**INSTRUCTIONS RELATED TO FOURTH PHASE EVALUATION PROJECT**

 The article should contain the following sub-topics:  
1. Problem Definition  
2. Data Analysis  
3. EDA Concluding Remarks  
4. Pre-processing Pipeline  
5. Building Machine Learning Models  
6.  Concluding Remarks

**THE PROBLEM DEFINITION:**

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). A set of reasonably clean records was extracted using the following conditions: ((AAGE>16) && (AGI>100) && (AFNLWGT>1) && (HRSWK>0)).

**The prediction task is to determine whether a person makes over $50K a year***.*

The problem statement here is that the data extracted from the 1994 Census bureau database shows insight on the yearly income of the people. The idea is that the annual income is not the same in the overall population and as a data scientist we are expected to determine the population that makes over 50k dollas in the US citizen. Some features of this data are their work class, age, final weight (Fnlwgt), educational status, capital gain, capital loss, nationality or the hour per weeks. The exploratory data analysis can also be perform using different features and analysing their respective outcome based on the result of the visualisation. The use of different graph like count plot, bar plot, dist plot, heatmap and others. Data preprocessing, data equalisation using SMOTE from sklearn and removing outliers using zscore. The data will also undergo through data train test split and passing it into different machine learning model like the Random Forest Classifier, Gradient Boosting Classifier etc to determine the accuracy, prediction and confusion matrix.

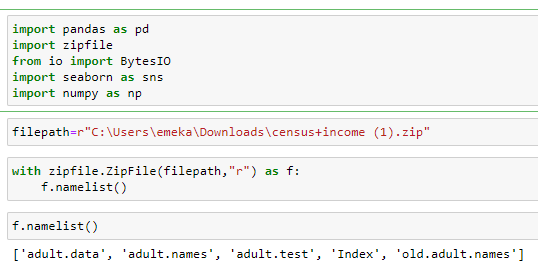
DATA ANALYSIS

There is presence of 15 features in the columns and 32560 in the row. The target variable are two category which are the income (<=50k and >50k). It has two data types which are the object and the int. The data has different work class which are the private, self emp not inc, local-got, state-gov, federal gov, without-pay,

never-work and the? The? was replaced using the

mode which is the highest in occurrence. The sex has only male and female. The data also has different race such as the white, black, asia-pac-Islander and Amer-India-Eskimo. The weight of the current survey are controlled by the dependent estimate of the populations

which are non-institutional population of the US.The population are controlled by a single cell estimate of the population which are above 16 for every state, the control for Hispanic origin by age and sex, control by race,age and sex.



The above libraries were used to extract the zip file.

A screenshot of a computer

Description automatically generated

The data extracted for the census income are depicted above.

It has 14 independent variables and 1 target variable which is for the income column. [AGE, Work class, Fnlwgt, Education, Educational Number, Marital Status, Occupation, Relation, Race, Sex, Capital Gain, Capital Loss, Hour Per Weeks, Native Country, Income]

The columns were changed for explicit purpose which are shown below.

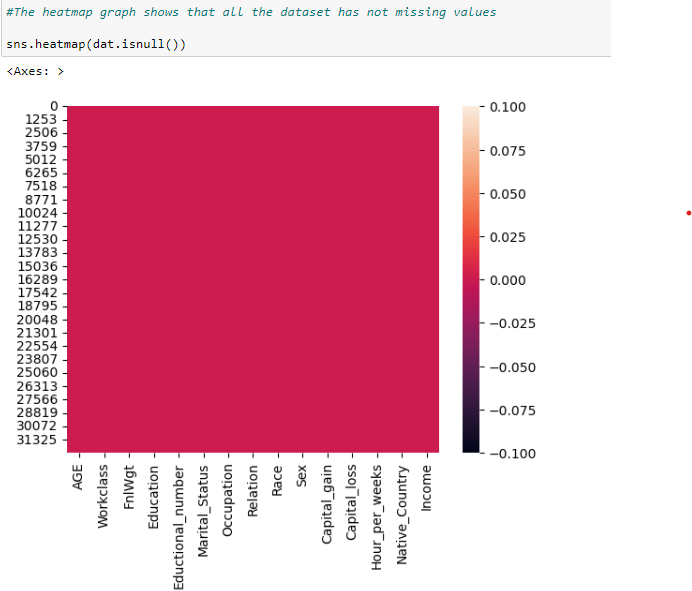
A screenshot of a computer

Description automatically generated

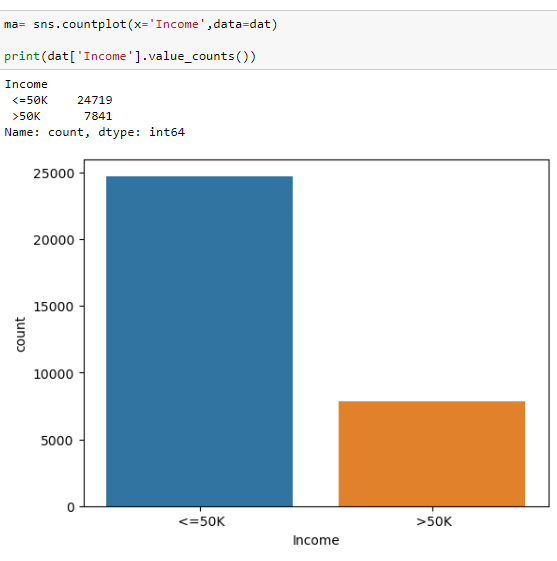


**EXPLORATORY DATA ANALYSIS CONCLUDING REMARKS.**

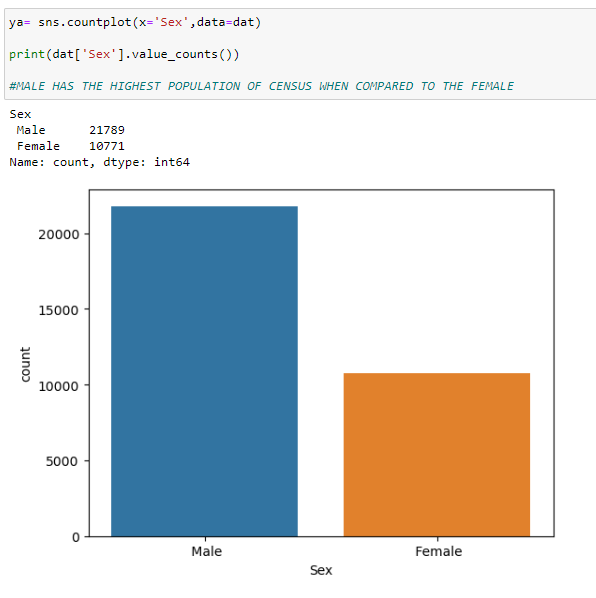
Using heatmap to determine the presence of missing or non-missing value. From this income project it shows that there are no missing values which is shown below. But the? is not known.



Checking for the income status in the state. The graph shows that the 24719 of the population of the census in the state make less than 50k dollar in a year while only 7841 of the population makes over 50k dollar per annual. The graph of this is depicted below.

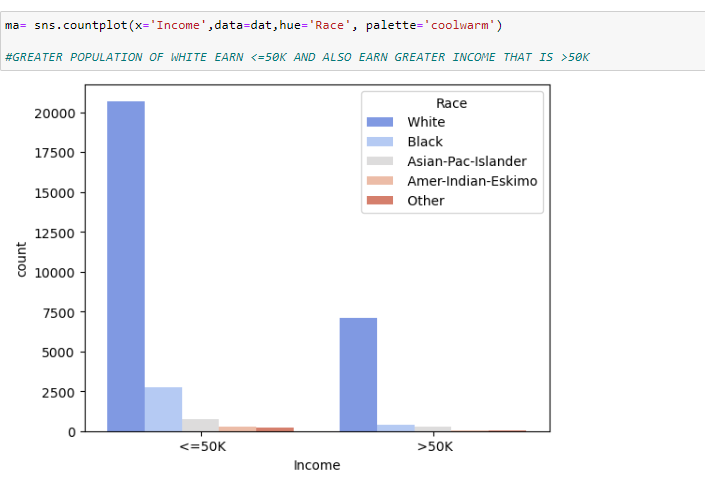


The population of male is over two times that of the female in the census. The population of male is exactly 21789 while that of the female is 10771 in the state which was captured during census.

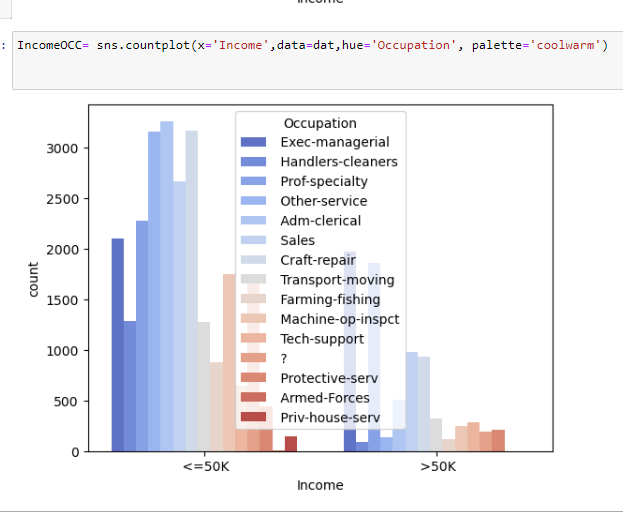


This graph shows that the population of the race and their respective income. It was clearly stated that people whose race are white has highest population is the people that earn less than or equal to 50k dollar. People from black race are the next in the population that earn <=50k. It was clearly shown that people earn the least annual income.

The graph also showed that greater number of the white people make over 50k dollar when compared to other ethnicity. The blacks and Asian-pac-Islander people who make over 50k dollar are low. But the black population that make over 50k dollar are more when compared to the Asia-pac-Islander.



The amount of data the population made varies based on the kind of job they do. It is very important to know that the amount of money depends on the kind of job they do. The Adm clerrical has the greater number of people who do earn <=50k dollar per annual. It is also noted from the graph that the exec-managerial in the US has the highest population of people that make above 50k dollar in a year. In the US the priv house serv has the least population of people who make <=50.



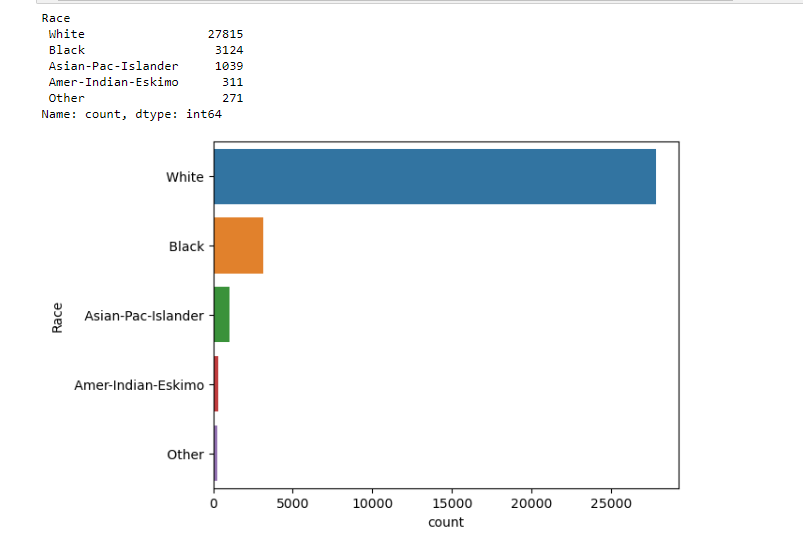
The graph shows that the income made varies based on the kind of work class. The private work class has the highest population among the work class who made <50k dollar per year whilst the private work class also has the highest population of people that earn more than 50k dollar.

The income made by people who never work in the US citizen are the least number of the people that made <=50k.

A screenshot of a graph

Description automatically generated

The graph shows that the count of race varies from white to other. The white people in the US are more than every other race in the country. The graph shows that the count goes from white to other in a negative form. The population decreases from white (27815) to other which is (271). The population of the Asian Pac Islander (1039) is three time that of the Amer Indian Eskimo (311).

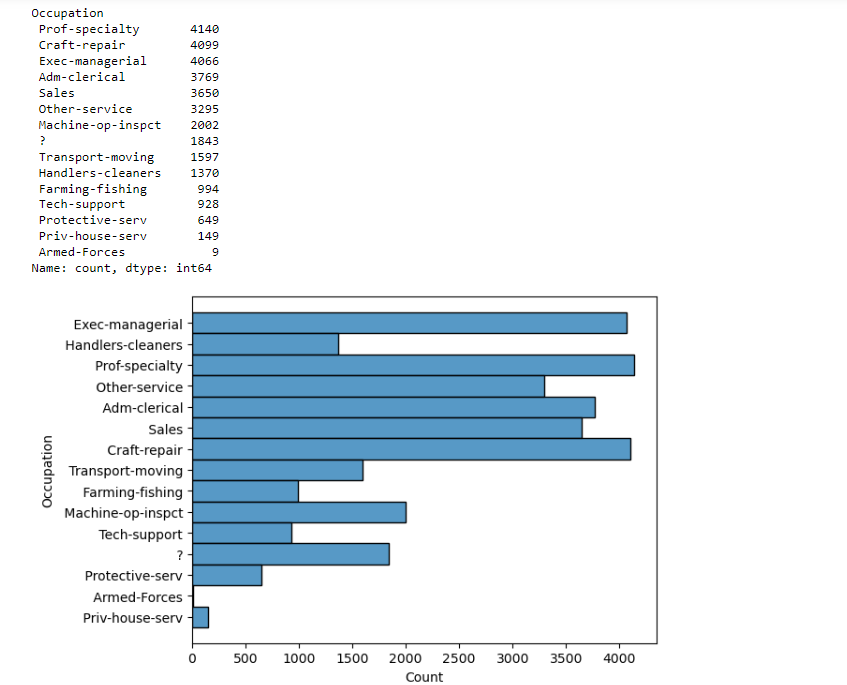


The relation in the census count varies from one relation to another. From the graph bellow 13193 are for husband which is the greatest. The population of wife is very small compared to the husband which will have negative effect to the men and can to men competition in the society. This could make the men to spend more of their income trying to please the wife.

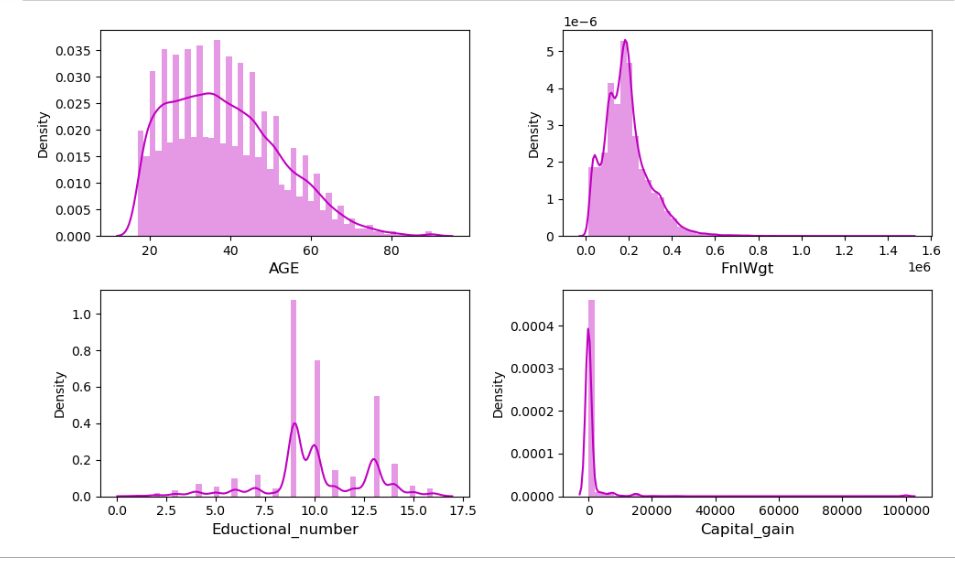
A graph with different colored bars

Description automatically generated

From the graph below it shows that in US population of occupation prof-specialty, Craft-repair and Exec-managerial are the top three population while the armed-force has the least population.



In the distribution plot below, the minimum age is 17 and the max age is 90. The Fnlwgt shows that majority of the population weight are within 0.2. The maximum capital gain is 29849. The education number varies 0 to 17.5 but the majority of the people in academic has number ranging from 7.5 to 10.

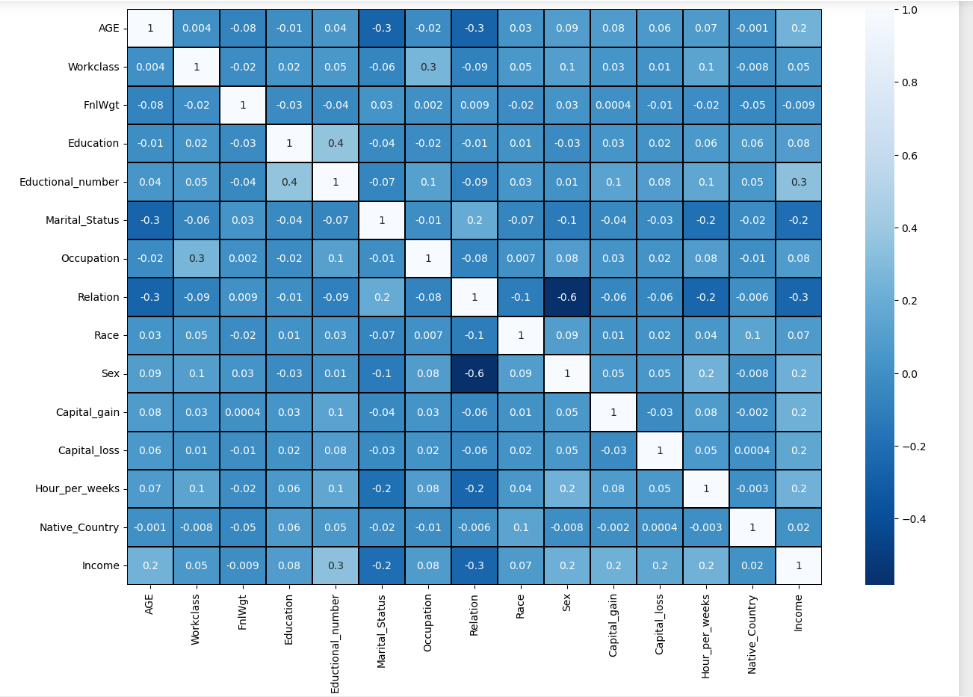


**PRE-PROCESSING PIPELINE**

The dataset extracted from the census bureau has 15 columns which are the features, and these are age, work class, education number, marital status, occupation, relationship, race, sex, capital gain, capital loss, hours per week, native country, fnlwgt and income. The income here is the target variable.

In this pre-processing, label encoder was used to convert all the categorical columns into numerical. The categorical columns are marital status, education, work class, occupation, relation, race, sex, native country and income. The target column is converted into 0 and 1.

The below graph is the heatmap of the correlation matrix which have both positive and negative correlation.



The above heatmap and the barplot of the graph shows that the educational number has the highest positive correlation while the fnlwgt has the least negative correction. The race and occupation has almost the same positive correlation.

A graph of purple squares

Description automatically generated

**REMOVING THE OUTLIERS:**

The outliers were removed using the zscore which help to cut off the outliers presence in the bureau census data which was collected. After which the target variable was not balanced, and it was passed into SMOTE which help to balance the imbalanced target variable (income).

A screenshot of a computer program

Description automatically generated

After balancing the oversampling data, the data was split by importing the train test split from the sklearn. The dataset is split into train (80%) and test (20%). The shape of the dataset to be passed into the machine learning is (27417,15). But after using SMOTE to balance the data is now 21279 for the 0 and 1(target).

**BUILDING THE MACHINE LEARNING MODEL**

The Random Forest Classifier is used to check for the best accuracy and the random state to be used for the machine learning. This shows that the income target has accuracy of 90 with random state of 31.

The two-machine learning used are the Random Forest Classifier and the Gradient Boost and Classifier.

The Random Forest Classifier is a large number of individual decision tree that operate as one of the Ensemble. Each individual tree is split to be able to predict the data given using machine learning appropriate coding. This help to show the level of accuracy, prediction, macro avg and the weighted avg.

A screenshot of a computer

Description automatically generated

From the above the accuracy is 89% and the precision for the population of the US citizen who makes greater than 50k dollar is 89%. The total population is 6138 in all the census collected.

Gradient Boost Classifier is one of the types of the Ensemble method used in the machine learning paradigm where the multiple models called the weak learners are trained to solve the problem and give better result. This builds the model in a stage wise pattern like other boosting method. Gradient boosting combines weak ‘learners’ into a single strong learner in the best fashion.

The gradient boost has the accuracy of 87% which is slightly lower that the accuracy gotten from the Randon Forest Classifier.

A screenshot of a computer

Description automatically generated

**CONCLUDING REMARK.**

The main purpose of this was successfully achieved by predicting the income of the US citizen who were able to make above 50k dollar every year by using different classification of the algorithms. The exploratory data analysis was also carried out to determine the relation and the insight of what happen in all the feature columns that was plotted using different types of graphs. Pre-processing pipeline was carried out where label encoder was used to convert all the categorical data to numerical data. After which the dataset was passed through the machine learning.

In this analysis different classification of algorithm was carried out. The Random Forest Classifier gives the best accuracy of 89%. This is fair enough to say that the prediction is correct.