BDOS_Report

December 27, 2020

Trina Sahoo (Matricola - 1901254)

Introduction : In this project we analyze a U.S. census data taken from the UCI (University of California at Irvine) Machine Learning Repository. The project is divided into four parts: Cleaning and Exploratory Data Analysis, Preprocessing the Data, Predictive Analysis and Theoretical Background. Our final goal is to build a model, which can predict whether the income of a random adult American citizen is less or greater than 50000$ a year based on given features, such as age, education, occupation, gender, race, etc. We fit six different predictive models – a logistic regression model, a random forest model, a K Nearest Neighbor model, decision tree classification model, Gaussian naive bayes model and a neural networks model. All models achieve approximately the same prediction accuracy.

In the first part of the project we clean and explore the dataset and use different visualization techniques to conduct a preliminary analysis of the impact of each predictor (called also independent varible or explanatory variable, or covariate) on the response variable (called also dependent variable) "income". We have performed hypothesis testing to observe the dependency among the variables. In the third part of the project we build predictive models using different algorithms. We apply logistic regression, random forests, k nearest neighbor, decision tree, gaussian naive bayes and neural networks. We test the accuracy of the built models both on the training dataset and on a test dataset. In the last part of the project we provide a theoretical overview of some of the methods that we use.

Dataset : By observing the dataset it is evident that the variables "age", "fnlwgt", "education_num", "capital_gain", "capital_loss" and "hours_per_week" are of type integer, whereas all the other variables are factor variables with different number of levels.

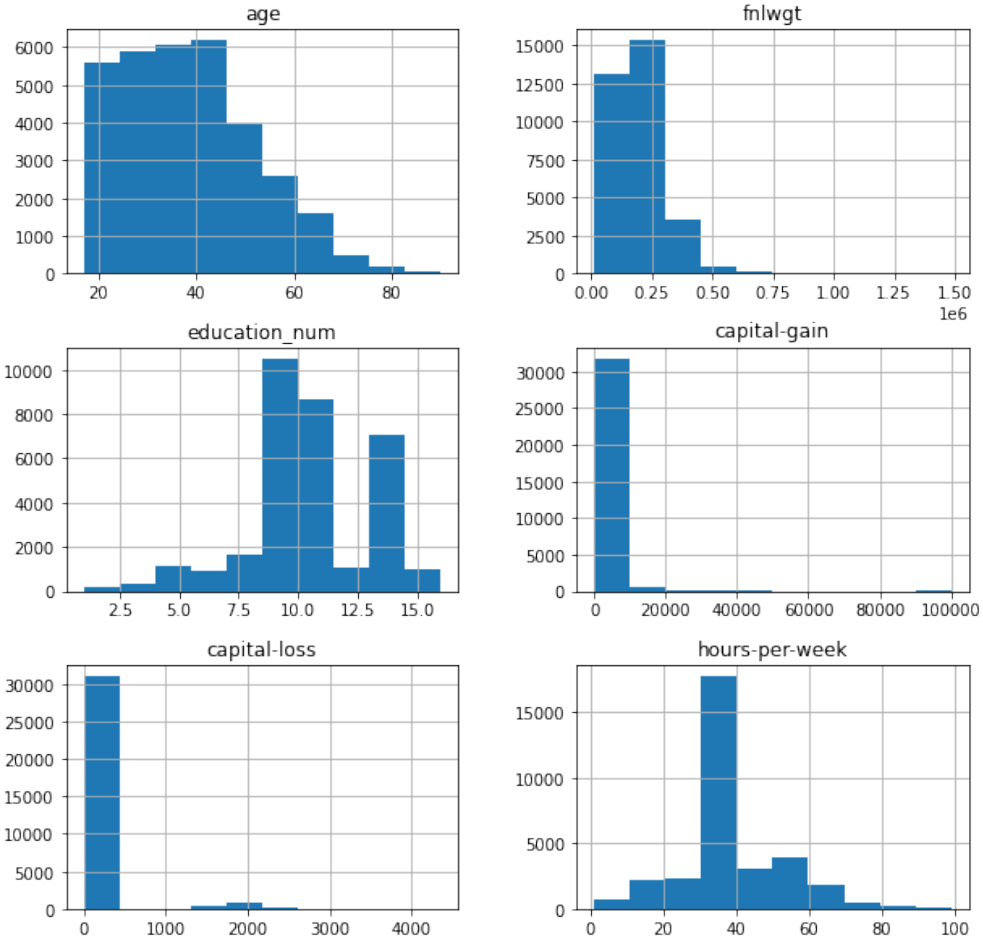
Exploratory Data Analysis:

1. **Missing Values :** Before we can proceed with the exploratory data analysis (EDA) and the predictive analysisi later, we have to get rid of the missing values. In order to do that, we obtain the value count of all the variables and observe the "?" and replace the "?" with the mode.

   From the output of the value count we observe that there are "?" in 3 columns, ie., workclass, occupation and native country. Then we replace the "?" of workclass with the category "Private" which has the highest count. Similarly, we have replaced "?" of occupation with the category "Prof-Specialty" and native-country with the category "United-States". We have performed the same for both train and test dataset.

2. **Visualization of numerical and categorical columns :**

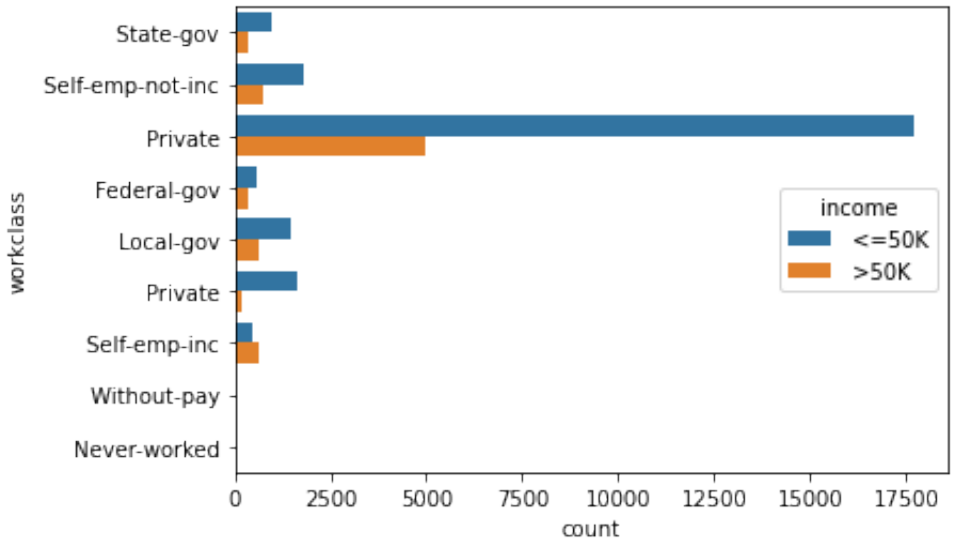
Visualization of numerical columns with histogram:



Inference : 1. Age : From the histogram plot of the variable age it is evident that the number of workers of age 30-50 is more than the other age group. 2. Capital gain : The histogram confirms once more what we have already established, namely, that the majority of people with positive capital gain have a capital gain between 0 and 25,000 dollars, and there are also about 150 people with capital gain of around 100,000 dollars. We also note that the biggest number of people with positive capital gain are those with about 5,000 dollars. 3. Capital loss : The non-zero capital gain is much bigger than the capital loss. From the histogram we can see that the biggest number of capital loss lies between 0-1000 US Dollar. 4. Hours per week : The majority of people work between 40 and 45 hours a week, but there is also a considerable amount of participants who work between 45 and 60 hours per week as well as less than 40 hours a week.

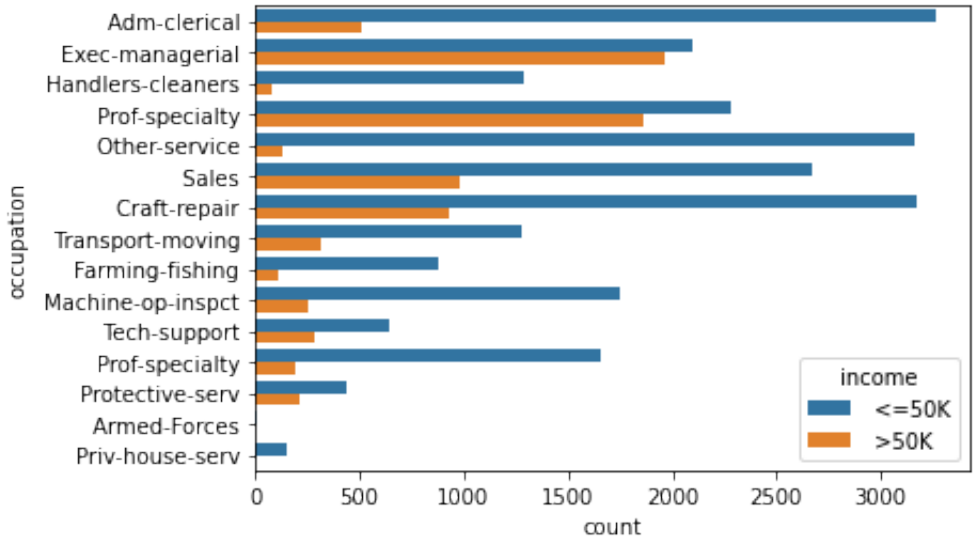
Visualization of categorical columns:

A. Effect of workclass on income:



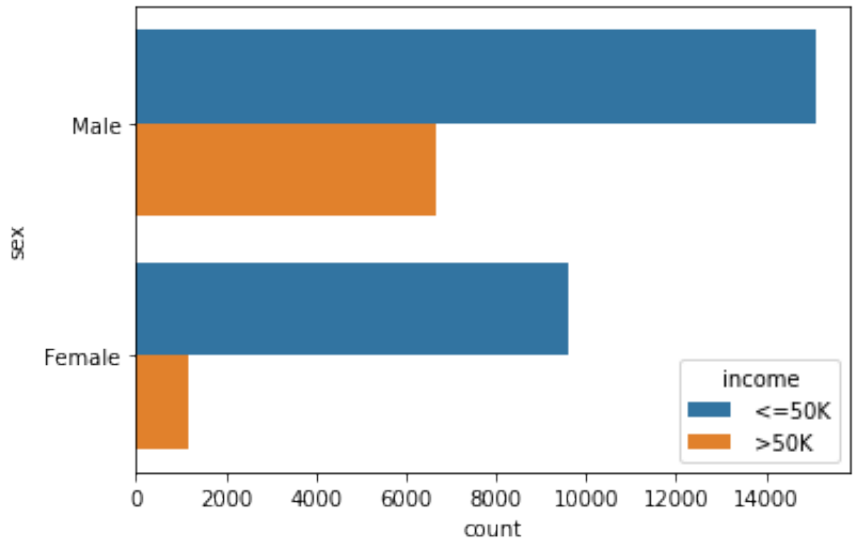
Inference : From the plot it is evident that the number of people worked in the private sector is more than the other sectors. At the same time we can observe that a good percentage of people have income more than 50K dollars.

B. Effect of occupation on income :



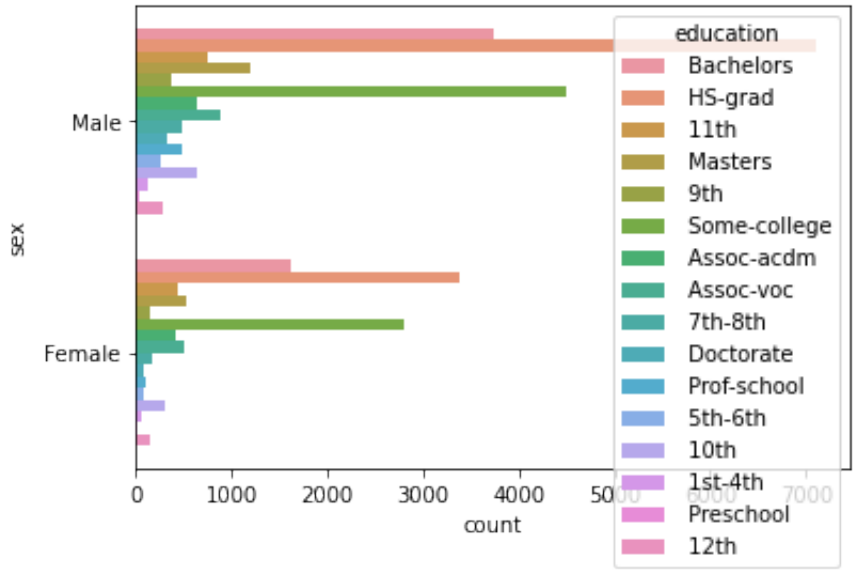
Inference : From the plot of occupation we can say that there is no such salary discrimination with respect to the occupation.

C. Effect of gender on income :



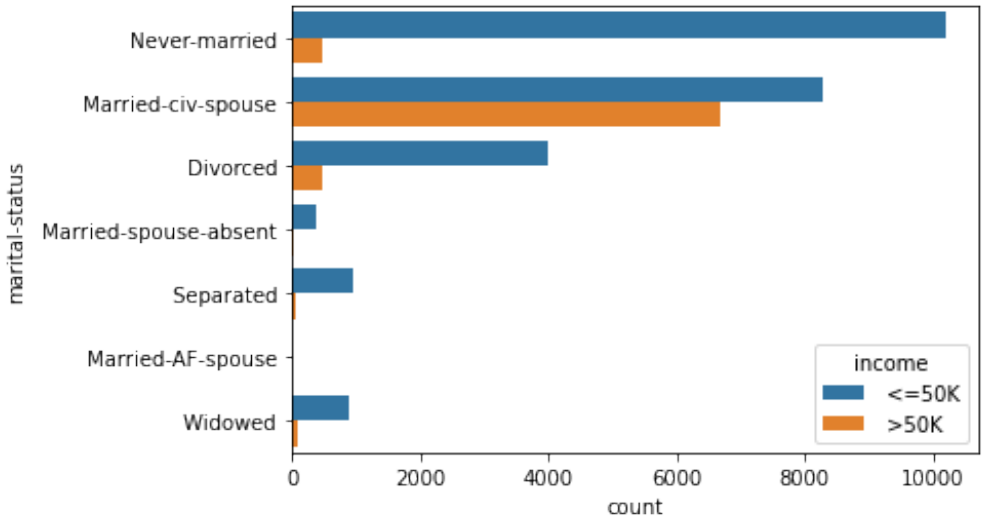
Inference : From the plot of gender and income we can say that the income does not depend on the gender. But it is evident that the number of male workers are more than the number of female workers in the dataset.

D. Relationship between gender and education :



Inference : The same analysis can be possible from this plot as of the effect of gender on income.

E. Effect of marital status on income:



Inference : The number of unmarried people have more earning than the other group of people.

3. **Hypothesis Testing :**

We can interpret the test statistic in the context of the chi-squared distribution with the requisite number of degress of freedom as follows:

If Statistic >= Critical Value: significant result, reject null hypothesis (H0), dependent.
If Statistic < Critical Value: not significant result, fail to reject null hypothesis (H0), independent.

Chi-square goodness of fit : A chi-square goodness of fit test allows us to test whether the observed proportions for a categorical variable differ from hypothesized proportions. The chi-square statistical test is used to determine whether there's a significant difference between an expected distribution and an actual distribution.

• To test the relationship between workclass and income: At first we compute the contingency table.

income	<=50K	>50K
workclass		
Local-gov	2	3
Private	38	8
Self-emp-inc	1	1
Self-emp-not-inc	3	0
State-gov	5	0
Private	4	0

The table determines whether one variable is depends on the other variable. We can also determine the dependency using a statistical method called Pearson's Chi-Squared test.
H0 : There is no relationship between workclass and income
H1 : There is a relationship between workclass and income

Output :
degrees of freedom=5
p_value 0.08096328825908083

[[4.07692308 0.92307692]
[37.50769231 8.49230769]
[1.63076923 0.36923077]
[2.44615385 0.55384615]
[4.07692308 0.92307692]
[3.26153846 0.73846154]]
probability=0.950, critical value=11.070, statistic=9.805
Result : Independent (fail to reject H0)

Inference :As we have accept the null hypothesis, that is, H0 we can conclude that there is no dependency between workclass and income

• To test the relationship between education and income:

H0 : There is no relationship between education and income

H1 : There is a relationship between educaion and income

Now compute the contingency table:

income	<=50K	>50K
education		
10th	1	0
11th	4	0
7th-8th	1	0
9th	3	0
Assoc-acdm	2	0
Assoc-voc	2	2
Bachelors	7	4
Doctorate	1	0
HS-grad	22	1
Masters	2	2
Prof-school	1	0
Some-college	7	3

Output :

degrees of freedom=11
p_value 0.20657766122750312
[[0.81538462 0.18461538]
[3.26153846 0.73846154]
[0.81538462 0.18461538]
[2.44615385 0.55384615]
[1.63076923 0.36923077]
[3.26153846 0.73846154]
[8.96923077 2.03076923]
[0.81538462 0.18461538]
[18.75384615 4.24615385]
[3.26153846 0.73846154]
[0.81538462 0.18461538]
[8.15384615 1.84615385]]
probability=0.950, critical value=19.675, statistic=14.499
Result : Independent (fail to reject H0)

Inference : As we have accept the null hypothesis, that is, H0 we can conclude that there is no dependency between education and income.

• To test the relationship between gender and income:

H0 : There is no relationship between gender and income

H1 : There is a relationship between gender and income

Contingency table:

income	<=50K	>50K
sex		
Female	17	0
Male	36	12

Output :

degrees of freedom=1
p_value 0.05494397325265139
[[13.86153846 3.13846154]
[39.13846154 8.86153846]]
probability=0.950, critical value=3.841, statistic=3.684
Result : Independent (fail to reject H0)

Inference : As we have accept the null hypothesis, that is, H0 we can conclude that there is no dependency between gender and income.

4. **Correlation :** To determine the relationship between the variables:

Correlation

Interpretation : 1. There is no strong correlation between the numerical attributes 2. There is neither strong positive correlation nor strong negative correlation between the variables. 3. The strongest correlation is 0.078, which is between capital-gain and hours_per_week, which is also less than 0.1.

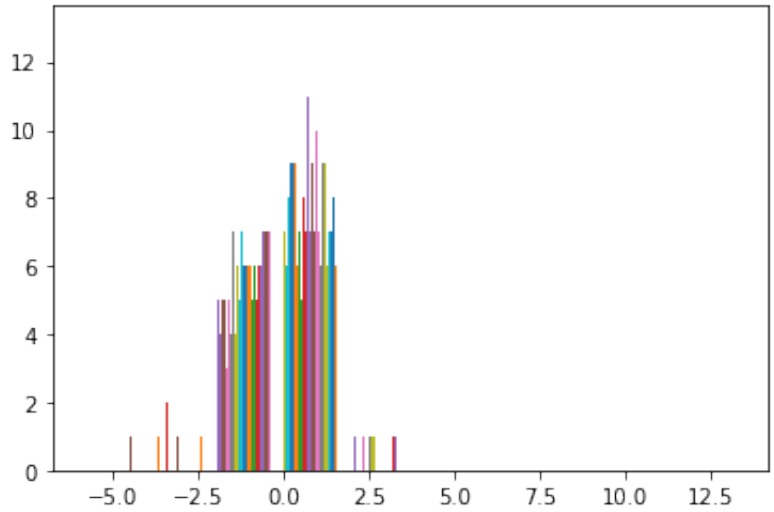
Data Pre-processing :

The data processing have three parts - encoding, splitting of train and test dataset and standardization.

1. **Encoding :** At first the income category <=50K and >50K is classified to 0 and 1 respectively for both train and test dataset. Then the categorical variables are considered and transformed into numerical values using LebelEncoding function.

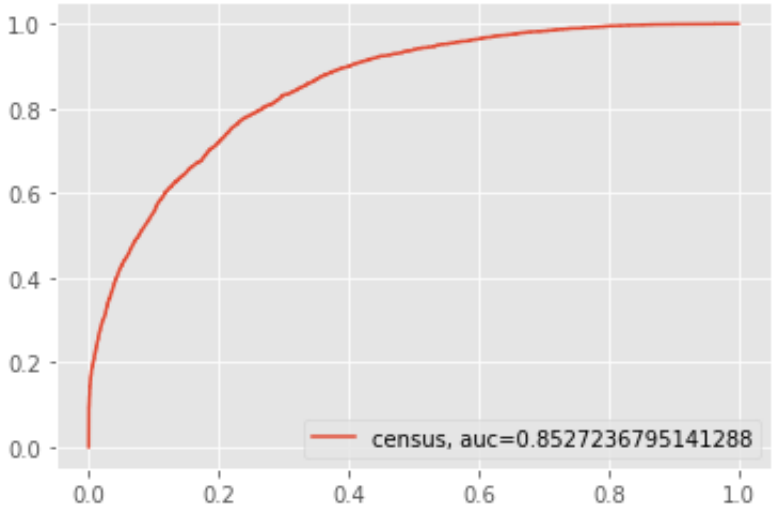
2. **Train and test dataset :** Splitting the train data as x_train and y_train. x_train contains all the features except income column and y_train contain the target variable, that is, the income column. Similarly, the test data is splitted into x_test and y_test where x_test contains all the features except the income column and y_test contains the target variable.

3. **Standardization :** Standardization allows to put different variables on the same scale. This process allows to compare scores between different types of variables. Typically, to standardize variables, wecalculate the mean and standard deviation for a variable. Then, for each observed value of the variable, we subtract the mean and divide by the standard deviation. The visualization of the standardized dataset :

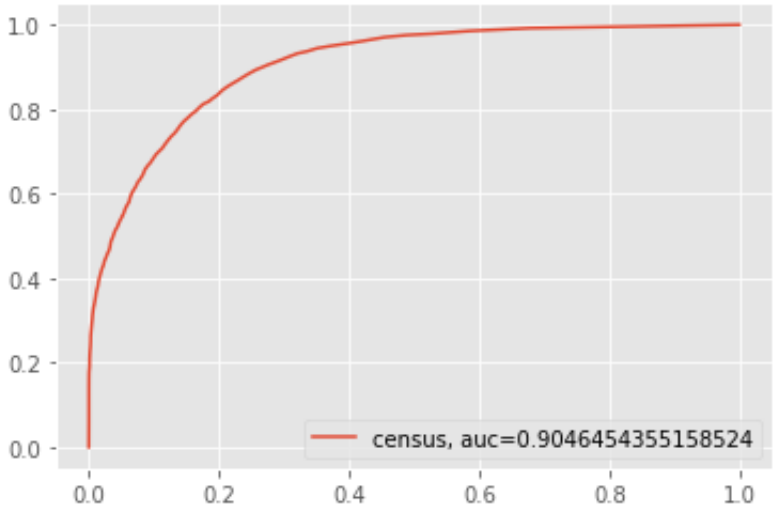


Building and Comparing ML Model : We have used 6 predictive model and compare the scores

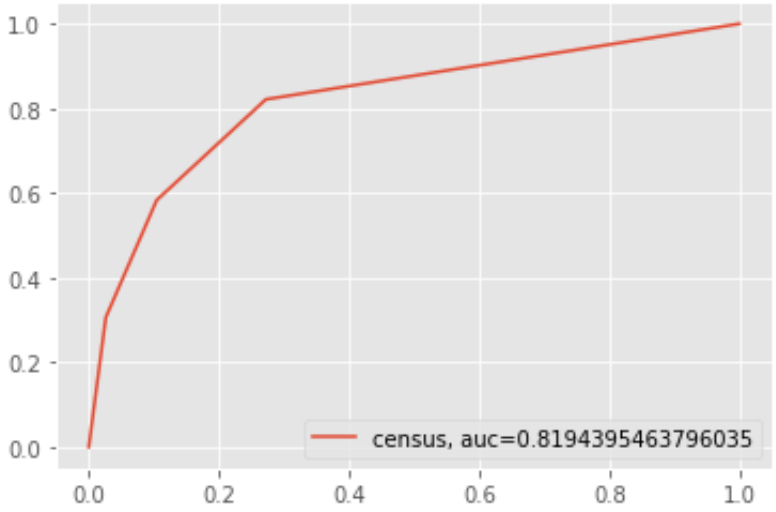
1. **Logistic Regression :** Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. The area under the ROC curve is obtained below:



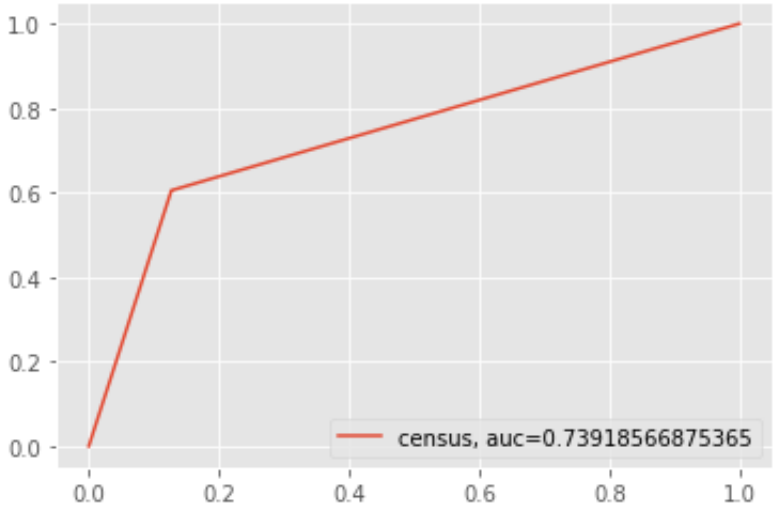
2. **Random Forest :** Random forest algorithm can be used for both classifications and regression task. It provides higher accuracy through cross validation. The area under the ROC curve is obtained below:



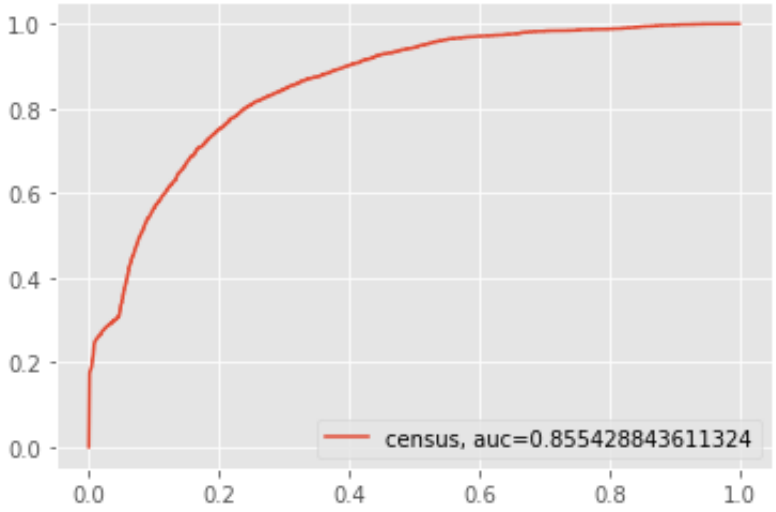
3. **K Nearest Neighbor :** The KNN algorithm can compete with the most accurate models because it makes highly accurate predictions. The area under the ROC curve is obtained below:



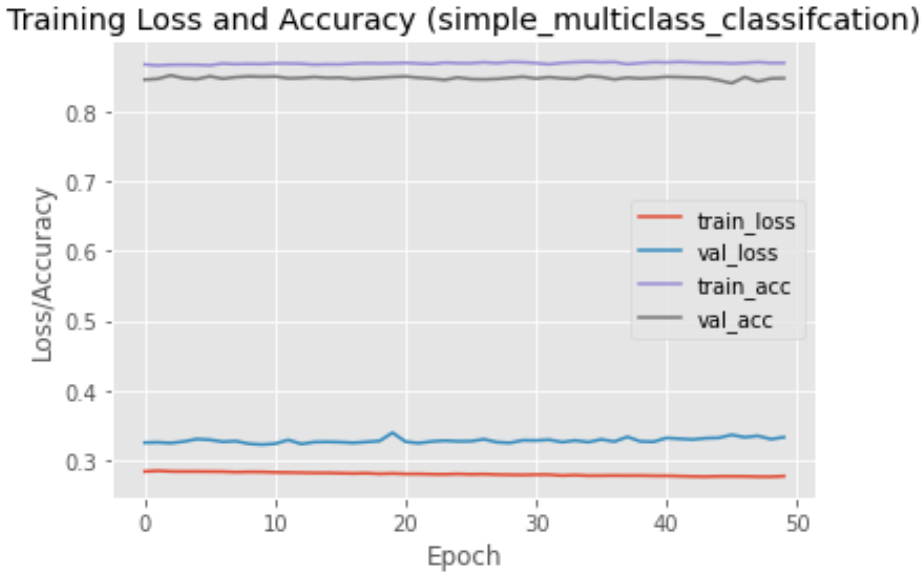
4. **Decision Tree :** Decision trees are extremely useful for data analytics and machine learning because they break down complex data into more manageable parts. They're often used in these fields for prediction analysis, data classification, and regression. The area under the ROC curve is obtained below:



5. **Gaussian Naive Bayes :** The Gaussian (or Normal distribution) is the easiest to work with because you only need to estimate the mean and the standard deviation from your training data. The area under the ROC curve is obtained below:



6. **Neural Network :** Neural network helps to cluster and classify. They help to group unlabeled data according to similarities among the example inputs, and they classify data when they have a labeled dataset to train on. The accuracy and loss curve is showns below:



Comparing the models:

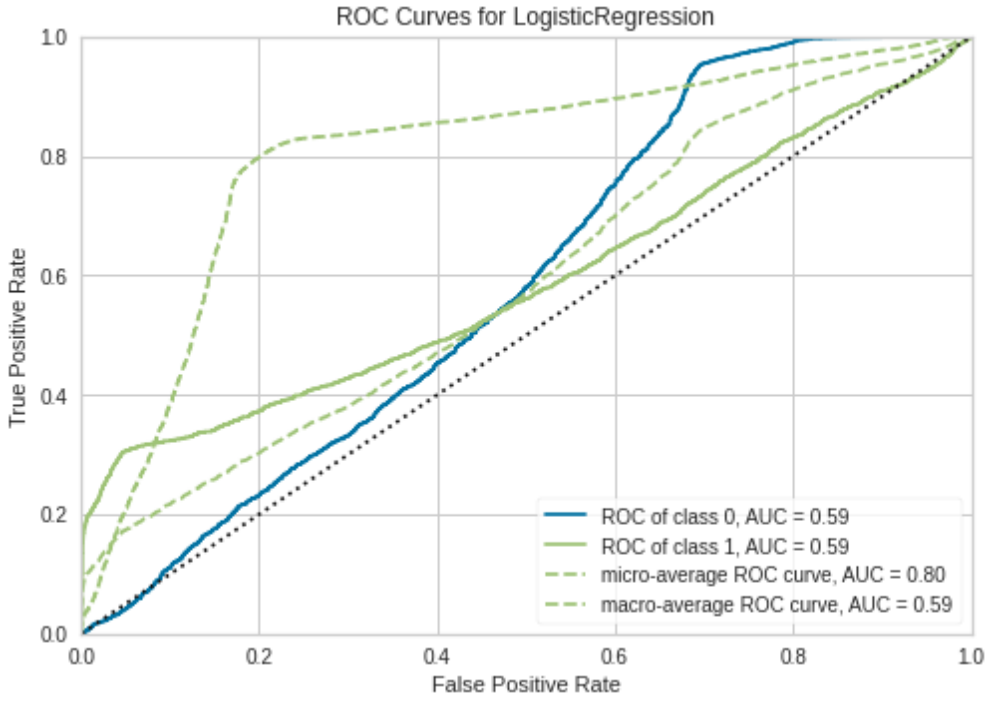
Model	Train_Score	Test_Score
Decision Tree	0.999969	0.808366
Random Forest	0.999939	0.852896
KNN	0.897515	0.821817
Neural Network	0.872117	0.846877
Logistic Regression	0.824913	0.825318
Gaussian Naive Bayes	0.803999	0.805110

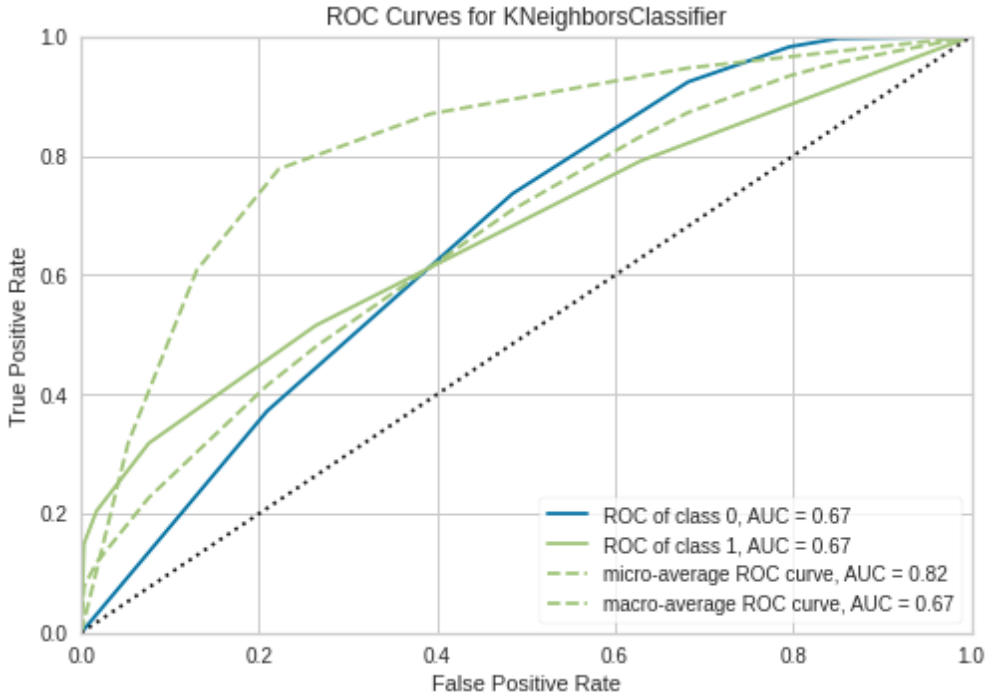
Inference : From the table we can see that the accuracy for the decision tree is the highest.

Although the lowest accuracy, that is, the accuracy for gaussian naive bayes is not that low. Moreover, we can see overfitting in the Logistic regression model and Gaussian naive bayes model. To overcome the overfitting we have done hyperparameter tuning for both the model. The hyperparameter tuning for both the models will be discussed on the later part of the project.

Pycaret Library :

With the help of the pycaret library we have create Logistic regression model and K nearest neighbor model. For both the model it has shown 10 iterations. The model shows the accuracy, auc, recall, precision, kappa and mcc. The ROC curve for the logistic regression and KNN are shown below:





We have also show the model accuracy, auc, recall, precision, kappa and mcc after hyperparameter tuning. And we have compare different models using the library. The accuracy score for CatBoost Classifier is the highest for this dataset.

Metrics :

1. **Classification Metric :**
   The choice of metrics to evaluate the machine learning model is very important. Choice of metrics influence how the performance of the machine learning is measure and compared.

   A. **Confusion matrix :** The Confusion matrix is one of the most intuitive and easiest (unless of course, you are not confused)metrics used for finding the correctness and accuracy of the model. It is used for Classification problem where the output can be of two or more types of classes.
   Output :
   array([[10813, 1622], [1498, 2348]])

   The Model classifies the salary of the person more than 50K US dollar and less than 50K US dollar (case of False positive). Now, in this situation, this is pretty bad than classifying a person earning less than 50K US dollar as similar as the person earning more than 50K US dollar. So in case of Salary classification, minimising False positives is more important than False Negatives.

   B. **Precision and recall :** Precision is a measure that tells us what proportion of people earn more than 50K US dollar. The predicted positives (People predicted in the salary group more than 50K US dollar are TP and FP) and the people actually earn more than 50K US dollar are TP.
   Recall is a measure that tells us what proportion of people actually earn more than 50K. The actual positives (People earning more than 50K are TP and FN) and the people identified by the model earning more than 50K are TP. (Note: FN is included because the Person actually earn more than 50K even though the model predicted otherwise).

Output :
Precision: 0.5914357682619648
Recall: 0.6105044201768071

C. **F1_score :** If one number is really small between precision and recall, the F1 Score kind of raises a flag and is more closer to the smaller number than the bigger one, giving the model an appropriate score rather than just an arithmetic mean. *Output* :
F1_score : 0.6008188331627431

2. **Cross validation :**
Cross-validation is a statistical method used to estimate the skill of machine learning models. We split our data into K parts, suppose we have used K=4 for knn model. We have 32561 instances in our dataset, We split it into four parts, part 1, part 2, part 3 and part 4. We then build four different models, each model is trained on three parts and tested on the fourth.

Hyperparameter Tuning :

1. **Hyperparameter tuning for logistic regression model :** We have tuned the hyperparameter by using the GridSearchCV. Before using GridSearchCV, lets have a look on the important parameters.

• estimator: In this we have to pass the models or functions on which we want to use GridSearchCV
•param_grid: Dictionary or list of parameters of models or function in which GridSearchCV have to select the best. We have set C and Alpha of Logistic regression classifier model with different set of values. And we have set different type of penalty, solvers and different values for max_iter.

The cross validation generator is taken as Repeated Stratified K Fold with 5 splits, 3 repeats and random state as 999. Then the logistic model is obtained by tuning the parameters in the stated way.

Output :
Fitting 15 folds for each of 1200 candidates, totalling 18000 fits
Accuracy on training data - : 0.82501
Accuracy on test data - : 0.82624

Inference : The accuracy for the training dataset has increased slightly and the accuracy on the test dataset has increased too. By observing the accuracy score we can say that there is still some kind of overfitting. To handle them we can further rearrange the train and test dataset by removing the less important columns like “fnlwgt”, “education_num”.

2. **Hyperparameter tuning for Gaussian naive bayes model :** We have fitted Gaussian Naive Bayes model and optimize its only parameter, var_smoothing, using a grid search. Variance smoothing can be considered to be a variant of Laplace smoothing in the sense that the var_smoothing parameter specifies the portion of the largest variance of all features to be added to variances for calculation stability. We have also changed the linespace and keep the scoring as “accuracy”. The cross validation generator is taken as Repeated Stratified K Fold with 5 splits, 3 repeats and random state as 999. Then the logistic model is obtained by tuning the parameters in the stated way.

Output :

Fitting 15 folds for each of 100 candidates, totalling 1500 fits
Accuracy on training data - : 0.81923
Accuracy on test data - : 0.81801

Inference : The accuracy score for both training and test data has increased. There is no more overfitting after hyperparameter tuning

Conclusion : It is important to note that the model is predicting whether someone earns more than 50K US dollar. However, decision tree model work best and then random forest model also works better. The highest accuracy is 0.99 considering all the features of the training data. Moreover, the accuracy on the test data is the highest for random forest model.