Adult Census Income - Final report

Team Details:

1. Asha Nagireddy (00682057)

2. Saikiran Rao Nellimarla (00686772)

Code link:

https://colab.research.google.com/drive/1bsC0rzqgqKJJO80ge4EWngXtZ5nu4oWu

Objective:

To predict the annual income range (<=50k or >50k) of US population, by analyzing various parameters of adult census Income data and building a predictive model.

- ✓ What race have the higher income?
- ✓ Does women earn less money than men?
- ✓ In what age do we have better chances to earn more?

Background Research:

The prominent inequality of wealth and income is a huge concern especially in the United States. The likelihood of diminishing poverty is one valid reason to reduce the world's surging level of economic inequality. 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.

Dataset:

Data was extracted from the 1994 Census bureau database by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics).

https://www.kaggle.com/uciml/adult-census-income#adult.csv

This dataset has 32561 data points and 15 features.

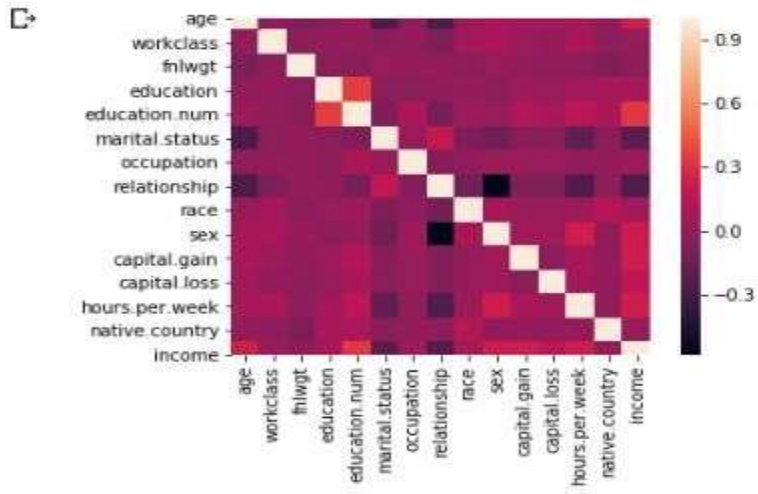
Attributes:

1. income: >50K, <=50K
2. age: continuous
3. workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
4. fnlwgt: continuous
5. education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool education-num: continuous
6. marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
7. occupation: 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
8. relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
9. race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
10. sex: Female, Male
11. capital-gain: continuous
12. capital-loss: continuous
13. hours-per-week: continuous
14. native-country: 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

Procedure:

1. Created a workspace on Google Colaboratory.
2. Imported all the necessary Libraries.
3. Loaded the ‘adult.csv’ dataset
4. Listed all Categorical and Numerical columns.
5. Visualized correlation between the featured and identified similar columns. Dropped education column as it represents same data as numerical column education_num.

Heat map :



We see there is a high correlation between education and education.num

6. <u>Data Cleaning:</u>

- Replaced '?' values with NAN
- Grouped categorical components.

    separated = ['Separated','Divorced']
- Dropped similar columns(education).
- Renamed columns.

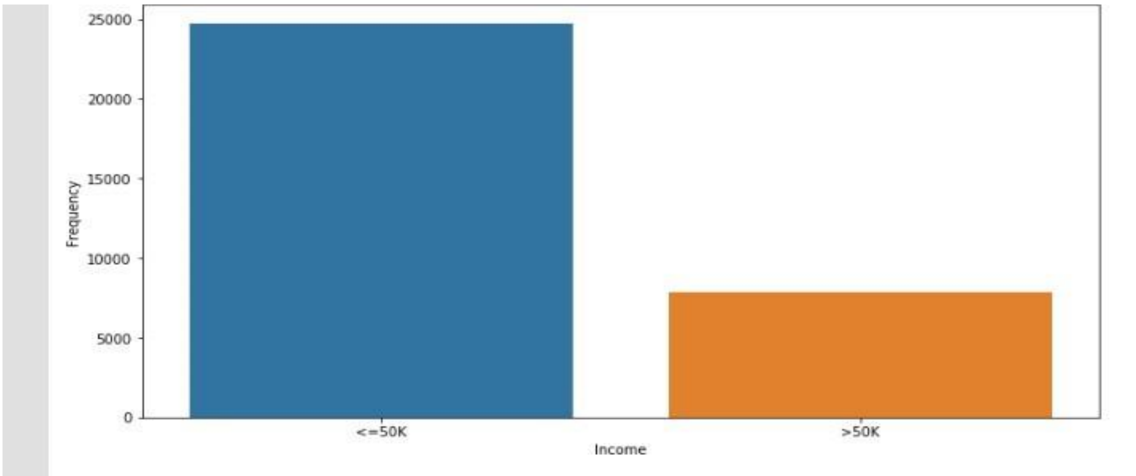
Before cleaning –

age	workclass	fnlwgt	~~education~~	education.num	marital.status	occupation	relationship	race	sex	capital.gain	capital.loss	hours.per.week	native.country	~~income~~
90	?	77053	HS-grad	9	Widowed	?	Not-in-family	White	Female	0	4356	40	United-States	<=50K
82	Private	132870	HS-grad	9	Widowed	Exec-managerial	Not-in-family	White	Female	0	4356	18	United-States	<=50K
66	?	186061	Some-college	10	Widowed	?	Unmarried	Black	Female	0	4356	40	United-States	<=50K

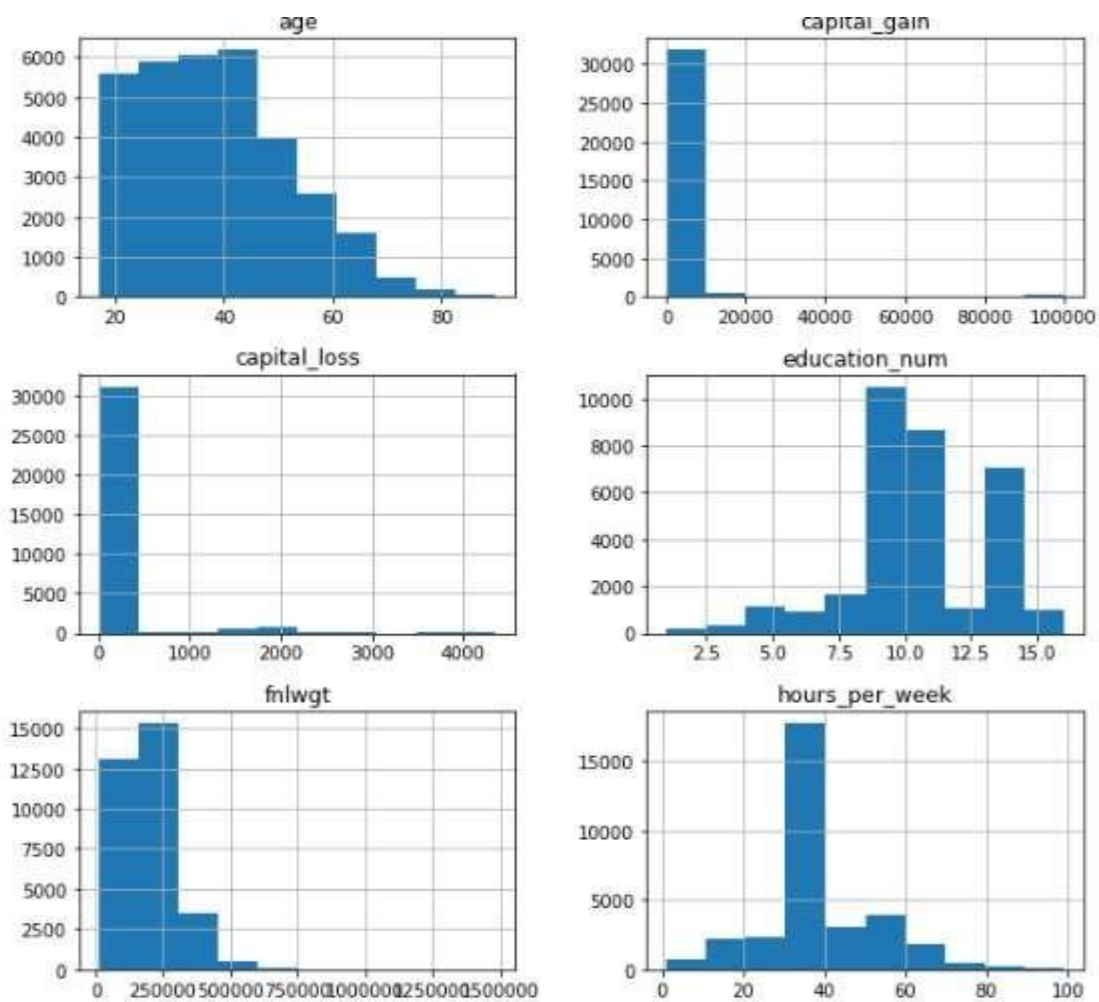
After Cleaning –

age	workclass	fnlwgt	education_num	marital_status	occupation	relationship	race	sex	capital_gain	capital_loss	hours_per_week	country	income_val
90	NAN	77053	9	Widowed	NAN	Not-in-family	White	Female	0	4356	40	United-States	0
82	Private	132870	9	Widowed	Exec-managerial	Not-in-family	White	Female	0	4356	18	United-States	0
66	NAN	186061	10	Widowed	NAN	Unmarried	Black	Female	0	4356	40	United-States	0

7. Data Visualization: Represented the cleaned data and frequency of income with Histograms.



This dataset contain 75% of people earning less than 50K and remainng 25% above 50K.



8. Analyzing the features effecting income using covariance metrics.
    - Appended a ‘income_val’ (numerical) column representing ‘income’ (categorical) column
    - Calculated covariance and correlation values for numerical columns and analyzed the outputs.

```
In [46]:  1  adult.cov()
```

Out[46]:

	age	fnlwgt	education_num	capital_gain	capital_loss	hours_per_week	income_val
age	186.061400	-1.103507e+05	1.281849	7.824819e+03	317.560742	11.580130	1.364997
fnlwgt	-110350.685300	1.114080e+10	-11729.527298	3.366625e+05	-436030.333167	-24460.426185	-427.056721
education_num	1.281849	-1.172953e+04	6.618890	2.330008e+03	82.856445	4.705338	0.368685
capital_gain	7824.818537	3.366625e+05	2330.007877	5.454254e+07	-94085.760688	7150.032029	705.230910
capital_loss	317.560742	-4.360303e+05	82.856445	-9.408576e+04	162376.937814	269.953755	25.935432
hours_per_week	11.580130	-2.446043e+04	4.705338	7.150032e+03	269.953755	152.458995	1.212651
income_val	1.364997	-4.270567e+02	0.368685	7.052309e+02	25.935432	1.212651	0.182826

Except the fnlwgt, the rest of the features are directly proportional to income.

As the income class increases, the frequency of population falling under the particular income class decreases.

```
adult.corr()
```

	age	fnlwgt	education_num	capital_gain	capital_loss	hours_per_week	income_val
age	1.000000	-0.076646	0.036527	0.077674	0.057775	0.068756	0.234037
fnlwgt	-0.076646	1.000000	-0.043195	0.000432	-0.010252	-0.018768	-0.009463
education_num	0.036527	-0.043195	1.000000	0.122630	0.079923	0.148123	0.335154
capital_gain	0.077674	0.000432	0.122630	1.000000	-0.031615	0.078409	0.223329
capital_loss	0.057775	-0.010252	0.079923	-0.031615	1.000000	0.054256	0.150526
hours_per_week	0.068756	-0.018768	0.148123	0.078409	0.054256	1.000000	0.229689
income_val	0.234037	-0.009463	0.335154	0.223329	0.150526	0.229689	1.000000

education_num feature has the highest correlation with income, followed by age, hours_per_week and capital_gain in order. fnlwgt has negative and least correlation.

9. Separated adult dataset into train and test data – 80% train data and 20% test data.

10. Building machine learning models with train dataset

- ***Logistic regression*** - Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). In statistics, the logistic model (or logit model) is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc... Each object being detected in the image would be assigned a probability between 0 and 1 and the sum adding to one. Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

- ***Navie Bayes*** - Naive Bayes is a probabilistic machine learning algorithm that can be used in a wide variety of classification tasks. it assumes the features that go into the model is independent of each other. Naïve Bayes has been studied extensively since the 1960s. It was introduced (though not under that name) into the text retrieval community in the early 1960s, and remains a popular (baseline) method for text categorization, the problem of judging documents as belonging to one category or the other (such as spam or legitimate, sports or politics, etc.) with word frequencies as the features. With appropriate pre-processing, it is competitive in this domain with more advanced methods including support vector machines. It also finds application in automatic medical diagnosis.That is changing the value of one feature, does not directly influence or change the value of any of the other features used in the algorithm.

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

- ***K-nearest Neighbors*** - k-nearest neighbors can be used in classification or regression machine learning tasks. It is non-parametric, which means that it does not make any assumptions about the probability distribution of the input. In pattern recognition, the k-nearest neighbors algorithm (k-NN) is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression:

Euclidean $$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

11. Cross – validated the models accuracy using test data set.
Used error correction/accuracy score and confusion matrix metrics to validate the models

a) Logistic Regression

```
Score accuracy:  0.810379241516966
confusion matrix:
           [[4961   16]
            [1252  284]]
```

b) Navie Baye's

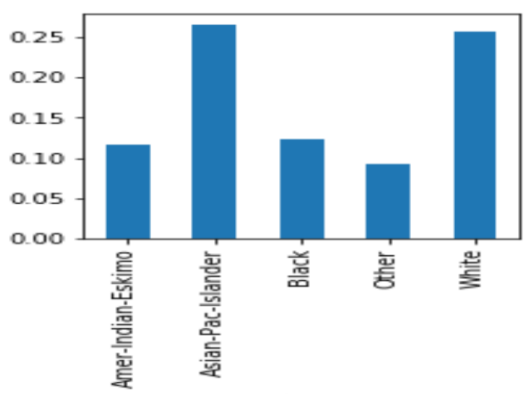
```
Score accuracy:  0.8109933978197451
confusion matrix:
              [[4766  211]
               [1020  516]]
```

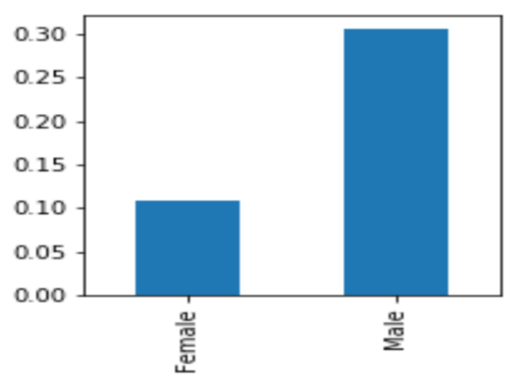
c) K-nearest Neighbors

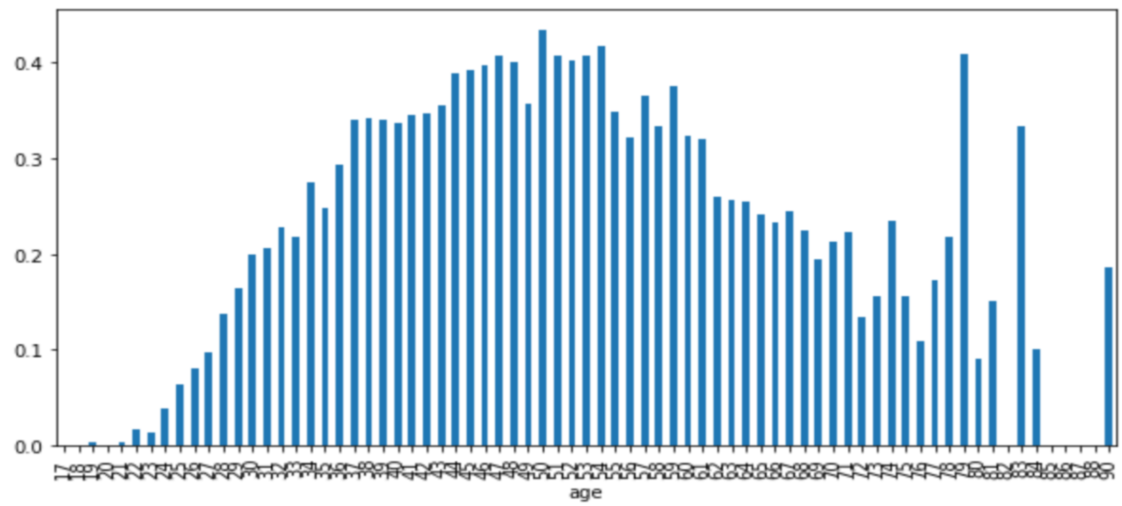
```
Accuracy:  78.44311377245509
confusion matrix:
              [[4583  394]
               [1010  526]]
```

12. Conclusion**:** Navie Baye's model has highest accuracy of 81.09% followed by Logistic regression model.

	Accuracy %
Navie Bayes	81.099340
LogisticRegression	81.037924
KNN	78.443114







- Men have more chances to have a higher income
- White and Asian Pacific Islanders have more chances than other races
- Income sort of follows the normal deviation, with a peak at 50 years old

References:

https://github.com/pooja2512/Adult-Census-Income
https://github.com/JcFreya/Adult-Census-Income
https://www.kaggle.com/ccentola/adult-census-income-analysis
https://towardsdatascience.com/logistic-regression-classifier-on-census-income-data-e1dbef0b5738
https://yanhan.github.io/posts/2017-02-15-analysis-of-the-adult-data-set-from-uci-machine-learning-repository.ipynb.html

Relevant papers

Ron Kohavi, "Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid"