

|  |
| --- |
| Coventry University  6006CEM Machine Learning and Related Applications  Adult-Census-Income |

**ID: 7982202**

**Document Version: 2**

**Due Date: 05/12/2020**

# Introduction

This goal of this project is to build machine learning models that will predict whether income exceeds $50K/year based on the following data sets. With the proper use of data and machine learning this can help people improve their income by giving them a prediction. A similar study done in 2017 checked level of education, income, unemployment rates to determine wealth (Wolla & Sullivan, 2017).

The dataset has 48842 instances, mix of continuous and discrete (train=32561, test=16281). The attributes in the dataset are:

* age: continuous.
* workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
* fnlwgt: continuous.
* education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
* education-num: continuous.
* marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
* occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
* relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
* race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
* sex: Female, Male.
* capital-gain: continuous.
* capital-loss: continuous.
* hours-per-week: continuous.
* native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinidad & Tobago, Peru, Hong, Holland-Netherlands.

# EDA (Exploratory Data Analysis)

Data was plotted against charts and graphs to find some key insights

* Sex: Out of total male 30% of them earn salary more than 50K while less than 15% female earn more than 50K. 89% female earn less than 50K
* Race: White and asain-pac-Islander earn salary more than 50K
* marital\_status: 41% of married people seem to earn salary greater than 50K.
* Those with a degree which includes a doctorate, prof-school, masters all make a salary more than 50K.
* Out of all the workclass only 59% self-employed people are making salary more than 50K.
* If we check by occupation, Proportion of people making salary less than 50K is higher.

# Data and Pre-processing

The first step was to read the data by using the pandas function read\_csv(). This allows us to open the data file so it can then be placed in graphs, charts, histograms etc. The data was placed in a table and grouped together by category so they can be easily distinguished. The data was also checked for any missing fields, if there are any, they were removed through a process known as imputation. To impute means that any missing attributes are either substituted or removed, this needs to be done because the code cannot process non numerical data in the pipeline. The numerical data was also described by selecting the columns and their numerical attributes so they can be calculated by count, mean, standard deviation, minimum, maximum, etc. The tables purpose helps to illustrate the average data that is being worked upon (Bhandari, 2020).

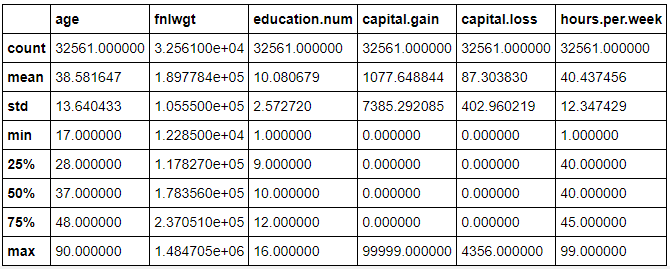


Figure 1: Numerical data

Much like the numerical pipeline, the categorical pipeline also needs to be checked for any missing attributes, if there are any missing then we will need to impute the data so it can be processed. The categorical data was group together by sections, counted, and given a percentage so the makeup of each column was known (Perez, 2020).

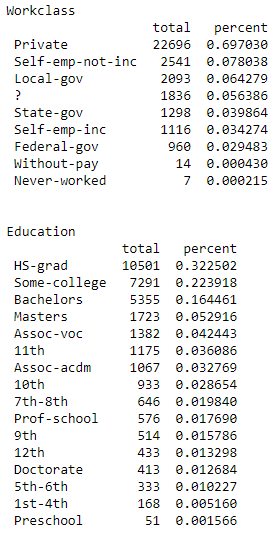


Figure : Categorical data

Now that the data has been checked for null values; values which could be missing, and imputation has taken place it can now be applied to machine learning algorithms. Since this data set falls under the machine learning classification technique the logistic regression algorithm and the random forest algorithm were selected.

## Implementation

Logistic regression and random forest are applied in three ways:

* Using default parameters
* Using Grid Search CV to find best parameters
* Using RFE (recursive feature elimination) to find best features

## Logistic regression

Logistic regression was chosen because the data set chosen falls under classification, a technique in machine learning. Logistic regression is an algorithm which is used to predict outcomes of a target usually between 0 and 1. Firstly before we begin implementing the logistic regression the data set must be checked using a correlation matrix, this is done by plotting the numerical data against each other to see if any of them have a direct affect greater than 0.90. Outliers must also be detected as they can skew the algorithm into producing incorrect results. Once all this has been done the logistic regression algorithm can be implemented. Normal logistic regression achieved an accuracy score of 84.7803%, the accuracy score describes the percentage of correct predictions to incorrect predictions. The precision and recall scores also did quite well each being able to precisely estimate the data and to identify the data. The logistic regression algorithm was also further optimized by using grid search. Grid search is a method of creating and evaluating a model for all the possible combinations. After applying grid search to the logistic regression algorithm, it was determined to have a very small impact. Recursive feature elimination was also used to eliminate features from the data set which would not be needed for training. After applying recursive feature elimination, it was found that the accuracy score for the logistic regression algorithm was increased to 84.8174% (*Machine Learning - Logistic Regression*).

# Random Forest

The second algorithm chosen was the random forest algorithm, much like the logistic regression algorithm it was chosen because the data set falls under the classification technique. The random forest algorithm usually has a higher accuracy when compared to other algorithms (Donges, 2020). The first iteration of the random forest algorithm achieved an accuracy of 84.8174, which is slightly better than the first iteration of the logistic algorithm by 0.000371%. Random forest’s precision score was a 70.4110% which is worse than the logistic algorithm, this means that it was less precise. The recall score however did better than the logistic algorithm by about 3.1%. Grid search cross validation was also applied to the random forest algorithm to further optimize the algorithm using hyperparameters. After applying grid search cross validation, the new accuracy and precision percentages set a new benchmark, however the algorithms ability to recall took a significant decrease. Recursive feature elimination was also used in attempt to further optimize the algorithms accuracy, precision, and recall. After RFE was applied the accuracy and precision percentages declined, unlike the algorithms ability to recall.

# Evaluation

The exhibitions/performances are assessed by accuracy, confusion matrix, precision, recall and f1-score.

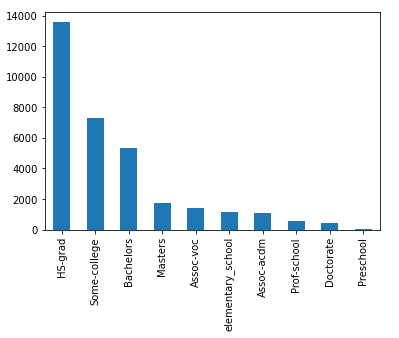


Figure 3: Education

After filtering the data, it by selecting education levels be seen that most adults questioned have a high school diploma followed by some college education.

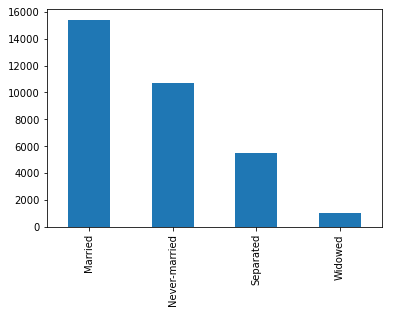


Figure 4: Relationship status

Most participants in this data set are in a relationship seconded by individuals which have never been married. Those separated and widowed combined make up less than half the participants.

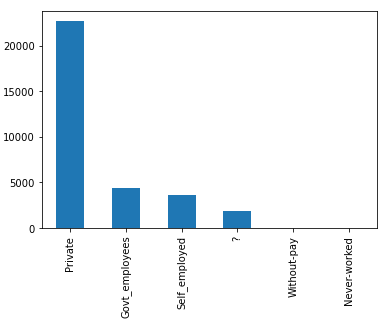


Figure 5: Workforce

An outstanding majority of people work in the private sector followed by government employees and the self-employed. The (?) symbolizes missing information, which has not been provided in the data set.

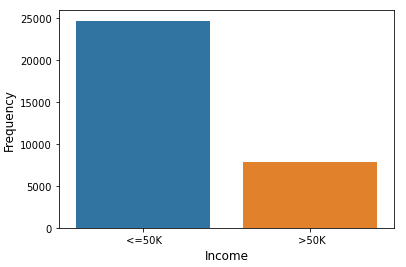


Figure 6: Income Frequency

Less than half of all participants surveyed earn an income greater than 50,000 dollars a year.

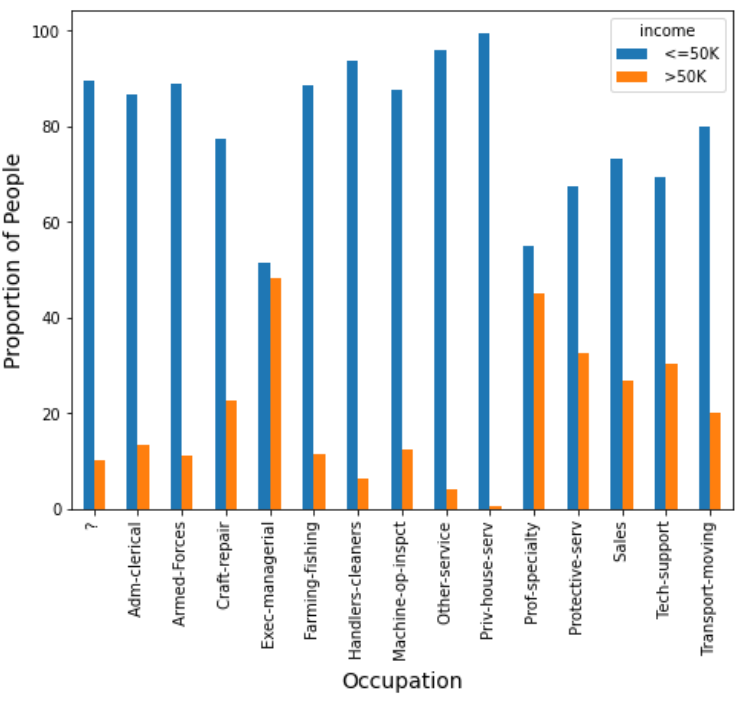


Figure 7: Occupation

According to this graph, those who work in Exec-managerial Are most likely to be earning more then 50,000 dollars a year, it can be assumed that an exec-managerial is an in-demand position. Those working in private house services are the least likely to be earning more then 50,000 dollars a year, it can also be seen that most people work in the priv-house-serv.

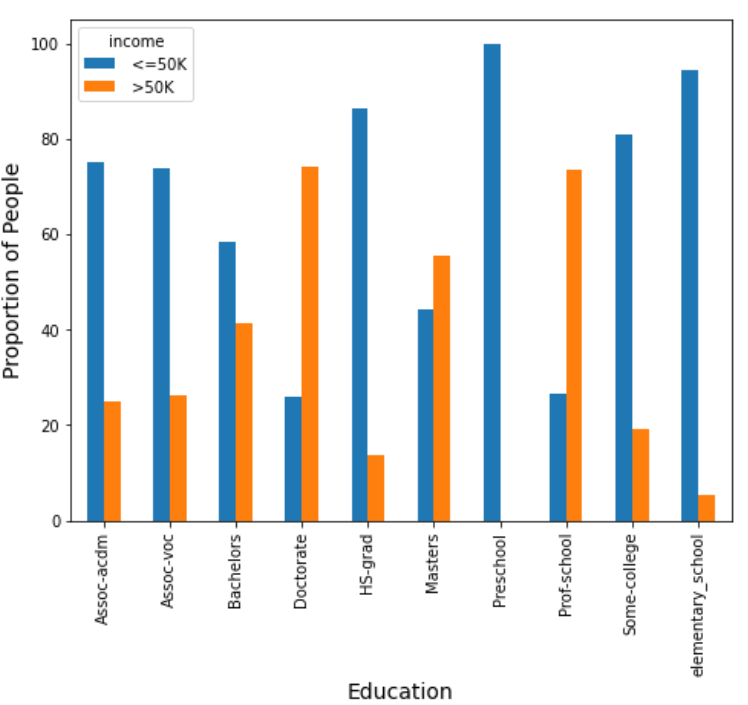


Figure 8: Income based on Education

Those who have a doctorate degree are most likely to be earning more than 50,000 dollars a year, closely followed by those who attended profession school. Those who only have a preschool education do not make more than 50,000 dollars a year.

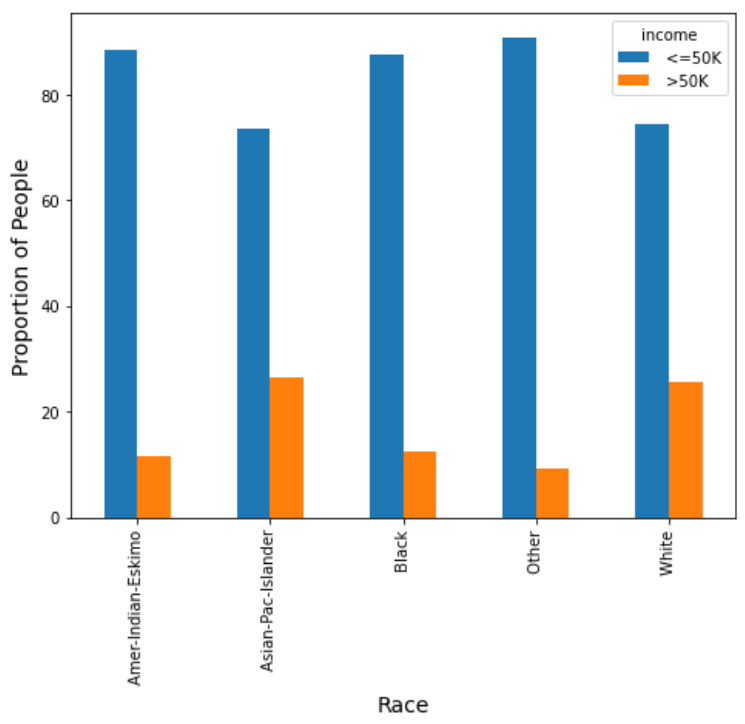


Figure 9: Race

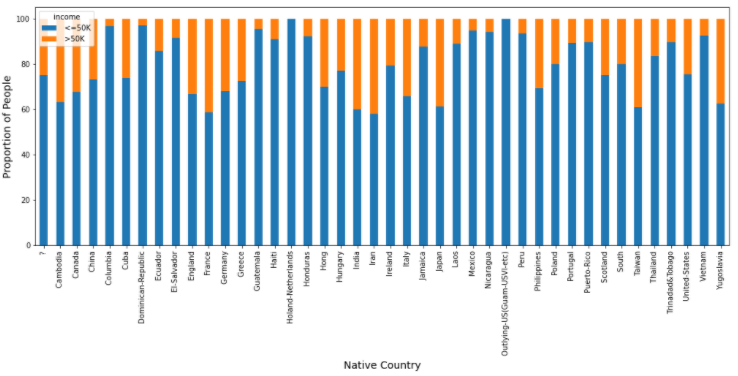
Most participants are classified as Other, meaning their race was not on the survey. Those who are of Asian-Pacific-Islander descent are slightly more likely to earn more than 50,000 dollars a year compared to those who are of White decent. Those who are of Other decent are the least likely to be breaking the 50,000 dollars a year.

Figure 10: Income by Country

Participants from the Netherlands and from an Outlying-US territory did not earn more then 50,000 dollars a year. Those who come from Iran or France were the most likely to earn more than 50,000 dollars a year.

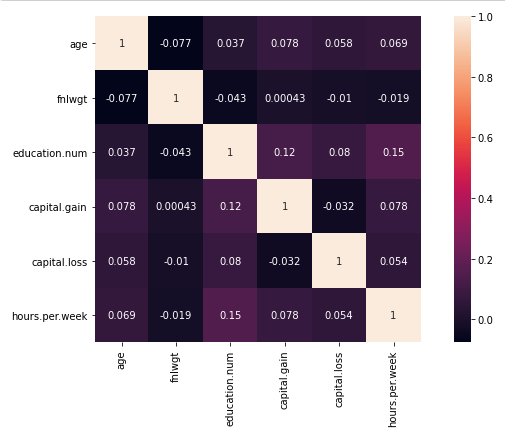


Figure 11: Correlation

Age and working hours per week have a strong correlation as well as capital gain and hours per week. This means that one is almost certainly connected to the other. Capital loss and age has a moderate correlation to one another. Fnlwgt and age have no correlation.

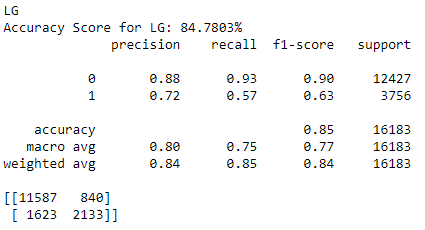


Figure 12: Logistic Regression

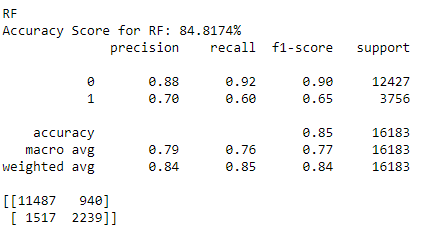


Figure 13: Random Forest

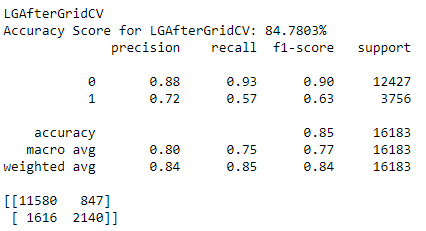


Figure 14: Logistic Regression after GridCV

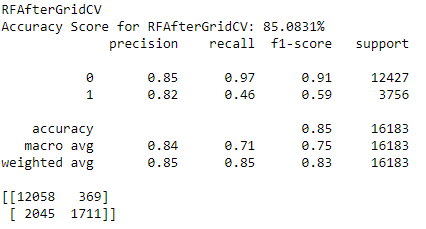


Figure 15: Random Forest after GridCV

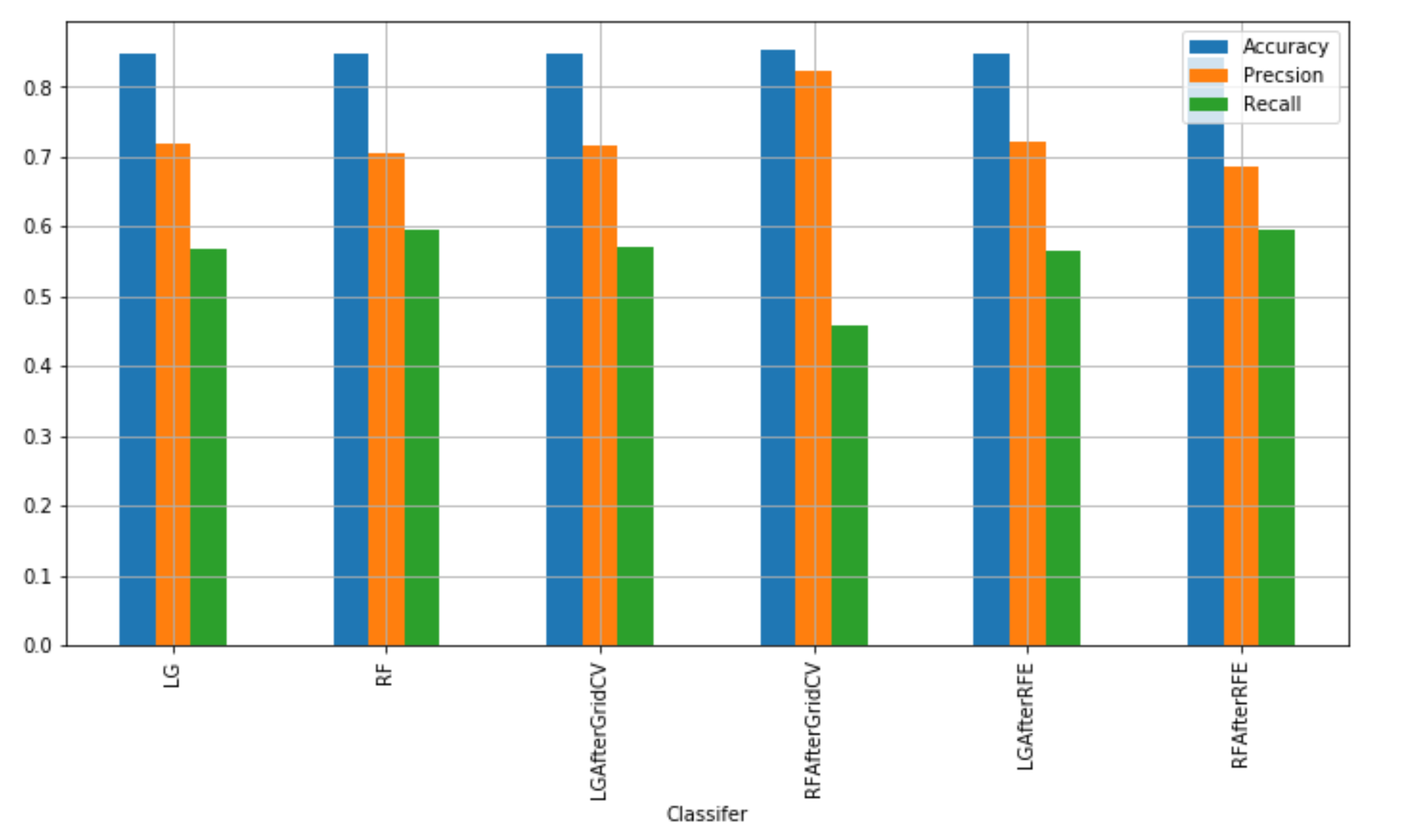


Figure 16: Matrix Evaluations

After applying and evaluating different models, it was found out that all the models performed well. Logistic regression after applying Grid Search CV and RFE performs fractionally better than other models. The models have an accuracy around 85%, which can be improved. The dataset is somewhat unbalanced, i.e. the dataset contains a greater number of rows for the people that have income less than or equal to 50k (50k <= almost 25000 and 50 > almost 8000). In future if the dataset is more balanced, we might have better models.

[Word Count: 1545]

# Bibliography

Data collected from: <http://archive.ics.uci.edu/ml/datasets/Adult>

GitHub repository: <https://github.coventry.ac.uk/akhnoukp/6006CEM_2021s1_7982202_PA>

Anon. (n.d.) *Machine Learning - Logistic Regression* [online]available from <https://www.tutorialspoint.com/machine\_learning\_with\_python/machine\_learning\_with\_python\_classification\_algorithms\_logistic\_regression.htm> [19 November 2020]

Bhandari, A.A. (2020) *Feature Scaling: Standardization Vs Normalization* [online]available from <https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/> [27 November 2020]

Donges, N. (2020) *A Complete Guide to the Random Forest Algorithm* [online]available from <https://builtin.com/data-science/random-forest-algorithm> [21 November 2020]

Perez, G. (2020) *Pre-Process Data Like a Pro: Intro to Scikit-Learn Pipelines* [online]available from <https://towardsdatascience.com/clean-efficient-data-pipelines-with-pythons-sklearn-2472de04c0ea> [27 November 2020]

Wolla, S.A. and Sullivan, J. (2017) *Education, Income, and Wealth* [online]available from <https://research.stlouisfed.org/publications/page1-econ/2017/01/03/education-income-and-wealth/> [27 November 2020]