CLASSIFYING ADULT CENSUS DATASET USING PYTHON AND WEKA – IBM WATSON

SUBHASREE VADUKOOT

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).

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.

ATTRIBUTES

Viewer

Relation: adult-weka.filters.unsupervised.instance.Resample-S1-Z15.0-no-replacement

No.	1: age Numeric	2: workclass Nominal	3: fnlwgt Numeric	4: education Nominal	5: education.num Numeric	6: marital.status Nominal	7: occupation Nominal	8: relationship Nominal	9: race Nominal	10: sex Nominal	11: capital.gain Numeric	12: capital.loss Numeric	13: hours.per.week Numeric	14: native.country Nominal	15: income Nominal
1	48.0	Federal-gov	3286...	HS-grad	9.0	Divorced	Handlers-...	Not-in-family	White	Male	0.0	0.0	40.0	United-States	(=50K
2	41.0	Self-emp-...	6663...	Some-col...	10.0	Divorced	Craft-repair	Not-in-family	White	Male	0.0	0.0	40.0	United-States	(=50K
3	19.0	Private	4055...	HS-grad	9.0	Never-married	Handlers-...	Own-child	White	Male	0.0	0.0	40.0	United-States	(=50K
4	28.0	Self-emp-...	1157...	Some-col...	10.0	Never-married	Exec-man...	Not-in-family	White	Male	0.0	0.0	50.0	United-States	(=50K
5	49.0	Private	2661...	HS-grad	9.0	Married-civ-s...	Machine-o...	Wife	White	Fem...	0.0	0.0	40.0	United-States	(=50K
6	23.0	Private	5954...	7th-8th	4.0	Never-married	Craft-repair	Not-in-family	White	Male	0.0	0.0	40.0	Mexico	(=50K
7	51.0	Private	1769...	HS-grad	9.0	Married-civ-s...	Machine-o...	Husband	White	Male	0.0	0.0	48.0	United-States	(=50K
8	23.0	Private	2887...	Some-col...	10.0	Never-married	Adm-clerical	Not-in-family	White	Fem...	0.0	0.0	30.0	United-States	(=50K
9	42.0	Self-emp-...	9652...	Bachelors	13.0	Married-civ-s...	Sales	Husband	White	Male	0.0	0.0	40.0	United-States	)50K
10	50.0	Self-emp-...	2836...	Doctorate	16.0	Married-civ-s...	Prof-speci...	Husband	White	Male	15024.0	0.0	60.0	United-States	)50K
11	17.0	Private	1830...	10th	6.0	Never-married	Other-serv...	Own-child	White	Fem...	0.0	0.0	25.0	United-States	(=50K
12	25.0	Private	2427...	Some-col...	10.0	Never-married	Transport-...	Not-in-family	White	Male	10520.0	0.0	50.0	United-States	)50K
13	23.0	Private	5324...	Bachelors	13.0	Never-married	Sales	Not-in-family	White	Male	0.0	1602.0	12.0	United-States	(=50K
14	26.0	Private	2470...	HS-grad	9.0	Married-civ-s...	Exec-man...	Husband	White	Male	0.0	0.0	52.0	United-States	(=50K
15	35.0	Self-emp-...	1113...	Assoc-ac...	12.0	Married-civ-s...	Sales	Husband	White	Male	0.0	1887.0	45.0	United-States	)50K
16	75.0	Self-emp-...	3059...	Masters	14.0	Married-spou...	Prof-speci...	Not-in-family	White	Fem...	0.0	0.0	50.0	United-States	(=50K
17	47.0	Private	1768...	HS-grad	9.0	Divorced	Craft-repair	Not-in-family	Black	Male	8614.0	0.0	44.0	United-States	)50K
18	36.0	Private	1207...	Masters	14.0	Married-civ-s...	Prof-speci...	Husband	Asia...	Male	0.0	0.0	40.0	China	(=50K
19	49.0	State-gov	5593...	Bachelors	13.0	Married-civ-s...	Adm-clerical	Husband	White	Male	0.0	0.0	40.0	United-States	)50K
20	22.0	Private	1139...	Some-col...	10.0	Never-married	Handlers-...	Own-child	White	Male	0.0	0.0	40.0	United-States	(=50K
21	41.0	Local-gov	5111...	Bachelors	13.0	Widowed	Adm-clerical	Not-in-family	White	Fem...	0.0	0.0	40.0	United-States	(=50K
22	57.0	Private	3191...	Bachelors	13.0	Never-married	Sales	Not-in-family	White	Male	0.0	0.0	40.0	United-States	(=50K
23	33.0	Private	1125...	Some-col...	10.0	Divorced	Sales	Not-in-family	White	Fem...	0.0	0.0	25.0	United-States	(=50K
24	20.0	Private	2383...	HS-grad	9.0	Never-married	Sales	Not-in-family	White	Fem...	0.0	0.0	30.0	United-States	(=50K
25	52.0	Local-gov	7478...	Masters	14.0	Married-civ-s...	Prof-speci...	Wife	White	Fem...	0.0	0.0	40.0	United-States	)50K
26	27.0	Private	3186...	10th	6.0	Never-married	Other-serv...	Not-in-family	White	Male	0.0	0.0	60.0	Mexico	(=50K
27	26.0	Self-emp-...	1024...	HS-grad	9.0	Never-married	Craft-repair	Own-child	White	Male	0.0	0.0	50.0	United-States	(=50K
28	35.0	Private	4602...	Assoc-ac...	12.0	Divorced	Other-serv...	Unmarried	White	Fem...	0.0	0.0	50.0	United-States	(=50K
29	22.0		2914...	12th	8.0	Never-married		Own-child	Black	Male	0.0	0.0	40.0	United-States	(=50K
30	22.0	Private	3850...	HS-grad	9.0	Never-married	Craft-repair	Own-child	White	Male	2907.0	0.0	40.0	United-States	(=50K
31	36.0	Self-emp-...	2873...	Assoc-ac...	12.0	Divorced	Sales	Unmarried	White	Fem...	0.0	0.0	35.0	United-States	(=50K
32	39.0	Private	2410...	Some-col...	10.0	Married-civ-s...	Sales	Husband	White	Male	0.0	0.0	40.0	Philippines	)50K
33	27.0	Private	1350...	HS-grad	9.0	Married-civ-s...	Craft-repair	Husband	White	Male	0.0	0.0	40.0	United-States	(=50K
34	34.0	Private	1335...	Some-col...	10.0	Divorced	Transport-...	Not-in-family	White	Male	2174.0	0.0	40.0	United-States	(=50K
35	17.0	Private	3276...	10th	6.0	Never-married	Other-serv...	Own-child	White	Male	0.0	0.0	15.0	United-States	(=50K

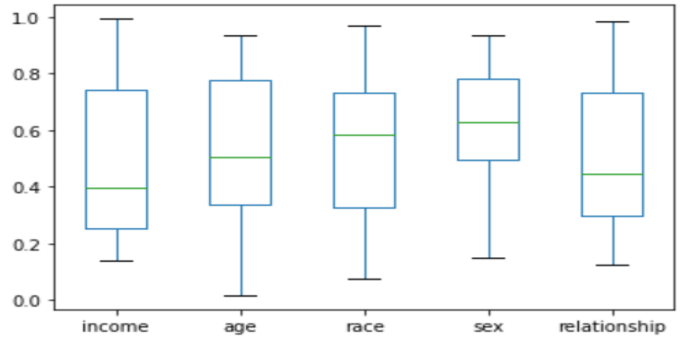
Add instance | Undo | OK | Cancel

EXPLORATORY DATA ANALYSIS

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>

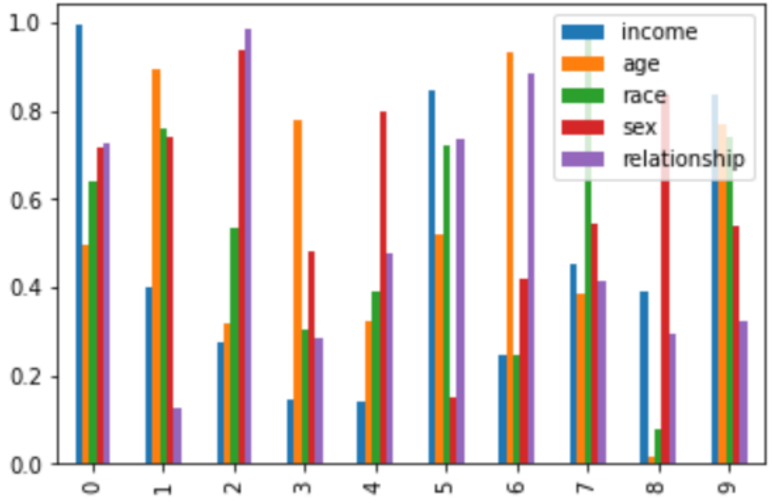


There are no outliers in the dataset

Bar plot

```
In [19]: df1.plot.bar()

Out[19]: <matplotlib.axes._subplots.AxesSubplot at 0x7fd02302f610>
```



CLASSIFICATION – MACHINE LEARNING

- DECISION TREE
- KNN
- LOGISTIC REGRESSION
- SVM

CLASSIFICATION CODE

```
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
```

CONFUSION MATRIX

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
```

RESULTS AND CONCLUSION

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