**Census\_Income Project**

**Problem definition**

A census is way of getting information about every member of population, whether the person is married or not or how much he/she earns and many things related to it. In Census\_Income\_dataset we have been provided with the data from 1994 after going through their complete details we have to make a model that predicts that whether a person earns more than 50k over a year or not .

**About Dataset**

This data was extracted from the 1994 Census bureau database 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)). 

To help the bureau to take such dynamic decision instantly we will Analyse the data, then we will do Exploratory Data Analysis and write its remark, then we will Train the model and conclude it what factors play an essential role that a person can earn more than 50k in a year

**Data Analysis**

The Census\_Income dataset is a dataset with 32,560 and 15 coulmns the columns are-

1. **Age**

The Age column describes the age of population. The minimum age recorded is 17 and the maximum age recorded is 90.

1. **Workclass**

The workclass column conatins where the people works like 22696 people belongs to Private Workclass, 2541 people are self\_emp\_not\_inc workclass, 1836 belongs to Unknown workclass, 1297 are in State-gov workclass , 1116 are of self-emp-inc workclass , 906 are from Fedral-gov-workclass , 14 are on without pay , and 7 have never-worked.

1. **Fnlwgt**

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 by Population Division here at the Census Bureau. with 3 sets of controls. As follows

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

ii) Controls for Hispanic Origin by age and sex.

iii) Controls by Race, age and sex.

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 . 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.

The maximum weight recorded is 1484705

1. **Education**

It describes the education of the people

like 10501 people are HS-graded , 7291 people belongs to the some college, 5354 people are Bachelors, 1723 have studied Masters , 1832 people have done associated vocational, 1175 are 11th passed , 1067 have done associated academics, 933 are 10th passed , 514 people are 9 th passed , 646 are 7th or 8th passed people, 576 have completedtheir prof-school , 433 people have completed their 12th , 413 people have done Doctorate and 333 people are 5th - 6th passed , 166 people have done their studies upto 1st to 4 th and 51 have completed only their pre-school .

1. **Education\_num**

the maximum education num is 16 and minimum education num is 1

1. **Maritial\_status**

The column describes the marital status of person like 14976 are Married-civ-spouse , 10682 are never-married , 4443 are Divorced , 1025 are separated , 993 are widowed , 413 are Married-spouse-absent and 23 are Married-AF-spouse

1. **Occupation**

The column specifies the occupation of the people like 4140 people are Professional-specialty , 4099 do the occupation of Craft-repair , 4066 do the occupation Executive -managerial , 3796 are doing Administration-clerical work ,3650 people are in Sales Occupation , 3295 are engaged in Other-service , 2002 people are in Machine-op-inspct 1843 are doing some unknown occupation ,1597 people are doing Transport-moving occupation , 1370 are Handlers-cleaners , 994 do the Farming-fishing , 928 are in Tech-support, 649 do Protective-services , 149 are private -house-servants and 9 are in Armed-Forces

1. **Relationship**

The column describes the relationship of people like 13193 are Husbands , 8304 people are not in family , 5068 are Own-child , 3446 are unmarried , 1568 are wife and 981 are other-relative

1. **Race**

The column specifies the race of person like 27815 people are White people , 3124 are black people , 1039 are Asian-Pac-Islander , 311 are American-Indian-Eskimo and 271 are of other Race

1. **Sex**

The column describes the gender of people like 21789 are Male candidates and 10771 are Female candidates

1. **Capital\_gain**

The column describes the capital\_gain of the people. The minimum capital\_gain starts from 0 and maximum capital\_gain is 99,999.

1. **Capital\_loss**

The maximum capital loss is of 4356 and minimum capital loss is 0

1. **Hours\_per\_week** –

The people work for minimum 1 hour of week to the maximum of 99 hours per week

1. **Native\_country** -

29169 people belong from United States and remaining are from various native countries

1. **Income** –

24719 people have income greater than or equal to 50 Thousand $ and remaining 7841 people have income less than 50Thousand $

**Exploratory Data Analysis**

Exploratory Data Analysis(EDA) is an approach of analysing data sets to summarize their main characteristics, often with visual methods, a statistical model can be used or not , but primarily EDA is for seeing what the data can tell us beyond the formal modelling or hypothesis testing task. we can say that EDA is statisticians way of story telling where you explore data , find patterns and tell insights. EDA is a phenomenon under data analysis used for gaining a better understanding of data aspects like :- main features of data - variables and relationships that hold between them - identifying which varaibles are important for our problem

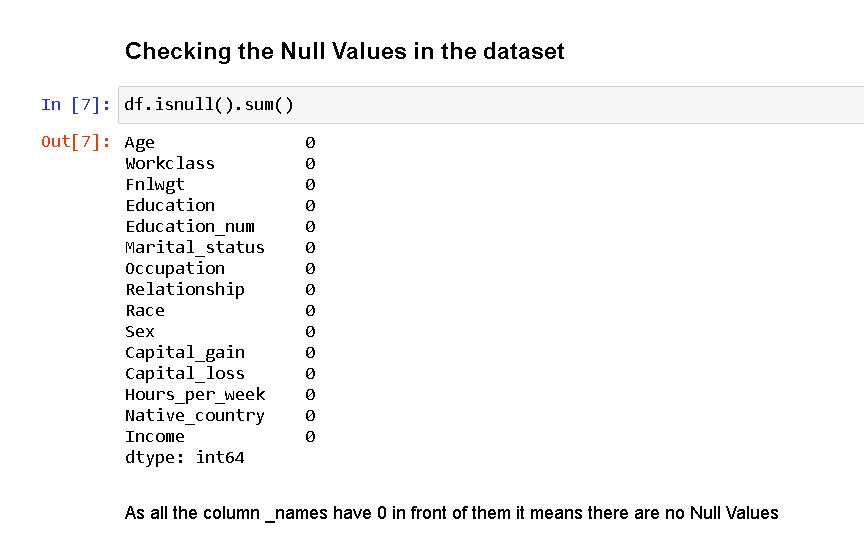
we shall look at various exploratory data analysis methods like –

1. **Variable Identification-**

Looking the input data, what will be the output variable

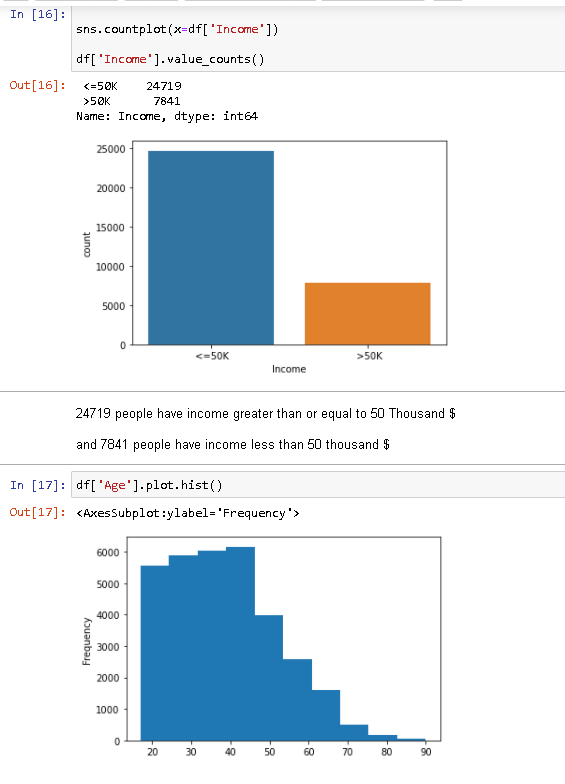
1. **Handling/ Removing null values/** missing data –

It is possible that there may be null values in the dataset. missing values usually occurs when an entry is left empty .It can be filled with the mean ,median or mode of the column through fillna() function. Or sometimes SimpleImputer() function is used to replace the nulll values along with label encoding them In this dataset there are no Null values , but there are “ ? ” in the Workclass, Occupation , Native\_country so they are replaced with most\_frequent values of respected column using SimpleImputer but they are not counted as Null values you can see in the following code snipet



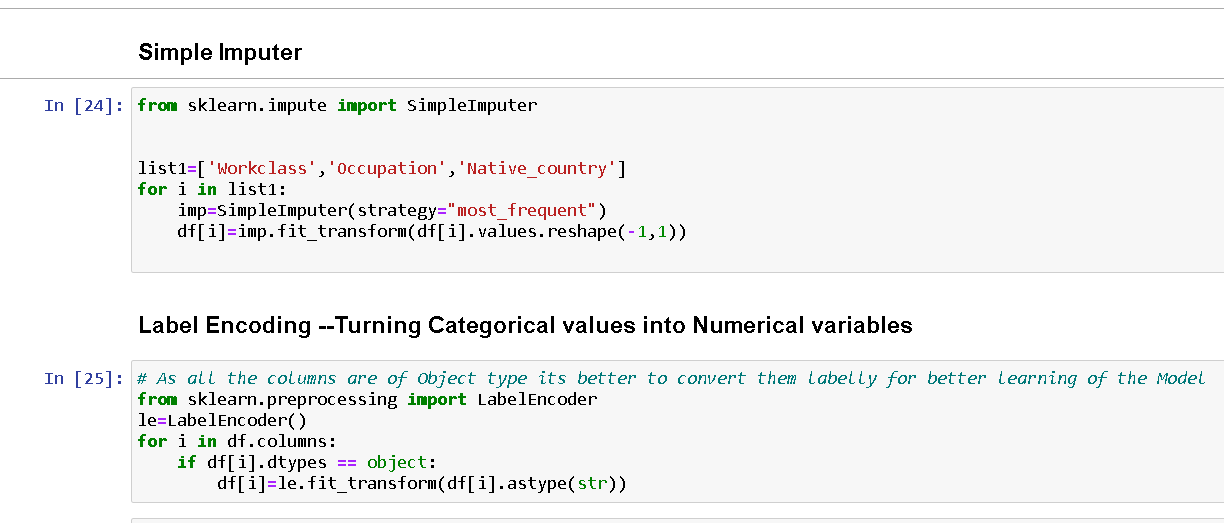
1. **Visualization of dataset**

It consist of exploring the dataset visually through histogram, bar chart usally for single variables, Like its better to use countplot for the categorical value and histogram or distplot for continuous variable like for all the categorical values like Workclass, Education , Marritial\_status, Occupation, Relationship , Race , Sex , Native\_Country and Income we used countplot and for the columns with continuous values like Age, Fnlwgt , Education\_num , Capital\_gain, Capital\_loss we used histogram



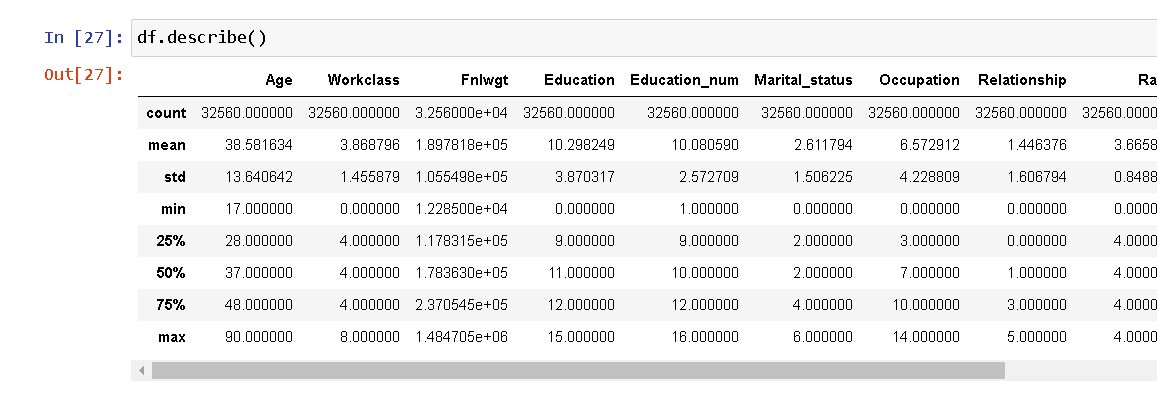
1. **Handling Categorical variables**

Most statistical model cannot take Objects / Strings as input they only takes numbers as inputs , with LabelEncoder() it is possible to categorize the string into Numbers as 1,2,3 and so on like Workclass, Education , Marritial\_status, Occupation, Relationship , Race , Sex , Native\_Country and Income have string input they all are converted through LabelEncoder().fit\_transform function, but as there are ‘ ? ’ in some values they are first replaced with most\_frequent value of that column as you can refer to the following code



1. **Descriptive Statistics** –

It is a way of giving a brief overview of the dataset we are dealing with , including some measures and features of the sample.   
 For this we can use the describe() function the describe() function tells the mean() , min(), max(), 25th percentile , median 50th percentile and 75th percentile of the data.

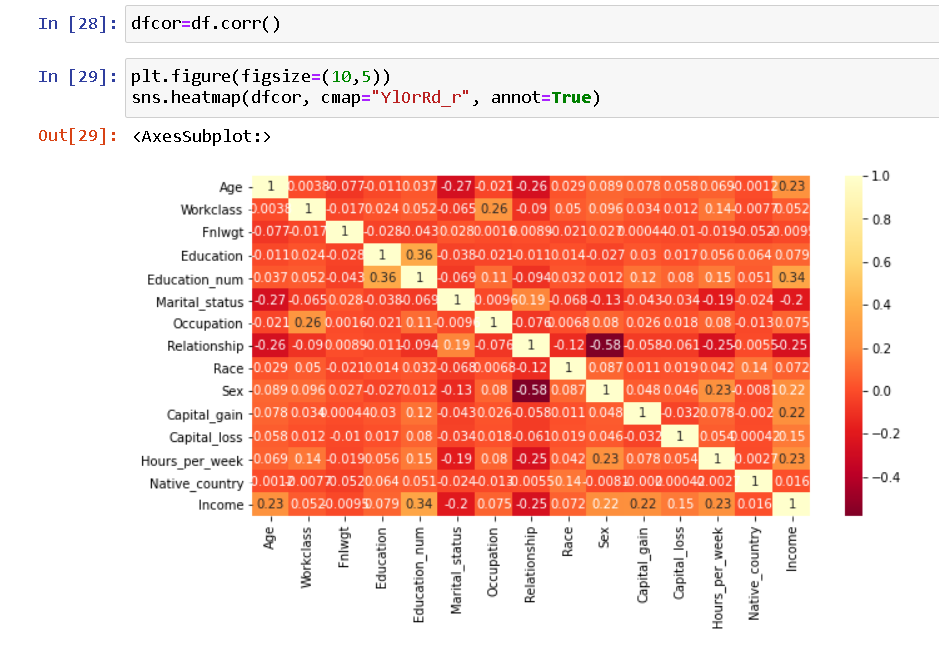


If There is a significant differenec between 75 % and max it means outliers are present. Like in Age , Capital\_gain , Hours\_per\_week and Income.

And if The value of standard deviation is so high it means the skewness in present in columns like capital\_gain and capital\_loss and as the std() range is high it means the data is highly spread and skewness is present.

1. **Correlation**

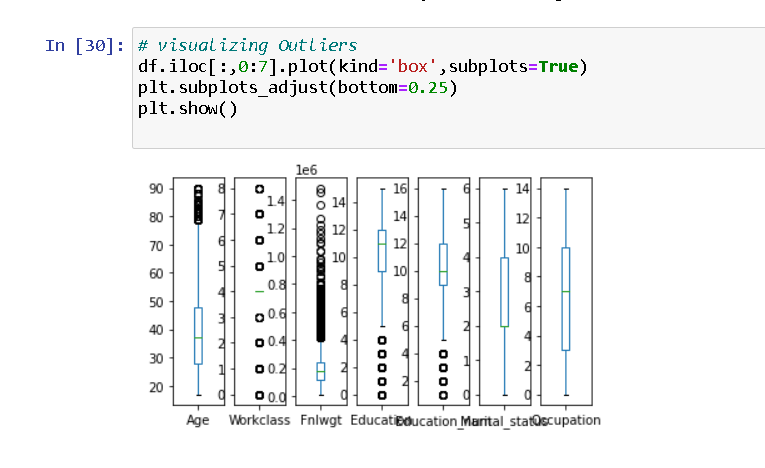
Correlation is the statistical metric for measuring to what extent Different variables are interdependent, like if one variable changes how it affects the change in other variables. corr() function is used to see the correlation among the dependent variable and independent variable you can see correlation in the following figure



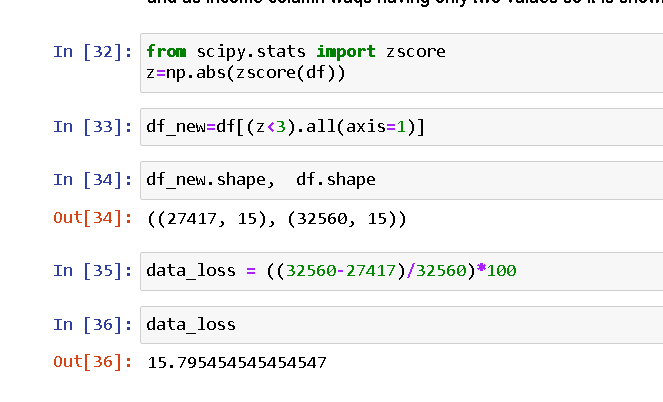
Like Income is positively correlated with Education\_num , Capital\_gain , Hours\_per\_week and Negatively correlated with Maritial\_status and Relationship and all other features like Workclass, Fnlwgt, Education , Native\_country have week relation with Income.

1. **Outliers**

Outliers are the value that lies outside the range of the data, Outliers can be identified through zscore() function from scipy.stats library or sometimes through IQR method, To remove the Outliers we have to select the data that has zscore less than 3 .



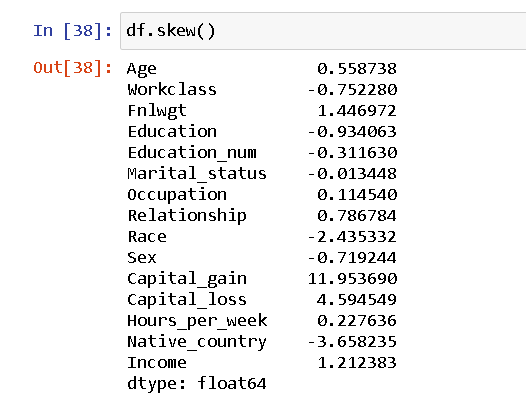
In our case as Workclass, Education and Native\_country is a conversion of Categorical data into numerical so it is showing the outliers point but very high amount of outliers are present in FnlWgt capital\_gain , Capital\_loss and Hours\_per\_week columns.



15% of the total data contains outliers, Loosing 15% is not at all affordable so dropping the idea of removing outlier.

1. **Skewness**

The data can be right skewed or left skewed if the median or mean is high and data is highly spread it can be observed through the skew() method, if the skew score is negative and greater than 5 it means data is negatively skewed on left side and if the data is more than +5 it means the data is skewed on right side.



In our dataset Fnlwgt, Capital\_gain , Capital\_loss have the skew score is greater than +0.5 and Workclass and Income has negatively scored .It is necessary to cure skewness for this PowerTranform () function is useful but it should be done after separating features and target

1. **Preprocessing Pipeline**

Pipelines are the special way to simplify the code , Pipeline is generally used if we have to perform the code repeatedly usually when there is different train and Test data

Here Our EDA process is completed now moving towards next step

**Building Machine Learning Models**

A machine Learning Model can be a thought of a mathematical equation used to predict a given one or more other values

1. **Separating Features and Target column** –

It is necessary to separate the independent/Features column into a variable ( x ) and target column into a variable ( y ) . here we have to separate all columns in x DataFrame (variable) and income variable in y DataFrame (variable)

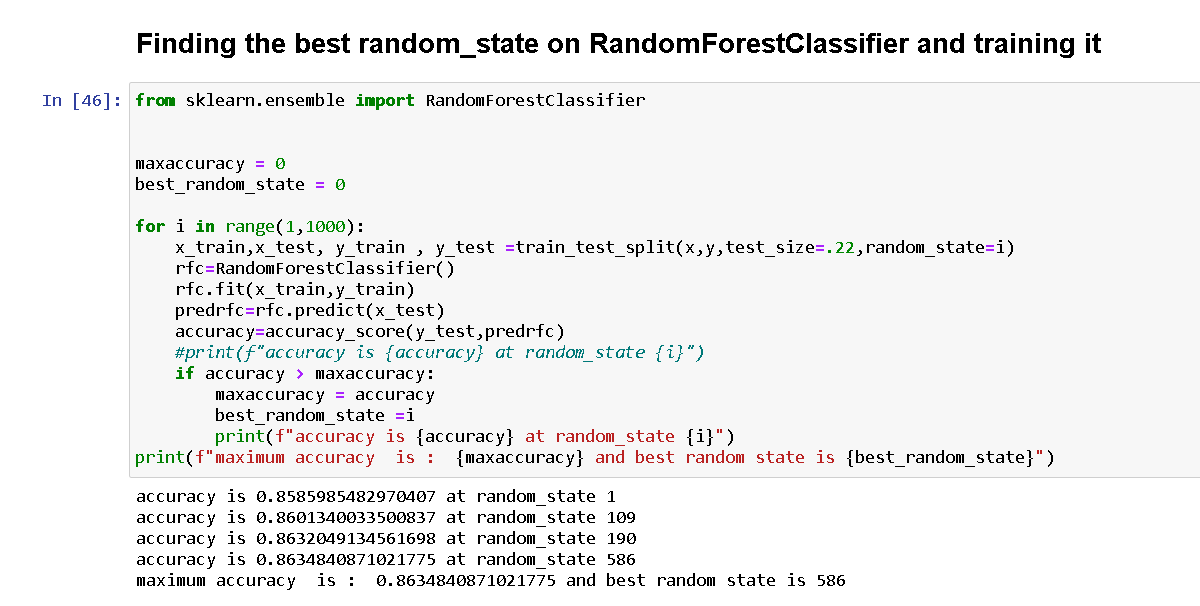
1. **Splitting the Data for Training and Testing**

In ML the separated data is splitted into 4 parts for Training and Testing of features ( x ) and for Training and Testing of Target ( y ) like x\_train , x\_test , y\_train , y\_test.

It is possible through a inbuilt library of sklearn’s train\_test\_model , it separated the data into test\_size suggested by the DataScientist

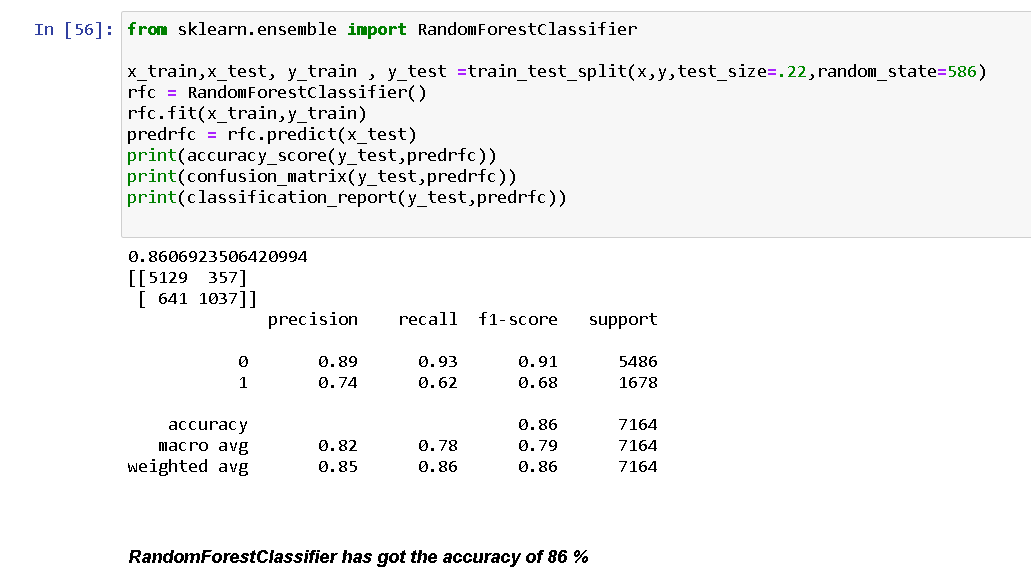
1. **Finding the Best Random State**

The random\_state is one of the important parameter to be submitted and inserting best random state gives the best accuracy. It can be done through the for loop , In it we can compare the accuracy at each random state and find the best accuracy , we can find the best random state for each Models



1. **Training the Models**

To find the best model it is necessary to train 3-4 models,



In the same way I have trained LogisticRegressionModel , DecisionTreeClassifier, RandomForestClassifier, and SupportVectorClass to check which model is giving the best accuracy

i) LogisticRegression Model is giving the accuracy of 83%

ii) DecisionTreeClassifier Model is giving the accuracy of 81%

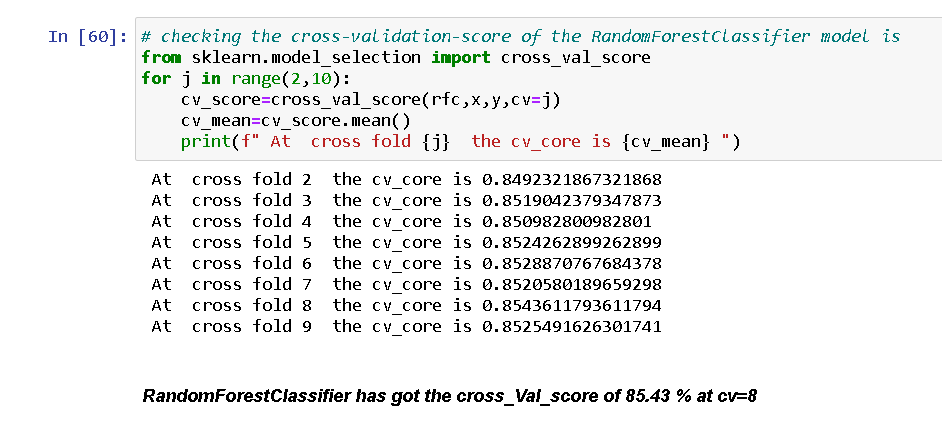
iii) RandomForestClassifier Model is giving the accuracy of 86%

iv) SupportVectorClass Model is giving the accuracy of 84%

Here the above models are giving quite good accuracy, so it is necessary to recheck the model’s learning

**5)Cross-Validation of the Model**

Sometimes it is possible that model may overfit for that cross validation is used to check model’s learning ,so we have done cross-validation of all the models along with 9 cv values and compared the difference to check for best model and select it, you can go through the image below



Like this we find the cross\_val\_score of all four models

i) cross\_val\_score of LogisticRegression Model is 82.51 % at cv= 8

ii) cross\_val\_score of DecisionTreeClassifier Model is 80.37 % at cv= 9

iii) cross\_val\_score of RandomForestClassifier Model is 85.43 % at cv= 8

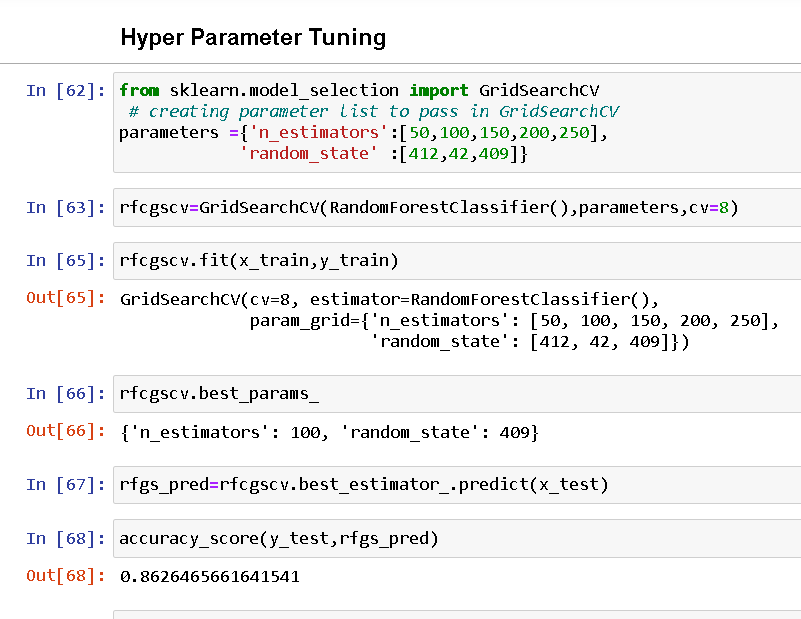
iv) i) cross\_val\_score of SupportVectorClass Model is 83.67 % at cv= 9

6**) Selecting the Best Model** -

After getting the accuracy and cross\_val\_score of all the models the comparision is made and the concusion can be made that we can see that all models have diffetence of 1 between their Accuracy and Cross\_Val\_Score , but RandomForestClassifier has the highest accuracy score of 86 % and highest cross\_val\_score of 85 % it means that RandomForestClassifier has learnt well

1. **Hyper Parameter Tuning Using GridSearchCV** –

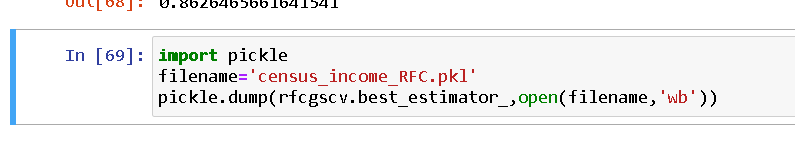
After selection of model it is necessary to Give the best parameters to the model and it is possible with GridSearchCV method which chooses the best parameters and makes the model fit for exporting



As you can see that we have Trained the RandomForestClassifier Model and we have got the accuracy of 86 %

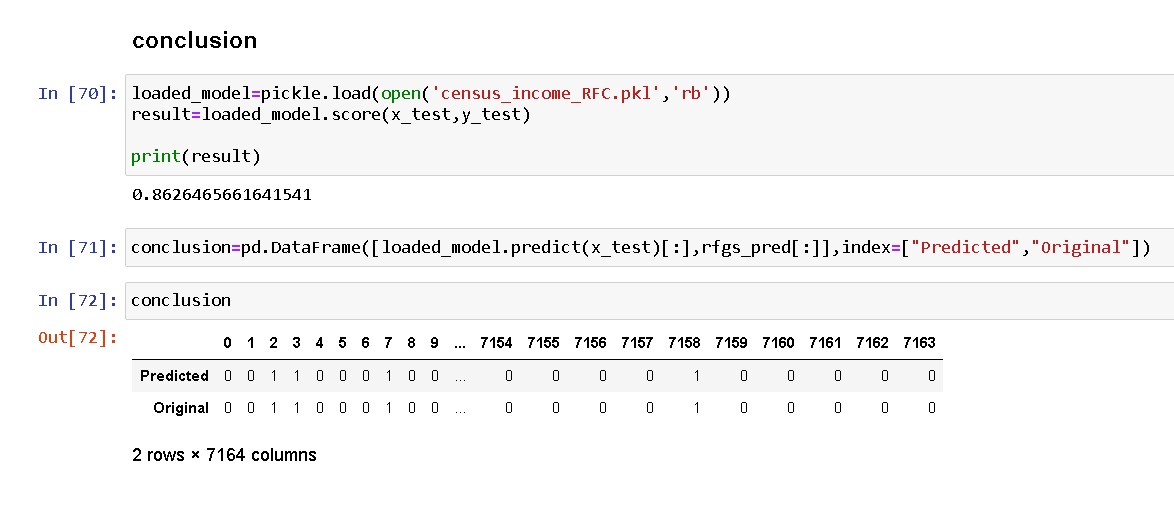
7)**Exporting the Model**

After GridSearchCV the model is exported into .obj file or .pkl file it can be done through Joblib Library or Pickle Library



**Conclusion**

As we have gone through a good portion of the problem, we have done data analysis, data cleaning, EDA, we have built the model and exported it too.After we have exported the model we can cross check the performance of our model by making DataFrame of predicted and original values. We can also do a cross check to test our model’s performance you can see below in the image.



Now we can easily predict whether the people is earning greater than $50k or less than $50k by loading the model and providing the values.

In this way we have successfully created the model and exported it .