**Group Code: ML063B4 Date of submission: 13/08/2020**

**Project Report**

**Project Overview**

Human intellect is reaching new heights with every passing day, and the basic instinct to outperform and compete is whose core motivation is survival in the best way possible. Income is the consumption and saving opportunity gained by an entity within a specified timeframe, which is generally expressed in monetary terms. Many factors contribute to people having a higher income, including education, globalization, and favorable political circumstances such as economic freedom and peace. Developed countries (defined as countries with a "developed economy") have higher incomes as opposed to developing countries tending to have lower incomes. Through this project the aim is to analyze the key factors leading to various incomes, and the prediction capacities with the help of machine learning models and data analytics.

**Dataset**

In pursuit of successful completion of the major project work as a part of the online internship in machine learning, the dataset chosen is about income classification and the web link for the same is provided at the end of this report.

**Problem Statement**

* Take up three classification algorithms of choice and build three respective Machine learning models. Compare the Accuracy of all three and suggest which ML algorithms suits best for the given problem.
* Ask any two questions on the dataset taken and provide answers for the same.

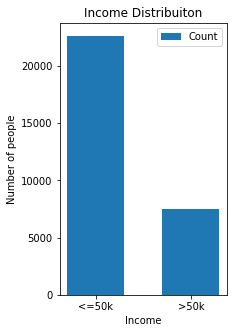
**Data Exploration**

In this process the column head names were extracted and the ones which will be useful for applying classification models are picked. The significant column heads wereage, education, education-num, marital-status, occupation, relationship, race, sex, capital gain, capital-loss, hours-per-week, native-country and income.

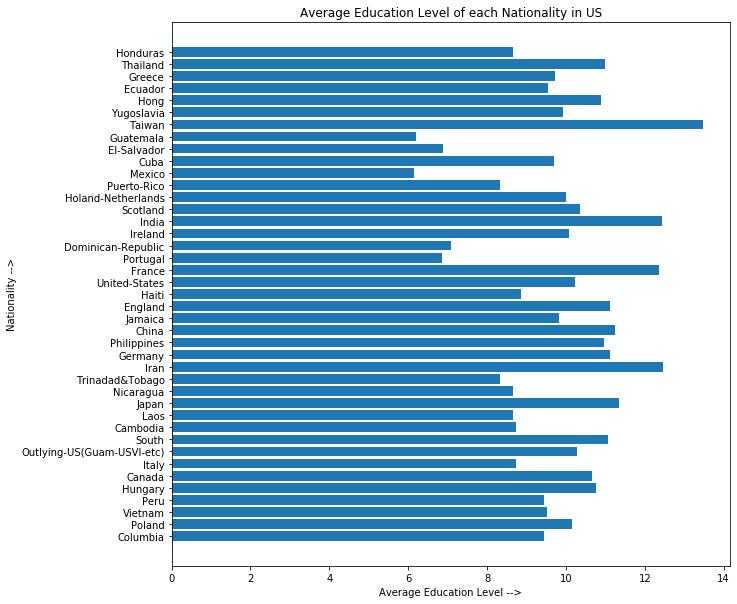
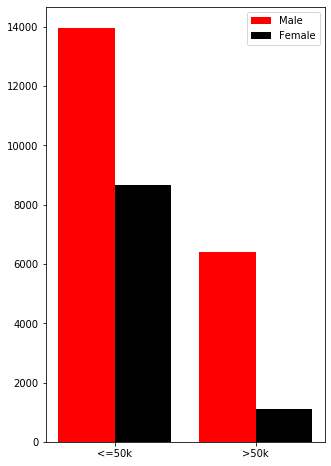
### The shape of the data was assessed and duplicate data was removed. The primary concerns over the dataset found in data exploration were the null values and the skewed structure of the dataset with ratio of income less than fifty thousand to the income more than equal to fifty thousand being around three is to one. About the null values, these did not seem to exist but unknown values denoted by “?” sign were found in the data, these values are removed for reducing complications.

**Exploratory Visualization**

After the data is cleaned, the visualization techniques were applied. Firstly a broad display of income classification is showcased with dividing the income groups with the threshold of income as fifty thousand.



Further for finding the solutions of the question posed on the dataset visual aids were used for better comprehension of the data



If you have a data set with many columns, a good way to quickly check correlations among columns is by visualizing the correlation matrix as a heat map. With this method we can easily trace the dependent variables applicable for income classification.



**Algorithms**

The three algorithms chosen are:

* Logistic Regression
* Support Vector Method
* K-Nearest Neighbor

**Questions posed on the dataset**

1. What percentages of women are earning above 50k and what percentages of men are earning above 50k?

## What is the nationality of people with highest and lowest average education level?

**Results**

1. **Questions on dataset**

* Percentage of women earning above 50k is **11.378287117568812 %**

Percentage of men earning above 50k is **31.395463026612983 %**

* Country with highest average education level is **Taiwan**

Country with lowest average education level is **Mexico**

1. **Accuracy values**

|  |  |
| --- | --- |
| **Name** | **Accuracy** |
| **Logistic Regression**  **(From Scratch)** | 0.7665560716655607 |
| **Support Vector Method** | 0.844591904445919 |
| **K- Nearest Neighbor** | 0.8449900464499005 |
| **Logistic Regression**  **(in-built)** | 0.762043795620438 |

Note: - The in-built logistic regression algorithm has 0 score, i.e. it predicts all values as 0

**Conclusion**

From the application of three algorithms and a comparative analysis of the accuracy values, the inference received is that “K-Nearest Neighbors” has the best accuracy and the score is reliable for the concerned dataset.

**Link for dataset**

https://www.kaggle.com/lodetomasi1995/income-classification