ST. CLAIR COLLEGE OF APPLIED ARTS AND TECHNOLOGY

Ace Acumen Campus: Mississauga

Final Project Report

Adult income dataset by using Machine Learning Algorithm

Supervisor’s Name: Savita Shehrawat

Submitted by: Karanpreet Singh (St. Id: 0788823)

**Contents**

**1.Abstract…………………………………………………………………………………...3**

**2.Research questions……………………………………………………………………….4**

**3.Introduction……………………………………………………………………………4-5**

**4.Literature review……………………………………………………….……………...5-7**

**5.Methodology…………………………………………………………………………...7-9**

**6.Data Details……………………………………………………………………………….9**

**7.Data Pre-processing………………………………………………………………….9-11**

**8. visualization of data……………………...……………..…………………………11-16**

**9.Conclusion……………………………………………………………………………..17**

**10.Reference…………………………………………………………………………..17-18**

Abstract

Income plays a significant role in determining a person's level of living and financial standing in society. It is crucial in deciding the nation's growth. Our goal is to find relevant insights that may be used to make better decisions. Our database has 32561 records with various parameters such as occupation, age, relationship, weekly hours, education, and income. Between the dependent and independent variables, exploratory analysis will be performed. For our purposes, not all of the qualities are relevant. The useful features will be chosen based on the results of the various algorithms. There are numeric variables as well as numerous factors. Based on every observation's attributes, predict whether a person, income exceeds $50,000 annually. Also known as "Census Income" dataset. It's Classification Task, with categorical and numerical features, some of the instance have missing values, the missing value was denoted as "?"

Total have 15 columns of features, the last column: income is the classification label: >50k, <=50k

Literature Review

Barry Becker (1994) 32561 separate records and 14 attributes for 42 countries are included in the data collection. Age, education, nationality, marital status, relationship status, occupation, work classification, gender, race, working hours per week, capital loss, and capital gain are among the 14 attributes. The income level is the binomial label in the data set, and it predicts whether a person makes more than $50,000 per year based on a given set of attributes.

S.Deepajothi and Dr. S.Selvarajan: (8, October2012) present a comparative study of the classification accuracy provided by different classification algorithms like Naïve Bayesian, Random forest, Zero R, K Star on census dataset and provide a comprehensive review of the above algorithms on the dataset.

Vidya Chockalingam, Sejal Shah and Ronit Shaw (2017) moved on to a classification task of predicting whether the income is >=50k/year from a person’s attributes, by using important features. For the classification task, we implemented various machine learning models, that after the initial task would prove useful. After implementing various machine learning models, we compared their results on the training and the test set to arrive at a model that works best for the predictive task on both test and training data set with 87% accuracy

Chet Lemon, Chris Zelazo and Kesav Mulakaluri: (1994) 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. The Gradient Boosting Classifier Model was deployed which clocked the highest accuracy of 88.16%, eventually breaking the benchmark accuracy of existing works

ckalingam, Sejal Shah and Ronit Shaw:(2017) employed different Machine Learning Models, including Logistic Regression, Stepwise Logistic Regression, Naive Bayes, Decision Trees, Extra Trees, k-Nearest Neighbor, SVM, Gradient Boosting, and 6 configurations of Activated Neural Network, to explore and analyse the Adult Dataset. They also conducted a comparison investigation of their prediction abilities.

Sisay Menji Bekena: 30th July, 2017 Important features prediction shows marital status, capital gain, education, age and hours per week are the top features which account for larger shares of the model accuracy. Using decision tree classifier also shows that these variables are the top 5 features in importance.

Bricker, J., Ramcharan, R. and Krimmel, J. (2014) used Logistic Regression as a Statistical Modelling Tool and four different Machine Learning Techniques: Neural Network, Classification and Regression Tree, Random Forest, and Support Vector Machine.

C. Jayavarthini, Ishu Todi, Kshitij Kumar Agarwal (2018) suggestions could be given based on the predictions to students who are in need to pursue higher education and people who are spending less time in the workplace. We also aim to measure the accuracy of different models using Logistic Regression, Naive Bayes classifier etc

Hong, M.,Z.(2019) The dataset extracted from Adult Census Income in 1994 by Ronny Kohavi and Barry Becker, the dataset includes 15 variables. One predication goal is to determine if a person makes over $50K a year based on the biography and background.

Wang Tongwen and Guan Lin (2007) In initial stages, we preprocessed the data and developed understanding of the data and its useful features that explain the variances by doing various types of exploratory analysis. Later, we moved on to a classification task of predicting whether the income is >=50k/year from a person’s attributes, by using important features. For the classification task, we implemented various machine learning models. After implementing various machine learning models, we compared their results on the training and the test set to arrive at a model that works best for the predictive task on both test and training data set with 87% accuracy.

Alina Lazar: (2004) A systematic study is performed on the influence of this statistical narrowing on the grid parameter search, training time, accuracy, and number of support vectors. Accuracy values as high as 84%, when compared against a test population, are obtained with a reduced set of parameters while the computational time is reduced by 60%. Tailoring computational methods around specific real data sets is critical in designing powerful algorithms

**Methodology:**

**Github source :** [**https://github.com/karana**](https://github.com/karana)

First solve the three business question after that clean the dataset remove the null value and duplicates check the dataset attributes by bar chat and histogram then use minmax scaling , smote and other models . in the last add the conclusion.

Summary of Categorical Attributes:

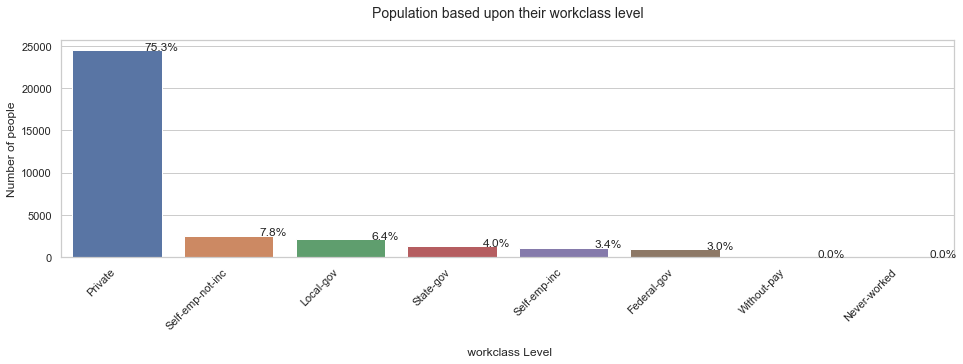
|  |  |  |  |
| --- | --- | --- | --- |
| Attribute | Description | No. of Levels | Categories: Counts |
| Work class | Class of work  Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked | 8 | Private 22286  Self-emp-not-inc 2499  Local-gov 2067  State-gov 1279  Self-emp-inc 1074  Federal-gov 943  Without-pay 14 |
| education | Education of the individual  Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool | 16 | HS-grad 9840  Some-college 6678  Bachelors 5044  Masters 1627  Assoc-voc 1307  11th 1048  Assoc-acdm 1008 |
| MaritalStatus | Marital status of the individual  Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse | 7 | Married-civ-spouse 14065  Never married 9726  Divorced 4214  Separated 939  Widowed 827  Married-spouse-absent 370  Married-AF-spouse 21 |
| occupation | Occupation of the individual  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 | 14 | Prof-specialty 4038  Craft-repair 4030  Exec-managerial 3992  Adm-clerical 3721  Sales 3584  Other-service 3212  Machine-op-inspct 1966  Transport-moving 1572 |
| relationship | Present relationship  Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried | 6 | Husband 12463  Not-in-family 7726  Own-child 4466  Unmarried 3212  Wife 1406  Other-relative 889 |
| race | Race of the individual  White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black | 5 | White 25933  Black 2817  Asian-Pac-Islander 895  Amer-Indian-Eskimo 286  Other 231 |
| sex | Sex of the individual  Female, Male | 2 | Male 20380  Female 9782 |
| Native.country | Average number of hours spent by the individual on work  United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India | 40 | United-States 27504  Mexico 610  Philippines 188  Germany 128  Puerto-Rico 109  Canada 107  El-Salvador 100  India 100  Cuba 92 |
| Income | <=50k , >50k | 2 | >50k 22654  <=50k 7508 |

Summary of Numerical Attributes:

|  |  |  |
| --- | --- | --- |
| Attribute | Description | Numeric: Range |
| Age | Age of the individual | 36- 86 |
| fnlwgt | Final Weight Determined by Census Org | 145522-203488 |
| Education .num | Number of years of education | 1-14 |
| Capital.gain | Capital gain made by the individual | 0-15024 |
| Capital.loss | Capital loss made by the individual | 0-2457 |
| hours.per.week | Average number of hours spent by the individual on work | 40-87 |

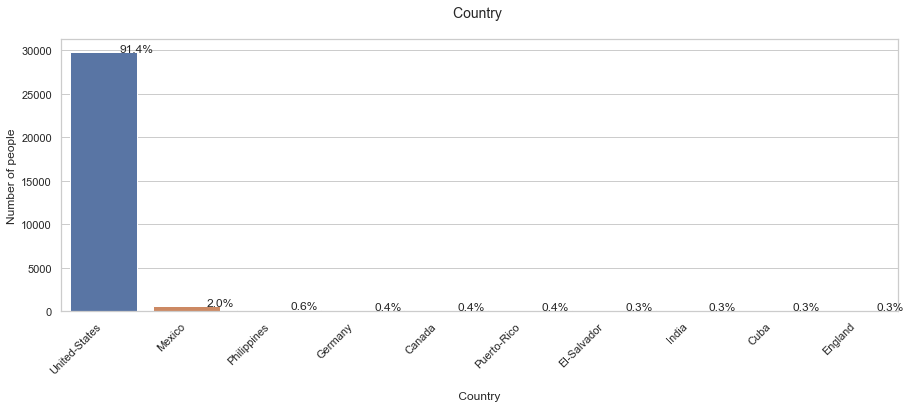
Keywords: Logistic Regression, Naive Bayes, Random Forest, KNN.

QUESTION 1 :- Which profession people hitting the target more as compare to other profession ?



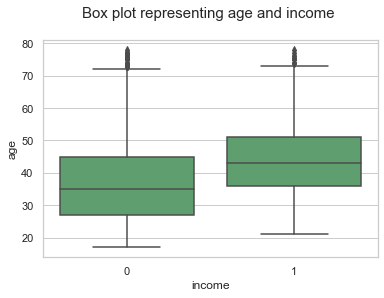
This outline that individuals accomplishing private work calling making more as contrast with rest of others.

Question 2:- Detect those people which part of the country making more money as compare rest of other ?



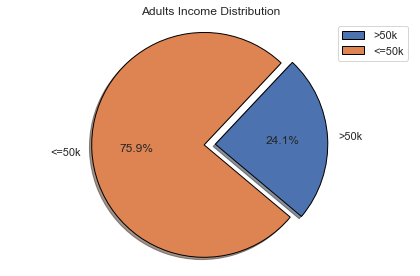
This shows that most of the people which making more than 50 k belongs to united states .

Question 3:- Detect those age group people which make less than 50k and more than 50k ?

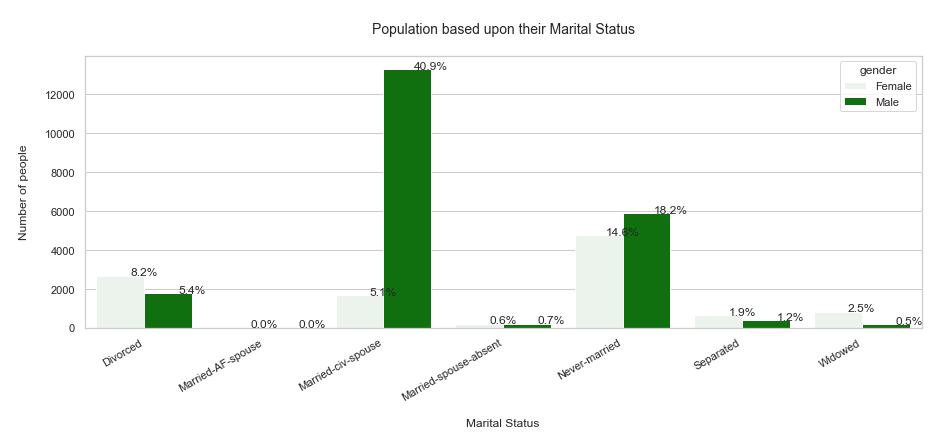


Here 1 indicate as the more than 50k and 0 represent less than the 50k age group between 30 to 40 making less than 50k. on the other hand, age group between 40 to 50 make more than 50k.

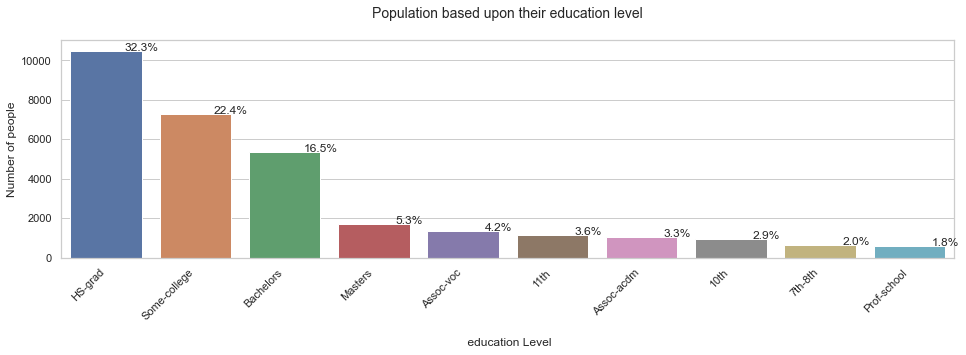
exploratory data analysis



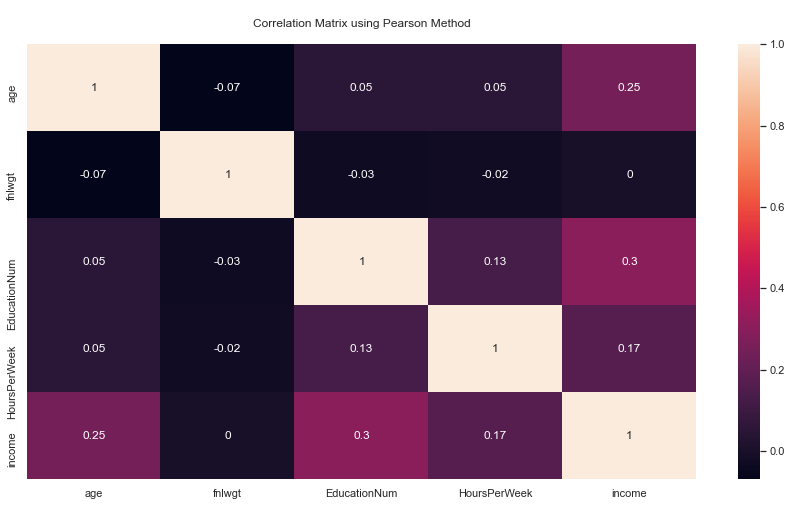
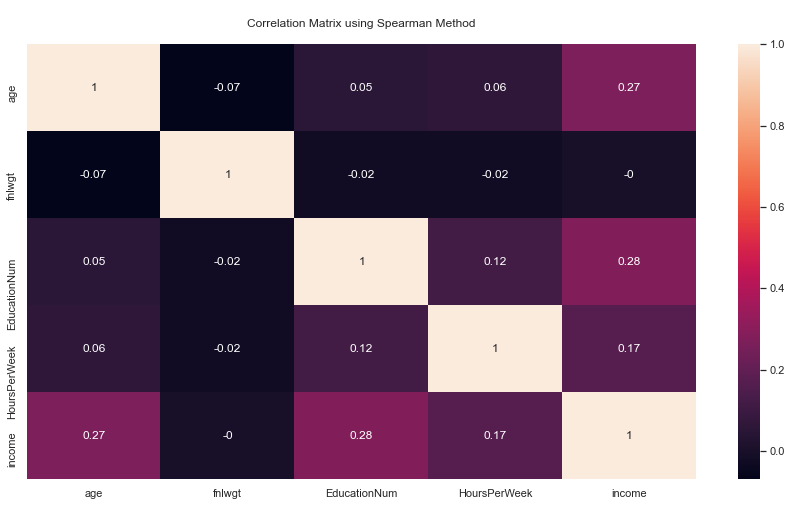
From the above illustration, 75.9% of the population have less than equal to $50000 income while the people having greater than $50000 income is 24.1%.



The majority of working males are married, followed by singles. However, the majority of working females are single, followed by divorced and married women.



The bargraph illustrates that the majority participants in the dataset are High School Graduates, followed by some college, then Bachelors degree holders.



Graphical user interface, text, email

Description automatically generatedText

Description automatically generatedThe above graphs show the relation between two variables using two different methods and shows how the change in one variable impact other. The value varies between -1 to 1. No strong correlation has been observed between any of the variables.

As demonstrated in the graph above, the majority of people earn more than or equal to $50,000. from united states . here show the first 10 top country which are earning more than the 50k as the rest of compare with others united states have the majority of people which are making 50k.

Chart, bar chart

Description automatically generated

Adults with a Prof-school and Doctorate educational background will have a greater salary, and it is likely that they will earn more over $50,000.

Chart, bar chart

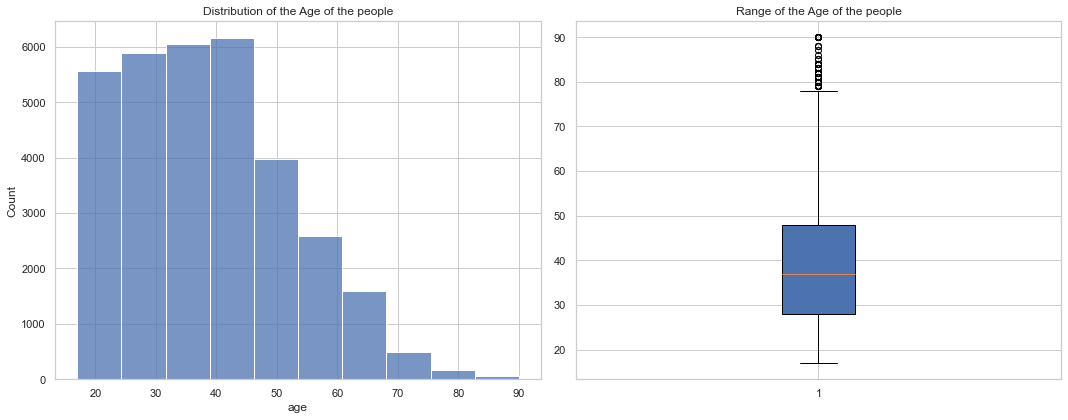
Description automatically generated

According to our statistics, people with the jobs Prof-specialty and Exec-managerial have a better chance of earning more than $50,000.

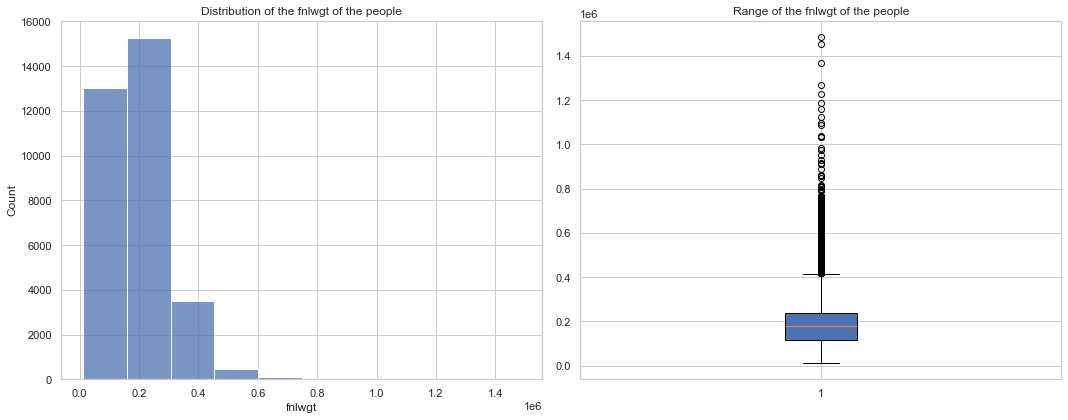
Chart, line chart

Description automatically generated

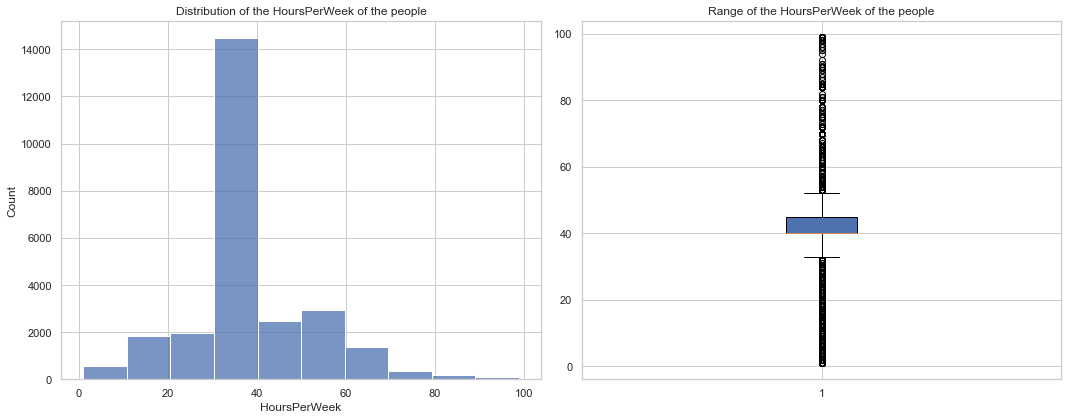
This histogram illustrates that the majority of the people are between the ages of 20 and 40, with the age group beyond 80 having the least number of people.



This histogram illustrates that the dataset containing people age most of the people age between 40 to 50 years 80 year people are least number. Here get the upper bound outliers that remove in the boxplot



The final weight attributes histograms illustrates that the people distribution between 0.0 to 0.2 is the highest and 0.6 is the least occurrence here also get the upper bound outliers . further in the boxplot that removed.



Hours per week illustrates that the people are doing how many hours work in the week and here majority people work in between the 35 to 40 hours in the week as 80 to 90 least people are doing work . here get the both outlier upper and lower bound.

Now using the minmax scaling applying one hot encoding on the dataset use the smote to balancing the dataset . use LogisticRegression , GaussianNB , KNeighborsClassifier RandomForestClassifier. Here get RandomForestClassifier accuracy is higher as compare to rest of other

Logistic Regression 80.3966301

KNN 75.3246752

Naive Bayes 48.4380483

Random Forest 81.326781

K-Folds Cross Validation : To reduce the model's bias, the K-Folds method is applied. Because each data record has a chance to appear in both the training and test data sets. The K-Folds approach partitions the dataset into k-folds. At random, I divided the data into five folds. The four folds are then applied to the model, and the fifth fold is used to test it. Repeat until each fold has been utilised as a test set. After that, the average is calculated by averaging all of the findings.

Table

Description automatically generated

In this strategy, k-fold cross-validation is paired with typical train-test splits. utilise a cross-validation approach to generate random divisions of the data in the training test set, and then split and test the algorithms multiple times. The data was used to produce five Repeated Random Test-Train Splits.

Table

Description automatically generated

Chart, treemap chart

Description automatically generated

A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing here get the true positive value 68.11%.

Chart, line chart

Description automatically generated

An ROC curve shows the relationship between clinical sensitivity and specificity for every possible cut-off. The ROC curve is a graph with: The x-axis showing 1 – specificity (= false positive fraction = FP/(FP+TN)) The y-axis showing sensitivity (= true positive fraction = TP/(TP+FN))

Table

Description automatically generated

Now looking for Most important features are age, fnlwgt, MaritalStatus\_Married-civ-spouse, HoursPerWeek, EducationNumu.

Conclusion :- illustrate that the random forest provide the best accuracy here mostly males are earning more than 50 k as the basis of the education and occupation people with study less and doing some private work are making more than 50k with in a year and majority of the population belong to the united states with making more than 50 k .Most important features are age in the dataset .

References

[1] Barry Becker (1994) Machine Learning Repository at the University of California, Irvine (UCI) [8]. <https://archive.ics.uci.edu/ml/datasets/adult>

[2] S.Deepajothi and Dr. S.Selvarajan: 8, October2012 “A Comparative Study of Classification Techniques On Adult Data Set”, International Journal of Engineering Research Technology (IJERT), ISSN: 2278-0181 Vol. 1 Issue

[3] ckalingam, Sejal Shah and Ronit Shaw:(2015) “Income Classiﬁcation using Adult Census Data”, [https://cseweb.ucsd.edu/classes/wi17/cse258- a/reports/a120.pdf](https://cseweb.ucsd.edu/classes/wi17/cse258-%20a/reports/a120.pdf)

[4] Chet Lemon, Chris Zelazo and Kesav Mulakaluri: (1994)“Predicting if income exceeds $50,000 per year based on 1994 US Census Data with Simple Classification Techniques”, https://cseweb.ucsd.edu/ jmcauley/cse190/reports/sp15/048.pdf

[5] ckalingam, Sejal Shah and Ronit Shaw:(2017) “Income Classiﬁcation

using Adult Census Data”, https://cseweb.ucsd.edu/classes/wi17/cse258-

a/reports/a120.pd

[6] Sisay Menji Bekena: 30th July, 2017 “Using decision tree classifier to predict income levels” https://mpra.ub.uni-muenchen.de/83406/ MPRA Paper No. 83406

[7] Bricker, J., Ramcharan, R. and Krimmel, J. (2014), Signaling status: the impact of relative income on household consumption and financial decisions, Federal Reserve Board, Finance and Economics Discussion Series (FEDS) Working Paper no. 2014-76.

[8] C. Jayavarthini, Ishu Todi, Kshitij Kumar Agarwal (2018) International Journal of Pure and Applied Mathematics Volume 118 No. 22, 587-590

[9] Hong,M.,Z.(2019),Income Prediction using CART Retrieved from <https://manzih/income-prediction-using>

[10] Wang Tongwen and Guan Lin(2007)“ A Data Mining Technique Based on Pattern Discovery and k-Nearest Neighbor Classifier for Transient Stability Assessment” pp-118 - 123

[11] Vidya Chockalingam, Sejal Shah and Ronit Shaw:(2017) “Income Classification using Adult Census Data”, [https://cseweb.ucsd.edu/classes/wi17/cse258- a/reports/a120.pdf](https://cseweb.ucsd.edu/classes/wi17/cse258-%20a/reports/a120.pdf)

[12] Alina Lazar:(December 2004)“Income Prediction via Support Vector Machine”, International Conference on Machine Learning and Applications - ICMLA 2004,page no. 16-18.