Introduction

The dataset contains information about US citizens census data such as their occupation, education, gender, race, salary, etc. There are 16 attributes and 48842 rows of information in the dataset. Various python libraries will be used to perform the analysis in this project. The objective of this project is to build a model and see how precisely we can detect weather a US citizen is a low income or high-income citizen based on their attributes. I'll start with the data cleaning steps followed by some visualizations to understand the relationship between these attributes in the exploratory data analysis step. Finally using k nearest neighbor classifier algorithm I'll try to predict if the citizen is low- or high-income type, this model will help us comprehend the characteristics that contribute to wealth and how to enhance American policies.

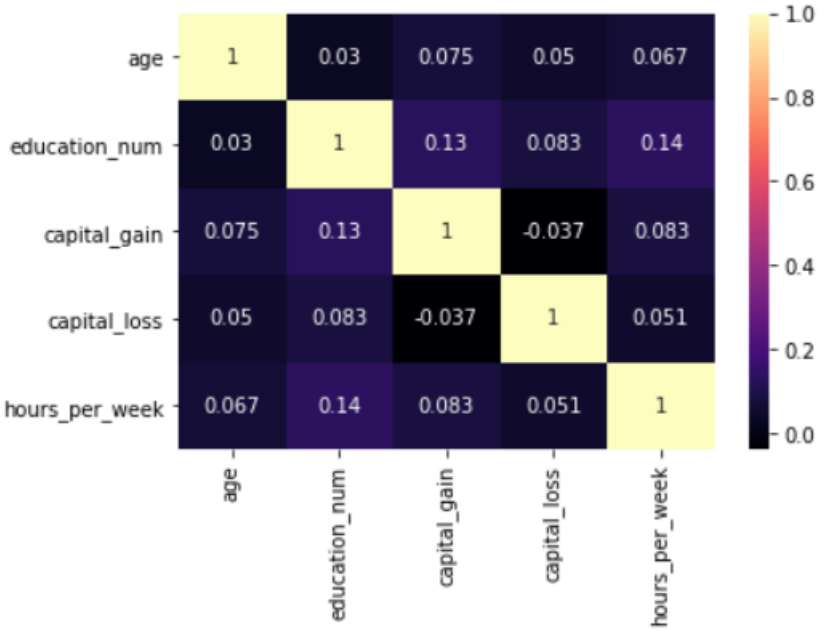
Data cleaning

The first step of the project is importing the dataset. As the csv data file does not contain the headers, I've created a headers list that contains the names of all the columns and passing this list as the column name in read csv function loads the data in python data frame format. Now next step is changing the column names into snake case format i.e., lower case with an underscore in place of spaces. The column "fnlwgt" is of not much significance so I've dropped this column. Next, after conducting further investigations I noticed some special characters in the dataset, so I've replaced these with null values in the data set and further I've dropped the null values from the dataset as the percentage of these null values is very low. Now, check for duplicates, if there is any, keep the first duplicate value and drop rest from the dataset. I also observed some values in the capital gain column, which could look like outliers such as '99999' but since the number of entries for those is higher than 5%, I've kept them in the dataset.

Exploratory data Analysis

There are a total of 39240 rows and 14 columns after completing the data cleaning steps. There are 7 different work class citizens in the census data record for US citizens, out of which private work class accounts for more than 70% of the citizens while 0.1% population is without pay. The number of citizens working increases as the age increases as the age increaases from 17 to 35 and drops down linearly as the age goes above 35 as shown in histogram in appendix. Prof-specialty, Exec-managerial and Adm-clerical is the most common occupations in US as per the census data while fish farming is the least likely occupation. In terms of race, White accounts for almost 85% of the population followed by black at 10%.

The dataset has both continuous and numerical variables data. For numerical attributes a correlation plot has been created, that shows the relation of each numerical variables to the other in the data set as displayed below.



As we can see clearly from the plot that hours per week and education have a direct significant relation with each other.

In every work class males are working more hours per week than females as displayed in the bar graph in appendix. On further exploring the capital gain and capital loss attributes, I noticed that the capital gain is significantly higher for self emp -inc citizens at 5000 followed by self emp non-inc at 2000. From capital loss bar graph, we can interpret that the higher the salary greater the loss and vice versa. Low-income people are more likely to work in the private sector.

KNN Models

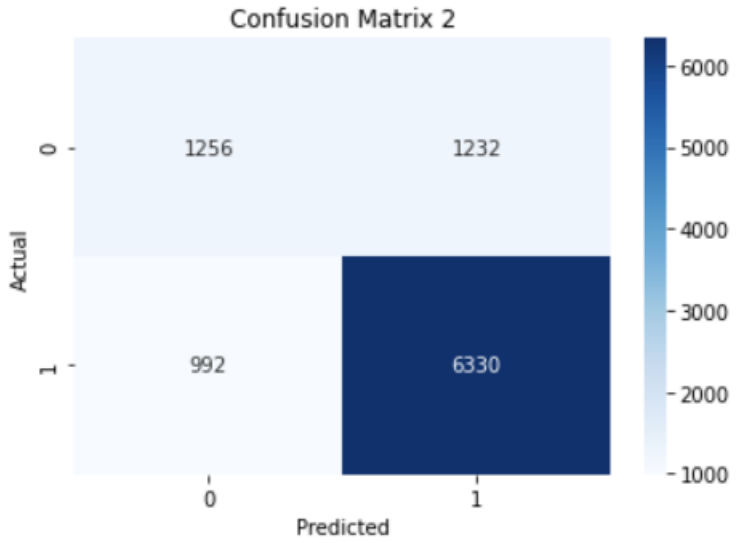
Now we we've got the data clean and ready for further model processing. Since the KNN algorithm does not work with the categorical data, we'll use encoding to provide numerical values for the categorical variables data. As education variable is already encoded as education_num in the dataset, I've dropped this education column. Next, replacing rest of the categorical data with numerical data using dummies functionality in python. This way we'll create n-1 columns having binary values i.e., 0,1 for each unique entry n in the categorical variables column. Now again we'll encode the salary column also as using label encoder from sklearn package. Now we'll be preparing three different models for our analysis using a combination of three different variables in out KNN model.

The dependent variable for our KNN model is salary as we're trying to predict weather the citizen will be a low income or high income citizen. The variables of interest i.e., independent variables in the first KNN model are age, capital gain, education num. Next step is to split the dataset into train and test data. For analysis in this report, I've divided the data in to 75:25 ratio for each model.

Now we'll be fitting the KNN model using K neighbors classifier and taking number of neighbors as 4 so that the algorithm checks the four nearest values to from the data point of interest. Next, we'll predict the y values for this model and check the accuracy of our model. The accuracy of model one is 76%. Now I'll create two more KNN model by changing the independent variables and compare them all later by their accuracy score.

For second model the variables of interest are i.e., independent variables in the first KNN model are hours per week, capital gain and education num. For this model the accuracy is 77%.

For the third model the variables of interest are i.e., independent variables in the first KNN model are age, capital loss and education num. The accuracy is 73% for this model. As the accuracy of model 2 is best among all three it is the best model from the analysis in this report.



Now further checking the confusion matric as shown above for model 2 it is quite evident that the model 2 is better than the rest two as it has least number of false positive and false negative values out of all three. As the accuracy is more than 77% we can say that we can correctly classify citizens based on income using KNN models.

Reference

1. Panda > Statsmodel: syntax errors implementing variance_inflation_factor. (n.d.). Stack Overflow. Retrieved February 26, 2023, from https://stackoverflow.com/questions/37124342/panda-statsmodel-syntax-errors-implementing-variance-inflation-factor
2. Panda > Statsmodel: syntax errors implementing variance_inflation_factor. (n.d.). Stack Overflow. Retrieved February 26, 2023, from https://stackoverflow.com/questions/37124342/panda-statsmodel-syntax-errors-implementing-variance-inflation-factor
3. random_state in Machine Learning | Data Science and Machine Learning. (n.d.). Kaggle. https://www.kaggle.com/questions-and-answers/49890

Appendix

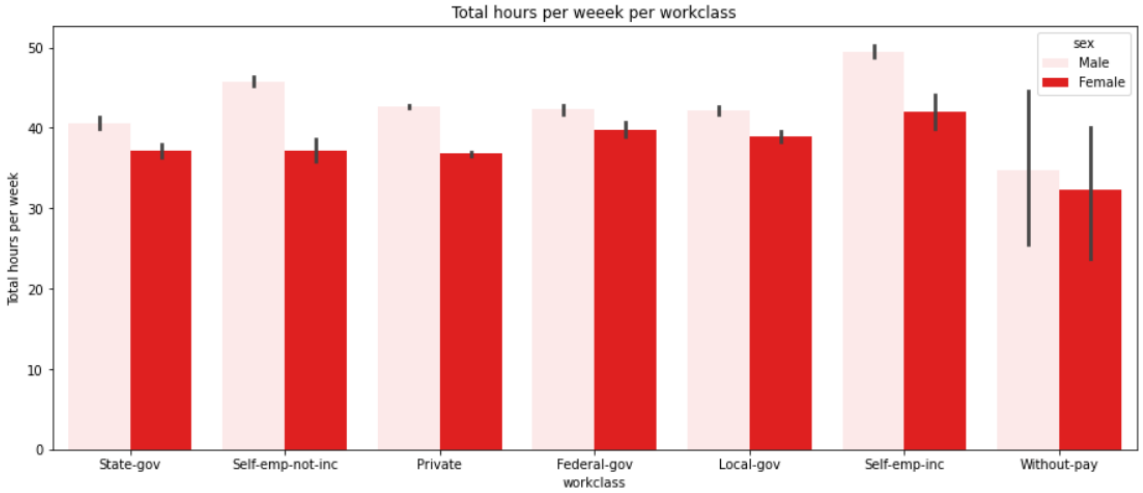
bam.glimpse(data)

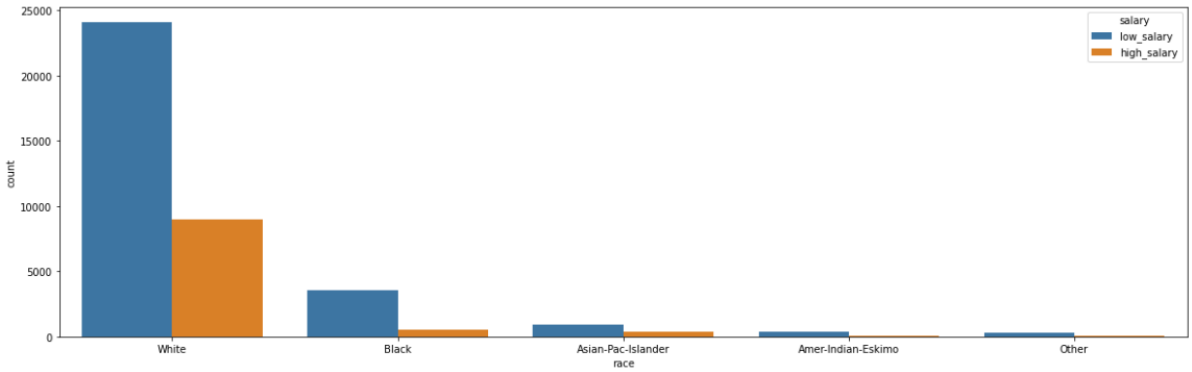
39,240 rows and 16 columns

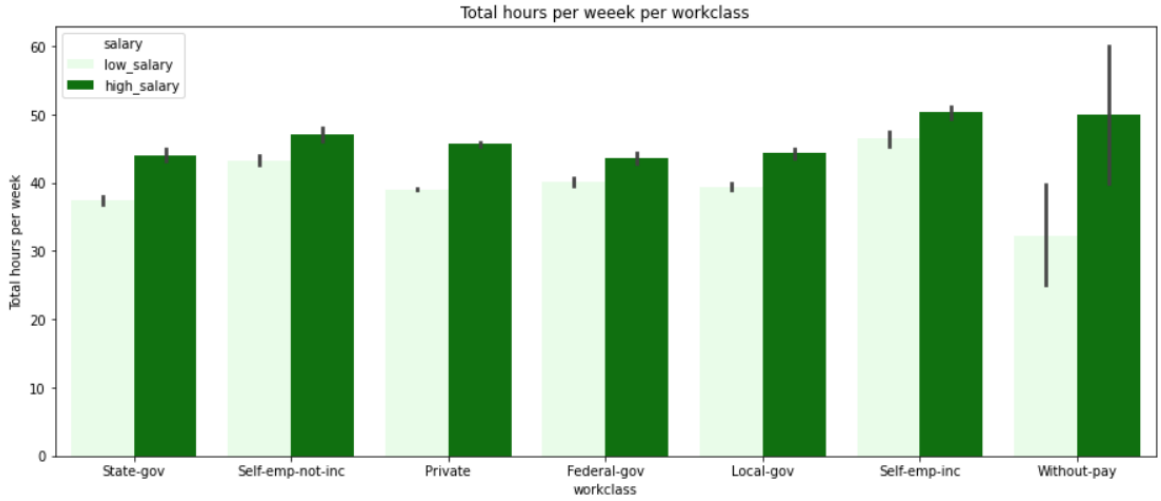
Show [All columns] Update

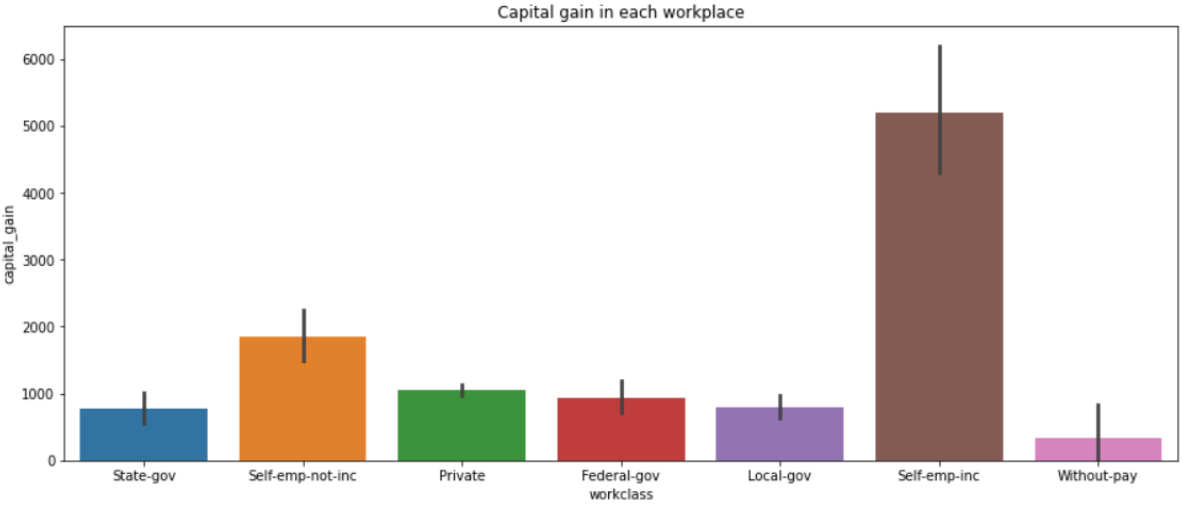
Column	Data type	Unique values	Missing values
age	int64	74	0
workclass	object	7	0
education	object	16	0
education_num	int64	16	0
marital_status	object	7	0
occupation	object	14	0
relationship	object	6	0
race	object	5	0
sex	object	2	0
capital_gain	int64	121	0
capital_loss	int64	97	0
hours_per_week	int64	96	0
native_country	object	41	0
salary	object	2	0
low_income	object	2	0
high_income	object	2	0

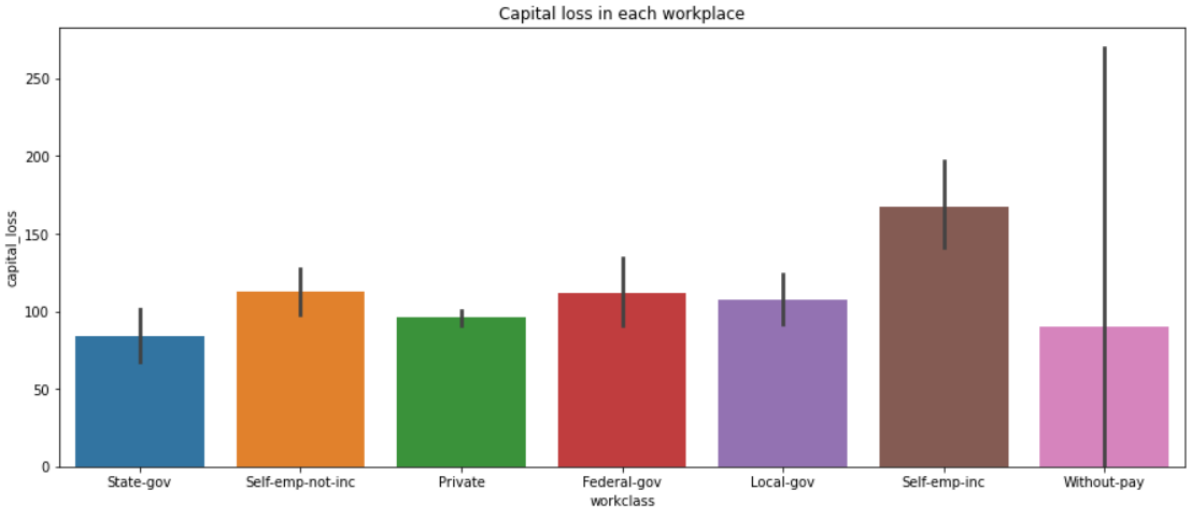
All value counts of column	Count	%	Cum. count	Cum. %
Private	27717	70.6	27717	70.6
Self-emp-not-inc	3669	9.4	31386	80.0
Local-gov	2975	7.6	34361	87.6
State-gov	1892	4.8	36253	92.4
Self-emp-inc	1595	4.1	37848	96.5
Federal-gov	1371	3.5	39219	99.9
Without-pay	21	0.1	39240	100.0

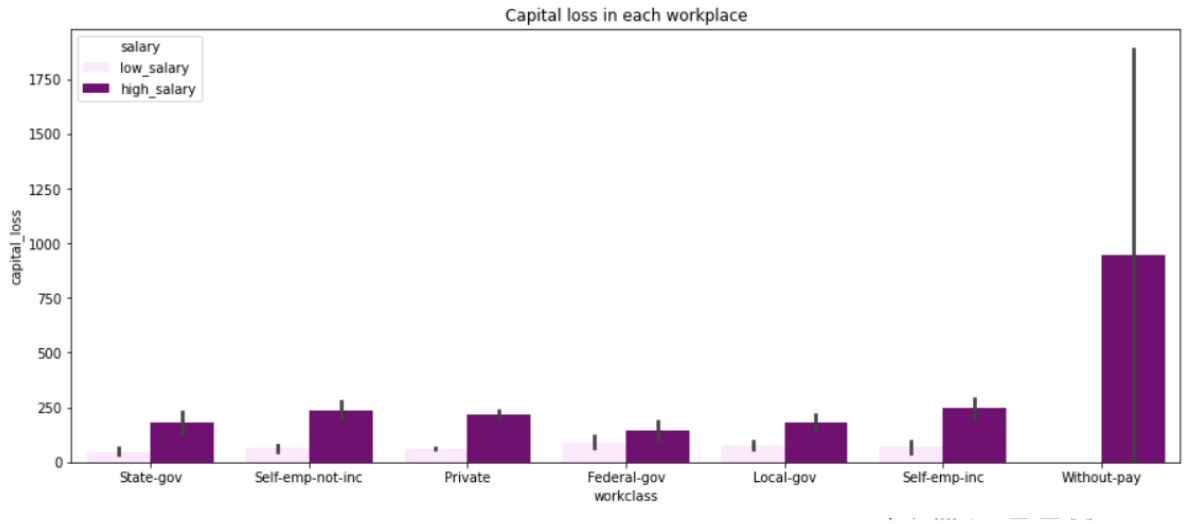


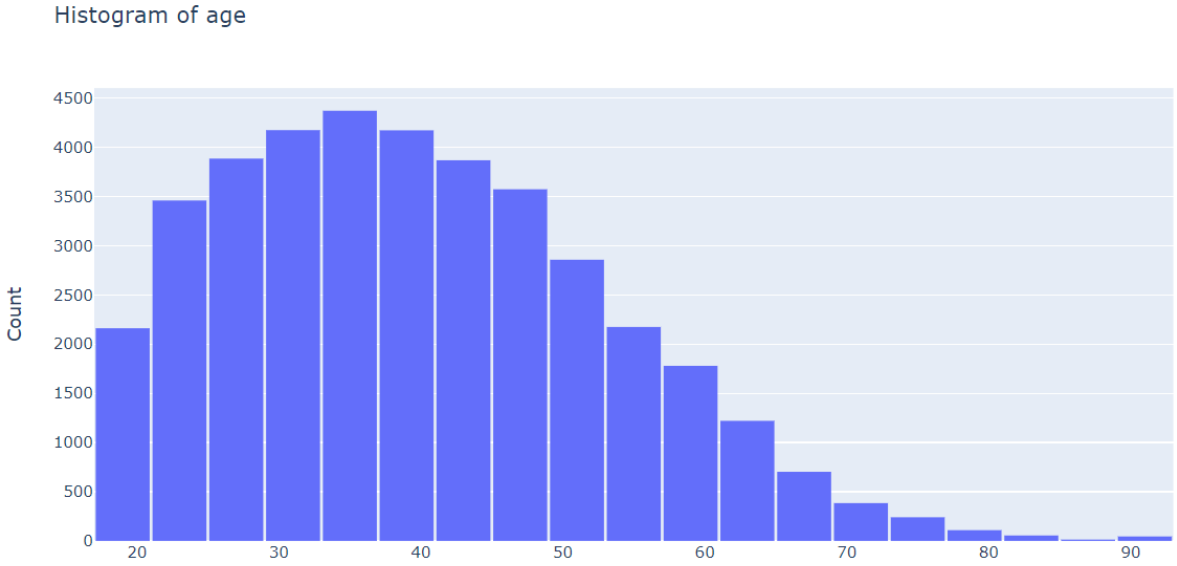


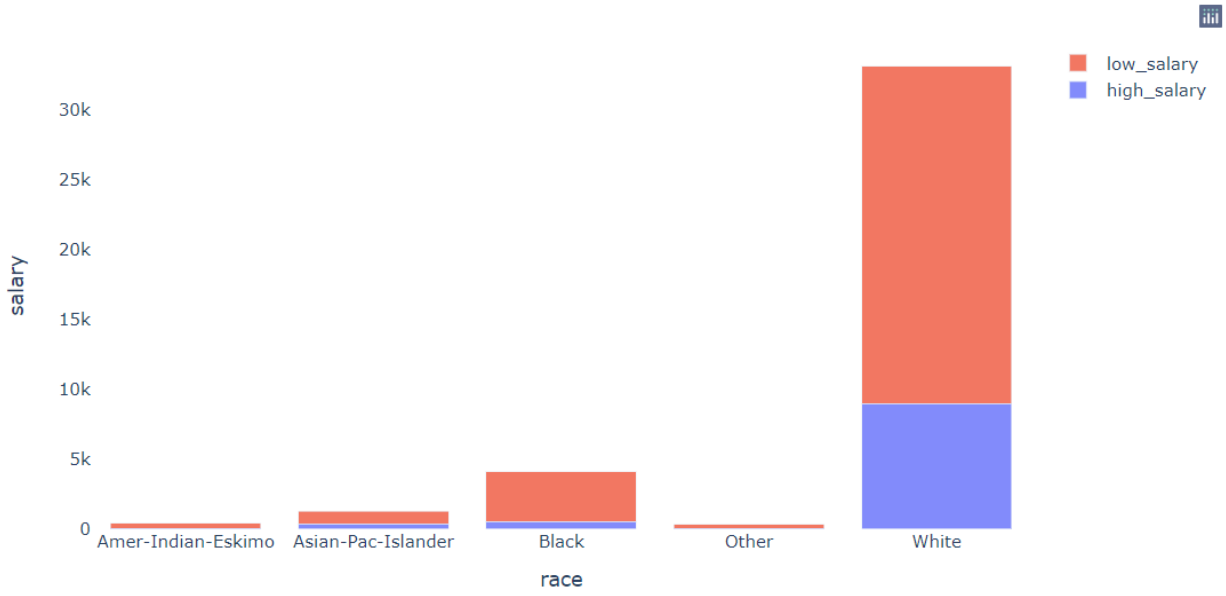

Total hours per weeek per workclass
salary
low_salary
high_salary
Total hours per week
0
10
20
30
40
50
60
State-gov
Self-emp-not-inc
Private
Federal-gov
Local-gov
Self-emp-inc
Without-pay
workclass

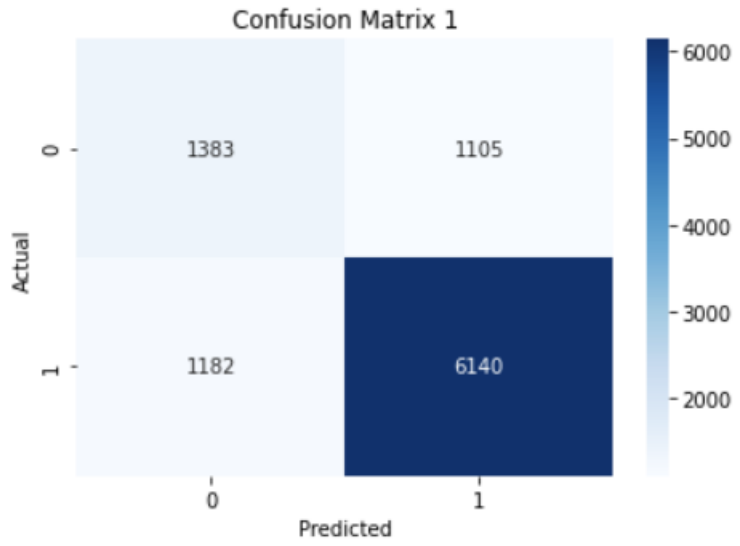

Capital gain in each workplace
capital_gain
0
1000
2000
3000
4000
5000
6000
State-gov
Self-emp-not-inc
Private
Federal-gov
Local-gov
Self-emp-inc
Without-pay
workclass

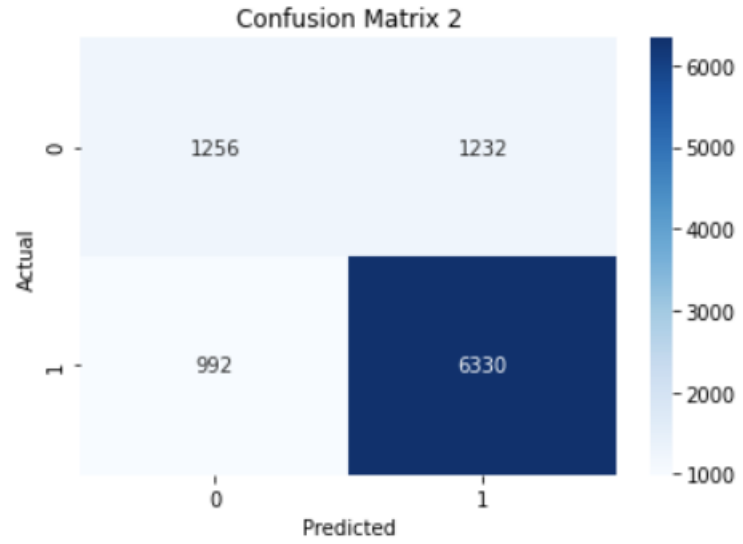

Capital loss in each workplace
capital_loss
0
50
100
150
200
250
State-gov
Self-emp-not-inc
Private
Federal-gov
Local-gov
Self-emp-inc
Without-pay
workclass

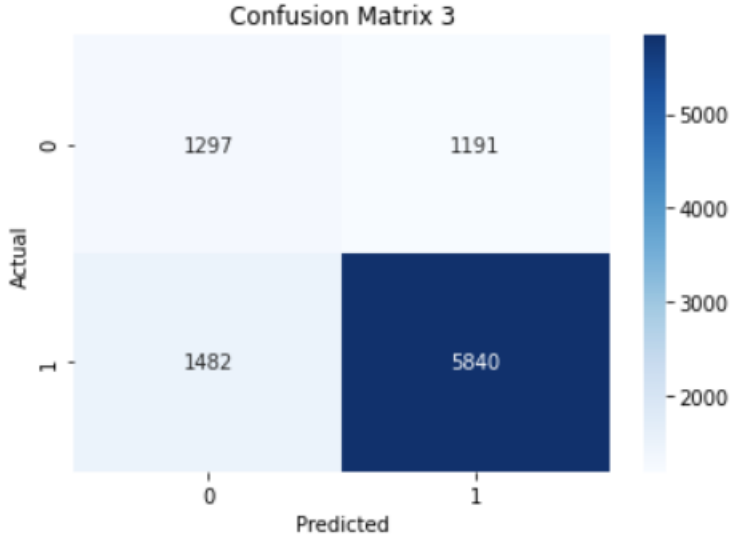

Capital loss in each workplace
salary
low_salary
high_salary
capital_loss
1750
1500
1250
1000
750
500
250
0
State-gov
Self-emp-not-inc
Private
Federal-gov
Local-gov
Self-emp-inc
Without-pay
workclass


Histogram of age
Count
4500
4000
3500
3000
2500
2000
1500
1000
500
0
20
30
40
50
60
70
80
90


low_salary
high_salary
30k
25k
20k
15k
10k
5k
0
salary
Amer-Indian-Eskimo
Asian-Pac-Islander
Black
Other
White
race


Confusion Matrix 1
1383
1105
1182
6140
0
1
Actual
0
1
Predicted
6000
5000
4000
3000
2000


Confusion Matrix 2
1256
1232
992
6330
0
1
Actual
0
1
Predicted
6000
5000
4000
3000
2000
1000


Confusion Matrix 3
1297
1191
1482
5840
0
1
Actual
0
1
Predicted
5000
4000
3000
2000