PROJECT 12 : INCOME CENSUS

Group Members:

Aman Kumar Singh

Anwesha Patnaik

Bapuji Sirla

Shagufta Anjum Azad

Sidhi Agarwal

Problem Statement*:* Predicting if a citizen has income above/below 50K from Demographic Census Data.

Objective:

- Data Pre-processing
- Exploratory Data Analysis
- Data Modelling
- Data Model Evaluation

Dataset Description:

Attribute Information:

Listing of attributes:

Income: >50K, <=50K.

1. Age: continuous.
2. Workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
3. fnlwgt: continuous.
4. 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.
5. 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.

Variable identification:

By checking the dataypes and the data in each column of the dataset we can identify the variables as :
Numeric - age, fnlwgt, education-num, capital-gain, capital-loss, hours-per-week
Categorical - work-class, education, marital-status, race, occupation, relationship, sex, income, native-country.

There are 32561 rows and 15 columns.
No missing values.
There are unknown values for many variables in the Data set.
Variables with unknown/missing values are: ‘work-class’, ’native-country’ and ‘occupation’.

Pre-Processing and Data Representation

In ‘native-country’, ‘occupation’ and ‘work-class’ column we have 583,1843 and 1836 unknown ‘?’ values respectively. Hence we have imputed all the unknown values '?' with the mode of that respective column.

We have outliers present in columns **-** age, fnlwgt, education-num, capital-gain, capital-loss, and hours-per-week.

- **Handling outliers in Age column:** Most of the outliers are above the upper whisker hence we are replacing the outliers with the upper whisker value.
- **Handling outliers in education-num column:** Most of the outliers are below the lower whisker hence we are replacing the outliers with the lower whisker value.
- **Handling outliers in fnlwgt column:** Most of the outliers are above the upper whisker hence we are replacing the outliers with the upper whisker value.
- **Handling outliers in hours-per-week column:** Outliers are present both above and below the upper whisker and lower whisker respectively hence we are replacing the below values with the lower whisker value and above values with the upper whisker value.
- **Handling outliers in capital-gain and capital-loss column:** We are dropping these columns as more than 93 percent of the value are 0, hence they are not giving any insights.

Aggregation of column values:

- **Aggregation of Marital Status:** We have replaced all the column values and aggregated them into 3 categories – ‘Married’,’ Separated’ and ’Single’.
- **Aggregation of Native-Country:** We have replaced all the column values and aggregated them into 2 categories – ‘United States’ and ‘others’.

- **Aggregation of work-class:** We have replaced all the column values and aggregated them into 4 categories – ‘Government’, ‘Private’, ’Self-employed’ and ‘Others’.

- **Aggregation of education:** We have replaced all the column values and aggregated them into 4 categories – ‘Below Matric’, ‘Below Intermediate’, ’Under-grad’ and ’Post-Grad and above’.

We are aggregating various column values so that we have to deal with lesser categories/class values and data computation will be easier.

Label Encoder- We are using label encoder for all categorical variables except occupation they have lesser number of categories.

Creation of Dummy Variables – We are creating dummy variables for occupation because this column has large number of categories. If we are using label encoder here then our model gets biased towards the higher value.

Data Modelling (including Improvisation)

- **Splitting Data set into Train and Test data:** We are splitting our dataset into train and test data in the ratio of 80:20.
- **Data Scaling:** We are standardizing X train and X test data using standard scaler because ‘fnlwgt’ has higher scaled values.

Data Models:

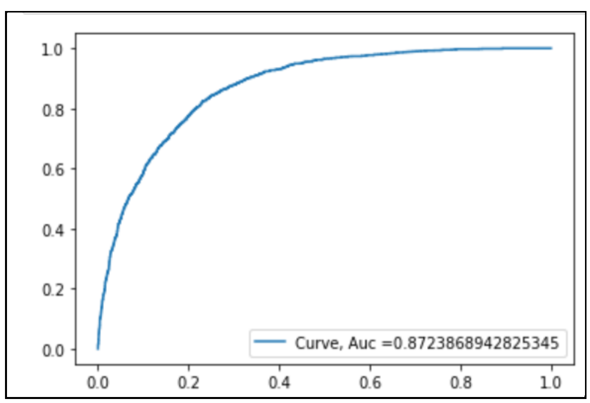
Logistic Regression:

For Test data

	precision	recall	f1-score	support
0	0.86	0.92	0.89	4918
1	0.68	0.54	0.60	1595
accuracy			0.82	6513
macro avg	0.77	0.73	0.74	6513
weighted avg	0.82	0.82	0.82	6513

For Train data

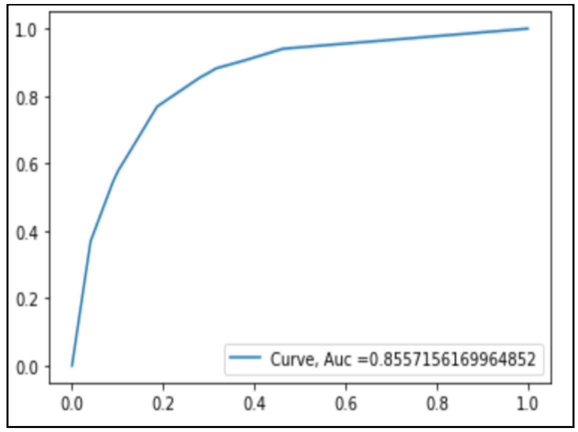
	precision	recall	f1-score	support
0	0.86	0.92	0.89	19802
1	0.69	0.54	0.60	6246
accuracy			0.83	26048
macro avg	0.78	0.73	0.75	26048
weighted avg	0.82	0.83	0.82	26048



Logistic Regression model is giving 0.82 percent accuracy and AUC curve value is 0.87 which states that it is quite a good model hence we are not going for hyper parameter tuning.

Decision Tree (After using hyper parameter tuning):

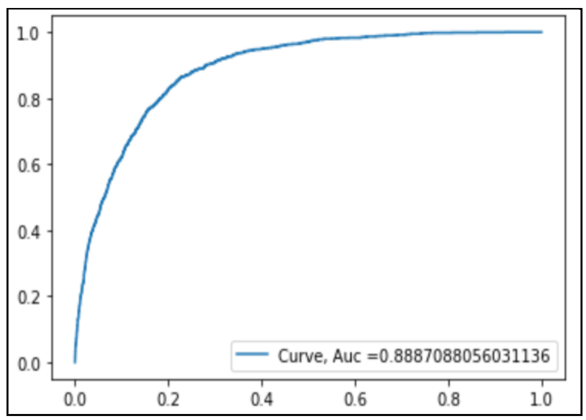
```
For Test data
              precision    recall  f1-score   support

           0       0.86      0.91      0.88      4918
           1       0.66      0.54      0.60      1595

    accuracy                           0.82      6513
   macro avg       0.76      0.73      0.74      6513
weighted avg       0.81      0.82      0.81      6513

For Train data
              precision    recall  f1-score   support

           0       0.86      0.92      0.89     19802
           1       0.68      0.54      0.60      6246

    accuracy                           0.83     26048
   macro avg       0.77      0.73      0.75     26048
weighted avg       0.82      0.83      0.82     26048
```



Decision tree model was suffering from over fitting hence we have opted for hyper parameter tuning using Grid Search in order to optimize the model to the dataset.

Random Forest Classifier(After using hyper parameter tuning):

```
For Test data
              precision    recall  f1-score   support

           0       0.87      0.92      0.89      4918
           1       0.70      0.56      0.62      1595

    accuracy                           0.83      6513
   macro avg       0.78      0.74      0.76      6513
weighted avg       0.82      0.83      0.83      6513

For Train data
              precision    recall  f1-score   support

           0       0.89      0.94      0.91     19802
           1       0.77      0.61      0.68      6246

    accuracy                           0.86     26048
   macro avg       0.83      0.78      0.80     26048
weighted avg       0.86      0.86      0.86     26048
```



Experimental Results and Comparison:

```
Accuracy from Logistic Regression: 0.8245048364808844
Accuracy from Decision Tree: 0.8206663595885153
Accuracy from Random Forest: 0.832181790265226
```

Random Forest Classifier is giving the best results in terms of accuracy.

Conclusion: Among all the models, Random Forest Classifier is giving the best accuracy.