**CENSUS INCOME DATA ANALYSIS**

***Tanmayi Vetukuri***



Table of Contents

[1 Background 2](#_Toc7990072)

[2 Objective 2](#_Toc7990073)

[2.1 Dataset 3](#_Toc7990074)

[2.2 Techniques Used 5](#_Toc7990075)

[3 Exploratory Data Analysis & Data Manipulation: 5](#_Toc7990076)

[3.1 Initial data loading and data manipulation: 5](#_Toc7990077)

[3.2 Exploratory Data Analysis 6](#_Toc7990078)

[3.2.1 Geographical distribution of people across different countries in the dataset 6](#_Toc7990079)

[3.2.2 Distribution of Income 7](#_Toc7990080)

[3.2.3 Correlation between different numerical variables using correlation matrix 7](#_Toc7990081)

[3.2.4 Distribution of Education 8](#_Toc7990082)

[3.2.5 Distribution of Gender & Race 9](#_Toc7990083)

[3.2.6 Distribution of Age 11](#_Toc7990084)

[3.2.7 Pairplot between weeks worked per year & Age 12](#_Toc7990085)

[3.2.8 Stacked Bar Graph of Race & Gender vs Income 12](#_Toc7990086)

[3.2.9 Economic Inequalities between male & female 13](#_Toc7990087)

[4 Model Evaluation 18](#_Toc7990088)

[4.1 Logistic Regression 19](#_Toc7990089)

[4.2 K- Nearest Neighbours 22](#_Toc7990090)

[4.3 Linear Discriminant Analysis 25](#_Toc7990091)

[4.4 Random Forest 27](#_Toc7990092)

[4.5 Decision Tree 29](#_Toc7990093)

[5 Model Comparison 31](#_Toc7990094)

[6 Conclusion 32](#_Toc7990095)

[7 How Analysed Data Can Be Used ? 32](#_Toc7990096)

# Background

The prominent inequality of income based on gender is a huge concern especially in the United States. The principle of universal moral equality ensures sustainable development and improve the economic stability of a nation. Governments in different countries have been trying their best to address this problem and provide an optimal solution. This study aims to show the usage of machine learning and data mining techniques in providing a solution to the income equality problem. The UCI Adult Dataset has been used for the purpose. Classification has been done to predict whether a person's yearly income in US falls in the income category of either greater than 50K Dollars or less equal to 50K Dollars category based on a certain set of attributes.

# Objective

The overall objective of the project is to determine whether the income level of a person exceeds over 50K per year, how income is controlled by race, age and sex and how the above factors discriminate female’s role in employment. This has been achieved by performing extensive exploratory data analysis on the Adult Census dataset taken from UCI Machine Learning repository and finding some interesting relations and insights from the data. I applied various classification and regression techniques which included Logistic regression, KNN, Linear Discriminant Analysis, Random Forest, Decision Tree models and compared the results of the techniques to find the best model.

## Dataset

**A. Data overview**

The dataset used in this project has 199,523 records and a binomial label indicating a salary of <50K or >50K USD. 94% of the records in the dataset have a class label of <50K. The data has been divided into a training set containing 133,680 records and a test dataset containing 65,843 records.

There are 40 attributes consisting of thirty three nominal and seven continuous attributes. The employment class describes the type of employer such as self-employed or federal and occupation describes the employment type such as farming, clerical or managerial. Education contains the highest level of education attained such as high school or doctorate. The relationship attribute has categories such as unmarried or husband and marital status has categories such as married or separated. The other nominal attributes are country of residence, gender and race. The continuous attributes are age, hours worked per week, education number (numeric representation of the education attribute), capital gain and loss.

1. **B. Attributes**

Response Variable - Income: <=50K (94%), >=50K (6%)

|  |  |
| --- | --- |
| AAGE | Age classification is based on the age of the person at his/her last birthday. The adult universe(i.e., population of marriageable age) is comprised of persons 15 years old and over for March supplement data and for CPS labor force data. |
| ACLSWKR - class of worker | This refers to the broad classification of the person's employer. On the March file, these broad classifications for current jobs are private, government, self-employed, without pay, and never worked. Private and government workers are considered "wage and salary workers;" this classification scheme includes self-employed, incorporated persons in with "private" workers. For the longest job held last year, this class of worker scheme includes private; government by level/Federal, State, and local; self-employed incorporated, self-employed unincorporated or farm; and without pay. The wage and salary category for longest job held includes private, government (all levels), and self-employed incorporated. |
| ADTIND | Industry code where the person worked |
| ADTOCC | Occupation code of the person |
| AGI | Adjusted gross income |
| AHGA | Highest education level |
| AHRSPAY | Wage per hour |
| AHSCOL | Enrolled in educational institute last week |
| AMARITL | Marital status |
| AMJIND | Major industry code where the person worked |
| AMJOCC | Major occupation code of the person |
| ARACE | Race of the person |
| AREORGN | Hispanic origin |
| ASEX | Sex/Gender |
| AUNMEM | Member of a labor union |
| AUNTYPE | Reason for unemployment |
| AWKSTAT | Full or part time employment status |
| CAPGAIN | Capital gains in dollars |
| CAPLOSS | Capital Losses |
| DIVVAL | Dividends from stocks |
| FEDTAX | Federal income tax liability |
| FILESTAT | Tax filer status |
| GRINREG | Region of previous residence |
| GRINST | State of previous residence |
| HHDFMX | Detailed household and family status |
| HHDREL | Detailed household summary in household |
| MARSUPWT | March supplement final weight |
| MIGMTR1 | Migration code change in MSA |
| MIGMTR3 | Migration code change in REG |
| MIGMTR4 | Migration code move within REG |
| MIGSAME | Lived in this house 1 year ago |
| MIGSUN | Migration previous res in sunbelt |
| NOEMP | Number of persons worked for employer |
| PARENT | Family members under 18 |
| PEARNVAL | Total personal earnings |
| PEFNTVTY | Country of birth for father |
| PEMNTVTY | Country of birth for mother |
| PENATVTY | Country of birth self |
| PRCITSHP | Citizenship |
| PTOTVAL | Total personal income |
| SEOTR | Own business or self employed |
| TAXINC | Taxable income amount |
| VETQVA | Fill questionnaire for veteran's admin |
| VETYN | Veterans benefits |
| WKSWORK | Number of weeks worked in year |

## Techniques Used

I used techniques like correlation matrix to find the correlation matrix between the numerical variables in the dataset, extensive exploratory data analysis is carried out with various plots like bar plots, correlation plots, strip plot, pairs plot etc. Various regression techniques like Logistic regression, KNN, Linear Discriminant Analysis, Random Forest, Decision Tree models and compared the results of the techniques to find the best model. Then a comparison is drawn based on the confusion matrix and accuracy of the model and the best model is then chosen.

# Exploratory Data Analysis & Data Manipulation:

## Initial data loading and data manipulation:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Columns** | **FATHER** | **MOTHER** | **SELF** | **GRINST** |
| No. missing values | 6713 | 6119 | 3393 | 708 |

* I removed following columns from the dataset - 'MIGMTR1', 'MIGMTR3', 'MIGMTR4','MIGSUN','\*(pred)','AGI’,'AHRSPAY'
* These columns had more than 95 thousand missing values per column
* Removing records corresponding to these columns would result in lesser number of dataset observations.
* Imputing these many records with mean, mode etc. would make data biased.
* Predicting these many records would make final income predictions unreliable
* Imputed mode values for above columns with categorical data
* Imputed mode values in place of irrelevant values already present in the data.
* Created dummy variables for all categorical fields for classification results.
* Split the data into training and testing for model evaluation

A screenshot of a cell phone

Description generated with very high confidence

## Exploratory Data Analysis

### 3.2.1 Geographical distribution of people across different countries in the dataset

A close up of a map

Description generated with high confidence

The above map shows population distribution of people in the dataset across the world. As we can see, about 180,000 of people are from the United States and the rest 20,000 people have migrated to the US and hence have been included in the dataset with migration variables which we further dropped.

### 3.2.2 Distribution of Income

The Pie Chart above depicts the distribution of income across people in the dataset. We can see that about 94% of the people in the dataset have income less than 50K and 6% of people in the dataset have income greater than 50K

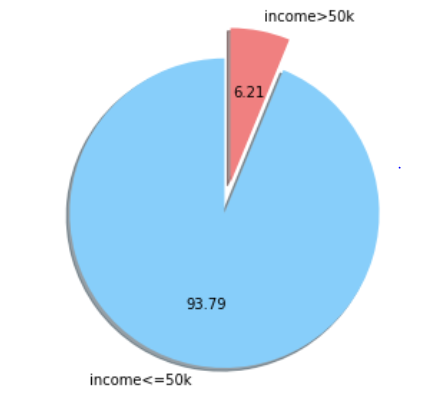


Figure 3.2.1 Pie Chart showing Income Distribution

### 3.2.3 Correlation between different numerical variables using correlation matrix

The Correlation is the association between two or more variables. The correlation matrix is used to investigate the dependence between multiple variables at the same time. The result is a table containing the correlation coefficients between each variable and the others.

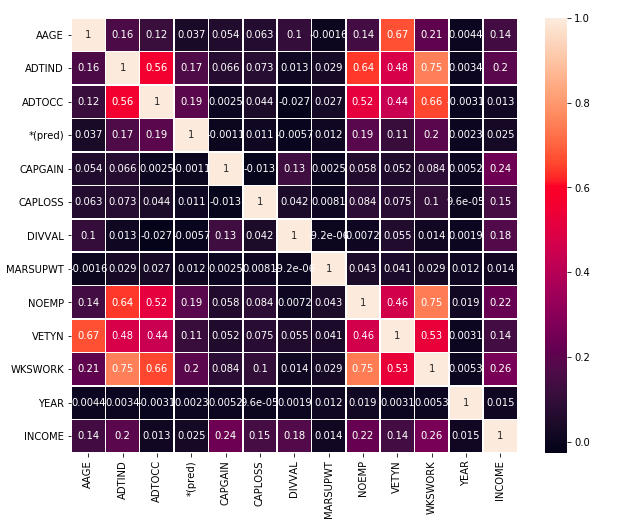


Figure 3.2.3 Correlation matrix using heat map

Each cell in the table shows the correlation between two variables. From the heat map we can see that not many variables are correlated except weeks worked in year (WKSWORK) with number of persons worked for employer (NOEMP) and Industry code (ADTIND)

### 3.2.4 Distribution of Education

The bar graph below shows that majority of people are high school graduates. There are almost 48k high school graduates in the dataset.

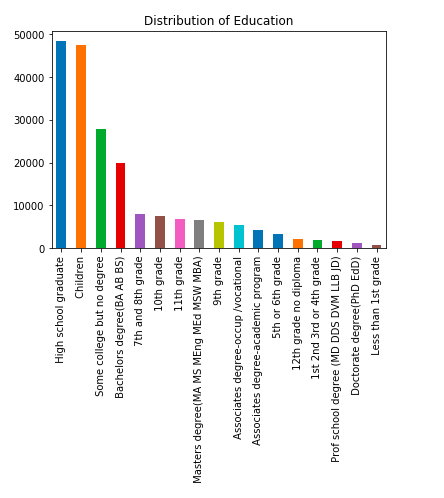
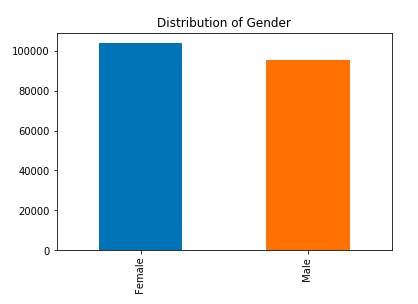


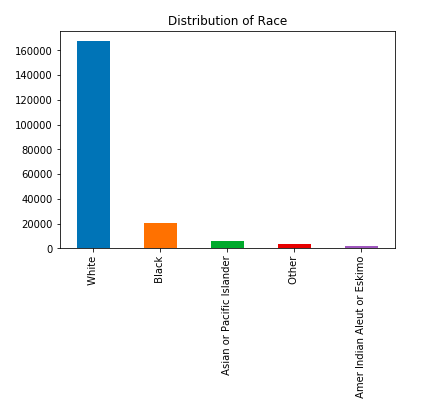
Figure 3.2.4 Distribution of Education

### 3.2.5 Distribution of Gender & Race

The bar graph shows the distribution of gender & race. Most of the people are whites and the distribution of gender is almost equal between females and males



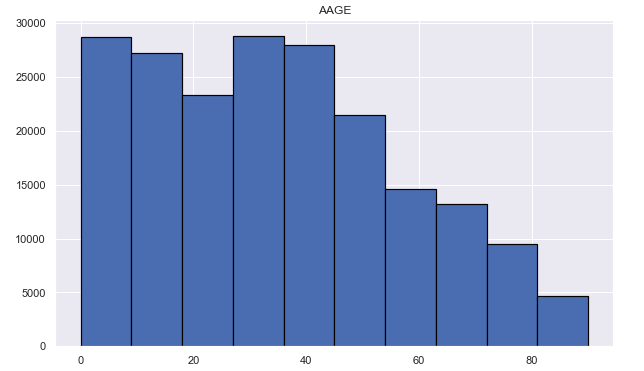
**Figure 3.2.5 Distribution of Gender**



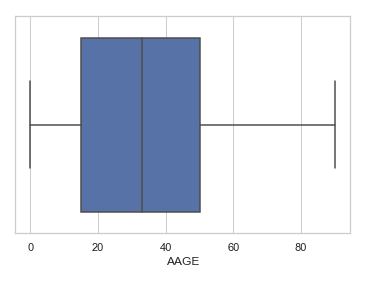
**Figure 3.2.5 Distribution of Race**

### 3.2.6 Distribution of Age

The below graphs show distribution of age. From the boxplot we can infer that 50 percentage of the people are between age 18 – 50 years. The median age in the dataset is around 36



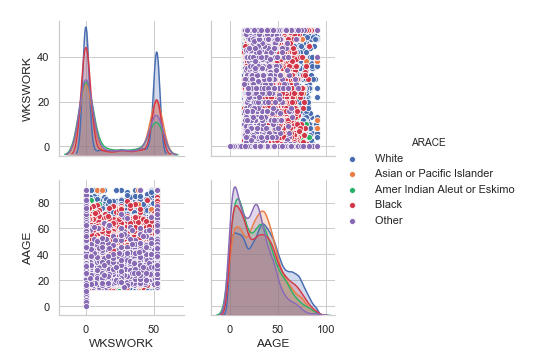
**Figure 3.2.6 Histogram of Distribution of Age**



**Figure 3.2.6 Boxlot showing Distribution of Age**

### 3.2.7 Pairplot between weeks worked per year & Age

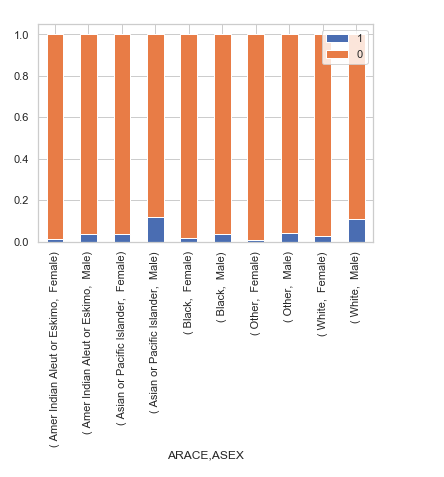
The pairs plot builds on two basic figures, the density and the scatter plot. The density on the diagonal allows us to see the distribution of a single variable while the scatter plots on the upper and lower triangles show the relationship between two variables. For example, the left-most plot in the second row shows the scatter plot of weeks worked per year versus age.



**Figure 3.2.6 Pairplots showing Distribution of Weeks worked per year & Age**

### 3.2.8 Stacked Bar Graph of Race & Gender vs Income

This is an example of creating stacked bar graphs which shows the income based on gender & race. From the graph we can see that males earn more than 50k $ per annum whereas most of the females earn<50K $ per annum.



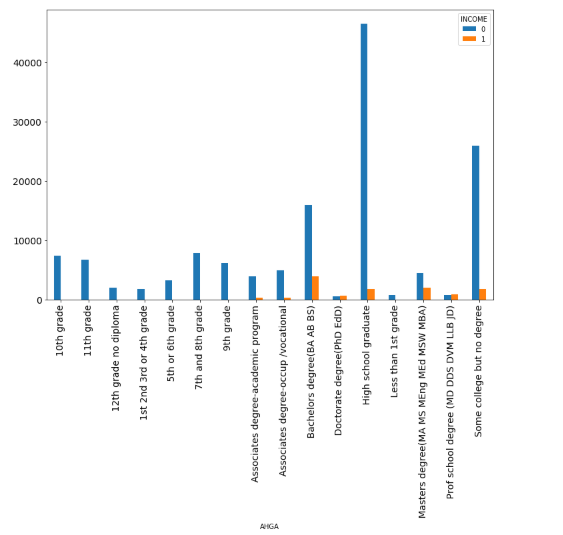
**Figure 3.2.6 Stacked Bar for Race and Gender Versus Income**

### 3.2.9 Economic Inequalities between male & female

In this entry we focus on visualizing data on economic inequalities between men and women during the years 1994 – 1995 with concise explanations.

**a. The Graph displaying income division across Education Levels**

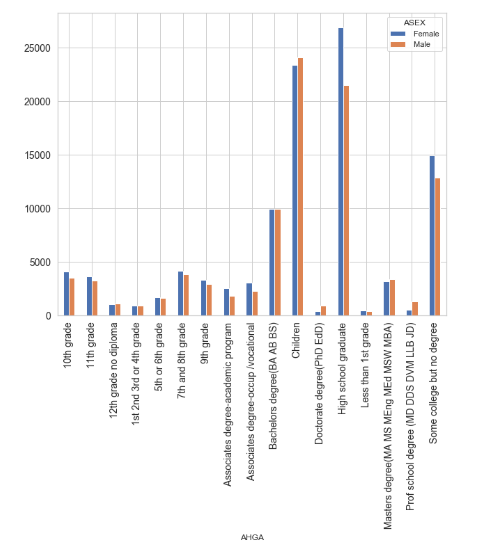
The graph shows that if a person has a Bachelor’s Degree, then there is a high probability of earning income more than 50k



**Figure 3.2.7 (a) Graph displaying Income across education levels**

**b. Graph displaying the Distribution of Male and Female Across the education levels**

The below plot gives us a clear picture on how the gender is distributed across school level education to doctorate level. We notice that women and men are distributed equally across certain categories like 1st,2nd and 3rd Grades, 12th grade diploma, bachelor’s degree

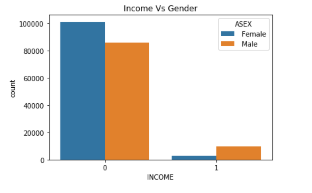


**Figure 3.2.7 (b) Graph displaying Gender across education levels**

**c. Graph displaying Income Distribution based on Gender**

The bar graph shows distribution of number of people who earn income between male and female and we notice that male has an upper hand with income> 50k.

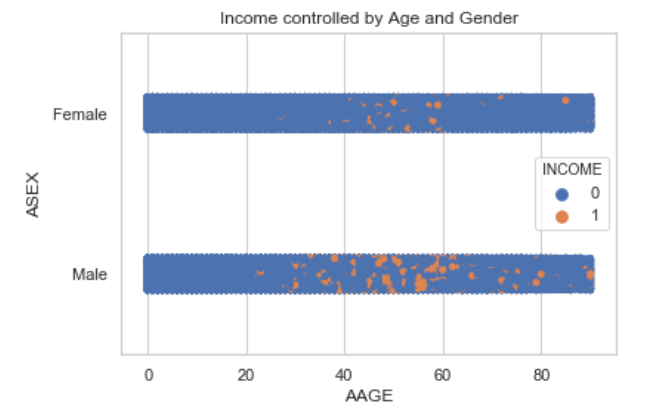
From the above graph, we notice that women and men are in equal numbers across various education levels and we still find a difference in wages which justifies the economic inequality by gender



**Figure 3.2.7 (c) Graph displaying Income distribution based on gender**

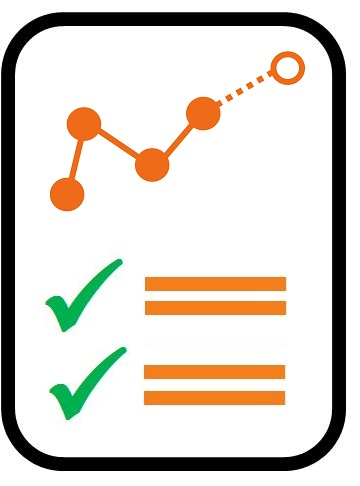
**d. Plot displaying how income is controlled by age and gender**

There is a greater distribution of income>50k among males between age 35 – 60 and there is a moderate distribution of income>50k among females between age 40 – 60 which clearly shows that gender pay gap is larger for older workers



**Figure 3.2.7 (d) Graph displaying Income across age and gender**

# Model Evaluation

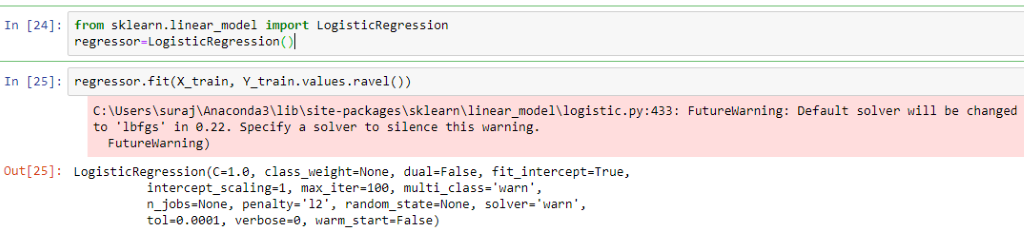
****

* **Logistic Regression**
* **K Nearest Neighbors**
* **Linear Discriminant Analysis**
* **Random Forest**
* **Decision Trees**

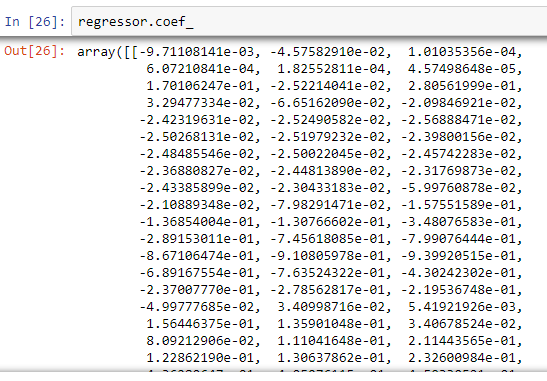
## Logistic Regression

After cleaning the data and splitting the data into training and testing, I used the training data to run a logistic regression on our model. Since the dependent variable is Binary and we are predicting the income whether it is greater than 50k or not. We are able to classify our data based on various features and using these features we are able to predict the income of the population in the data whether they earn more than 50k per year or less than that.

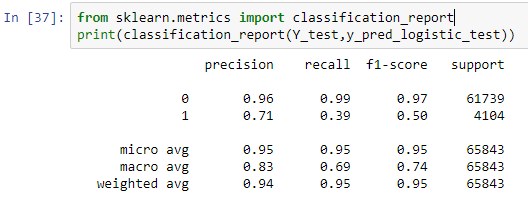
I used sklearn package in order to run the logistic model

****

I checked for regression coefficients, their shape and also the intercept values in order to derive meaningful insights.

****

Then I created predictions which will store the predicted values of our testing set.

****

The precision is the ratio tp / (tp + fp) where tp is the number of true positives and fp the number of false positives. The precision is intuitively the ability of the classifier not to label as positive a sample that is negative.

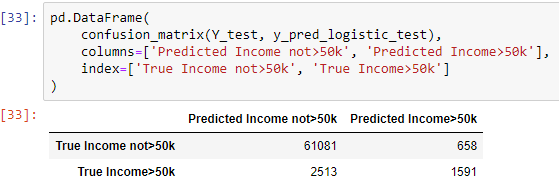
The recall is the ratio tp / (tp + fn) where tp is the number of true positives and fn the number of false negatives. The recall is intuitively the ability of the classifier to find all the positive samples.

The F-beta score can be interpreted as a weighted harmonic mean of the precision and recall, where an F-beta score reaches its best value at 1 and worst score at 0.

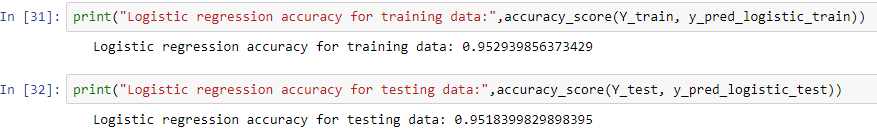
The F-beta score weights recall more than precision by a factor of beta. beta == 1.0 means recall and precision are equally important.

The support is the number of occurrences of each class in y\_true.

We can see the confusion matrix which represents the misclassifications as well as correctly classified observation count. In order to make this confusion matrix informative I created my own confusion matrix by creating a  dataframe.

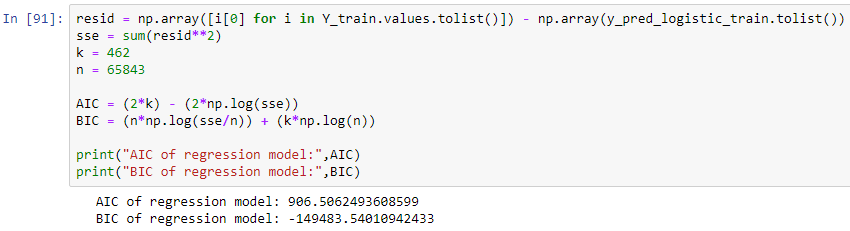
****

Here are the training and testing accuracies of the Logistic Regression model –

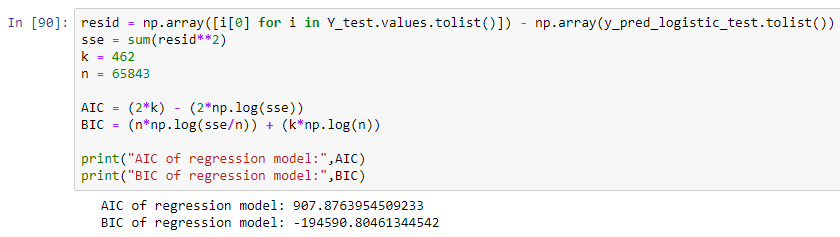
****

To further evaluate our model, I used AKAIKE INFORMATION CRITERION and BAYESIAN INFORMATION CRITERION.

Here are the AIC and BIC values of the Logistic Regression model for both training and testing data –

****

The lower the AIC & BIC values the better is the model and therefore we can say the model stands good for now and will compare these values when we run more models to get the best possible model.

****

## K- Nearest Neighbours

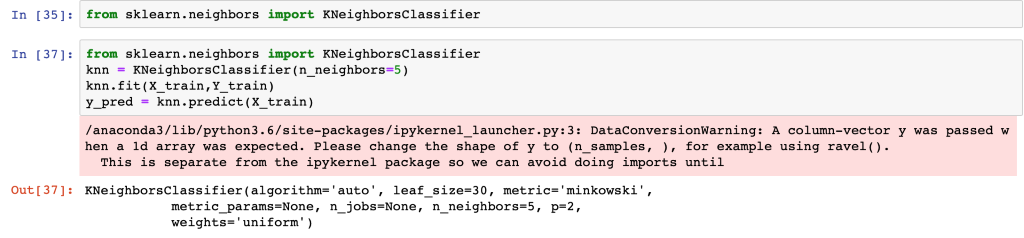
KNN is a non-parametric algorithm, which means that it does not make any assumptions on the underlying data distribution. KNN is a type of lazy learning where the function is only approximated locally and all computation is feferred until classification.

It is based on feature similarity, which means how closely out-of-sample features resemble our training set, determines how we classify a given data point.

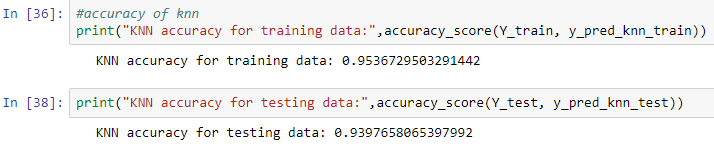
Why did we choose k=5 as we used hit and trial method to get the best possible results from our K-nearest neighbours model.

Small K values gave us higher influence on the results

Large K-values will make it computationally expensive.

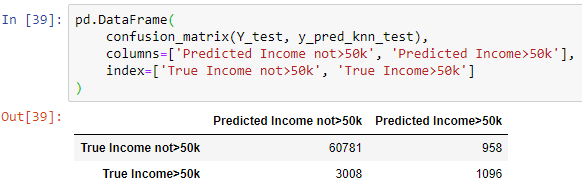
****

Here are the training and testing accuracies of our KNN model –

****

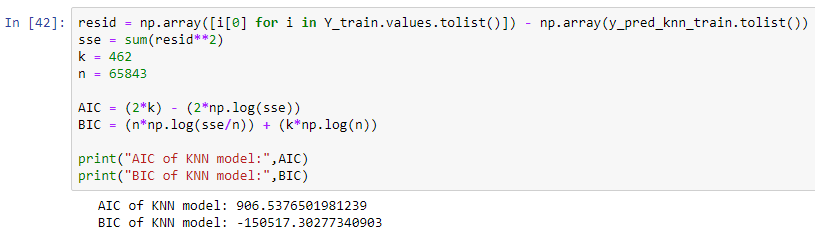
The Test set accuracy came out to be lesser than the training accuracy as the test set data was never seen before  and we can see that the test set results are significant.

Here is the confusion matrix of KNN model -

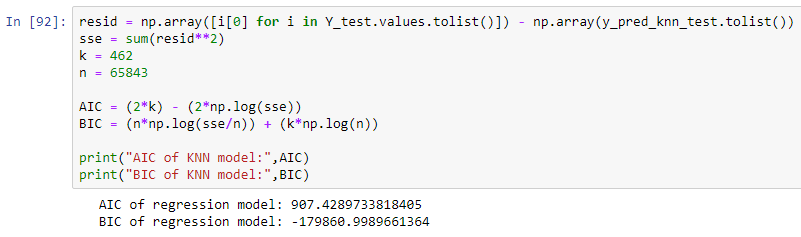
****

We can see that the overall error rate is less as there are very few misclassifications which leads to lesser overall error rate. We check for sensitivity and specificity values for knn model and compare it with other models.

Here are the AIC and BIC values of KNN model for both training and testing data –

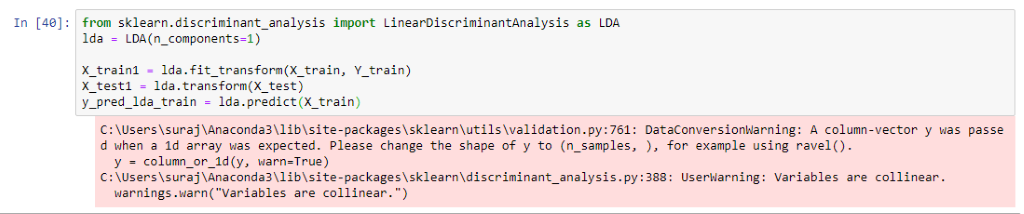
****

The lower the AIC & BIC values the better is the model and therefore we can say our model stands good for now and we will compare these values when we run more models to get the best possible model.

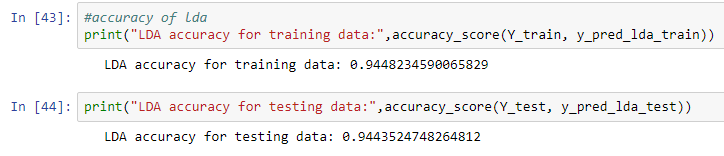
****

## Linear Discriminant Analysis

LDA is a very common technique used for supervised classification problems. LDA is a dimension reduction technique used as a preprocessing step in Machine Learning and pattern classification applications.

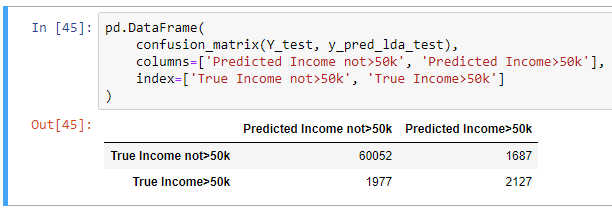
****

Here are the training and testing accuracies of LDA model –

****

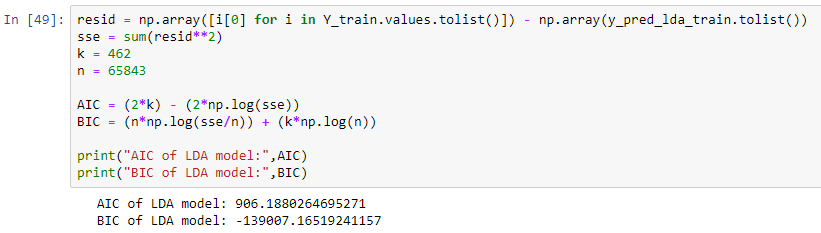
The Test set accuracy came out very close to the training accuracy and we can see that the test set results are significant.

Here is the confusion matrix of our LDA model –

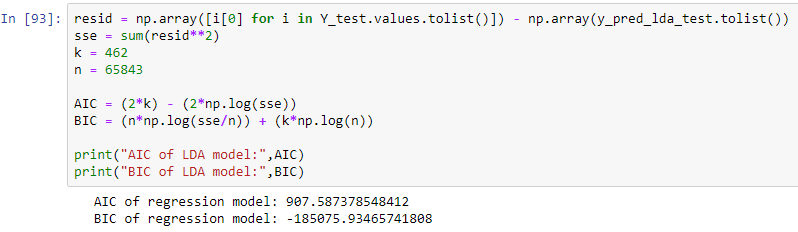
****

We can see that the overall error rate is less as there are very few misclassifications which leads to lesser overall error rate. I checked for sensitivity and specificity values for LDA model and compare it with other models.

Here are the AIC and BIC values of LDA model for both training and testing data –

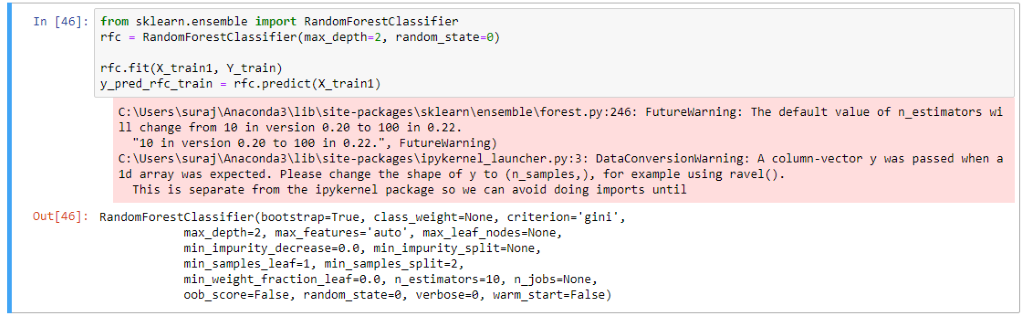
****

The lower the AIC & BIC values the better is the model and therefore we can say the model stands good for now and we will compare these values when we run more models to get the best possible model.

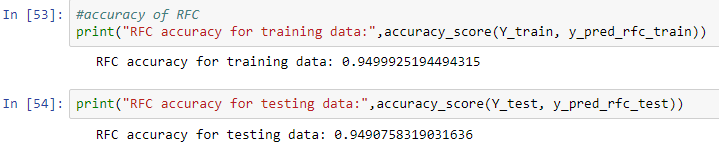
****

## Random Forest

Random Forest or Random Decision Forest is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

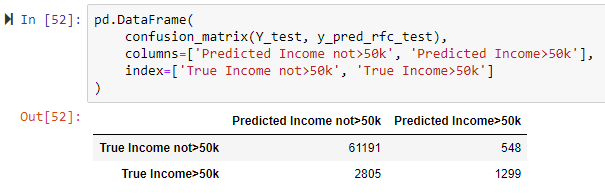
****

Here are the training and testing accuracies of our Random Forest model –

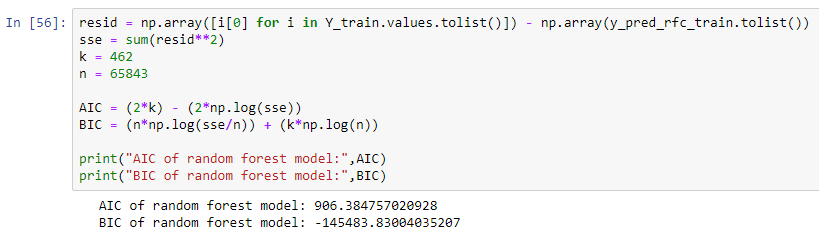
****

The Test set accuracy came out very close to the training accuracy and we can see that the test set results are significant.

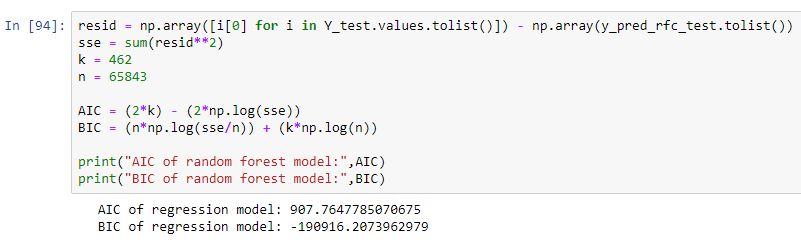
Here is the confusion matrix of Random Forest model –

****

Here are the AIC and BIC values of our Random Forest model for both training and testing data -

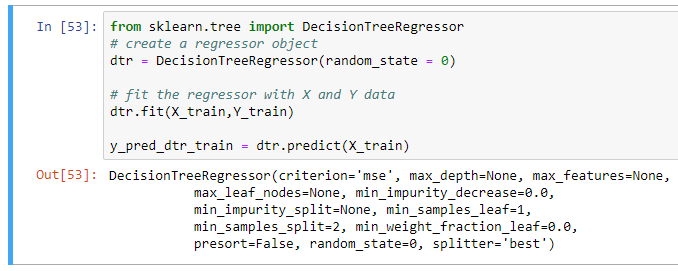
****

The lower the AIC & BIC values the better is the model and therefore we can say our model stands good for now and we will compare these values when we run more models to get the best possible model.

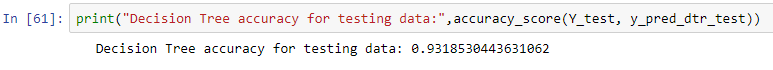
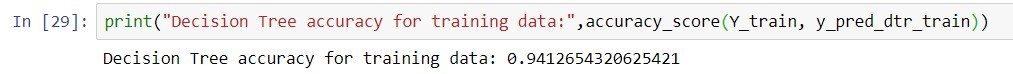
****

## Decision Tree

A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs and utility.

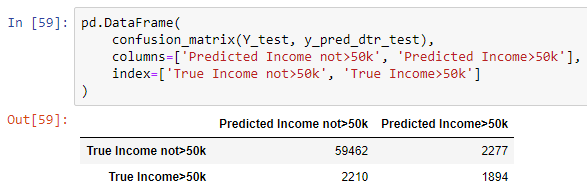
****

Here are the training and testing accuracies of Decision Tree model –

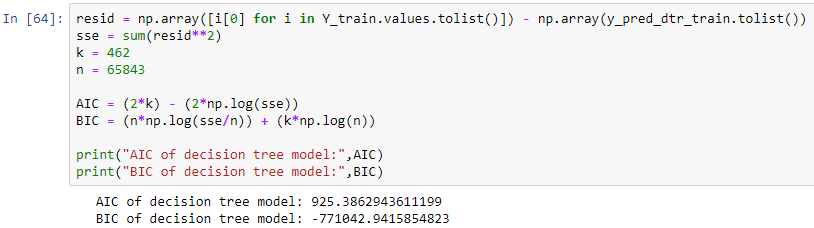
****

The Test set accuracy came out comparatively less close to the training accuracy and we can see that the test set results are significant.

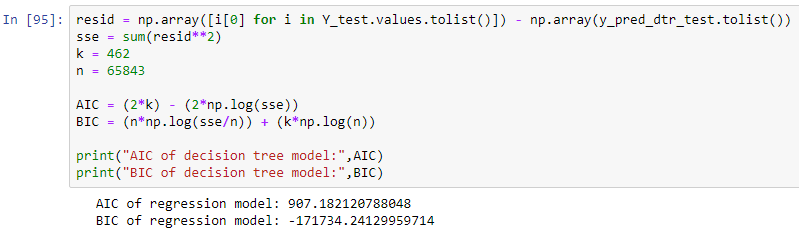
Here is the confusion matrix of Decision Tree model -

****

Here are the AIC and BIC values of Decision Tree model for both training and testing data –

****

The lower the AIC & BIC values the better is the model and therefore we can say the model stands good for now and we will compare these values when we run more models to get the best possible model.

****

# Model Comparison

|  |  |
| --- | --- |
| **MODEL** | **ACCURACY VALUES** |
| LOGISTIC REGRESSION | 95.1839 % |
| K NEAREST NEIGHBORS | 93.9765 % |
| LINEAR DISCRIMINANT ANALYSIS | 94.4352 % |
| RANDOM FOREST | 94.9075 % |
| DECISION TREES | 93.1853 % |

**Table 5.1 : Accuracy Comparison**

In terms of Accuracy we can see that Logistic provides the best accuracy amongst the models in **Logistic model.**

|  |  |
| --- | --- |
| **Models** | **AIC Values** |
| LOGISTIC REGRESSION | 907.8763 |
| K NEAREST NEIGHBORS | 907.4289 |
| LINEAR DISCRIMINANT ANALYSIS | 907.5873 |
| RANDOM FOREST | 907.7647 |
| DECISION TREES | 907.1821 |

**Table 5.2 : AIC Comparision**

Here I compared the AIC values and we can say that **Decision Trees** provides the lowest AIC values among all models

|  |  |  |
| --- | --- | --- |
| **Model** | **Sensitivity** | **Specificity** |
| LOGISTIC REGRESSION | 96.05% | 70.74 % |
| K NEAREST NEIGHBOR | 95.28% | 53.36 % |
| LINEAR DISCRIMINANT ANALYSIS | 96.81% | 55.77 % |
| RANDOM FOREST | 95.62% | 70.33 % |
| DECISION TREES | 96.42% | 45.41 % |

**Table 5.3 : Sensitivity & Specificity Comparison**

**Sensitivity**(also called the true positive rate, the recall, or probability of detection in some fields) measures the proportion of actual positives that are correctly identified as such.

**Specificity** (also called the true negative rate) measures the proportion of actual negatives that are correctly identified as such .

We calculate the sensitivity and specificity of the models in order to check which has the best optimum values and we found that Logistic Regression would be best compared to other models and the overall analysis.

# Conclusion

As far as AIC values are concerned, we see that there is not much difference within any of the models, hence we take other parameters like accuracy, sensitivity, specificity into consideration & we can certainly say that logistic model is better fit to predict the income of an individual based on the census data.

# How Analysed Data Can Be Used ?

* Spreading awareness among schools and colleges to maintain gender equality during on-campus placements
* We can target the 6% population with income > 50k for fundraising and starting campaigns for women empowerment
* Establish structures, policies, objectives in every organization to ensure gender balance
* This analysis is useful for loan lenders to know the american salary distribution and target relevant customers and segment them appropriately