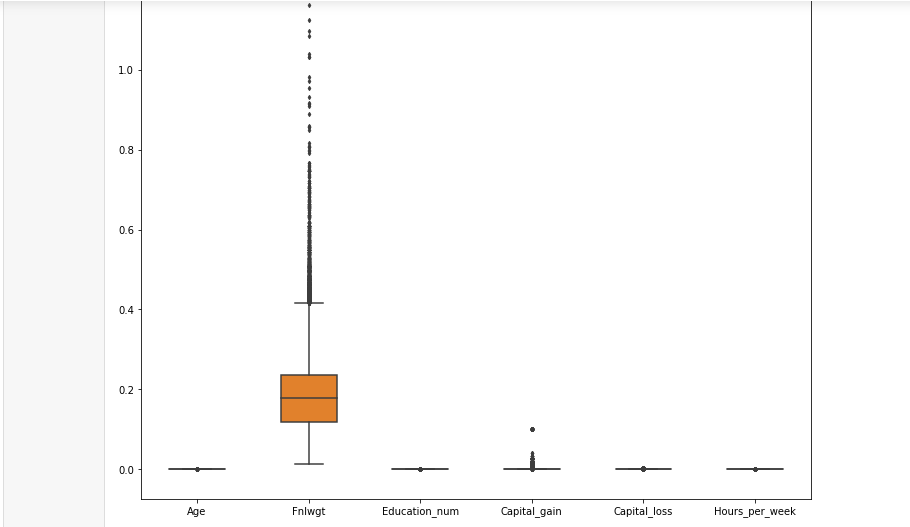
From the graph we have plotted it looks good with no correlation and good to proceed further.

Let me also check on the graph for the individual dataset how its been distributed, I am using histogram to get an overview on the data because it would give me the frequency of distribution where the width represents the class of interval and areas are proportion



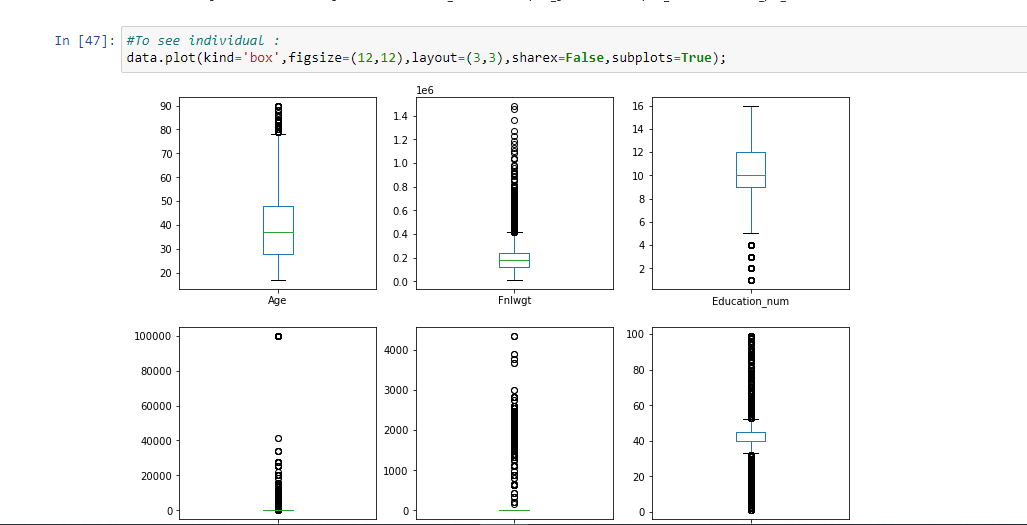
Let me check to see for the outliers :





So here anything greater than the max margin are the outliers that are present in those features.

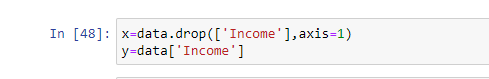
Let me see it individually for the features:



Here it is giving me the alongside of the outliers, I have used a box plot visualization.

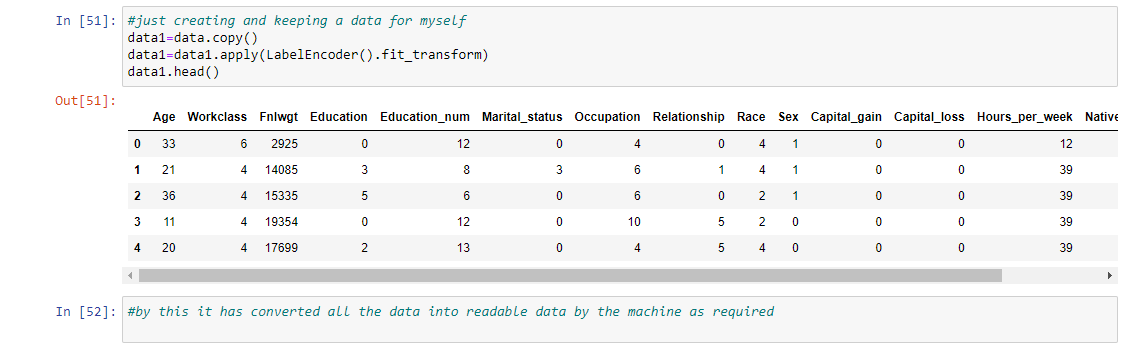
Before creating any model let me divide the feature and Label.

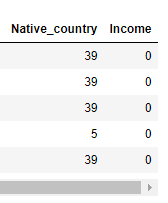
As we all know Income is the Label and others are all features, I would like to drop the Label from the main dataset, thereby the main dataset would only have features.



I would also create a copy of my dataset for my reference to keep up the original data and the copied data I would manipulate further to make all the standard since we are having strings in some. I would use **Label Encoder** to convert all this strings to a machine understandable code.

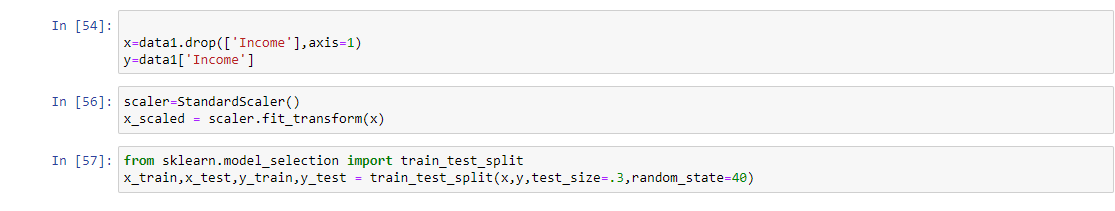






Now the data we see all are in integers or codes which can be easily understood by the machine.

Since the data we have occurred in to are all machine codes let me pull in the feature and Label from the converted data to do the future modelling.



Here I have declared train and testing data, I have chosen 30% as test size hence given **test\_size as 30 , random\_state** have given as 40,meaning to say the data I am having I am dividing 70% for the training and **30% for testing**

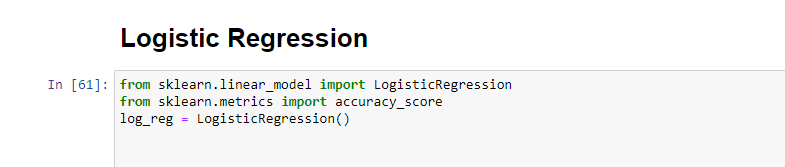
End of the Exploratory Data Analysis.

Let me try building the model and see which would be the best and give the best accuracy.

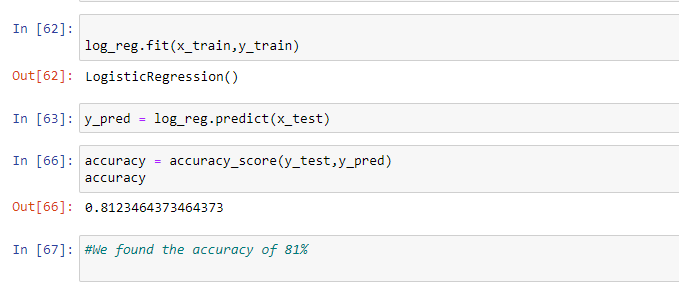
**MODEL BUILDING :**

1. **Logistic Regression :**

I would try my first model with the Logistic regression, since the Label is also a categorical data, since the response variable is binomial for the comparison.



I imported some important libraries for the Logistic regression I am performing.



After I assigned, I fit the Logistic regression to the training data for both the feature and the label.

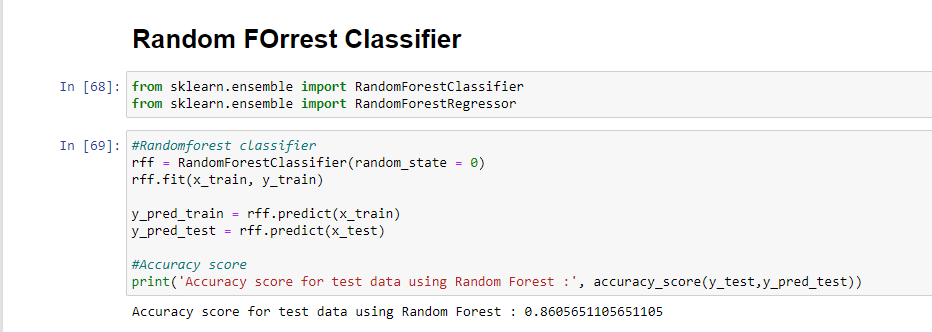
With the data that’s been trained I would try to predict only for the feature test data using the Logistic regression.

Now to find how well my model is or the accuracy is I would like to find it out with the Label test data with that of the Label predicted data.

I found the accuracy is 81.23% which is a good accuracy for the model we have built.

Let me check with other models as well and choose the best one that gives the best efficiency.

1. **Random Forrest Classifier:**

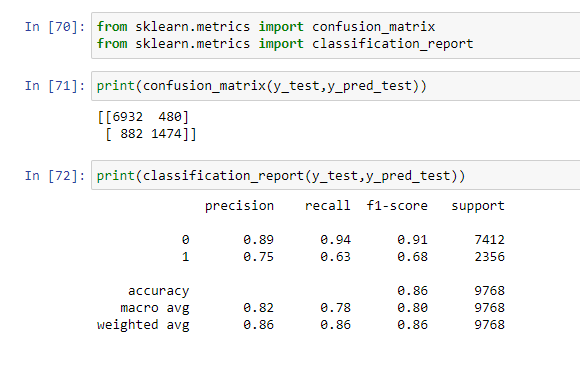


Here I am using the same data as well what we were doing for the Logistic regression and performing the train and test

In this case we got the accuracy as 86.05%

Let me check with the Confusion Matrix:

Here the confusion matrix uses True Positive ,True Negative , False Positive and False Negative format system and tries to find the accuracy.



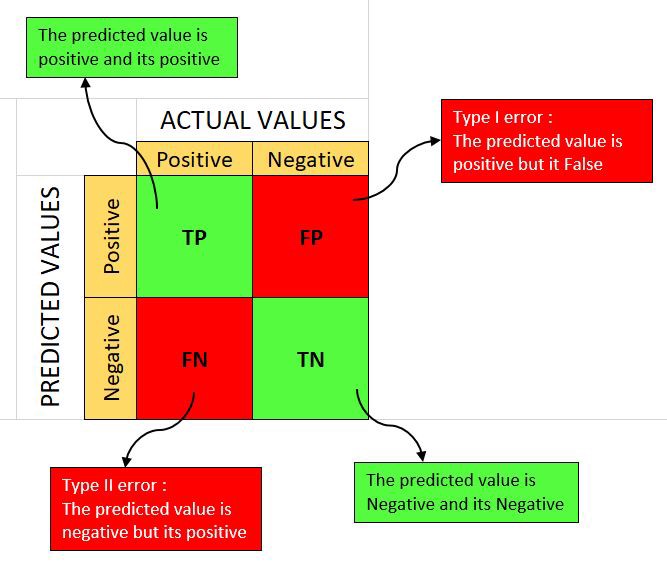
Confusion Matrix :

True Positive : 6932

True Negative: 1474

False Positive : 480

False Negative : 882



Source : Google on how the confusion matrix, please refer to the above image

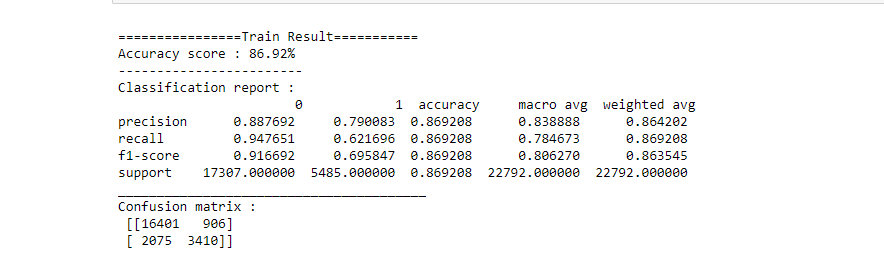
Based on the Confusion matrix we have received the output of all the Precision , Recall , f1-score , Support

1. **Gradient Boosting Technique :**

Now let me try to build up the model using Gradient Boosting technique, I would be following up in the same.

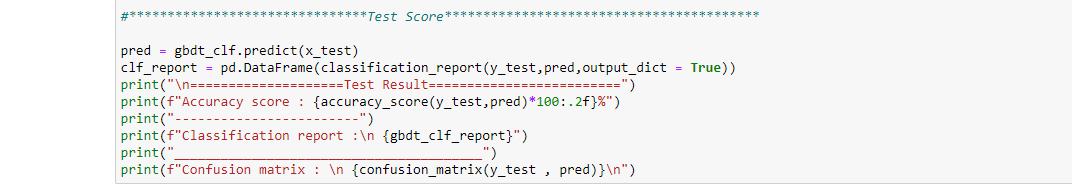


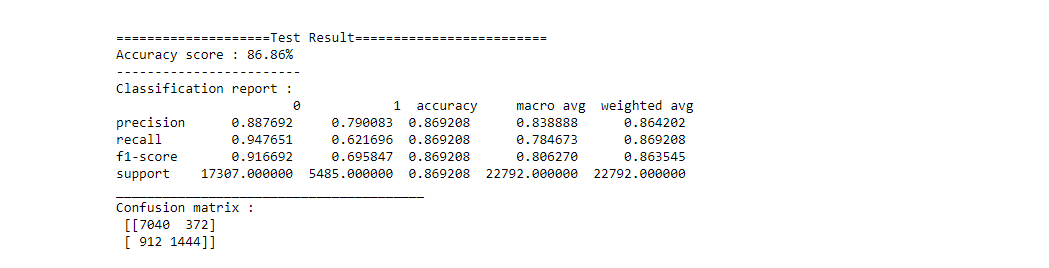
Here I am trying to create first with the training data and allocating to the confusion matrix to determine how well my data looks



Here for the training data the accuracy score 86.92%

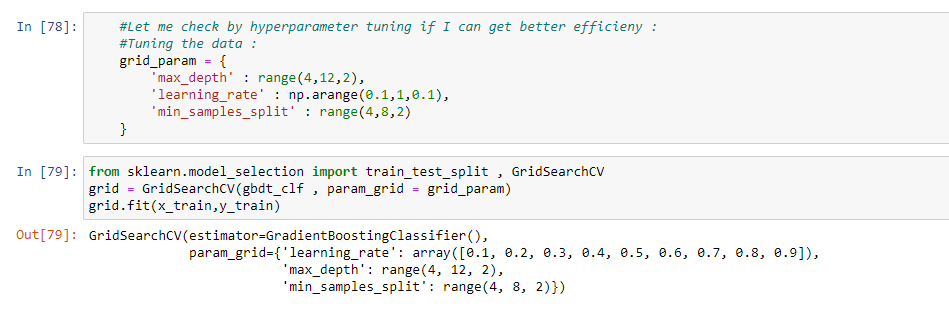
Now let me check for the test data :





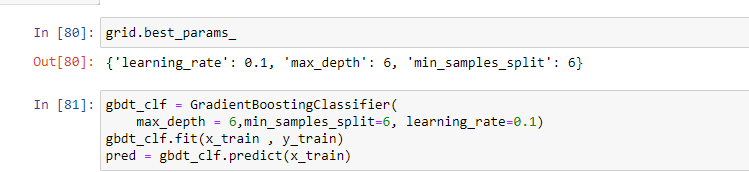
Well for the Test data we got an accuracy of 86.86%.

Let me try hyperparameter tuning to see if I can get a better efficiency? :



Here I have got the learning rates, maximum depth of the node and the minimum samples split.

I need to choose the best under all these to induce in my model to tune in more better to process.



Here by I got:

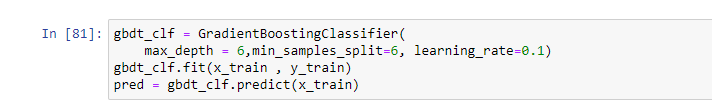
Learning Rate : 0.1

Max\_depth : 6

Min\_samples\_split :6

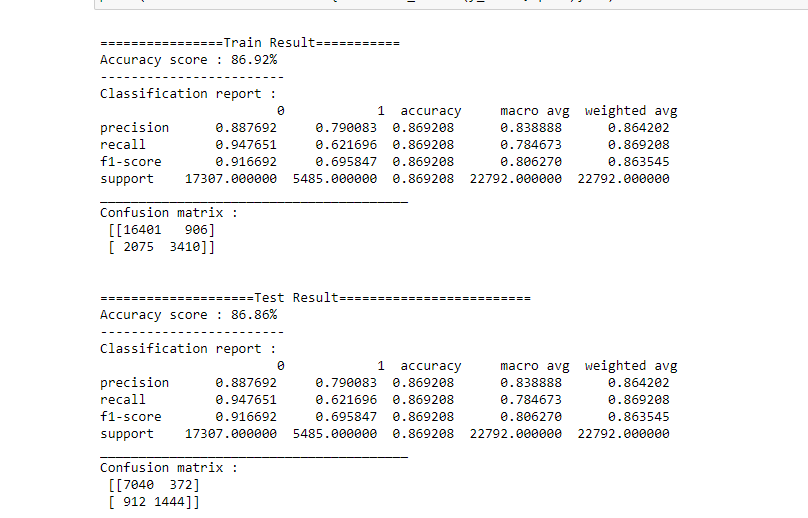
These are the best to be induced to my model .

Let me apply the same and see how the data would look :





Let me check for the accuracy on how the data is :

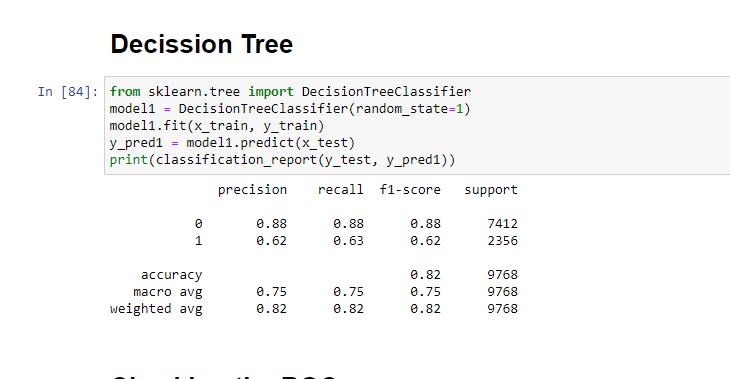


Well the data looks the same the efficiency seems to be the same no changes has occurred even after the hyperparameter tuning.

Therefore retaining with 86.86%

**3)Decission Tree :**

Now let me try with the decision tree, it would split a node into two or more sub nodes on all available variables and then selects the split which results in most homogeneous sub-nodes



Well we have built some of the models, to choose which among is the best we can get by the best accuracy we have received to visualize as well we can use another method that is the ROC (Receiver Operating Characteristic) curve by this we train with all the models and when we plot on graph we would choose the best that covers the maximum area on the graph we have plotted.

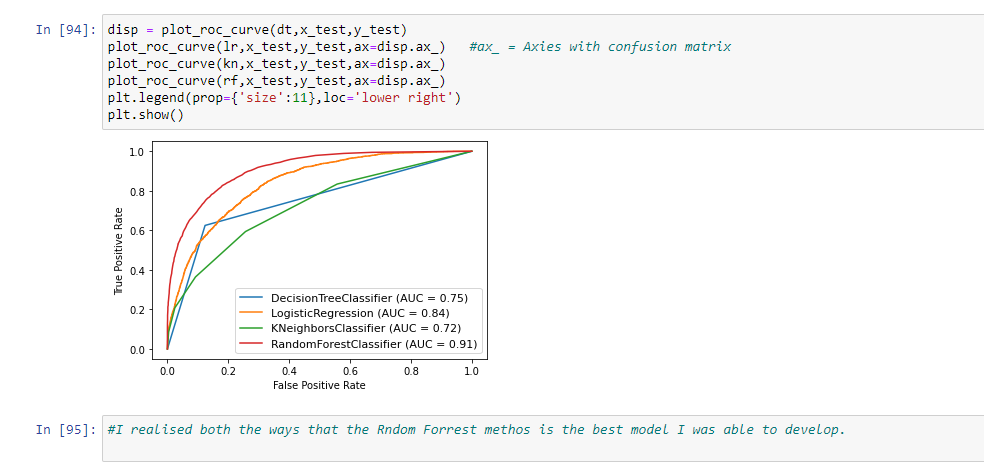


In the above I have declared and trained the dataset with the possible models and also applied for the training model.

I will now train the model to only for the feature test data and try to find the efficiency or the result output to see which among is the best in the numerical vice.

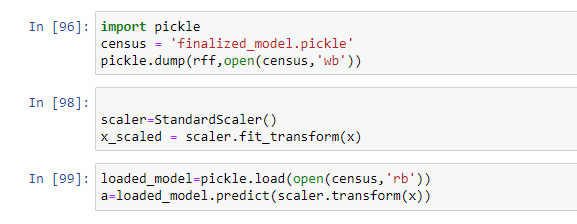


Since I have got the result for the feature test, now let me plot the graph with respect to the feature as well as the label and see which among the model has covered the maximum area under the plotted graph.



Even by visualizing I was able to relate that Random Forrest Classifier is the best model showing 91% and when we see the graph in red, it is the one that is covering the maximum area under the plotted graph.

There by I would choose the Random Forrest Classifier as the best model and save and dump the model.



I have saved the model in order to use it for the given features along with the Label.

**Thank You**

**Abhilash Ashok**