LOVELY
PROFESSIONAL
UNIVERSITY

Name – Ashutosh Kumar

Regd_no – 11804217

Roll_no – 03

Section – KM007

Dat-set – Adult

Lovely Professional University
Phagwara, India

Introduction:-

Adult Census Income is a categorical data-set in which there are details about a particular person like his or her gender, occupation, race, country and many more also, according to these details there is target whether the income of the person is less than , equal or more than 50K.

Now, on this data-set I have trained multiple features and tested on the target. Also I have used three different models(algorithms) to train and predict the output.

Problem statement:
I have been provided with an Excel dataset that has 15 columns and 48842 rows. My task is to analyse the dataset and predict whether the income of an adult will exceed 50k per year or not by developing a supervised machine learning model.

Cleaning and feature selection of Data-set

Firstly I started with renaming the columns with proper names after that I checked for null values in the data-set. Since there is no null values, I checked for special characters and I found that in some of the columns. After that I assigned NaN to all the special characters and later on I dropped them from the datset.

Now it's time to select the features to train and test. I implemented a loop to get all the unique values of all the columns. I saw there columns with multiple(2000+) unique values so, as it is categorical data- set it's become complicated and tuff to assign numeric values to all of them. I dropped some columns which were having noisy data and unique values more than 15.

Assigning Numeric Values

First I determined the target which is income and that is in categorical form so I assigned bool values in that. If income is less than or equal to 50k it is 0 or, income is greater than 50k it is 1.

As my features are also in categorical form , I need to assign numeric values to that so, I performed label encoding and assigned numeric values 1 to 15 to each feature.

	age	workclass	fnlwgt	education	education.num	marital.status	occupation	relationship	race	sex	ca
0	90	?	77053	HS-grad	9	Widowed	?	Not-in-family	White	Female	0
1	82	Private	132870	HS-grad	9	Widowed	Exec-managerial	Not-in-family	White	Female	0
2	66	?	186061	Some-college	10	Widowed	?	Unmarried	Black	Female	0
3	54	Private	140359	7th-8th	4	Divorced	Machine-op-inspct	Unmarried	White	Female	0
4	41	Private	264663	Some-college	10	Separated	Prof-specialty	Own-child	White	Female	0

Data Initially

	workclass	education	marital	occupation	relationship	race	gender	income
1	Private	HS-grad	Widowed	Exec-managerial	Not-in-family	White	Female	0
3	Private	7th-8th	Divorced	Machine-op-inspct	Unmarried	White	Female	0
4	Private	Some-college	Separated	Prof-specialty	Own-child	White	Female	0
5	Private	HS-grad	Divorced	Other-service	Unmarried	White	Female	0
6	Private	10th	Separated	Adm-clerical	Unmarried	White	Male	0
...	...	...	...	...	...	...	...	...
32556	Private	Some-college	Never-married	Protective-serv	Not-in-family	White	Male	0
32557	Private	Assoc-acdm	Married-civ-spouse	Tech-support	Wife	White	Female	0
32558	Private	HS-grad	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	1
32559	Private	HS-grad	Widowed	Adm-clerical	Unmarried	White	Female	0
32560	Private	HS-grad	Never-married	Adm-clerical	Own-child	White	Male	0

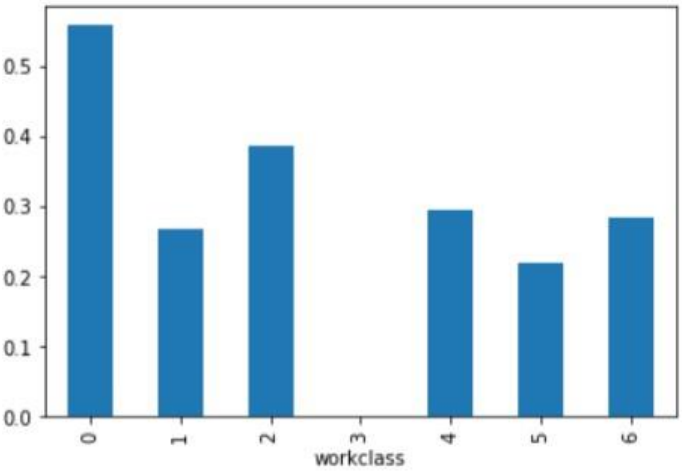
Data set after feature selection and dropping noisy data

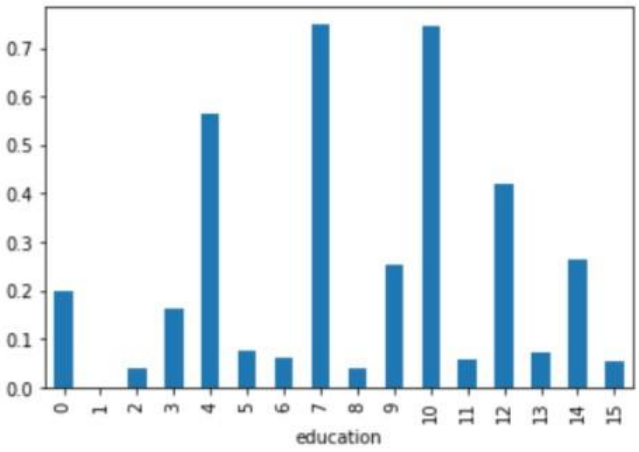
	workclass	education	marital	occupation	relationship	race	gender	income
1	5	3	1	7	0	3	1	0
3	5	6	4	6	3	3	1	0
4	5	0	3	5	5	3	1	0
5	5	3	4	13	3	3	1	0
6	5	13	3	3	3	3	0	0

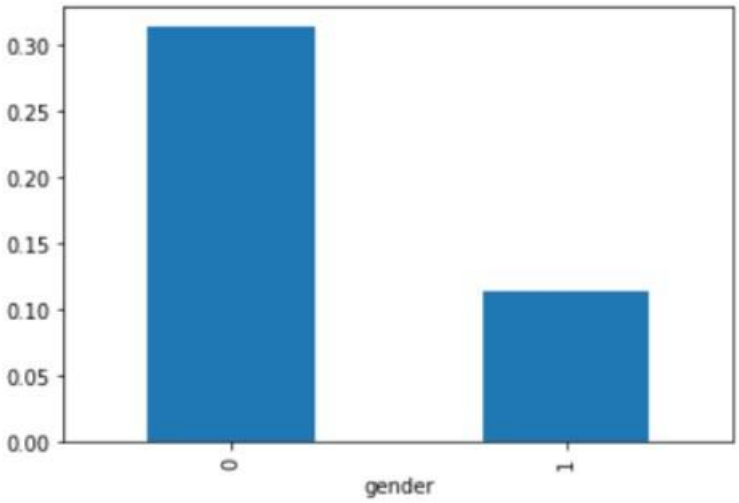
Final Data-set after performing label encoding.

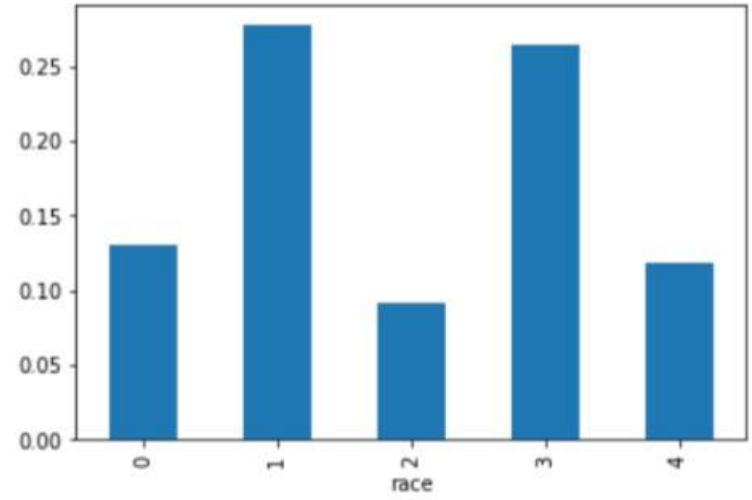
Data Visualization

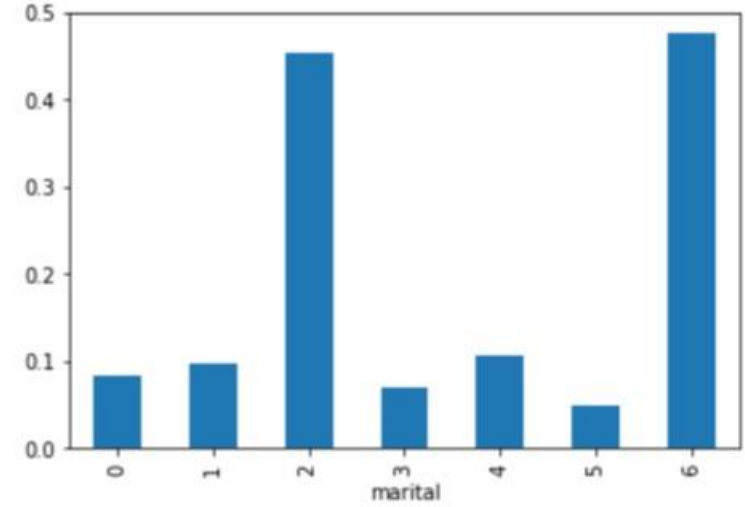
Here I plotted bar graph for each feature WRT target column.

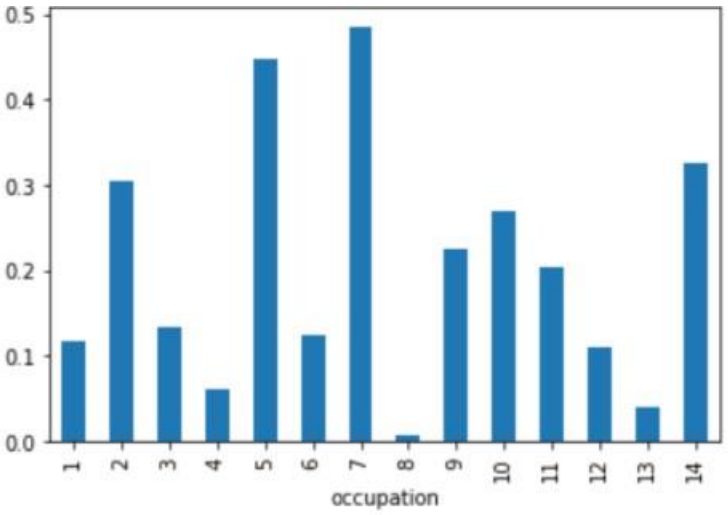


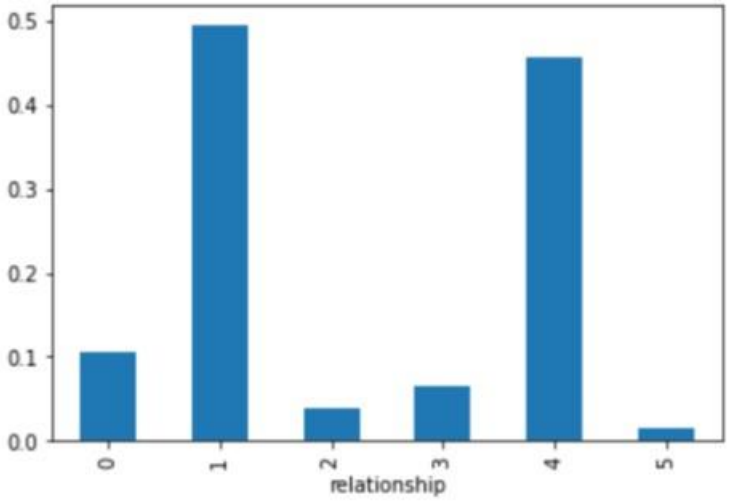



0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
1
gender


0.25
0.20
0.15
0.10
0.05
0.00
0
1
2
3
4
race


0.5
0.4
0.3
0.2
0.1
0.0
0
1
2
3
4
5
6
marital





By these visualizations, we can deduce some correlations between our independent attributes and dependent attribute:

1. Adults with an educational background of Prof-school (7) and Doctorate (10) will have a better income and it is likely possible that their income is higher than 50K.

2. Our data suggest that people with occupation Prof-specialty (5) and Exec-managerial (7) will have a better chance of earning an income of more than 50K.

3. The gender bar chart provides us some useful insight into the data that Men (0) are more likely to have a higher income.

4. relationship chart shows us that wife (1) and husband (4) has a higher income. A married couple would most likely earn >50K.

5. As per the data, an Asian-Pac-Islander (1) or a white (3) have more chances of earning more than 50K.

6. Self-emp-in (0), Federal-gov(2) workclass group have a higher chance of earning more than 50K.

Splitting Data and Model selection

After identifying feature that to be trained and tested I have splitted the data in the manner that 77% to be trained and 33% to be tested.

Model Selection

After splitting the data I have to select the model. I have used three different model. And those are:-

1. Random Forest Classifier
2. Logistic Regression
3. KNeighbors Classifier

HyperParameter Tunning

Performed Hyper Parameter Tunning for each of the three model so that I could get best parameter for the model.

In that I have used Greed Search technique

Confusion Matrix

Deployed confusion matrix for all the models and through that calculated the accuracy.

1. Random Forest Classifier

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an esemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction

In this model I have used parameters**[criterion = gini, max_depth = 10, max_features = auto, n_estimators = 60]**

I got the accuracy above 82%. In this model.

2. Logistic Regression

Logistic Regression is one of the easiest and most commonly used supervised Machine learning algorithms for categorical classification. The basic fundamental concepts of Logistic Regression are easy to understand and can be used as a baseline algorithm for any binary (0 or 1) classification problem.

It is a Statistical predicting model that can predict either a 'Yes'(1) or 'No'(0).

In this model I have use parameters**[C = 0.01, fit_intercept = False, multi-class = multinominal, penalty = 12, verbose = 0]**

I got the accuracy above 75%. In this model.

3. KNeighbors Classifier

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems.

The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other.

The KNN algorithm hinges on this assumption being true enough for the algorithm to be useful. KNN captures the idea of similarity (sometimes called distance, proximity, or closeness) with some mathematics we might have learned in our childhood— calculating the distance between points on a graph.

In this model I have used parameters**[leaf_size = 45, n_neighbors = 5, weights = uniform]**

I got the accuracy above 81%. In this model.

Result

After Traning and testing the data-set on three different models we got accuracy 82%, 75% and 81% respectively. Hence, our first model Random forest classifier is best for our data-set.

Prediction on Unknown data

After Doing all the things from scratch , I have tested My best model which is Random Forest Classifier on unknown data. In this I will take data from user, for better GUI I have created tkinter window.

Conclusion

In final words, firstly I fetch the data-set cleaned it and made that perfect, later on trained on three different models and got 82%, 75% and 81% accuracy respectively. Teste the best model on unknow data-set which I took as input on tkinter GUI.