

**EDO STATE UNIVERSITY UZUAIRE**

# REPORT

#### GROUP: GROUP C

#### DEPARTMENT: COMPUTER SCIENCE AND MATHEMATICS

#### COURSE TITLE: CSC 414

#### LEVEL: 400 LEVEL

#### TITLE: ADULT INCOME DATASET

#### GROUP MEMBERS: UYI PALMER (FSC/CSC/01801040)

**OBISHAI VALENTINE (FSC/CSC/01801027)**

**EMMANUEL FRED (FSC/CSC/01801014)**

**INU JOHN (FSC/CSC/01801020)**

**IDALU WENCELAUS (FSC/CSC/01801017)**

**OGHENEKARO EMMANUELLA (FSC/CSC/01801029)**

# Breakdown

This Project aimed at finding the best model to predict if the income of an individual (the respondents) is greater than 50K based on the several explanatory factors affecting the income in the dataset provided.

The predictor variables included information’s such as: age, final weight, education, occupation, work class, gender, native-country, etc.

Here is the breakdown of my Approach:

1. Data Pre-Processing which included missing values, conversion of data, normalization of data was handled
2. Visualization of Dataset
3. Feature Selection and dimensionality reduction techniques such as component analysis, recursive feature elimination was used to extract features for possible improvement in performance
4. Modelling Stage using some Algorithms (training and testing the dataset)
5. Cross Validation
6. Deployment

**Approach and Implementation**

1. **Pre-Processing Stage**:

In order to make the given data suitable for classification, pre-processing was performed.

All the features belonged to one of the following categories

1. **Numerical Data**: 6 out of 14 features were in numerical format. The features that belonged to this category are: age, final weight, education number, capital gain, capital loss, hours per week.
2. **Categorical Data**: 8 out of 14 features were in categorical format. The features that belonged to this category are: work class, education, marital status. Occupation, race, gender, native country, relationship

# Handling missing values

Some of the features (only categorical in this case) for certain data points contain unknown values which were replaced with “?”. It is necessary to substitute these unknown values with the value that best fits the context. The description of one technique used to handle the missing data is as under:

1. **Class independent mode imputation**: Since all the missing values in our dataset belong to the categorical features, mode is the best statistical substitute. Other measures like mean or other ways of data interpolation might not produce values which correspond to any of the categories. Hence, all missing value in a particular feature was substituted by the most frequently occurring category of that feature. This intuitively makes sense because the missing value along a particular feature is more likely to contain the category that occur the most. For instance, we replace the missing values in the feature ‘WorkClass’ by ‘Private’ as can be seen from the histogram plot for working class below. However, there is a certain degree of uncertainty attached to it (due to the assumptions made) and this might impact the final classification result either positively or negatively. This method was used for all the three features with missing data: work class, occupation and native country.

# Conversion of Data

all the categorical features wereconverted to numerical dataset using one hot encoder.

1. Data Visualization

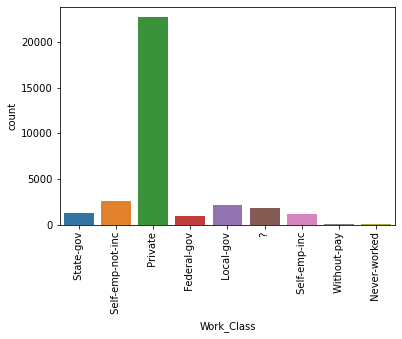
Below shows the graphical representation for most of the columns



**Figure 1: Gender**

**Gender of the respondents**

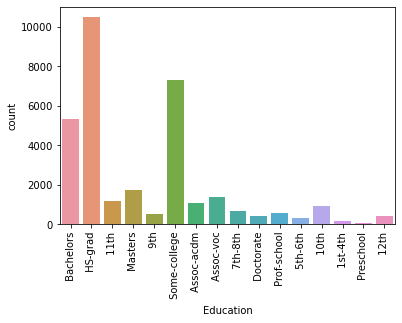
The result of the Analysis in figure 1 above showed the genders of the respondents. It can be observed that majority of the respondents were **males**. Followed by the respondents that were **females**. This implies that majority of the respondents that participated in the survey were **males**.

****

**Figure 2: Work Class**

**Working Class of the respondents**

The result of the Analysis in figure 2 above showed the working class of the respondents. It can be observed that majority of the respondents working class were **Private.** followed by the respondents which their working class were **Self-employed not in a company**. This implies that the least working class were coming from **without pay and never worked**.

****

**Figure 3: Education**

**Education of the respondents**

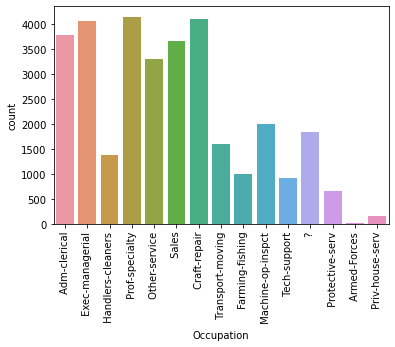
The result of the Analysis in figure 3 above showed the level of education of the respondents. It can be observed that majority of the respondent’s level of education was **HS-grad.** followed by the respondents which their level of education was **Some college**. This implies that the least level of education was **pre-school**.

****

**Figure 4: Native Country**

**Native country of the respondents**

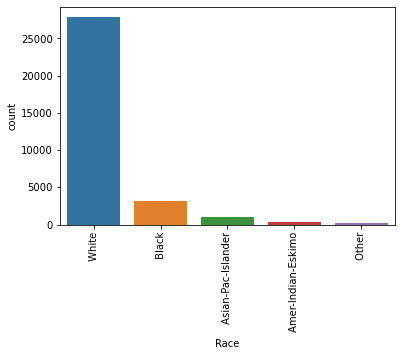
The result of the Analysis in figure 4 above showed the native country of the respondents. It can be observed that majority of the respondent’s native country was **United States.** followed by the respondents which their native country was **Mexico** and some who had no idea of where they came from.

****

**Figure 5: Occupation**

**Occupation of the respondents**

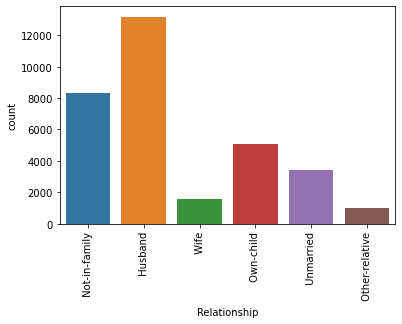
The result of the Analysis in figure 5 above showed the occupation of the respondents. It can be observed that majority of the respondent’s occupation was **Prof-Specialty.** followed by the respondents which their occupation was **Craft-repair**. This implies that the least occupation came from the respondent whose occupation was **Armed-forces**.

****

**Figure 6: Race**

**Race of the respondents**

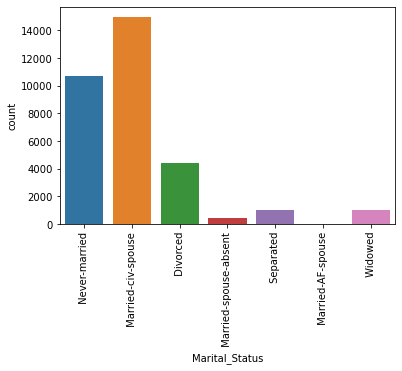
The result of the Analysis in figure 6 above showed the race of the respondents. It can be observed that majority of the respondent’s race was **White.** followed by the respondents which their race was **Black**. This implies that the least occupation came from the respondent whose occupation was **Others.**

****

**Figure 7: Relationship**

**Relationship of the respondents**

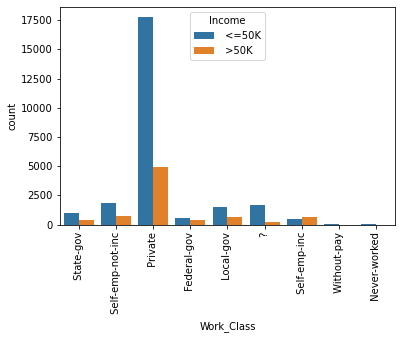
The result of the Analysis in figure 7 above showed the relationship of the respondents. It can be observed that majority of the respondent’s relationship was **Husbands.** followed by the respondents which their race was **Not-in-family**. This implies that the least occupation came from the respondent whose occupation was **Other-relative.**

****

**Figure 8: Marital Status**

**Marital status of the respondents**

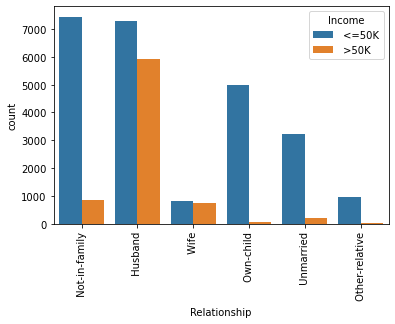
The result of the Analysis in figure 8 above showed the relationship of the respondents. It can be observed that majority of the respondent’s relationship was **Married-civ-spouse.** followed by the respondents which their race was **Never-married**. This implies that the least occupation came from the respondent whose occupation was **Married Spouse absent.** also **Married-A-Spouse** had no respondent.



**Figure 9: Income and Work Class**

**Relationship between Income and Work Class of the respondents**

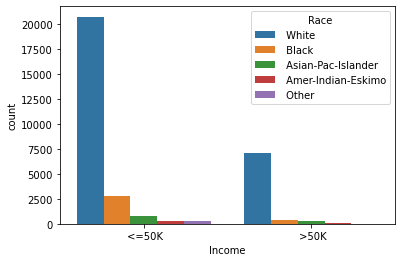
The result of the Analysis in figure 9 above showed the Relationship between Income and Work Classof the respondents.It can be observed that majority of the respondent’s income from the **private** work class had the highest income of **<=50k and >50k.** followed by the respondent’s income from the **self-employed not Inc.** work class had the income of **<=50k and >50k** This implies that the least occupation came from the respondent whose occupation was **Without pay and never worked.**

****

**Figure 10: Income and Relationship**

**Relationship between Income and Relationship of the respondents**

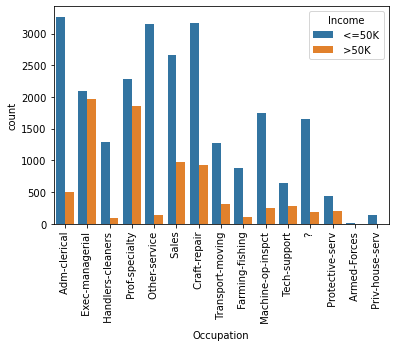
The result of the Analysis in figure 10 above showed the Relationship between Income and Relationship of the respondents**.** It can be observed that majority of the respondent’s income from the **Not-in-family** relationship had the highest income of **<=50k** and one of the **lowest incomes** for **>50k.** followed by the respondent’s income from the **Husband** which had the highest >50k. work class had the income of **<=50k and >50k** This implies that the least relationship came from the respondent whose relationship was **Other-relative.**

****

**Figure 11: Income and Race**

**Relationship between Income and Race of the respondents**

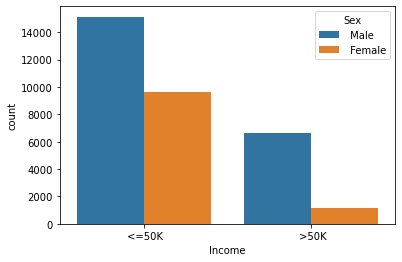
The result of the Analysis in figure 11 above showed the Relationship between Income and Race of the respondents**.** It can be observed that majority of the respondent’s income from all races had the highest values for **<=50k.** followed by the respondent’s income from all races which had >50k income**.**

****

**Figure 12: Income and Occupation**

**Relationship between Income and Occupation of the respondents**

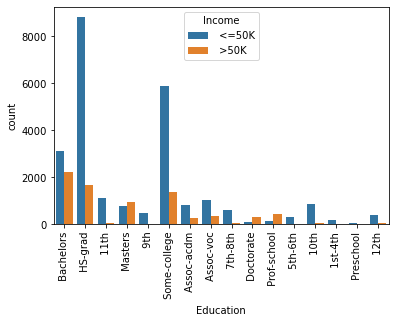
The result of the Analysis in figure 12 above showed the Relationship between Income and Occupation of the respondents**.** It can be observed that majority of the respondent’s income from the **Adm-clerical** occupation had the highest income of **<=50k** and one of the **lowest incomes** for **>50k.** followed by the respondent’s income from the **Exec-managerial** which had the highest >50k. This implies that the least relationship came from the respondent whose occupation was **Armed forces.**

****

**Figure 13: Income and Gender**

**Relationship between Income and Gender of the respondents**

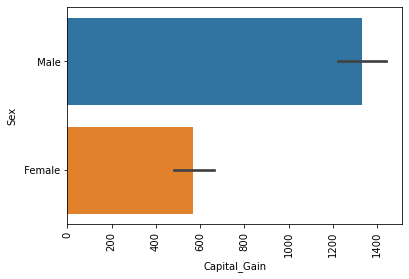
The result of the Analysis in figure 13 above showed the Relationship between Income and Gender of the respondents**.** It can be observed that majority of the respondent’s income from gender came from Male. it had the highest income of both **<=50k and >50k.** followed by the respondent’s income which came from the **Female** gender**.**

****

**Figure 14: Income and Education**

**Relationship between Income and Education of the respondents**

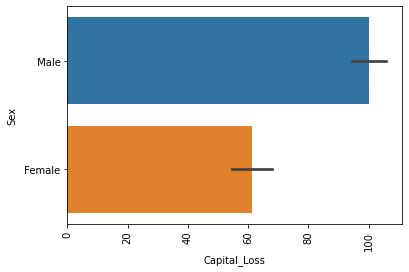
The result of the Analysis in figure 14 above showed the Relationship between Income and Education of the respondents**.** It can be observed that majority of the respondent’s income came from the **HS-grad** Education which had the highest income of **<=50k** and had a **lower income** for **>50k.** followed by the respondent’s income from the **Bachelors** which had the highest >50k and second highest n <=50k. This implies that the least income came from the respondent whose education was **Pre-School.**

****

**Figure 15: Gender and Capital Gain**

**Relationship between Gender and Capital Gain of the respondents**

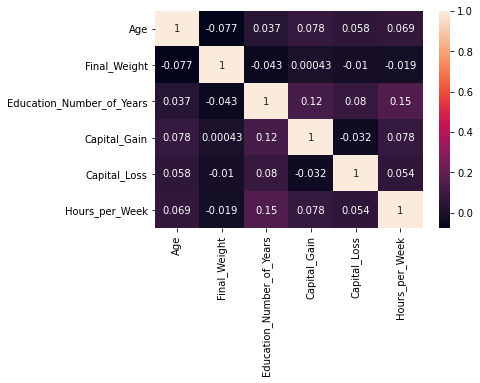
The result of the Analysis in figure 15 above showed the Relationship between Gender and Capital Gain of the respondents**.** It can be observed that majority of the respondent’s capital gain came from gender came from Male. followed by the respondent’s capital gain which came from the **Female** gender**.**

****

**Figure 16: Gender and Capital Loss**

**Relationship between Gender and Capital Loss of the respondents**

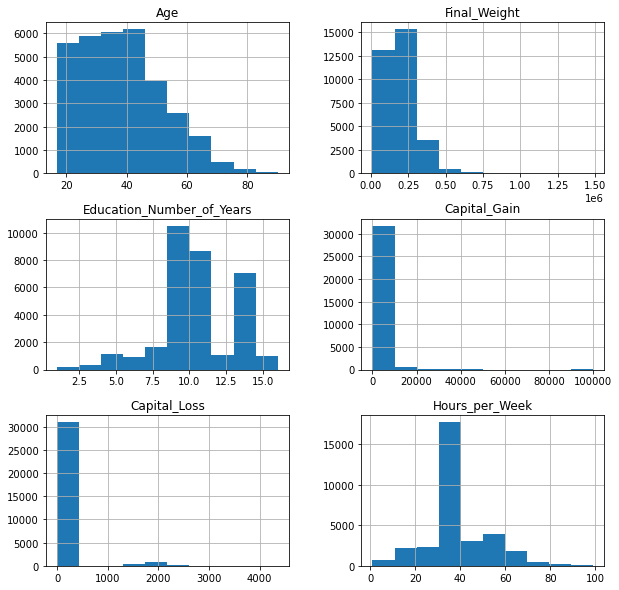
The result of the Analysis in figure 15 above showed the Relationship between Gender and Capital Loss of the respondents**.** It can be observed that majority of the respondent’s capital loss came from gender came from Male. followed by the respondent’s capital gain which came from the **Female** gender**.**

****

**Figure 17: Correlation between Numerical Features**

**Correlation between Numerical Features**

The result of the Analysis in figure 17 above showed the Correlation between Numerical columns of the data from the respondents**.** It can be observed that Correlation between Numerical columns like Income has 34% correlation with ‘Education\_num’, 23% correlation with ‘hours\_per\_week’ and ‘age’, and 22% correlation with ‘Capital\_gain’. This shows that the correlations between columns are moderate.

****

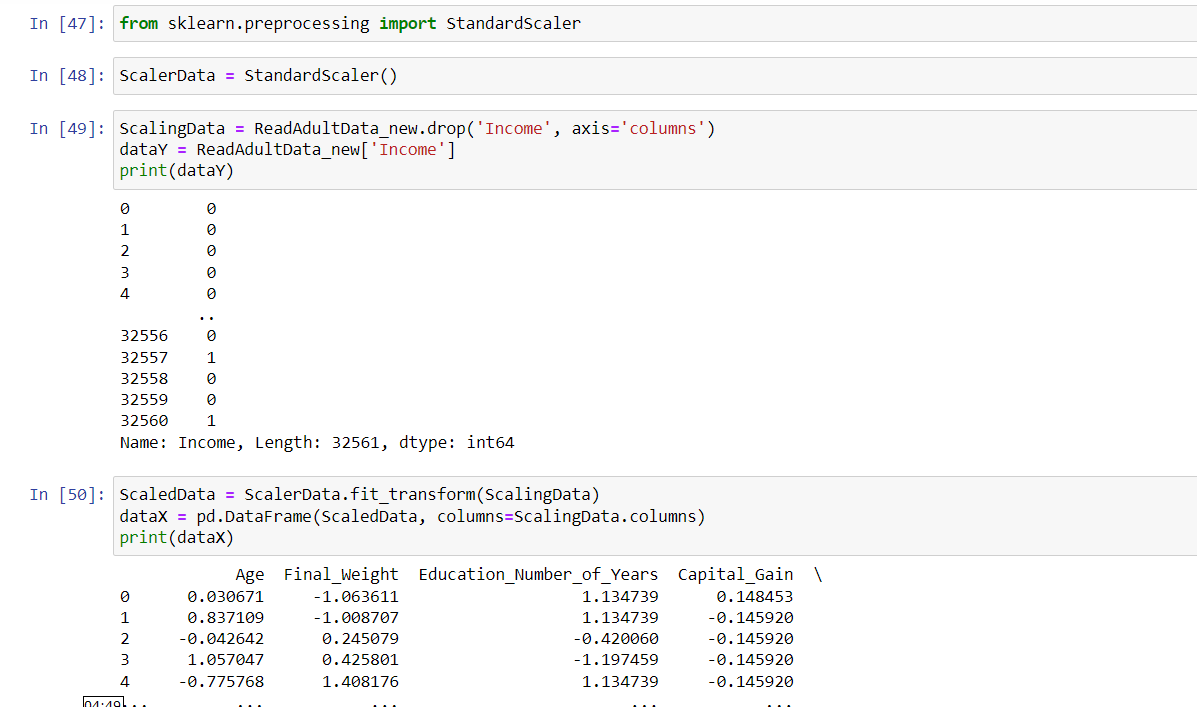
**Figure 18: Histogram plot for all the Numerical Columns**

**Histogram plot for all the Numerical Columns**

The result of the Analysis in figure 18 above showed the distribution between Numerical columns of the data from the respondents**.** It can be observed that the range of ages from our respondents is between the ages of 17 -85. Followed by the Histogram plot of final weight was between the range of 0-0.75, etc.

# Normalization of data

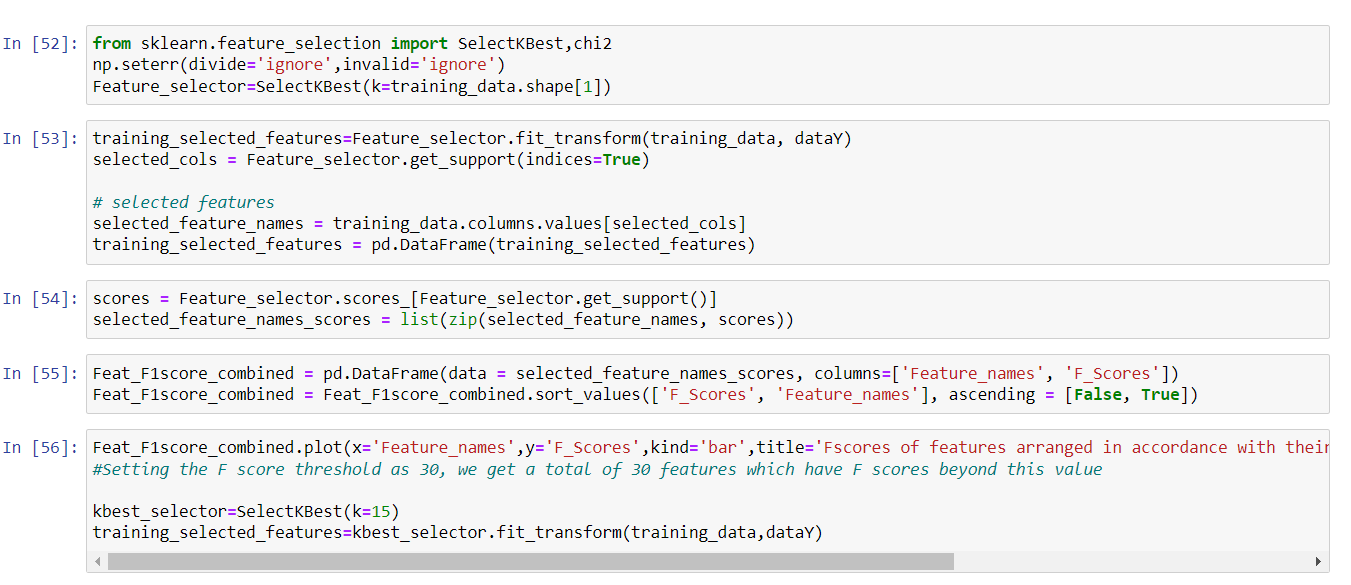
The numerical features in the data obtained from previous steps need not be along the same scale. Hence, all the numerical features are scaled to zero mean and one standard deviation before further processing. The StandardScalar() function offered by scikit learn library was used in order to achieve this.

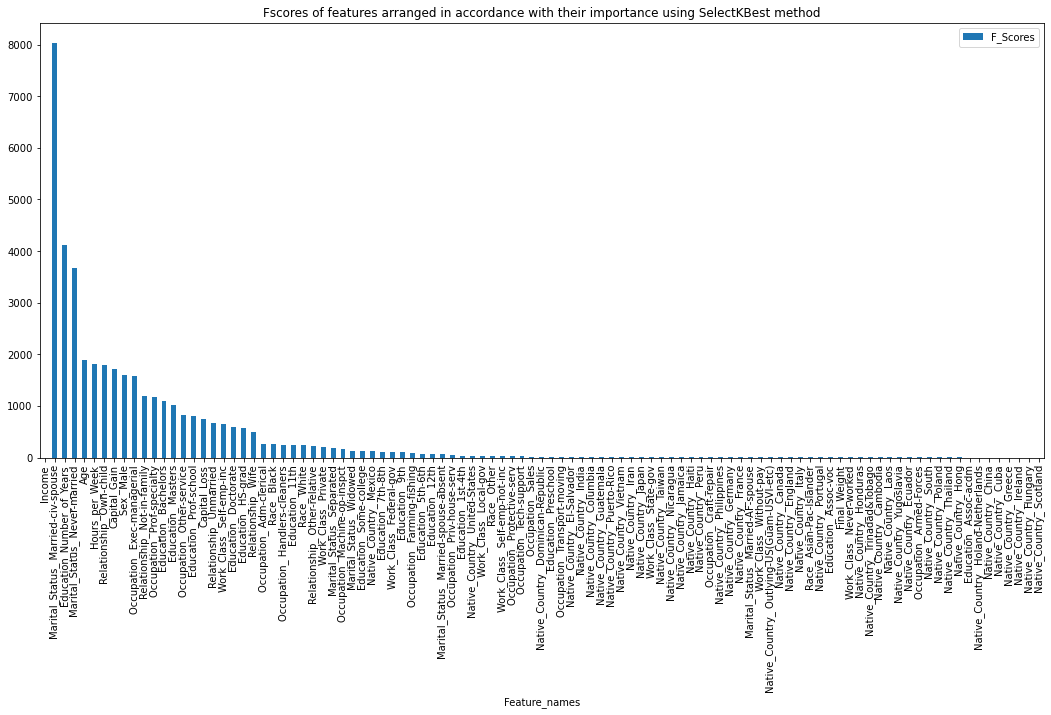


# Feature selection

is a category of methods that help choose features that contribute the most to the performance of the classifier and to remove the unwanted or redundant features that might unnecessarily increase the computational complexity. A description of the feature selection techniques used is as under:

Select ‘k’ best features: This is a feature selection method SelectKBest() from scikit learn library was used in order to achieve this. The value of k was chosen such that all features with F-score values greater than 15 were selected.

****

****

1. **TRAINING AND CLASSIFICATION**
2. **Naive Bayes**: This is a statistical classification technique focussing on determining the probabilities of the features belonging to a particular class. This is done by using the Bayes (conditional) theorem as given below:



This method assumes that all the features of a given dataset are independent of each other and hence their joint probabilities factor into individual probabilities. The prior probabilities i.e. the probability of occurrence of each class is obtained by using the number of data points belonging to a particular class divided by the total number of data points. This method is known to suit high dimensional data and has a better performance when the independence condition holds true.

Function used: GaussianNB() offered by the scikit learn naive\_bayes module- this function assumes Normal distribution for class conditional densities.

1. **Decision trees**: This is a supervised learning technique which resembles a tree like structure in its implementation. It starts with all training samples as a root node at the top and at every subsequent layer it considers the feature that provides the maximum information gain. By recursively doing this process it constructs multiple sub-trees over subset of training samples which are classified along the path.

Function used: DecisionTreeClassifier() offered by scikit learn-tree module

1. **Random forest classifier**: A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

Function used: RandomForestClassifier() offered by scikit learn-ensemble model

1. **Logistic Regression**: This is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

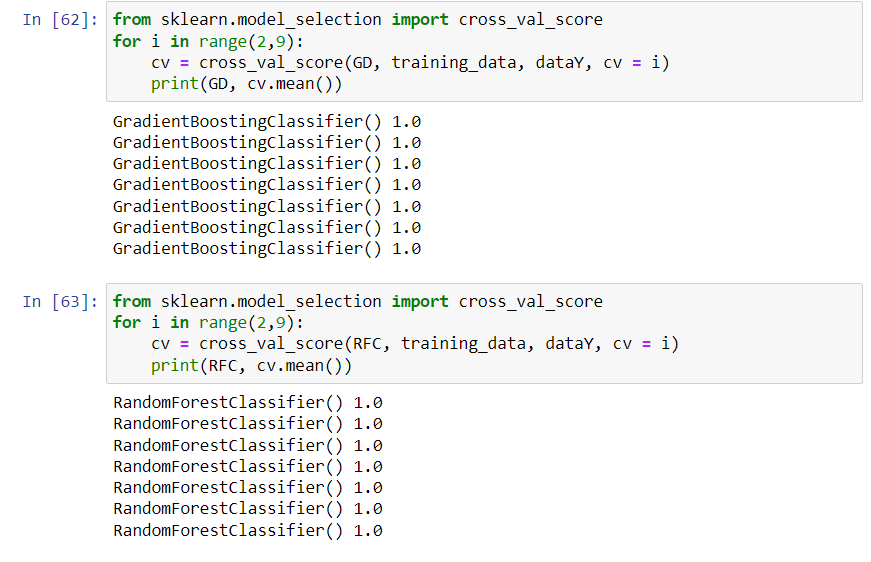
Function: LogisticRegression() offered by scikit learn – linear model module

**COMPARISON, RESULTS AND INTERPRETATION**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Classifier** | **Training accuracy** | | | **Test accuracy** | **Precision** | | | **Recall** | | **F1 score** | |
| Naive Bayes | | 1.0 | 1.0 | | 1.0 | 1.0 | 1.0 | | 1.0 | 1.0 | 1.0 |
| Decision Tree | | 1.0 | 1.0 | | 1.0 | 1.0 | 1.0 | | 1.0 | 1.0 | 1.0 |
| SVC | | 0.987 | 0.994 | | 0.99 | 0.99 | 1.0 | | 0.95 | 0.99 | 0.97 |
| Logistic Regression | | 1.0 | 1.0 | | 1.0 | 1.0 | 1.0 | | 1.0 | 1.0 | 1.0 |

# Cross Validation

The goal of cross-validation is to test the model’s ability to predict new data that was not used in estimating it, in order to flag problems like overfitting or selection bias and to give an insight on how the model will generalize to an independent dataset (i.e., an unknown dataset, for instance from a real problem). Made use of Gradient boosting classifier and random forest classifier for the cross validation



**Random Forest classifier:** The number of estimators is varied in the range (2,9) and the optimal value was obtained through 7-fold cross validation. The optimal number of estimators is found to be 1.0.

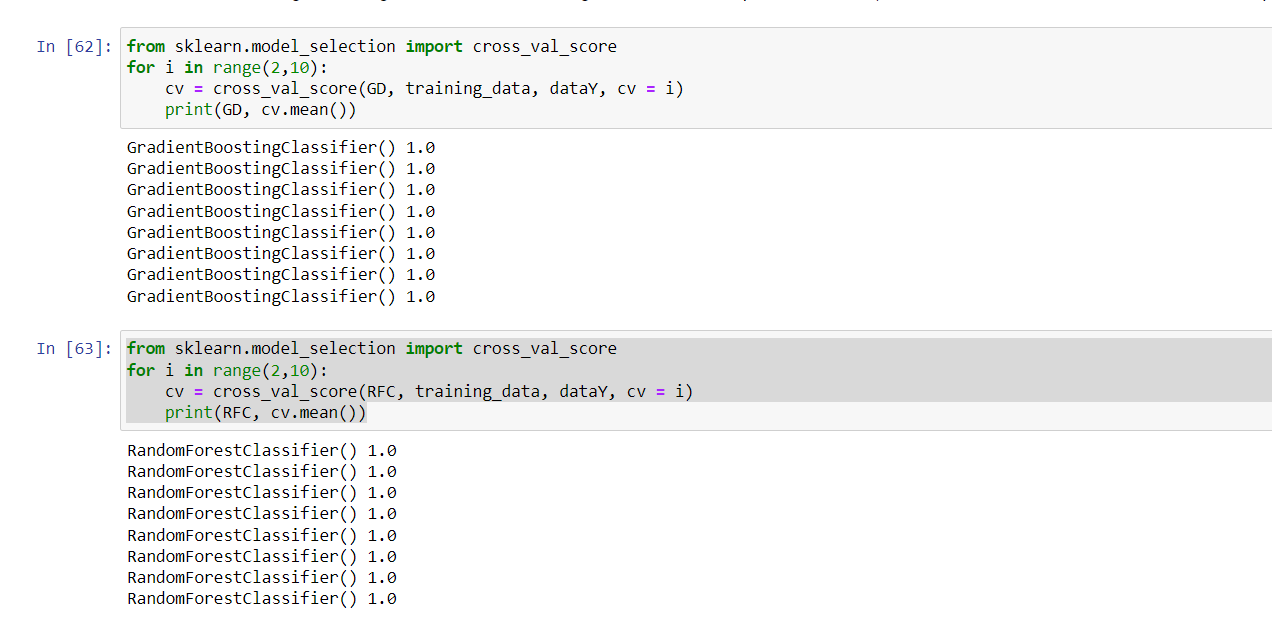
**Description of features in adult dataset**

|  |  |  |  |
| --- | --- | --- | --- |
| **Predictor variable** | **Description** | **Type** | **Values contained** |
| Age | Age of the individual | Numerical | Integer values between 17 and 90 |
| Work Class | Class of work | Categorical | Federal-gov, Local-gov, Never-worked, Private, Self-emp-inc, Self-emp-not-inc, State-gov, Without-pay |
| Fnlwgt | Final weight of how much population it represents | Numerical | Integer values |
| Education | Educational qualification of an individual | Categorical | 10th, 11th, 12th, 1st-4th, 5th-6th, 7th-8th, 9th, Assoc-acdm, Assoc-voc, Bachelors, Doctorate, HS-grad, Masters, Preschool, Prof-school, Some-college |
| Education Num | Numerical representation of educational qualification | Numerical | Integer values between 1 and 16 |
| Marital Status | Marital status of the individual | Categorical | Divorced, Married-AF-spouse, Married-civ-spouse, Married-spouse-absent, Never-married, Separated, Widowed |
| Occupation | Occupation of the individual | Categorical | Adm-clerical, Armed-Forces, Craft-repair, Exec-managerial, Farming-fishing, Handlers-cleaners, Machine-op-inspect, Other-service, Priv-house-serv, Prof-specialty, Protective-serv, Sales, Tech-support, Transport-moving |

|  |  |  |  |
| --- | --- | --- | --- |
| Relationship | Type of relationship | Categorical | Husband, Not-in-family, Other-relative, Own-child, Unmarried, Wife |
| Race | Race of an individual | Categorical | Amer-Indian-Eskimo, Asian-Pac-Islander, Black, Other, White |
| Sex | Gender of an individual | Categorical | Male, Female |
| Capital gain | Capital gain | Numerical | Continuous integer values |
| Capital loss | Capital Loss | Numerical | Continuous integer values |
| Hours per week | Average number of working hours per week | Numerical | Continuous integer values |
| Native Country | Country of origin | Categorical | 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, Trinadad &Tobago, Peru, Hong, Holand-Netherlands |
| Income | Income level | Binary | >50K, <=50K |

# Interpretation of the obtained results

1. Since all the missing values in the training dataset belong to categorical features, it makes logical sense to use the mode imputation.
2. The histogram plots help us understand the distribution of various categories in a feature and hence they play a crucial role in guiding towards handling different features.
3. In all the experiments conducted above we see that the training and the test dataset have their accuracies in almost the same range. This proves that the classifier is not over fitting the data.
4. Further, improvement in the performance of the classifier can be seen with more number of data points (observation made on comparing the results of the smaller training dataset with the bigger one).
5. For the cross validation of the project, I made use of GradientBoostingClassifer() and RandomForestClassifier(), the mean result turned out to be 1.0



# Conclusion

The Decision tree classifier,Gaussia DecisionTreeClassifier and logistic regression is found to perform well on adult dataset. However, performing the right set of pre-processing steps and selecting the right set of features plays a significant role in the performance of the classifier.

The whole project was implemented on Python. The scikit learn library was used to implement most of the classifiers and the feature reduction techniques.

# Deployment

I created a github repository and deployed it using streamlit.io

**REFERENCES:**

[1] https://archive.ics.uci.edu/ml/datasets/adult [Online]

[2] <https://medium.com/analytics-vidhya/machine-learning-application-census-income-prediction-868227debf12> [Online]

https://scikit-learn.org/stable/auto\_examples/svm/plot\_separating\_hyperplane\_unbalanced.html

[3] EE 559 discussion and lecture notes

[4]https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/ [Online]

[5] https://machinelearningmastery.com/imbalanced-classification-with-the-adult-income-dataset/ [Online]

[6] <https://rstudio-pubs-static.s3.amazonaws.com/235617_51e06fa6c43b47d1b6daca2523b2f9e4.html> [Online]