***ADULT INCOME ANALYSIS USING MACHINE LEARNING TECHNIQUES***

**PROBLEM STATEMENT:**

Explore the given data set with EDA techniques and build a suitable model for predicting whether the salary of the person is >50k or not and visualize the results.

**SOLUTION:**

The objective of the problem was to investigate the income to which group of people they must earn more. Our goal was to compare >50k based on age,workclass, education,marital status,occupation ,etc. The dataset includes an target/label variable(target) ,so it comes under supervised learning. Data preprocessing(EDA) , it includes importing data, Checking for missing values, splitting data ,feature scaling , encoding categorical data to numeric data and data visualization. Model Selection: Since, it is a classification problem. Logistic regression and Decision tree classifier are used to compare their results. Reasons for the model:

* **Decision tree Classifier**: Tree based methods are considered to be highly accurate. It is efficient for both linear and non-linear data.It can handle both categorical and continuous attributes. Decision tree uses multiple techniques for selecting the attributes at different levels. The most commonly used techniques are Gini Index, Information Gain etc.
* **Logistic regression:** It can handle both the dense and sparse input.

**PACKAGES USED:**

* NumPy
* Pandas
* Matplotlib
* Scikit – learn

**ASSUMPTIONS:**

**Hypothesis H1:** People of higher age group are earning more than lower age group.

**Hypothesis H0:**  The complement of H1.

**DATA PREPROCESSING:**

**Importing dataset:**

train = pd.read\_csv('train.csv', names=columns)

test = pd.read\_csv('test.csv', names=columns, skiprows=1)

Data is imported using pandas.

**INFO:**

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 age 32562 non-null object

1 workclass 30726 non-null object

2 fnlwgt 32562 non-null object

3 education 32562 non-null object

4 education.num 32562 non-null object

5 marital.status 32562 non-null object

6 occupation 30719 non-null object

7 relationship 32562 non-null object

8 race 32562 non-null object

9 sex 32562 non-null object

10 capital.gain 32562 non-null object

11 capital.loss 32562 non-null object

12 hours.per.week 32562 non-null object

13 native.country 31979 non-null object

14 target 32562 non-null object

dtypes: object(15)

**Shaping and reshaping data:**

print ("The train data has",train.shape)

print ("The test data has",test.shape)

The train data has (32562, 15)

The test data has (16281, 15)

**Null values:**

Checking for null values in the data.

ge 0

workclass 1836

fnlwgt 0

education 0

education.num 0

marital.status 0

occupation 1843

relationship 0

race 0

sex 0

capital.gain 0

capital.loss 0

hours.per.week 0

native.country 583

target 0

dtype: int64

The above output contains null values. So fill the null values using fillna()

After filling null values:

age 0

workclass 0

fnlwgt 0

education 0

education.num 0

marital.status 0

occupation 0

relationship 0

race 0

sex 0

capital.gain 0

capital.loss 0

hours.per.week 0

native.country 0

target 0

dtype: int64

**Count of target variable:**

<=50K 0.759167

>50K 0.240802

**Data visualization:**

To further analyse, let's first plot a histogram for the 'Age' variable.

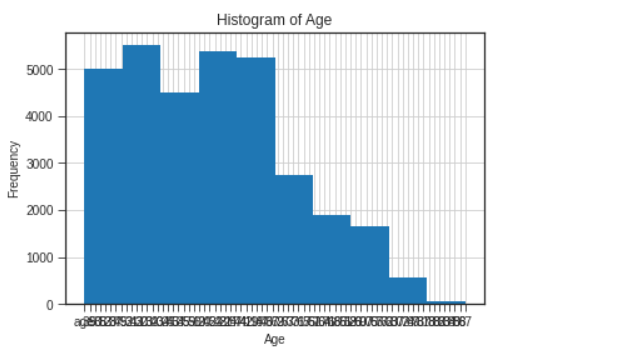
train.age.hist()

plt.title('Histogram of Age')

plt.xlabel('Age')

plt.ylabel('Frequency')

plt.savefig('hist\_age')

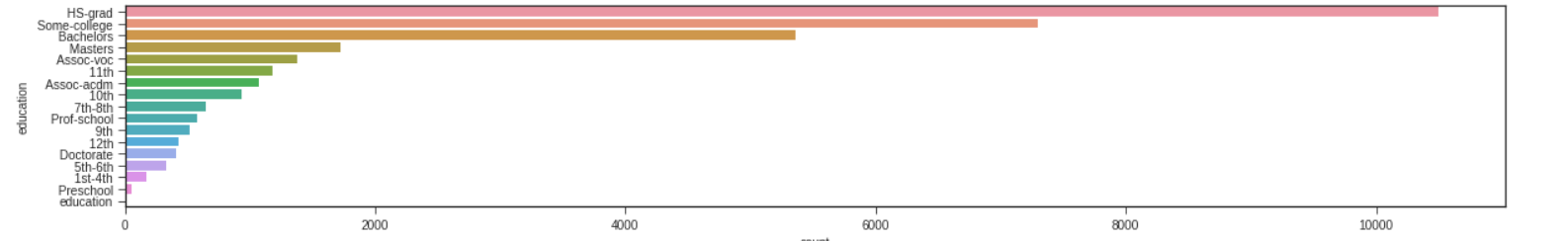
****

**The above graph that the variable 'Age' has a binomial distribution of data.**

**Categories of Education :**

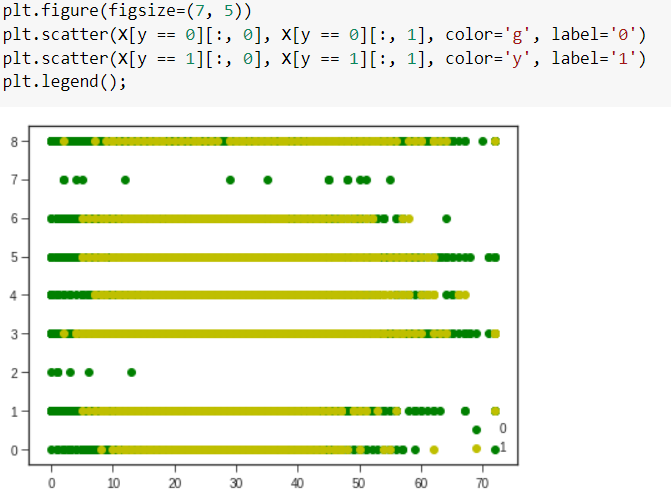
pd.crosstab(train.education, train.target,margins=True)/train.shape[0]

****

****

The above graph shows that the HS\_grad is more than other education.

**Plotting the data:**



**MODEL PREDICTION:**

**Model 1** : Logistic regression

**OUTPUT**

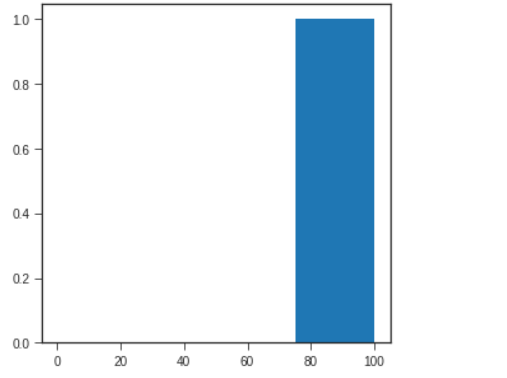
precision:0.7591131038294997

recall:0.759136416682022

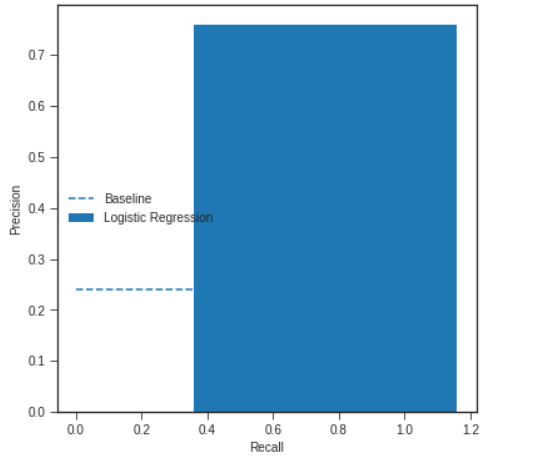
f1:0.7591247600767753

Accuracy: 75.91057060377126

**VISUALIZATION OF RESULT:**

****

The above graph shows the accuracy of the model.

****

The above graph is comparison between the recall scores and precision cores.

**Model 2:** Decision tree classifier

**OUTPUT:**

Accuracy: 80.80663322755656

classification report

precision recall f1-score support

0 0.88 0.87 0.87 7373

1 0.61 0.62 0.61 2396

accuracy 0.81 9769

macro avg 0.74 0.75 0.74 9769

weighted avg 0.81 0.81 0.81 9769

**CONCLUSION:**

From model 1, the accuracy score of logistic regression is 76% ,which was not good. From model 2, the accuracy of the Decision tree classifier is 81% accuracy is good.

The obtained results showed the use value of both machine learning models. Decision tree classifier model showed slightly better performance than the logistic regression model, but definitely, both models have shown that they can be very successful in solving classification problems.

*Finally, The assumption(hypothesis h1)* becomes *true.* *By exploring this particular dataset that*, higher education is no guarantee to high income. *People of higher age group are earning more than lower age group.*