NAME - ROHIT KUMAR
BATCH - DS2401
PHASE - 04

CENSUS INCOME PREDICTION

INTRODUCTION –

This data was extracted from the 1994 Census Bureau Dataset. This data contains the age, Work class, Education, Gender, Native Country, etc.

PROBLEM STATEMENT –

The problem of Census Income Prediction can be formulated as a Supervised Learning task, where the goal is to predict an Individual Income Level based on set of demographic and socioeconomic characteristics. The dataset used in this project is the Census Income from FlipnRobo's GitHub, which contains 32560 samples with 15 Columns including Age, Education, Occupation, Gender and Marital Status.

DATA ANALYSIS –

The data analysis part is to find the information, Null Values, Unique Values of the Object and Numerical Columns. The Observation tells -

- There are 15 Columns and 32560 Rows, where 6 columns are of Integer Datatype & 9 Columns are of Object Datatype.
- By checking Null Values – there are No Null Values.
- Information of Numerical Columns –
  By checking all unique values we find 'Capital Loss' & 'Capital Gain' columns have the most number of 0 into the columns.
  We fill that 0 value by Mean Method.
- Information of Object Columns –
  We observe that there are 3 columns in the dataset 'Work class', 'Occupation', 'Native Country' have the '?'.
  We fill it by using Lambda Technique and Replace '?' By 'Most Frequent' Value.

EXPLORATORY DATA ANALYSIS -

Now the EDA part start – we check the statistics first and find –

- There are huge difference in Minimum and Maximum of Capital Loss and Capital Gain which indicates the huge number of Outliers present in the columns.

DATA VISUALIZATION –

- Countplot of 'Sex vs Income', 'Native Country', 'Relationship vs Sex', 'Income vs Marital Status', 'Marital Status of a Male and Female'.
- Barplot of 'Sex vs Hours per Week', 'Race vs Hours per Week'
- Scatterplot of 'Native Country vs Education'
- Areaplot of 'Native Country vs Education'
- Age Distribution Plot.

DATA PRE-PROCESSING AND PIPELINE

- The data pre-processing part includes -
  - Checking Skewness, Removing Skewness of Capital Gain by cube root method.
  - Checking Outlier, Removing Outliers of Z-Score
  - Grouping Education and Education num columns, and Dropping Education Columns because both are same.
  - Label Encoding of all Object Column and make them Numerical column for prediction

- **Checking Correlation: -**
    - There are Strong Positive Correlation between ‘Age and Capital Loss’,’ Capital Loss and Hours Per week’ and ‘Hours per week and Native Country’
    - There are Strong Negative Correlation between ‘Relation and Sex’.

- Separating X (Independent) & Y (Target) columns.
- Scaling the X Variable using Standardization.
- Checking Multicollinearity by Variance Inflation Factor.
- Splitting the Data into training 75% and 25% testing.

BUILDING MACHINE LEARNING LIBRARY -

We use KNeighbours Classifier, Decision Tree Classifier, Logistic Regression Classifier, Random Forest Classifier, Gaussian NB, Extra Tree Classifier, Support Vector Classifier model was trained to predict income levels based on the 13 features.

The Performance of each model was evaluated using Accuracy Score. The model showes that –

Ther RandomForest Classifier model outperformaed the other models with an Accuracy Score of 86.34%.

HYPERTUNING

Checking the best parameter for prediction of Random Forest Classifier along using GridSearchCV.

We find – Max depth – 7
Max_Feature – None
Min_Samples_split – 7
N-Estimator – 34

Now Predicting the model by Hypertuning Parameters.

CONCLUDING REMARKS –

In this project, we demonstrated a machine learning approach to census income prediction using the Census Income dataset. The results showed that a Random forest model outperformed other models in predicting income levels based on demographic and socioeconomic characteristics. The project highlights the potential of machine learning in predicting income levels and informing policy decisions