Census Income Project Prediction

Dated: 09.06.2021

**Prepared By**: Bhumik P. Shah

**Data Trained Batch No**: 1828

**Problem Statement:**

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).  The prediction task is to determine whether a person makes over $50K a year or not.

**Dataset:**

The data has 32560 rows and 15 columns.

**Dataset Description:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **Data type** | **Values type** | **Description** |
| age | int64 | Continuous | Age of person |
| work class | object | Discrete | Work class of person |
| fnlwgt | int64 | Continuous | Final weight |
| education | object | Discrete | Education Degree of person |
| Education\_num | int64 | Continuous | Number of years of education |
| Marital status | object | Discrete | Marital status of person |
| occupation | object | Discrete | Occupation of person |
| relationship | object | Discrete | Relationship of person |
| race | object | Discrete | Race of person |
| sex | object | Discrete | Sex of person |
| Capital gain | int64 | Continuous | Capital gain of person |
| Capital loss | int64 | Continuous | Capital loss of person |
| Hours.per.week | int64 | Continuous | Number of hours per week |
| Native Country | object | Discrete | Native country of person |
| income | object | Discrete | Income category of person |

**Libraries used:**

* NumPy
* Scikit Learn
* MatplotLib
* Seaborn
* Pandas
* Github

**Data Types:**

**There are three formats of data types:**

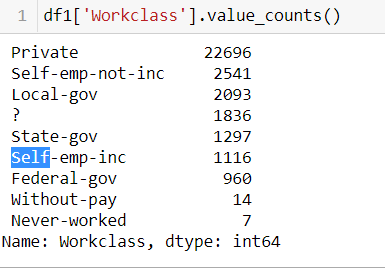
* **Object-** The variables are categorical in nature. The features named as work class, education, Marital status, occupation, relationship, race, sex, Native Country and income are of object datatype.
* **Int64-** It represents the integer variables. The features named as age, fnlwgt, Education\_num,

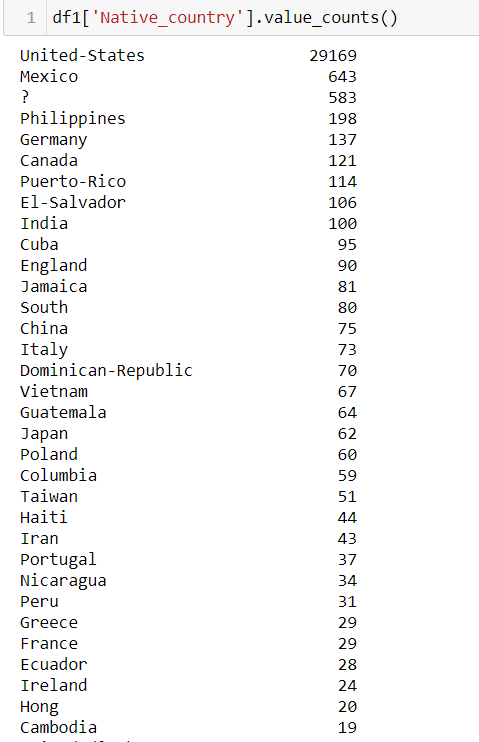
Capital gain, Capital loss, Hours\_per\_week are having int64 as a datatype.

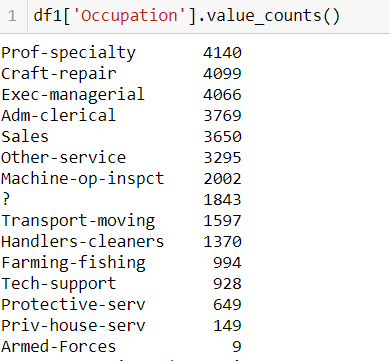
**Missing values in the Dataset:**

There are missing values in the features named as Work class, Occupation and Native country

represented by ‘?’.



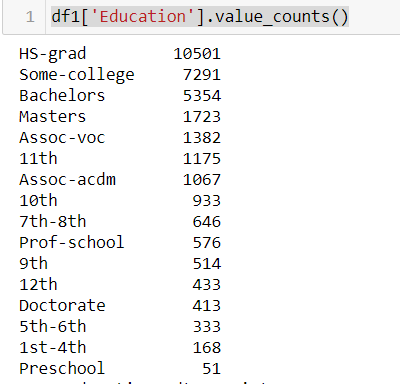




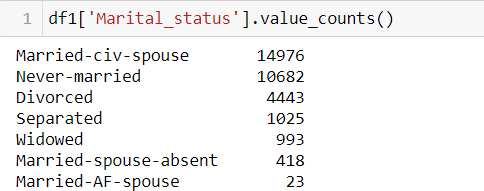
Since the missing values were represented by ‘?’ , they were replaced by the mode of that particular feature.



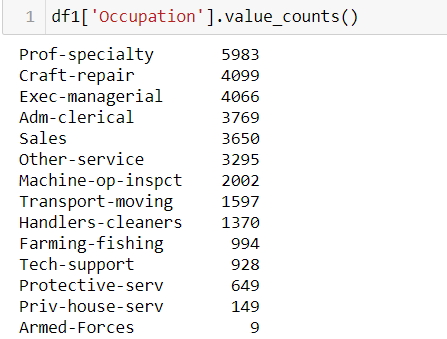
**Features Description**



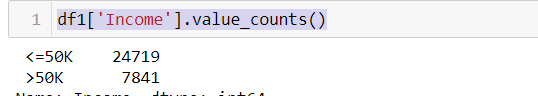
The highest records i.e. 10501 out of 32560 (33%) in the dataset are for those having high school graduation and the least records i.e 51 out of 32560 are for those having Preschool as an education qualification.



The highest records i.e. 14976 out of 32560 (46%) in the dataset are for those who are married and spouse is present.

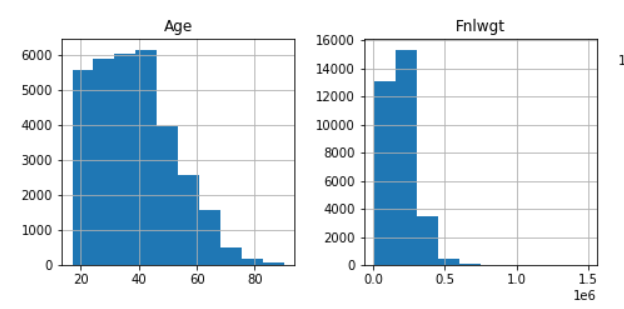


The highest records i.e. 5983 out of 32560(18%) in the dataset are for those who are professional having specialization and the least records i.e. 9 out of 32560 (0.0002%) are for those who are in Armed Forces.



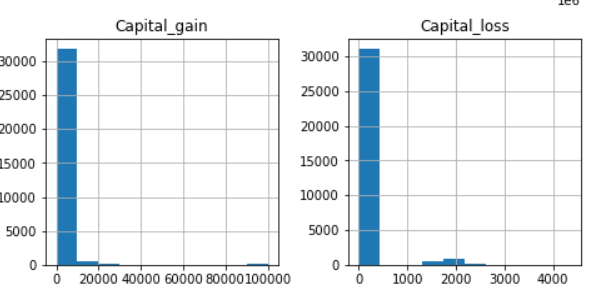
There are 75% of the records where the income lies in range of <=50 K and 25% of the records lies in the range of >50K.

**Understanding Distribution of Numerical Variables(Univariate Analysis):**



The age in the dataset is spreaded in the range of 17 years to 90 years. Further, most of the records in the dataset are having age in the bracket of 17 years to 50 years.

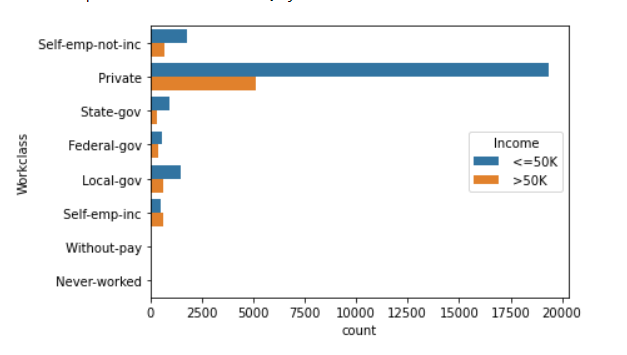
The Final Weight in the dataset is spreaded in the range of 0.0 to 0.8. Further, most of the records in the dataset are having final weight in the range of 0 to 0.3.



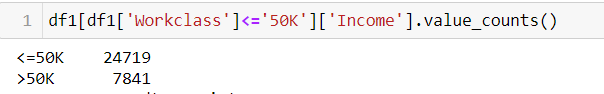
The Capital gain is lying in the range from 0 to 1500 for most of the records available in this dataset. Further, few of the records in this dataset are having Capital gain between 5000 to 23000 and 85000 to 100000 i.e very small blocks can be seen in the above histogram, the chances are they are either outliers or it can be a actual value for capital gain.

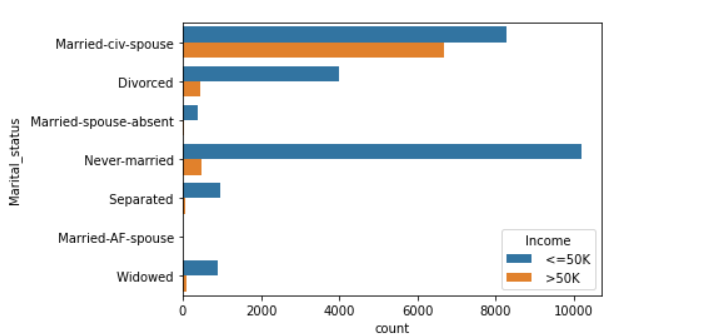
The Capital loss is lying in the range from 0 to 100 for most of the records available in this dataset. Further, few of the records in this dataset are having Capital loss between 1200 to 2300 i.e very small blocks can be seen in the above histogram, the chances are they are either outliers or it can be a actual value for capital loss.

**Bivariate Analysis:**

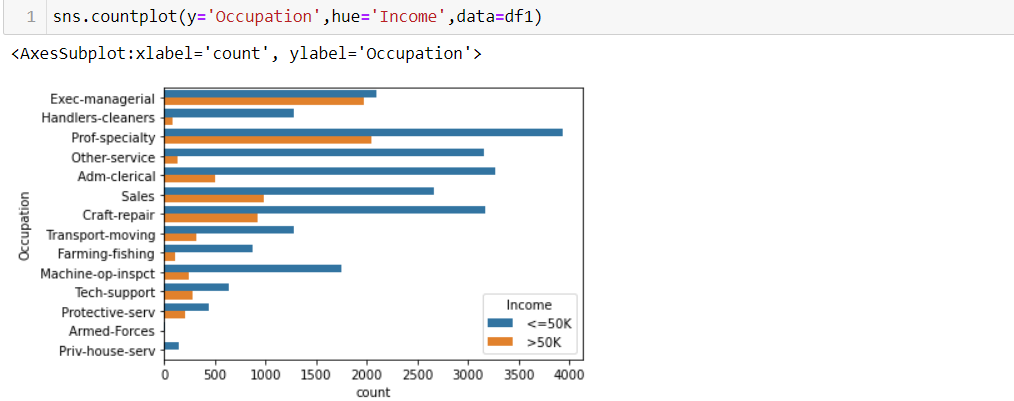


The highest records in the dataset belongs to private work class and most of them i.e 24719 records out of 32560(76%) are having income <=50 K.

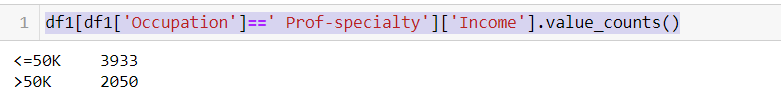


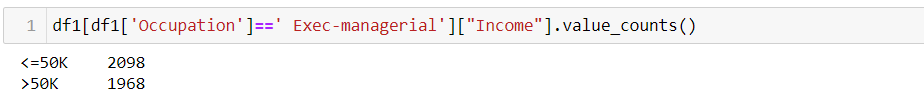


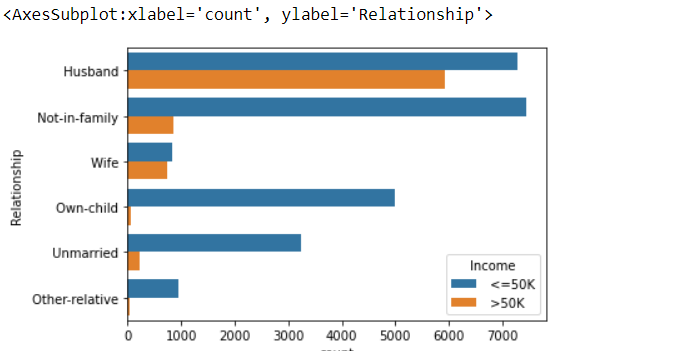
The most of the records in never married category i.e.10191 records out of 10682(95%) are having income <=50K.



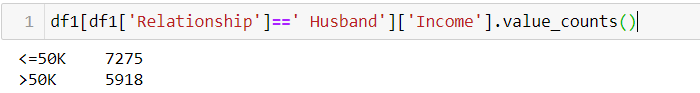
The professional with a specialization is having 3933 records with income <=50K and 2050 records with income >50K. i.e in the ratio of 65%(income<=50 K) and 35% (income>50K). The person employed as Exec-managerial is having 2098 records with income <=50K and 1968 records with income>50K.ie in the ratio of 52 %(income<=50K) and 48% (income >50K) so there are equal chances that a person employed as an Exec-managerial position can have income <=50 K and income >50 K.

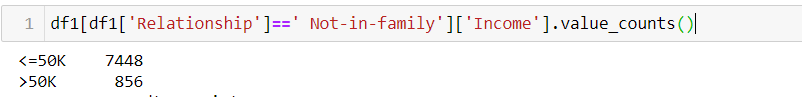


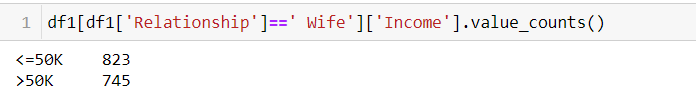


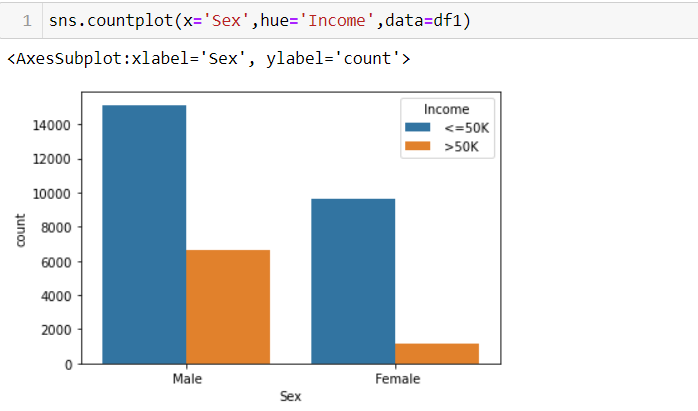


There are 7275 records related to husband having income<=50K and 5918 records having income >50 K. i.e. in the ratio of 55%(income <=50 K) and 45% (income >50K). There are 7448 records related to Not-in-family having income<=50K and 856 records having income>50K. i.e. in the ratio of 90% (income<=50K) and 10% (income>50K). There are 823 records related to wife having income <=50K and 745 records having income>50K. i.e. in the ratio of 52% (income <=50K) and 48% (income>50 K).





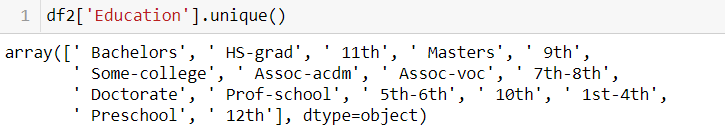


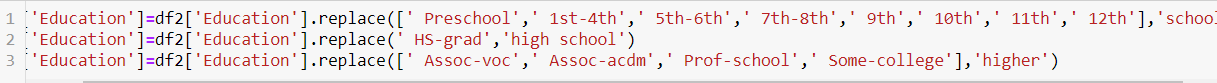


There are 15127 records from male category having income <=50 K and 6662 records having income >50K. i.e in the ratio 69% (income<=50 K) and 31% (income >50 K).

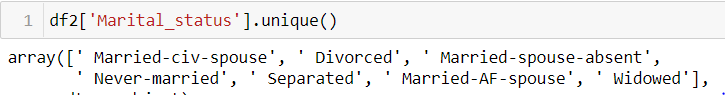
There are 9592 records from female category having income <=50K and 1179 records having income >50 K i.e. in the ratio 89% (income <=50K) and 11%(income >50 K).

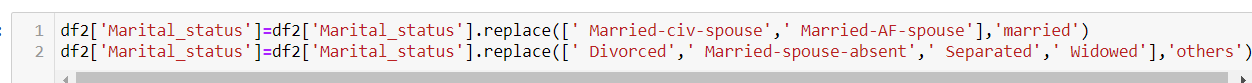
**Feature Engineering**





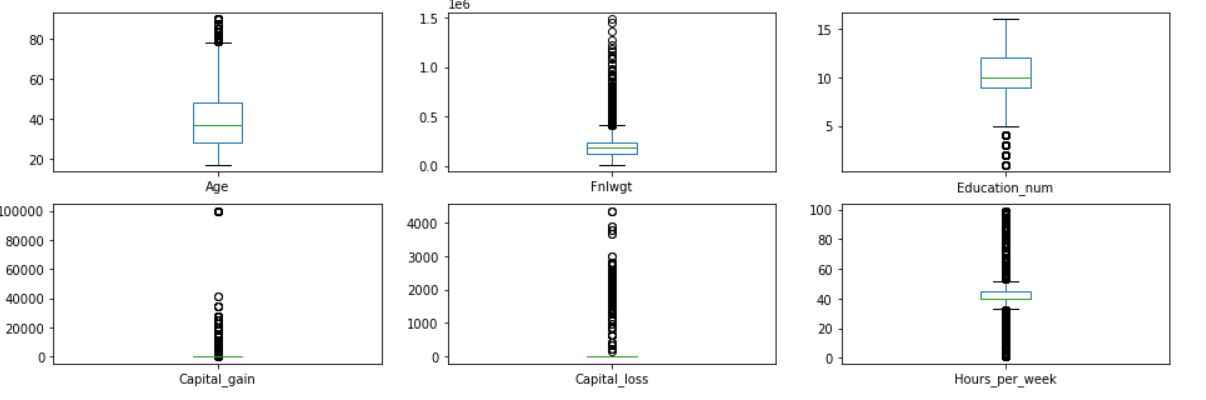
In Education feature I have replaced values ‘Preschool’,’1st-4th,5th-6th ,7th -8th,9th,10th,11th and 12th with a school as all these values are related to different grade while in school , similarly ‘HS-grad’ with ‘high school’ and ' Assoc-voc',' Assoc-acdm',' Prof-school',' Some-college' with higher as they are related to higher level of education.





In Marital status feature I have replaced values ‘Married-civ-spouse’,’Married-AF-spouse’ with married, similarly ‘Divorced’ ,’Married-spouse-absent’, ’Separated’, ’widowed’ with others.

**Outlier Detection**

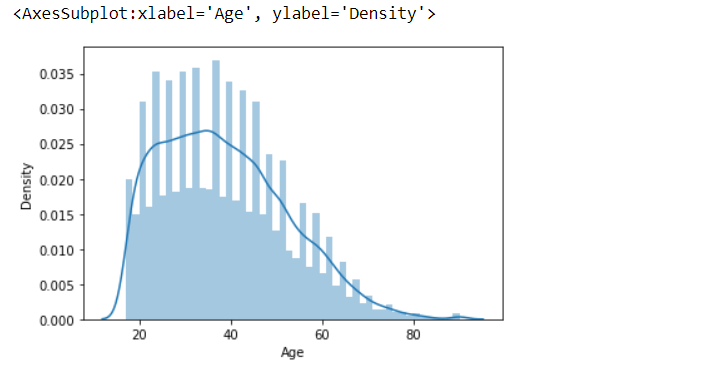


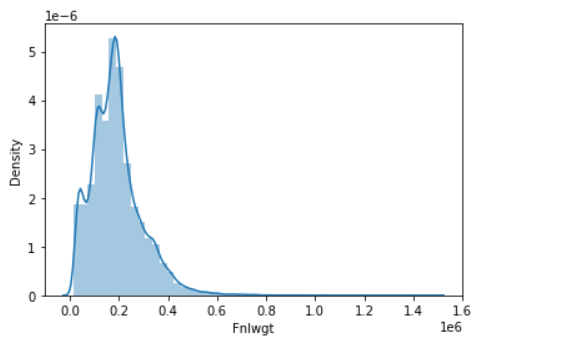
There are outliers present in features named as Age, Fnlwgt, Education num, Capital gain, Capital loss and hours per week as per the boxplot attached above.

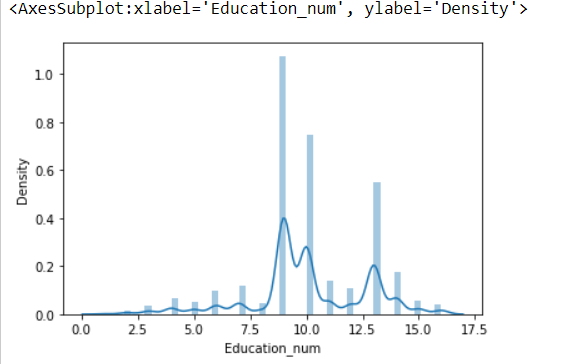
I am not treating any of the outliers, as the case may be that the values in Age, Fnlwgt,capital gain, capital loss features can be higher than the upper whiskers.

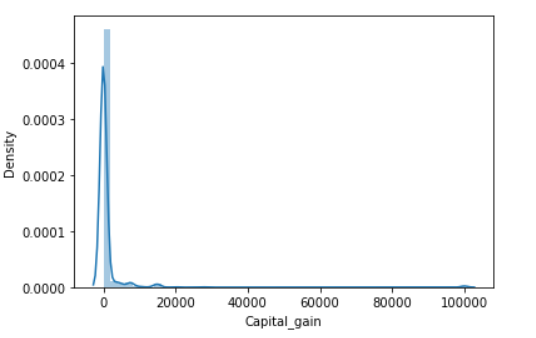
Hours per week can be higher or lower than upper/lower whisker depending on the person occupation, income, education and position.

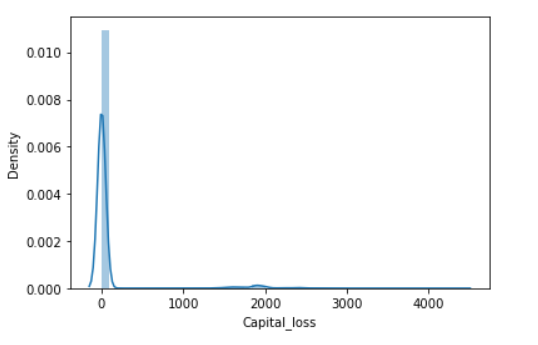
**Skewness**

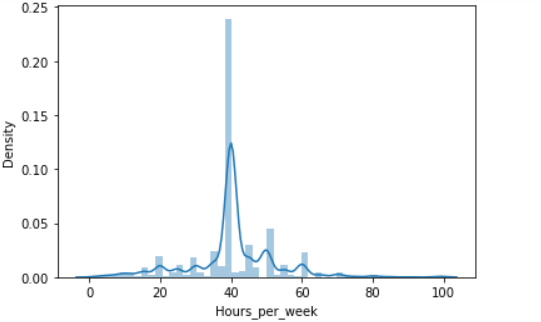


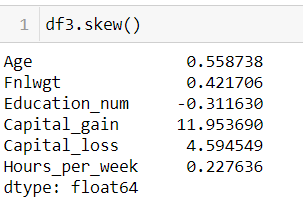






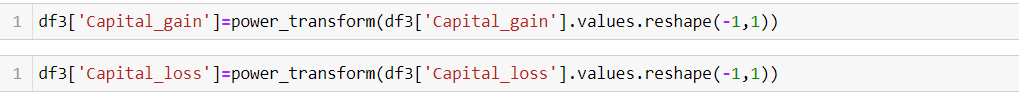






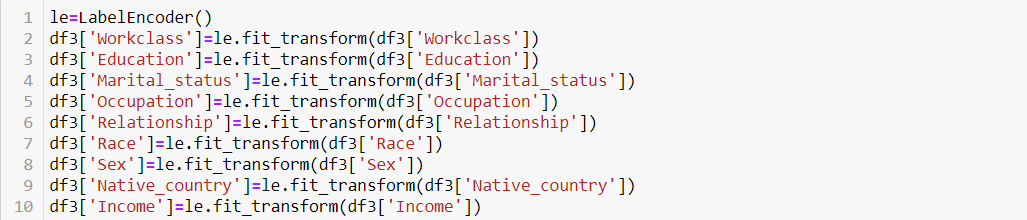
As the skewness in the features Capital gain and Capital loss is quite high which has to be treated**.**

There are different methods like log transformation, power transformation, square root transformation, cube root transformation etc.to treat skewness.



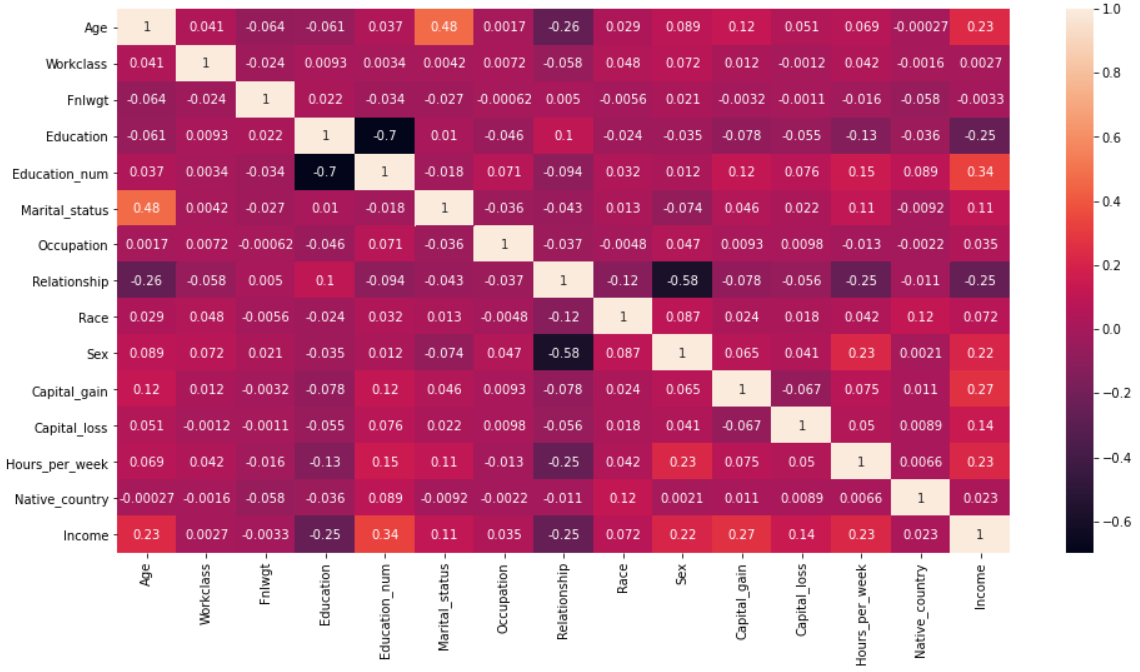
**Label Encoding**

The Label Encoding is applied on the categorical variables as the machine learning models only understand the numerical values i.e integer and float values.



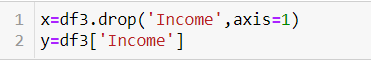
The Label Encoding is applied on above mentioned features.

**Correlation**



There is a positive correlation of the income with the age, Education\_num, Capital gain and hours per week.

**Model building**



The data is segregated into x which will have all independent features and y with a dependent feature i.e in other words it is the target variable.

Before splitting the data into train\_test\_split standard scaler is applied to independent variables(input variables). The idea behind Standard Scaler is that it will transform the data such that its distribution will have a mean value 0 and standard deviation of 1 so that the machine learning models will not be biased to any particular values.



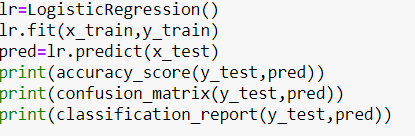
In Train Test Split validation we split the training data into training and validating data as to evaluate our model.

Splitting the data into test\_train\_split as per code attached below.



**Logistic Regression:**

The model is build using Logistic Regression as one of the supervised machine learning algorithms.

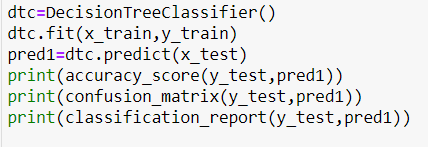




**The accuracy score is 0.813 for this model**.

**Decision Tree Classifier:**

The model is build using Decision Tree Classifier as one of the supervised machine learning algorithms.

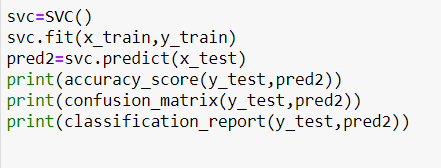




**The accuracy score is 0.801 for this model.**

**Support Vector Machine (SVC):**

The model is build using SVC as one of the supervised machine learning algorithms.

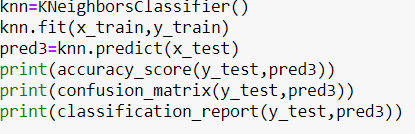




**The accuracy score is 0.834 for this model.**

**K Neighbours Classifier:**

The model is build using K Neighbours Classifier as one of the supervised machine learning algorithms.

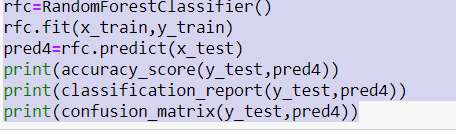




**The accuracy score is 0.819 for this model.**

**Random Forest Classifier:**

The model is build using Random Forest Classifier as one of the supervised machine learning algorithms.

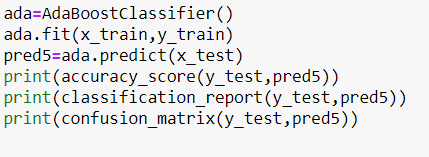




**The accuracy score is 0.850 for this model.**

**AdaBoost Classifier:**

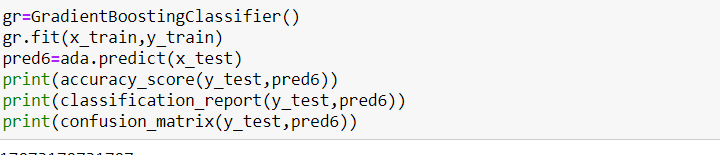
The model is build using AdaBoost Classifier as one of the supervised machine learning algorithms





**The accuracy score is 0.857 for this model.**

**Gradient Boosting Classifier:**

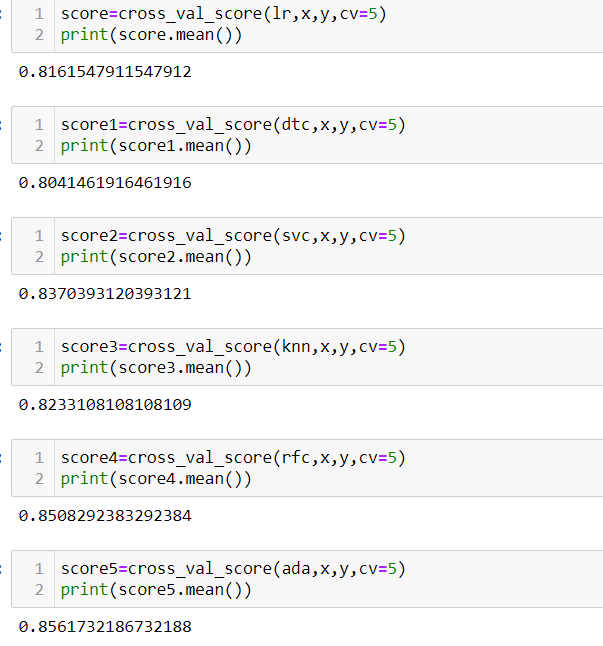


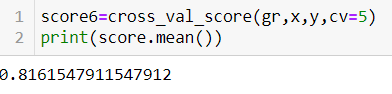


**The accuracy score is 0.862 for this model.**

**Cross validation score:**

The cross validation score for logistic Regression, Decision tree classifier, SVC, K Neighbours Classifier, Random Forest Classifier, Ada boost Classifier and Gradient Boosting Classifier is **0.816,0.804,0.837,0.823,0.850,0.856,0.816.**



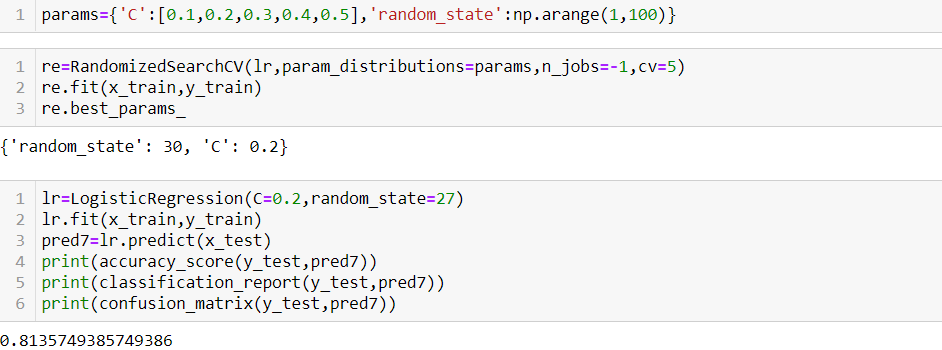


**Observations:**

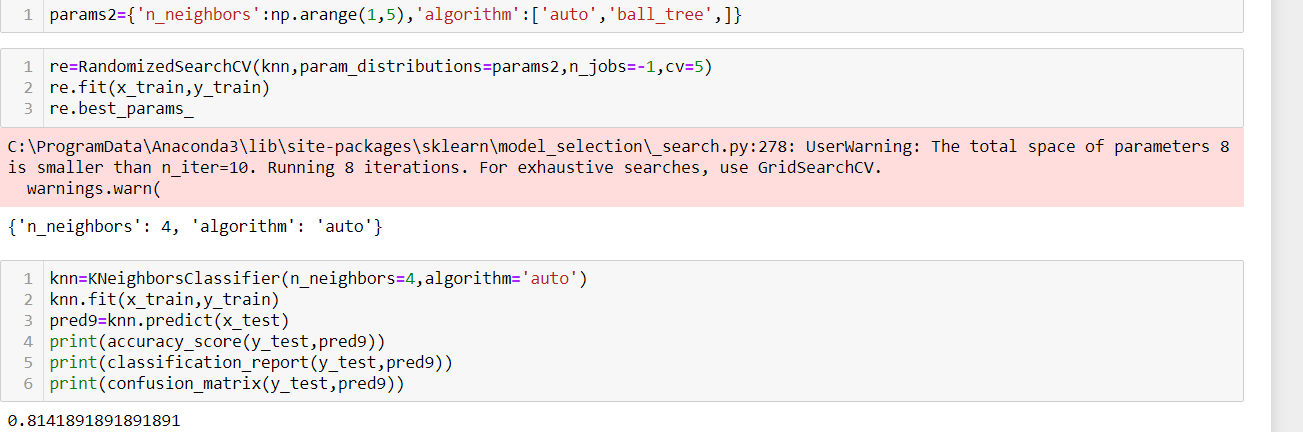
The difference between cross validation score and the accuracy score is minimum for Logistic Regression, Decision Tree Classifier, svc, K Neighbours Classifier Random Forest Classifier and ada Boost Classifier. To find the best model out of all these hyper tuning is required**.**

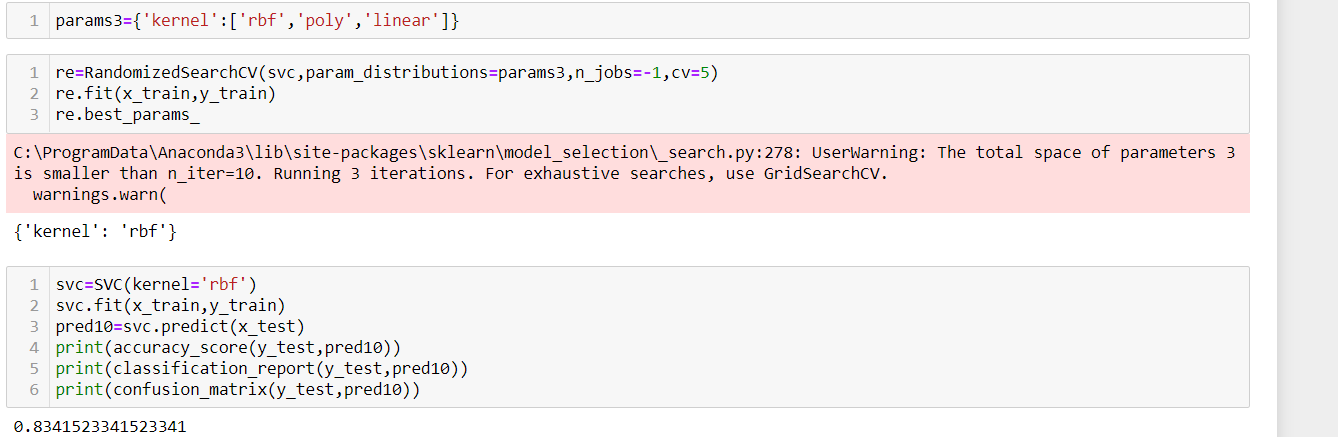
**Hyper Tuning:**

After performing hyper tuning on Logistic Regression, Decision Tree Classifier, K Neighbours Classifier, SVC, Random Forest Classifier and Ada boost Classifier the accuracy score is 0.813,0.839,0.814, 0.834,0.850 and 0.857

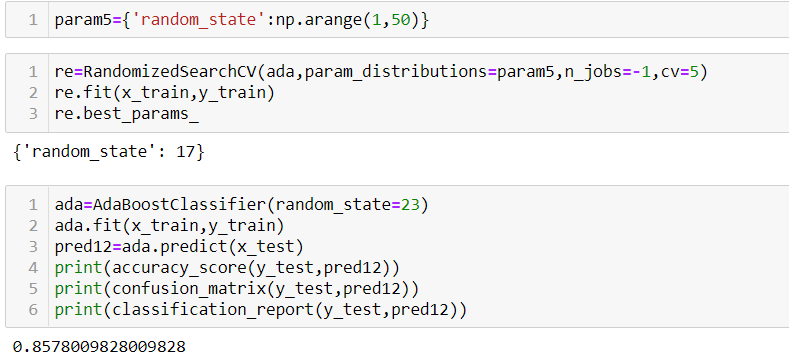








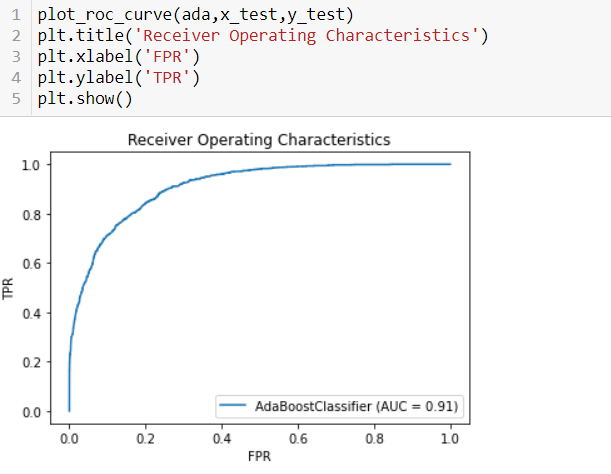




**Conclusion**

Out of all the classification algorithms used on the dataset, Ada Boost Classifier algorithm gives the best overall prediction accuracy i.e. accuracy score of 86%.

The AUC score for Ada Boost Classifier is 0.91



In near future this module of prediction can be integrated with module of automated processing system**.**