Data Mining using Classification and Clustering techniques on the Adult Census Dataset

Subhasree V

20-12-2020

INTRODUCTION

1. Description of the problem and a discussion of the background

The world we live in is controlled by the economies, which is extremely dependent on an individual's income. Cambridge defines income as "money that is earned from doing work or received from investments"(INCOME | meaning in the Cambridge English Dictionary, 2020). An individual's income is very much affected by his age, occupation and unfortunately factors like gender.

The dataset used here is originally the US Census data collected in 1994. However, all of the attributes which are factors that directly or indirectly affect the income of people, is valid even today. A study conducted on 'Annual Survey of Hours and Earnings'of the year 2016 by the UK Government gives insights about the factors that can affect earnings, which is also useful for income. Age, gender, sector, skill group etc. were found to be relevant factors (UK Government, 2020).

Understanding what affects an individual's income is important now more than ever when the world is trying hard to mitigate the effects of the Novel Coronavirus on the economy. The dataset is chosen after understanding how important income of individuals and a stable economy is for any country, especially during the time of a crisis. Another interesting article on the wall street journal suggests that the pandemic might balance wealth distribution by decreasing income equality. (Hannon, 2020) Thus the dataset that provides vital information about income and characteristics is always a relevant problem in any society.

DATA

2. Data description

2.1. The source of the data

The dataset originally comes from the US Census Bureau (Bureau, 2020). The United States Census Bureau is a principal agency of the U.S. Federal Statistical System, which is responsible for producing data about the people and economy of America. (United States Census Bureau, 2020)The organization releases the Census data for the public to use. Their censuses and surveys help in informed decision making and strategy building in the United States. The dataset is available in UCI Machine Learning Repository. It is the data that was extracted by Barry Becker from the 1994 Census database. The donors of this dataset are Ronny Kohavi and Barry Becker of Silicon Graphics. The data used here is downloaded from Kaggle which is named as Adult income dataset and this data is made available in Kaggle from UCI repository.

Extraction was performed by Barry Becker using some condition such as((AAGE>16) && (AGI>100) && (AFNLWGT>1)&& (HRSWK>0)). This ensures that the dataset doesn't have lots of unclean or noisy records. (UCI Machine Learning Repository: Adult Data Set, 2020)

2.1.1. The agencies working with the data

The Adult Census Dataset is cited widely and is used in various researches worldwide. The dataset was originally used by the US Census Bureau. This dataset was made publicly available and after the extraction performed by Barry Becker, the dataset continues to be used by various researchers at several academic institutions or organizations worldwide in order to study classification algorithms or to gain valuable insights. Some of the notable agencies or researchers working with the data are:

"Privacy Preserving OLAP" – Paper presented on SIGMOD Conference. 2005.

This research paper was presented on the Proceedings of the 2005 ACM SIGMOD International conference on Management of data. The authors of the paper are members of IBM Almaden and Stanford University. The paper presents the concept of data perturbation to protect from data breaches. This paper cited the Adult Census Dataset from the UCI Machine Learning Repository. The dataset was used for conducting experiments for the empirical evaluation of their algorithms.

"Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid"- Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996

The dataset was utilised in this very famous research paper that was presented in the second International Conference on Knowledge Discovery and Data Mining by Ron Kohavi. This paper introduced the Naïve Bayes Hybrid Algorithm, the NBTree which then became very popular amongst data scientists. (PDF) A Statistical Approach to Adult Census Income Level Prediction, 2020)The algorithm was performed on many many datasets available on the UCI repository, mainly the Adult Census Dataset.

2.2. The intended use of the data

The data was collected and prepared to understand the factors/attributes that are associated with the income of an individual and several other characteristics that relate to and individual's socio-economic life.The dataset is mainly used for Classification and Clustering Tasks. The classification task is to determine whether or not a person makes over 50K US Dollars a year.This information can be obtained by selecting the useful features and applying classification algorithms on them. The dataset is widely used and cited, hence the maximum achieved accuracy of the classification task is also available for reference. It is also useful to identify clusters among the population which helps us better understand the clusters of people and their characteristics that affect their income, even today.

Gender discrimination in wages is still an issue faced by countries worldwide. The dataset can also be used to understand if gender affected income in the US in the 1990s. According to the data released by the U.S Department of Labor and Fortune, women make up 63 percent of workers earning the federal minimum wage. Women are in the lowest rung of the income ladder even after their significant presence in health care and food industry. (Gender Economic Inequality - Inequality.org, 2020)This clearly shows a disparity which has to be studied. Thus, the dataset has much greater significance in helping us understand the social, political and economic context that narrates the story

of income in any society. New interventions with more data mining and analytics can help improve policy making and to reduce the concentration of wealth in a very small percentage of the society. Identifying patterns from this dataset is very beneficial to develop an understanding of the most crucial factors that can increase or decrease income level of an individual.

2.3. Attribute types of the data

The dataset is multivariate and consists of 14 attributes

Attribute Name	Description	Type	Distinct
age	Age of the individual	Numeric	67
workclass	Work class of an individual	Nominal	8
fnlwght	The weights on the Current Population Survey (CPS) files are controlled to independent estimates of the civilian noninstitutional population of the US. These are prepared monthly for use by Population Division here at the Census Bureau	Numeric	4508
education	Education of an individual	Nominal	16
education.num	Education number of an individual	Numeric	16
marital.status	Marital status of an individual. It can be Widowed,Divorced, Separated,Never-married, Married-civ-spouse, Married-spouse-absent, Married-AF-spouse	Nominal	7
occupation	Occupation of an individual	Nominal	14

relationship	Relationship of an individual. It can be Husband, Wife, Not-in-family, Unmarried, Own-child and other relative	Nominal	6
race	Race of an individual. It can be White, Black, Asian-Pac-Islander, Other and Amer-Indian-Eskimo	Nominal	5
sex	Gender of the individual. It is categorised as Female and Male	Nominal	2
capital.gain	Capital gain of an individual	Numeric	79
capital.loss	Capital loss of an individual	Numeric	56
hours.per.week	Hours per week of an individual	Numeric	78
native.country	Native country of an individual	Nominal	39
income	The class attribute. Income can be either greater than 50,000 or less than or equal to 50,000.	Nominal	2

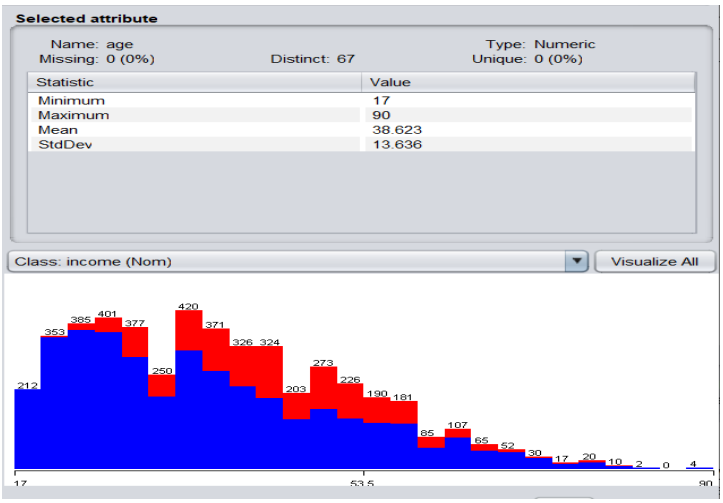


Figure 1 age

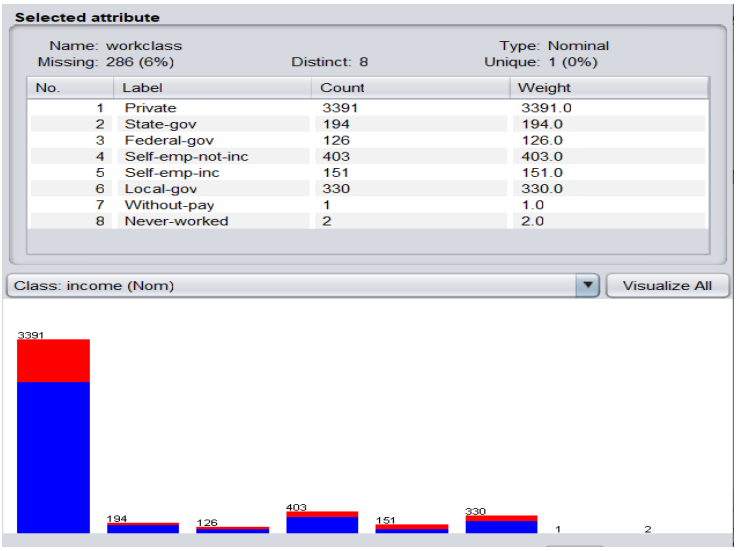


Figure 2workclass

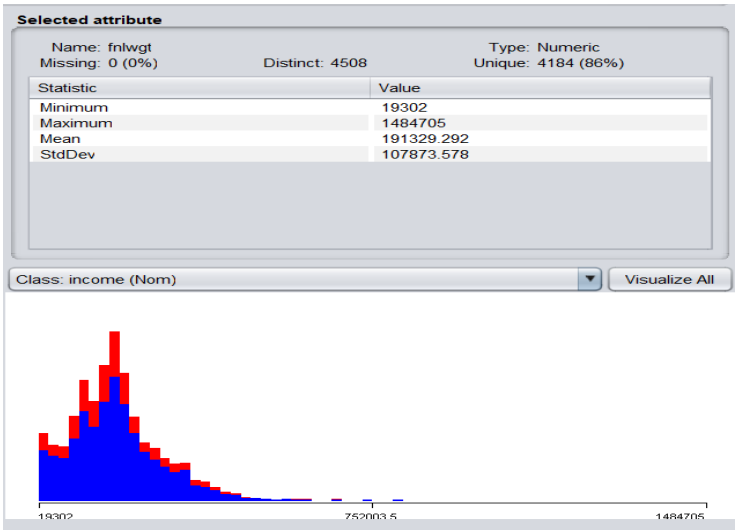


Figure 3fnlwgt

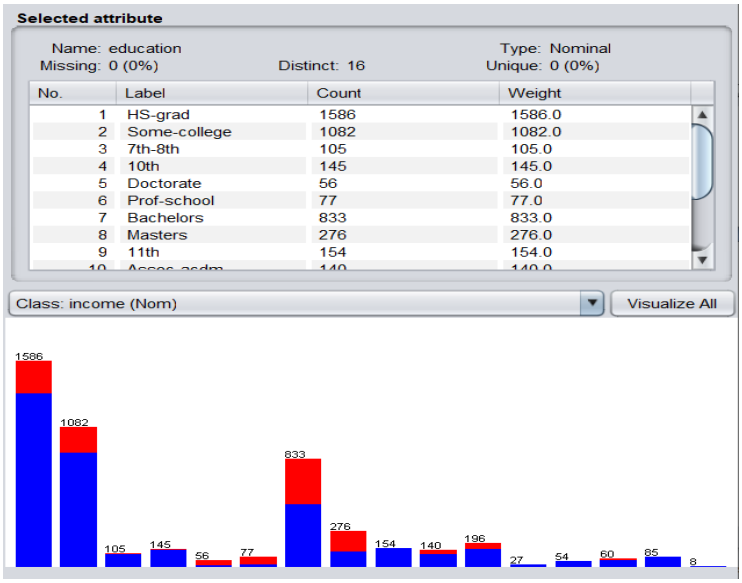


Figure 4 education

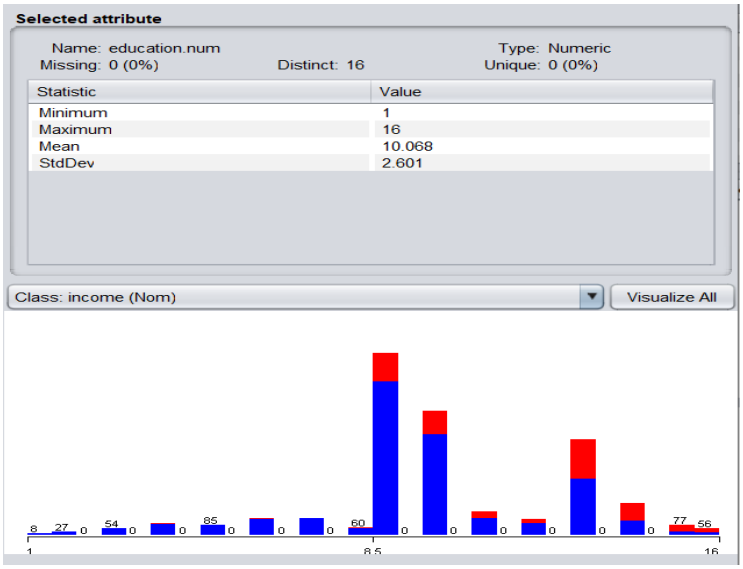


Figure 5education.num

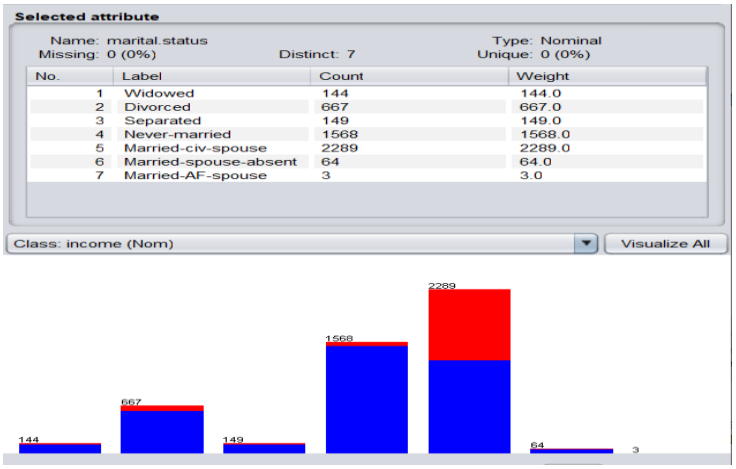


Figure 6marital.status

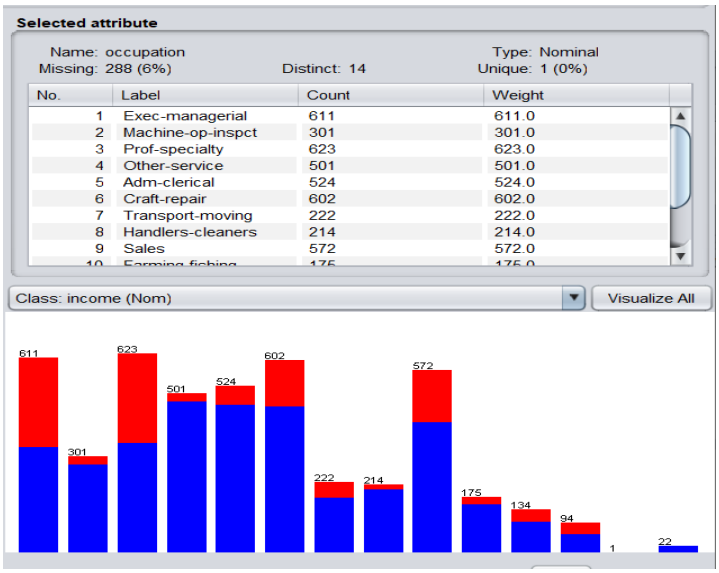


Figure 7 occupation

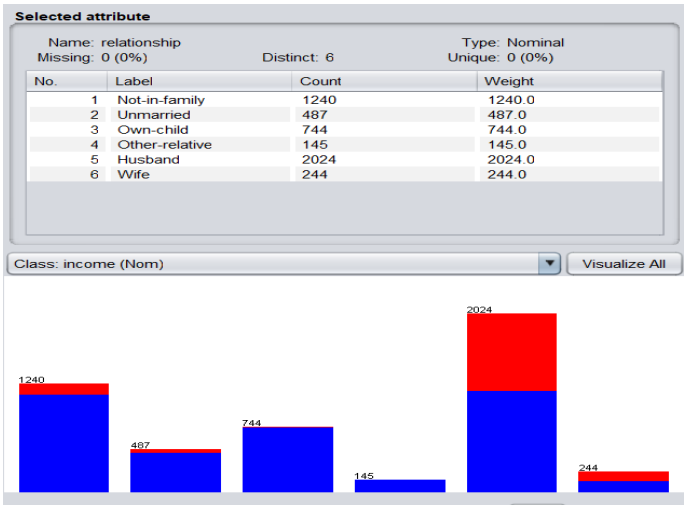


Figure 8 relationship

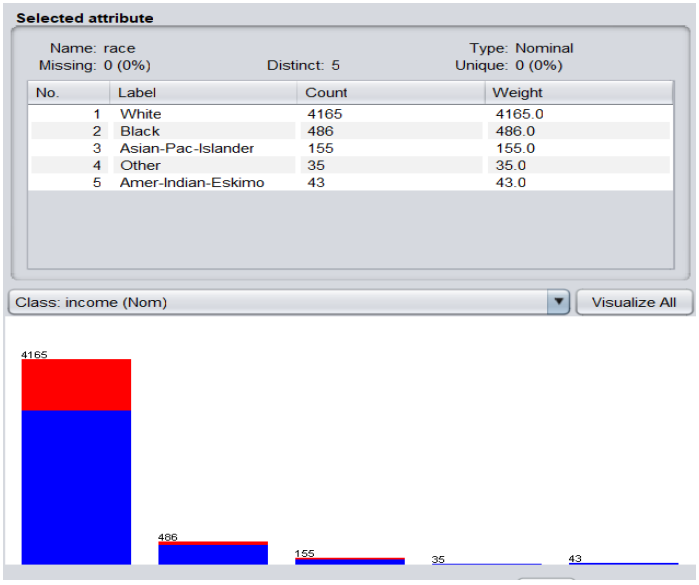


Figure 9 race

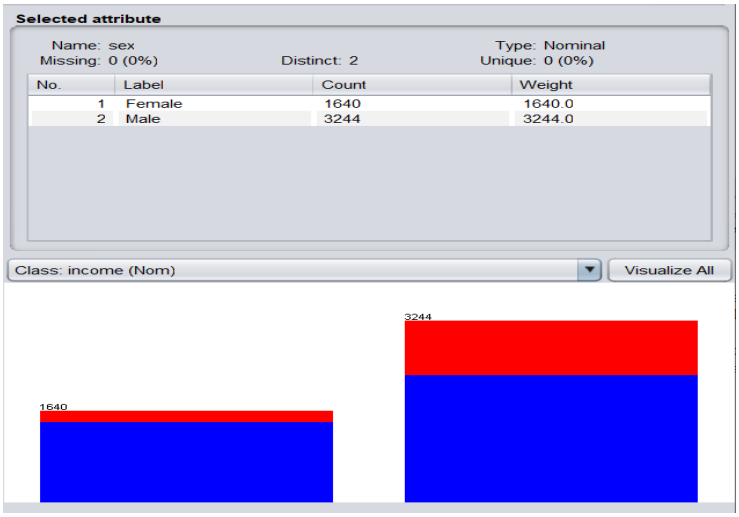


Figure 10 sex

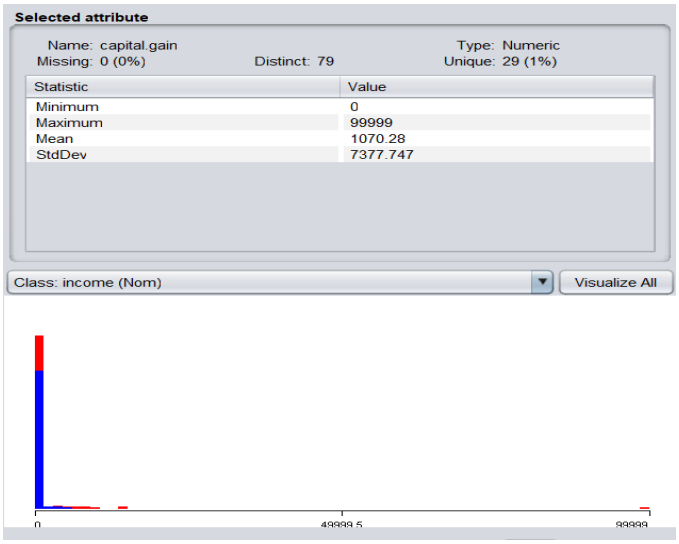


Figure 11capital.gain

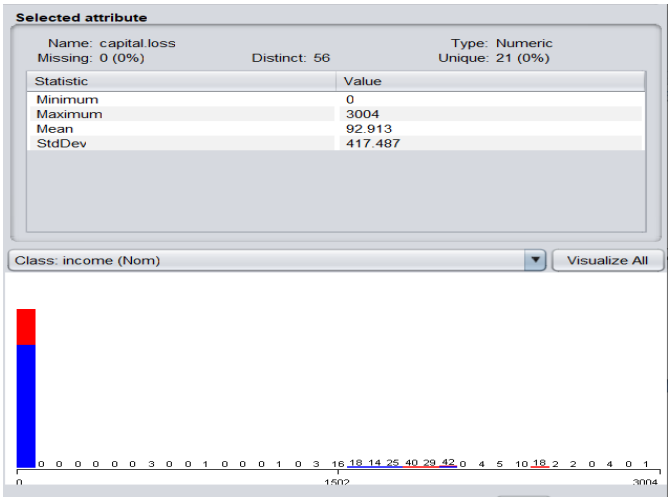


Figure 12capital.loss

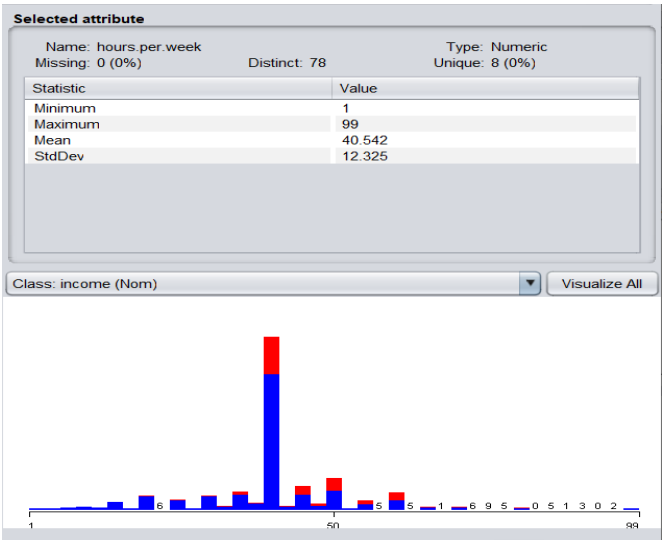


Figure 13hours.per.week

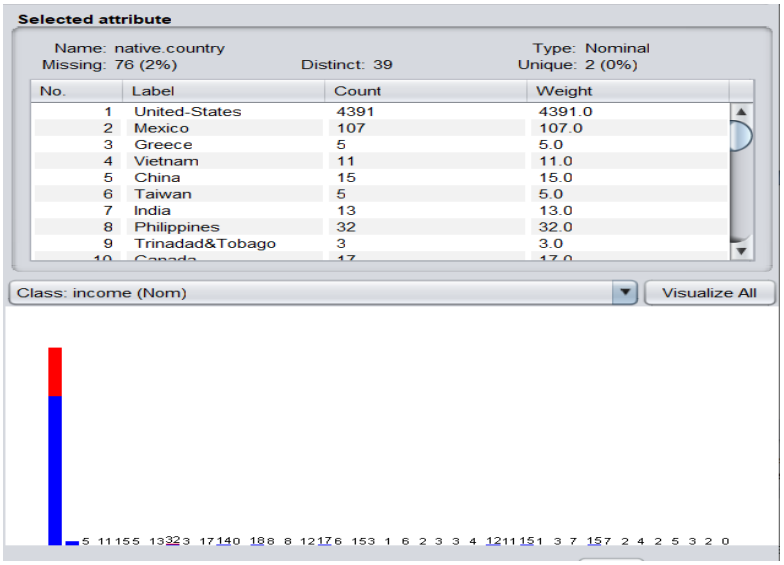


Figure 14native.country

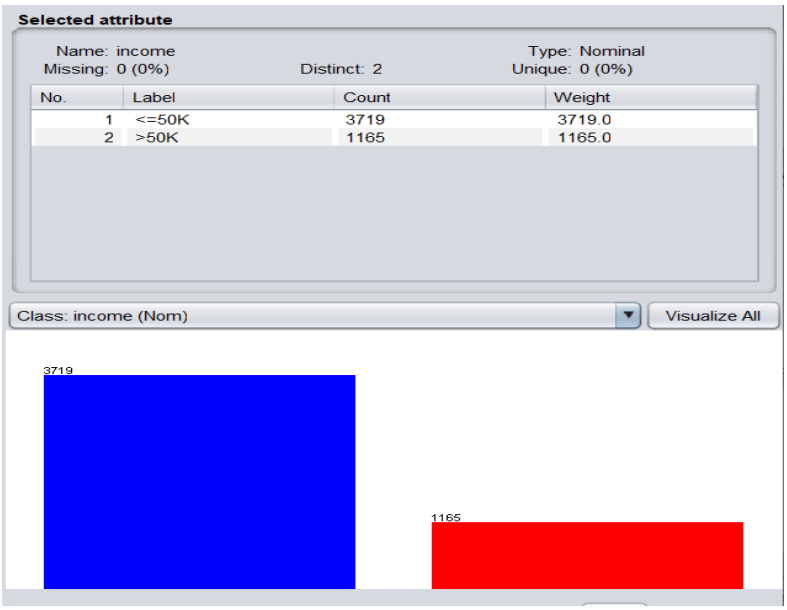


Figure 15income

Observations from attribute visualizations

- There are 15 multivariate attributes.
- Three attributes have very small percentage of missing values- 'workclass', 'occupation' and 'native.country'. None of the attributes have huge number of missing or noisy values.
- There are 6 numeric attributes and 9 nominal attributes.
- The dataset consists of 3719 instances with income equal to or less than 50,000 and 1165 instances with income exceeding 50,000. 'income' is the class that we intend to predict.

Objective:

1. To predict an individual's income through classification techniques such as decision tree and KNN, after understanding the effects of different factors such as age, gender,

occupation etc. through literature review and detailed study of the Adult Census Data Set and its attributes.

2. To apply various feature selection methods in Weka to identify the best features and to vary the hyperparameters to find and evaluate the models with better performance and accuracy after training the model and testing on the trained model.

3. To implement and evaluate ensemble methods such as Bagging, Boosting in Weka Experimenter to understand the effect of these ensemble methods on the model

Summary of Findings:

Preprocessing techniques such as feature encoding, feature selection, replacement of missing values, conversion of datatypes, outliers, normalization, discretization, class imbalance etc. were analysed for effectiveness on this dataset.

METHODOLOGY

1. Exploratory Data Analysis

```
In [10]: df.head(5)
```

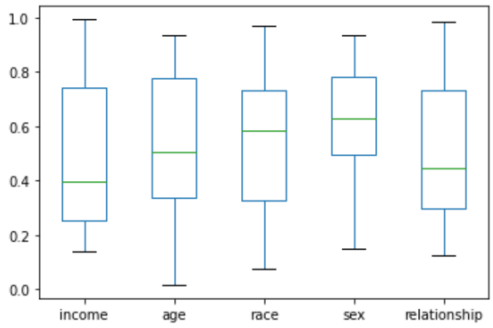
Out[10]:

m	marital.status	occupation	relationship	race	sex	capital.gain	capital.loss	hours.per.week	native.country	income
9	Widowed	?	Not-in-family	White	Female	0	4356	40	United-States	<=50K
9	Widowed	Exec-managerial	Not-in-family	White	Female	0	4356	18	United-States	<=50K
0	Widowed	?	Unmarried	Black	Female	0	4356	40	United-States	<=50K
4	Divorced	Machine-op-inspct	Unmarried	White	Female	0	3900	40	United-States	<=50K
0	Separated	Prof-specialty	Own-child	White	Female	0	3900	40	United-States	<=50K

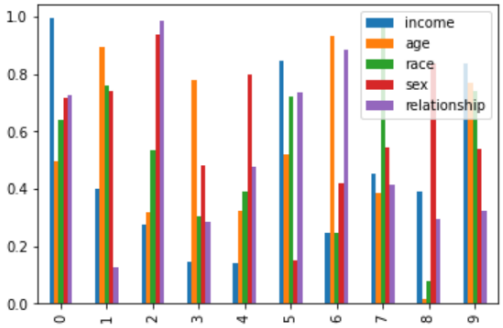
BOX AND WHISKER PLOT FOR CHECKING FOR OUTLIERS

```
In [18]: #checking for outliers using box and whisker plot
         df1= pd.DataFrame(np.random.rand(10, 5), columns=['income', 'age','race', 'sex', 'relationship'])
         df1.plot.box()
```

```
Out[18]: <matplotlib.axes._subplots.AxesSubplot at 0x7fd023106c90>
```



```
In [19]: df1.plot.bar()
Out[19]: <matplotlib.axes._subplots.AxesSubplot at 0x7fd02302f610>
```



2. Preprocessing

- 'fnlwgt' contains all 4508 unique values, thus they arenotusefulforclassification. It is thus not useful for data mining and is removedfrom the dataset.
- Normalization can be performed using Weka's normalize filter.
- However, in this dataset class imbalance is not a problem and hence class balancing need not be performed as it will reduce the performance.

3. Splitting

- The dataset is split into training and test dataset using 9:1 ratio.

4. Classification

For the classification task, decision tree, KNN, Logistic Regression and SVM is used. Decision trees are versatile and they can fit into complex datasets with a divide and conquer approach. J48 provides the best performance, the diagonal elements will be evenly distributed.

KNN is good for simple datasets, logistic regression works well when the attribute to be classified is non-numeric. SVM offers good performance with multidimensional data

In [28]:
```
from sklearn.linear_model import LogisticRegression
logreg = LogisticRegression()
logreg.fit(X_train, y_train)
print('Accuracy of Logistic regression classifier on training set: {:.2f}'
     .format(logreg.score(X_train, y_train)))
print('Accuracy of Logistic regression classifier on test set: {:.2f}'
     .format(logreg.score(X_test, y_test)))
```

```
Accuracy of Logistic regression classifier on training set: 0.81
Accuracy of Logistic regression classifier on test set: 0.81
```

DECISION TREES

In [29]:
```
from sklearn.tree import DecisionTreeClassifier
clf = DecisionTreeClassifier().fit(X_train, y_train)
print('Accuracy of Decision Tree classifier on training set: {:.2f}'
     .format(clf.score(X_train, y_train)))
print('Accuracy of Decision Tree classifier on test set: {:.2f}'
     .format(clf.score(X_test, y_test)))
```

```
Accuracy of Decision Tree classifier on training set: 0.85
```

Results

The best accuracy is offered by decision trees with 85% percentage.
Precision for income with <=50k is 84% and greater than 50k is 76%

In [33]:
```
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
pred = clf.predict(X_test)
print(confusion_matrix(y_test, pred))
print(classification_report(y_test, pred))
```

```
[[5944  249]
 [1161  787]]
              precision    recall  f1-score   support

       <=50K       0.84      0.96      0.89      6193
        >50K       0.76      0.40      0.53      1948

    accuracy                           0.83      8141
   macro avg       0.80      0.68      0.71      8141
weighted avg       0.82      0.83      0.81      8141
```

DISCUSSION

If we use Weka explorer for classification, some other observations are evident
Weka Explorer is used to test the trained model on a supplied test set. This is a very useful feature. Parameters are varied to see the effect in performance.

Results with confidence values, confusion matrix, tree, rules

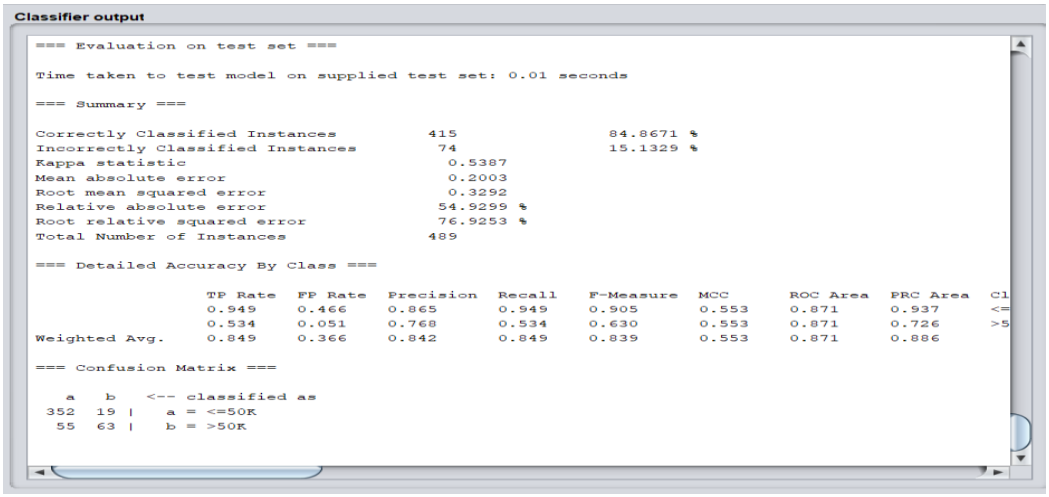


Figure 16 Experiment 1- J48 –Results

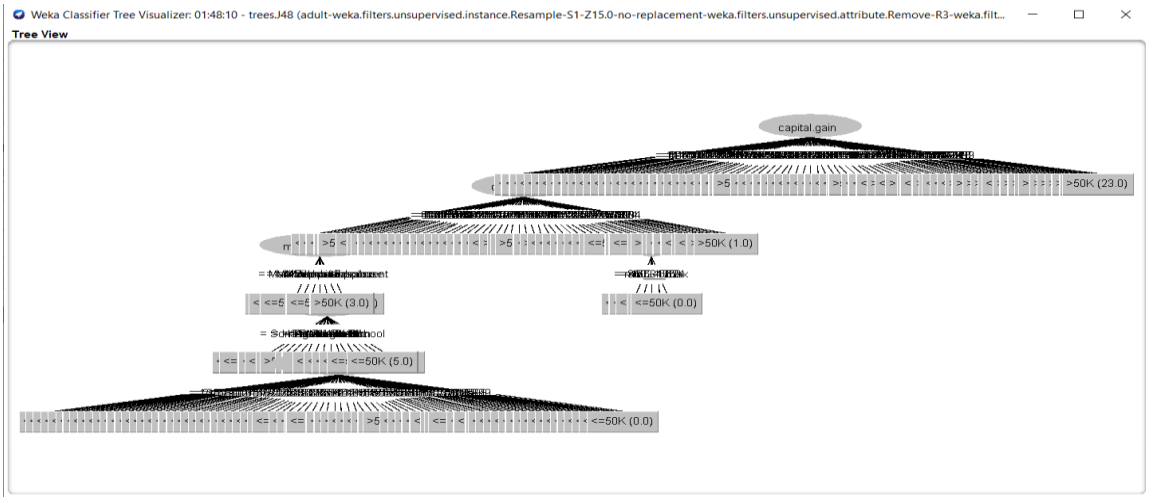


Figure 17 Experiment 1- J48 -Tree

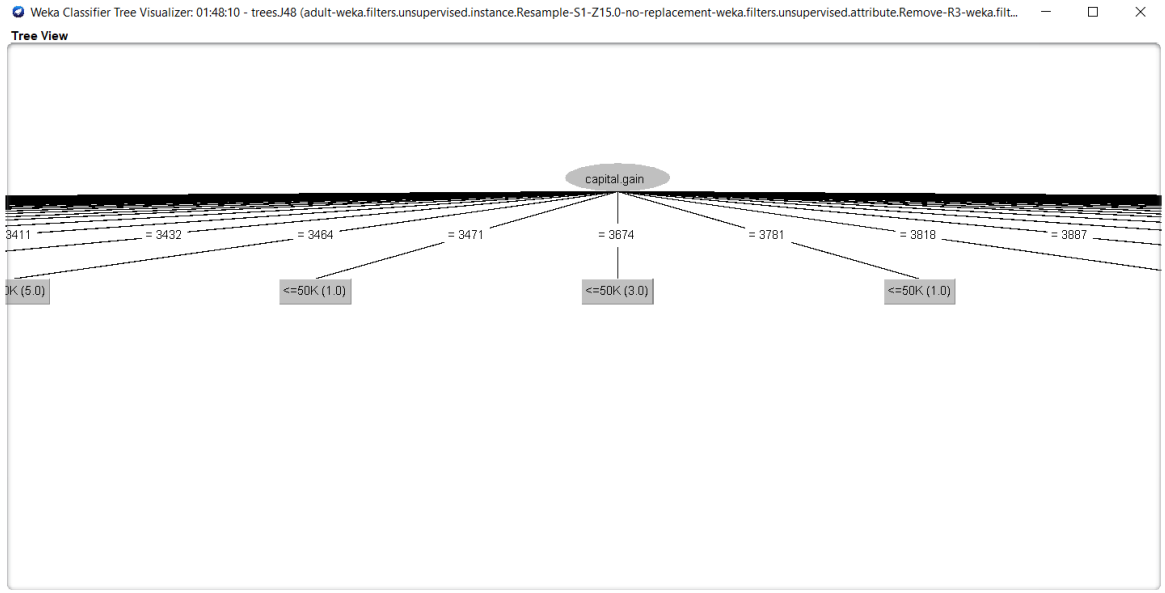


Figure 18 Experiment 1-J48-Tree

Findings

1. Using confidence factor 0.25 and minNumObj as 2, with pruned tree – 84.8671% of instances were correctly classified by this model.
2. Pruned tree increased the accuracy. The model used a pruned tree because pruning mostly reduces the complexity of the final classifier, reduce overfitting to increase predictive accuracy. (Wikipedia Decision Tree Pruning, 2020)
3. From, the confusion matrix it is understood that the model is more good at identifying the people with income less than 50k (a<=50k) – correctly classified 352 out of 371 and only incorrectly classified it as b >50k only 19 instances out of 371, whereas in identifying patients with income exceeding, it only classified 63 correctly and 55 were classified incorrectly.
4. From the visualization of the tree, it is evident that the tree gained more information from the 'capital.gain' attribute and further branching was made on this attribute. This also seems correct as this variable was the most ranked during feature selection. 'capital.gain' is the best feature on which the data is further split.

Conclusion

Income is still heavily dependant on factors such as sex, marital status and other attributes. We live in an era when people should be able to earn equally when compared to their counterparts of different gender, race, region etc. This has to change.
Analyses with data and machine learning can bring much needed change in this field

References

1. Archive.ics.uci.edu. 2020. *UCI Machine Learning Repository: Adult Data Set*. [online] Available at: <http://archive.ics.uci.edu/ml/datasets/adult> [Accessed 26 April 2020].
2. Bureau, U., 2020. *Census.Gov*. [online] Census.gov. Available at: <https://www.census.gov/> [Accessed 26 April 2020].
3. Dictionary.cambridge.org. 2020. *INCOME | Meaning In The Cambridge English Dictionary*. [online] Available at: <https://dictionary.cambridge.org/dictionary/english/income> [Accessed 24 April 2020].
4. Dl.acm.org. 2020. *Privacy Preserving OLAP | Proceedings Of The 2005 ACM SIGMOD International Conference On Management Of Data*. [online] Available at: <https://dl.acm.org/doi/10.1145/1066157.1066187> [Accessed 26 April 2020].
5. En.wikipedia.org. 2020. *United States Census Bureau*. [online] Available at: <https://en.wikipedia.org/wiki/United_States_Census_Bureau> [Accessed 23 April 2020].
6. Hannon, P., 2020. *How The Coronavirus Might Reduce Income Inequality*. [online] WSJ. Available at: <https://www.wsj.com/articles/how-the-coronavirus-might-reduce-income-inequality-11587304801> [Accessed 21 April 2020].
7. Inequality.org. 2020. *Gender Economic Inequality - Inequality.Org*. [online] Available at: <https://inequality.org/gender-inequality/> [Accessed 25 April 2020].
8. ResearchGate. 2020. *(PDF) A Statistical Approach To Adult Census Income Level Prediction*. [online] Available at: <https://www.researchgate.net/publication/328494313_A_Statistical_Approach_to_Adult_Census_Income_Level_Prediction> [Accessed 24 April 2020].
9. UK Government, 2020. [online] Available at: <https://www.youtube.com/watch?v=mo2dqHbLpQo> [Accessed 28 May 2020].
10. UK Government. 2020. [online] Available at: <https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/articles/analysisoffactorsaffectingearningsusingannualsurveyofhoursandearnings/01> [Accessed 20 April 2020].