****

**PROJECT REPORT**

|  |  |
| --- | --- |
| Name | Onta Shahriar |
| ID | 2031275643 |
| Course | Machine Learning (CSE445) |
| Faculty | Saif Ahmed |

**1.Description of the Dataset:**

The dataset was named, “Adult census income dataset”. This dataset was prepared by Ronny Kohavi and Barry Becker who used to work in Data Mining and Visualization Team of Silicon Graphics Company at that time. The data was extracted from the 1994 Census bureau database using the following conditions: ((AAGE>16) && (AGI>100) && (AFNLWGT>1) && (HRSWK>0)), and these conditions filtered out people who were above 16 years old and those who had average final weight of greater than 1 and those who used to work and those who had an average gross income of greater than 1. The dataset has 32,561 rows and 15 columns. The prediction task is to determine whether a person makes over $50,000 annually based on various parameters of their personal life and academic life/career. The parameters such as age, marital status, relationship, sex, race, and nationality provided a measure of each person’s social wellbeing. Features such as work-class, occupation, education level, education qualification(education.num), and others provided measure of each person’s progress in career. There is a special feature in this dataset, ‘fnlwgt’, final weight. The weights on the Current Population Survey (CPS) files are controlled to independent estimates of the civilian noninstitutional population of the US. Population Division at the Census Bureau monthly prepare these. There are 3 sets of controls. These are:

1.A single cell estimate of the population 16+ for each state.

2.Controls for Hispanic Origin by age and sex.

3.Controls by Race, age and sex.

They used all three sets of controls in our weighting program and "raked" through them 6 times so that by the end it comes back to all the controls they used. The term estimate refers to population totals derived from CPS by creating "weighted tallies" of any specified socio-economic characteristics of the population. People with similar demographic characteristics should have similar weights. There is one important caveat to remember about this statement. That is that since the CPS sample is actually a collection of 51 state samples, each with its own probability of selection, the statement only applies within state.

Some of the challenges of the dataset are described below and the methods I used to solve them are also mentioned.

Here, the continuous variables in the dataset are of varying ranges and some are much bigger than others. So, I had to normalize all the continuous variables. I used standard scaler to normalize all the continuous variables so that they all influence the model equivalently.

Also, most of the features of the dataset are categorical in nature. So, I used categorical codes to encode the different types involved in each category.

Here, in the dataset, ‘marital.status’ and ‘relationship’ are very similar data and had high correlation as shown below. ‘marital.status’ also had a good correlation with the ‘sex’ variable as can be seen below. Similarly, ‘education.num’ and ‘education’ accounted for the same thing which is the education qualification of a person. So, I dropped ‘marital.status’ and ‘education.num’ variables from the dataset.

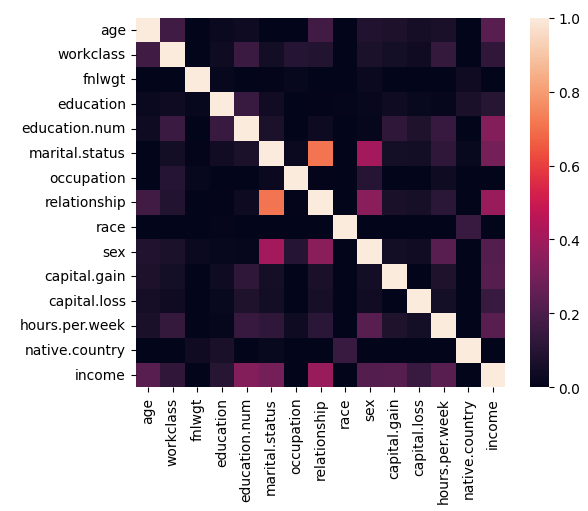


Fig: Figure showing correlation between marital.status and relationship and sex combined.

**2.Data Splitting:**

I have split the total dataset, at first into a main dataset containing 80% of total data, and into a test dataset containing 20% of the total data. Then, I broke down the main dataset into two parts. One part contained 80% of the main dataset and this is the train dataset and the other part contained 20% of the main dataset and this is the validation dataset. All the datasets were created by using random sampling.

**3.Data Analysis:**

In data analysis, I used Exploratory Data Analysis (EDA) to do analysis on my dataset. Staring with understanding the data, I loaded the dataset and looked into the complete picture of the dataset. Then, I looked into the shape of the dataset and the columns of the dataset, to get the idea about the features and the target. Then, I looked into the no. of unique values and classes each continuous and categorical data respectively has.

Then, I went onto cleaning the data. I dropped the features ‘marital.status’ and ‘education.num’ from the dataset. ‘marital.status’ has high correlation with ‘relationship’ variable and they both represented the same thing and similarly, ‘education.num’ and ‘education’ both represented the same thing which is education qualification of an individual, so I also dropped ‘education’ from the dataset. Then, I checked for any null values in the dataset, since, there was no null values in the dataset, so I did not need to handle any null values.

Then, I did statistical analysis of the continuous variables of the dataset. I found out the mean, maximum value, minimum value, standard deviation and three percentiles (25th,50th,75th) of all the continuous variables of the dataset. Then, I did some relationship analysis on my dataset where, I plotted several graphs depicting the relationships between different continuous features and the target of the dataset. I used boxplots, boxen plots, violin plots, point plots, and bar plots to depict several relationships between the continuous features (such as age, fnlwgt, capital.gain, capital.loss, and hours.per.week) and the target. All these plots depict the shape of the distribution of each of the continuous variable with respect to the target variable, ‘income’. Here all the plots were segregated by the two different values of ‘income’ variable which are ‘>50K’ and ‘<=50K’. Also, all the plots were further segregated by the different categorical types contained in each of the categorical features of the dataset.

**4.Hyperparameter Experiments:**

For this classification dataset, I used Random Forest Classification model, which is a renowned classifier and used for many classification problems. Here, I have done 4 hyperparameter experiments in order to find the highest mean accuracy score. Here, I have selected 4 hyperparameters namely, n\_estimators, max\_depth, max\_features, and max\_leaf\_nodes as these hyperparameters have large influence on the mean accuracy score of a Random Forest Classifier. n\_estimators denote the no. of trees used in the forest and I varied this hyperparameter from 110 to 125, increasing by 5. max\_depth denotes the maximum depth of the trees used in the model and I have varied this from 9 to 12, increasing by 1. max\_leaf\_nodes denotes the maximum leaf nodes a root-decision node/ sub-decision node of a decision tree can have and I have varied this hyperparameter form 9 to 12, increasing by 1. max\_features denotes the maximum number of features to consider when looking for the best split at any decision node and I have varied this hyperparameter from 3 to 6, increasing by 1.

**5.Comparison of Results:**

After running the 4 experiments, I have found that Random Forest Classification model having, n\_estimators=125, max\_depth=12, max\_leaf\_nodes=12, and max\_features=6 had the largest mean accuracy score among the 4 models I have experimented with.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Experiment No. | n\_estimators | max\_depth | max\_leaf\_nodes | max\_features | Mean Accuracy Score |
| 1 | 110 | 9 | 9 | 3 | 76 |
| 2 | 115 | 10 | 10 | 4 | 78 |
| 3 | 120 | 11 | 11 | 5 | 79 |
| 4 | 125 | 12 | 12 | 6 | 81 |

Fig: Table showing mean accuracy scores obtained for the 4 experiments.

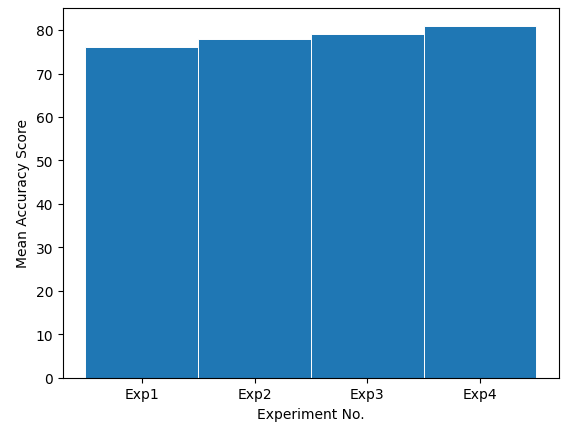


Fig: Bar plot showing the improvement in mean accuracy score.

As can be seen, the mean accuracy score increases with the increase in the value of the 4 hyperparameters. So, in conclusion, the mean accuracy score of a Random Forest Classifier increases with the values of the 4 hyperparameters.