Adult-Census-Income

# 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.
* People having degree doctorate, prof-school, masters are making 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 Processing

The first step is to convert all the data into a table, for this purpose pandas was used, the function read\_csv(), this function reads the data in comma separated files or csv and translates it into a data frame.

# Numerical Pipeline



The first step in this pipeline is to use impute. Imputation means that the missing values are substituted with associate degree estimate, then analyzing the total information set as if the imputed values were actual ascertained values. There are many methods to impute but, in this case, mean strategy will be used as it can use the same mean and sample size.

The second step in this pipeline is scaling the numerical columns. Feature scaling is a technique used to normalize the series of independent variables or options of knowledge in a column. In processing, it is additionally called data normalization and is mostly performed throughout the information preprocessing step. The data values were between a range of 0 to 1 using MinMaxScaler() function. As there are many values, not all algorithms will work properly so the functions will need to be standardized.. As an example, several classifiers calculate the space between 2 points by the distance. If one in all the options encompasses a broad range of values, the space is going to be ruled by this specific feature. Therefore, the vary of all options should be normalized so that every feature contributes proportionately to the distance (Bhandari, 2020).

# Categorical Pipeline



Categorical data is also handled in 2 steps the first step for missing values and second for encoding the categorical data for use in machine learning.

Handling missing values in categorical data pipeline is also done via imputing, however the method is different. In the case of categorical variables, using mean, median, or zero-imputation does not make any sense as it is not numerical or mathematical. One of the several ways to impute categorical features is to replace missing values with the most frequent class/category. However, it is generally most applicable to impute a missing numeric feature with zeros, generally a categorical feature’s missing-ness itself is effective data that should to be expressly encoded. If this is often the case, most-common-class imputing would cause this data to be lost. Instead, simply replace those values with a price like “Unknown” or “Missing” (Perez, 2020).

In this project since the nature of the categorical data or sense of this data is not clear our choice is to use the impute method that replaces the null values with most frequent classes.

The second step is encoding all the categorical data, since machine learning is predicated on mathematical equations or functions, it might cause a retardant once we keep categorical variables as is. Several algorithms support categorical values while not additional manipulation, however in those cases, it is still a subject of debate on whether to cipher the variables or not. The algorithms that do not support categorical values, in that case, square measure left with coding methods.

Our choice in this method is One Hot Encoder, in this strategy, we map every classification to a vector that contains 1 and 0 signifying the nearness of the component or not. The quantity of vectors relies upon the classes which we need to keep. For high cardinality includes, this technique creates a great deal of sections that hinders the adapting. There is a buzz between one hot encoding and sham encoding and when to utilize one. They are a lot of the same aside from one hot encoding produces the quantity of segments equivalent to the quantity of classifications and sham creating is one less.

# Transform data

After defining the pipelines for both categorical and numerical data we fit the functions defined in the pipelines and then transform this data and store it in the data. Sounds difficult but made simple by using a simple function fit\_transform(), this single function performs both these steps.

## Logistic regression

Logistic regression is the fitting regression investigation to direct when the needy/dependent variable is dichotomous meaning binary like 0, 1 etc. Like all regression analyses, the calculated regression is a predictive analysis. Calculated regression is utilized to depict information and to clarify the connection between one ward binary variable and at least one ostensible, ordinal, stretch or proportion level autonomous variables (*Machine Learning - Logistic Regression*).

## Random Forest

The fundamental goal of the production of a decision tree is to manufacture a training model. This training model is utilized to anticipate/predict the worth or class of the recipient variable. The degree of comprehension of the decision trees algorithms is a lot simpler than the other classification algorithms (Donges, 2020).

In the random forest classifier, each decision tree predicts a reaction for an event and the endmost reaction is chosen through voting.

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

# Evaluation

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

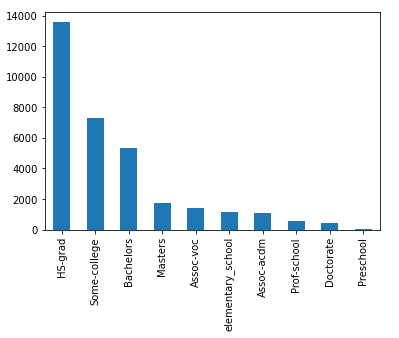


Figure 1: Education

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.

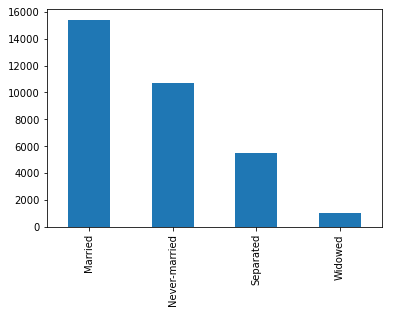


Figure 2: Relationship status

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.

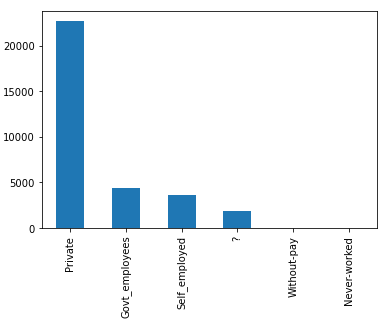


Figure 3: Workforce

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.

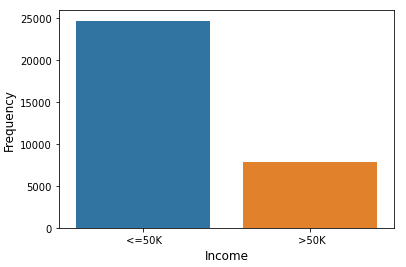


Figure 4: Income Frequency

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

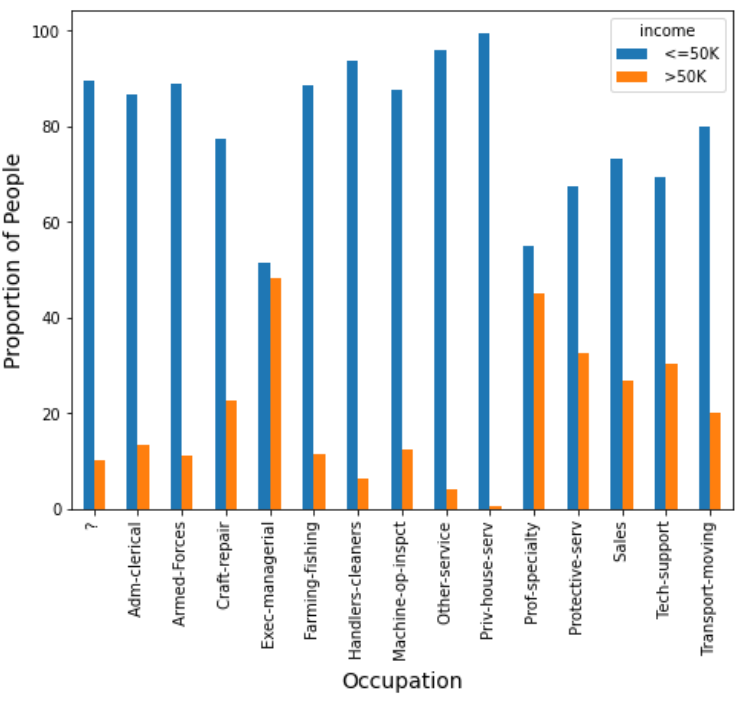


Figure 5: Occupation

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.

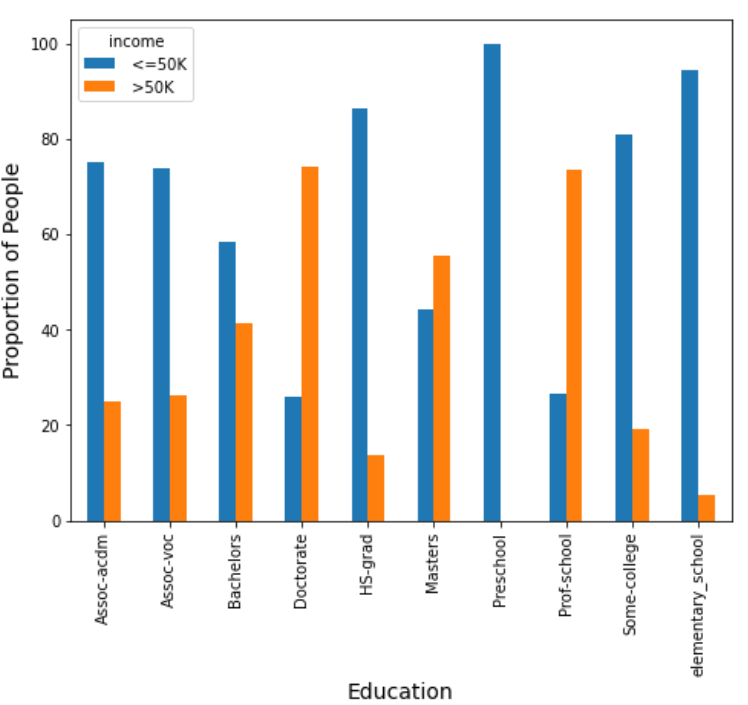


Figure 6: Income based on Education

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.

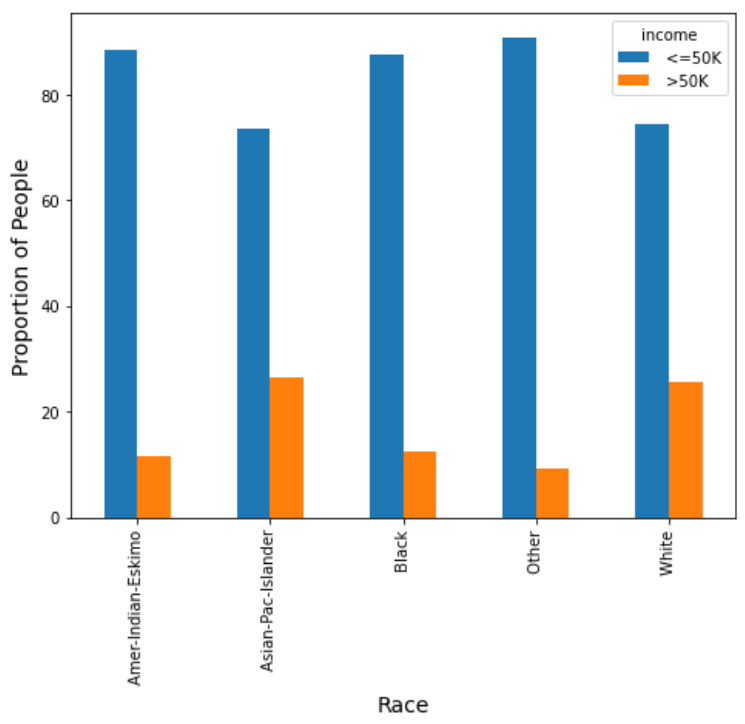


Figure 7: Race

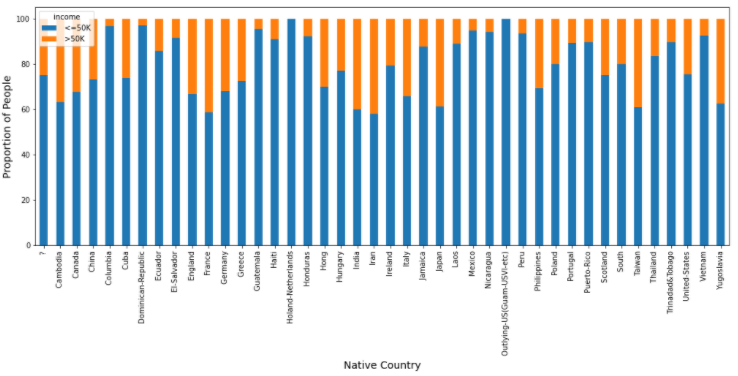
Most participants are classified as Other, meaning their race was not on the survey. Those who are of Asian-Pacific-Islander descent are slightly more likely to earn more than 50,000 dollars a year compared to those who are of White decent. Those who are of Other decent are the least likely to be breaking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

Figure 8: 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.

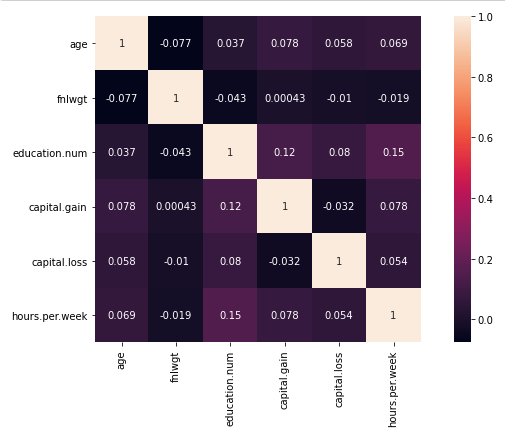


Figure 9: Correlation

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.

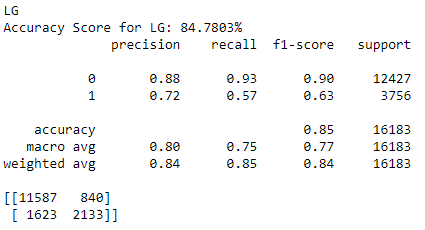


Figure 10: Logistic Regression

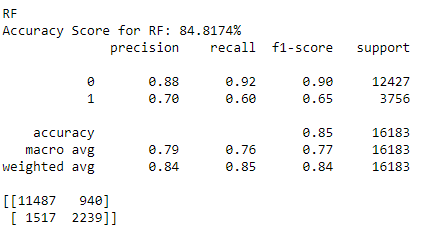


Figure 11: Random Forest

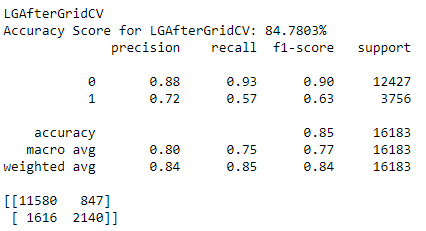


Figure 12: Logistic Regression after GridCV

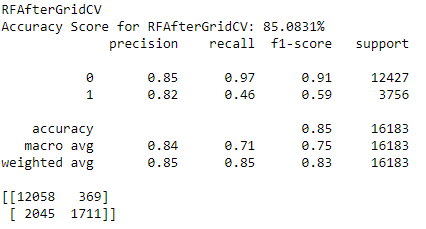


Figure 13: Random Forest after GridCV

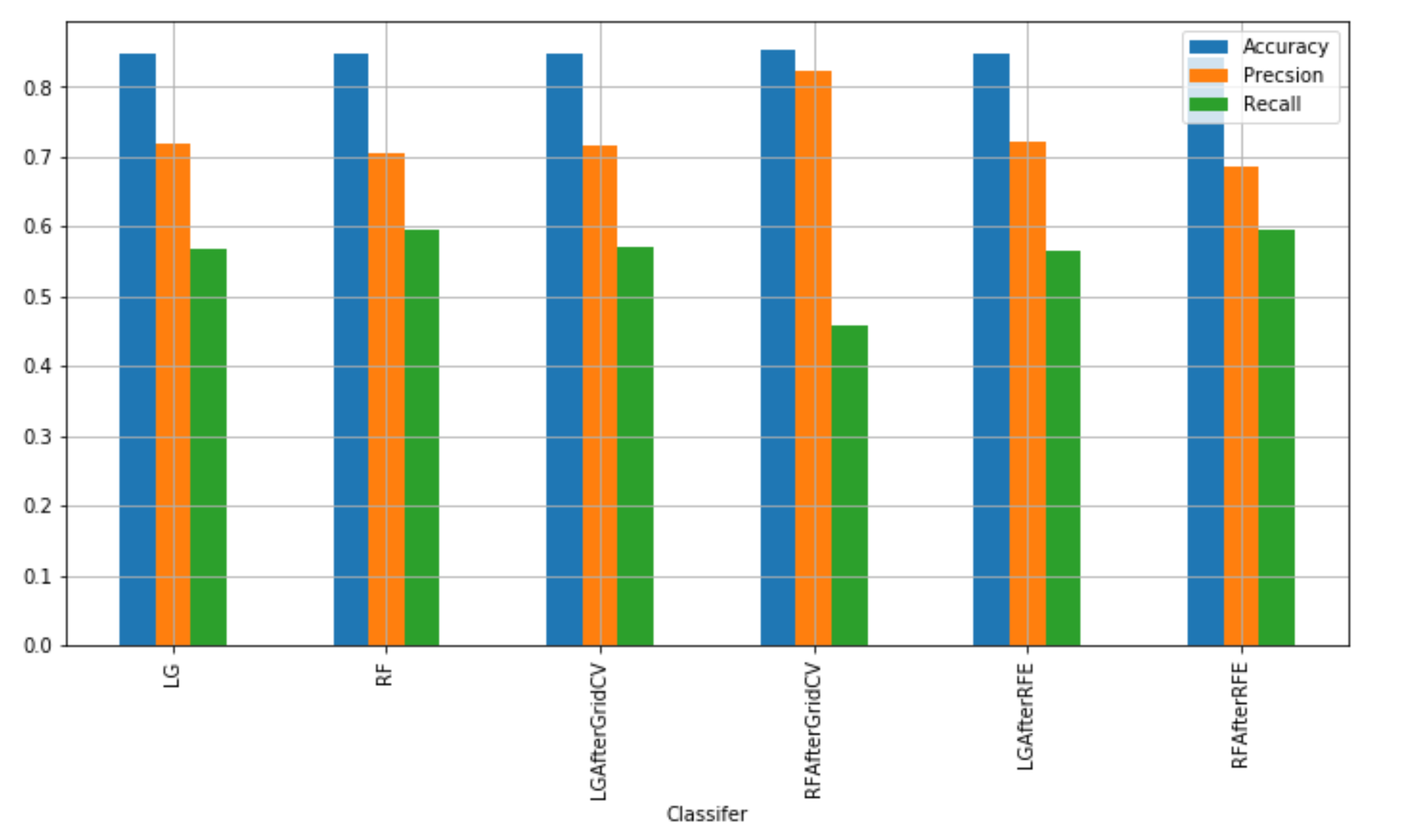


Figure 14: Matrix Evaluations

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: 1650]

# 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]